MOBIMEDIA 2020
13th EAI International Conference on Mobile Multimedia Communications
Cyberspace
27 - 28 August 2020

EDITORS
Lin Yun
Tu Ya
Wang Meiyu
Proceedings of the 13th EAI International Conference on
Mobile Multimedia Communications

27-28 August 2020, Cyberspace

MobiMedia 2020

General Chairs
Zheng Dou, Harbin Engineering University, China

Technical Programme Chair
Yun Lin, Harbin Engineering University, China
Lin Ma, Harbin Institute of Technology, China
Joey Tianyi Zhou, Institute of High Performance Computing, A*STAR, Singapore
Conference Organization

Steering Committee

Steering Committee Chair
Imrich Chlamtac  Bruno Kessler Professor, University of Trento, Italy

Steering Committee Co-Chair
Honggang Wang  University of Massachusetts Dartmouth, USA

Steering Committee Members
Chonggang Wang  InterDigital Communications, USA, EiC of IEEE IoT Journal
Yonggang Wen  Nanyang Technological University, Singapore
Wei Wang  San Diego State University, USA
Shaoen Wu  Ball State University, USA
Qing Yang  University of North Texas, USA
Shiwen Mao  Auburn University, USA
Dapeng Wu  Chongqing University of Posts and Telecommunications, China

Organizing Committee

General Chair
Zheng Dou  Harbin Engineering University, China

General Co-Chairs
Weixiao Meng  Harbin Institute of Technology, China
Honggang Wang  University of Massachusetts Dartmouth, United States
Shiwen Mao  Auburn University, United States
Guan Gui  Nanjing University of Posts and Telecommunications, China
Qingjiang Shi  Tongji University, China

Sponsorship and Exhibit Chair
Zhigang Li  Harbin Engineering University, China
Liang Ye  Harbin Institute of Technology, China

Local Chair
Ya Tu  Harbin Engineering University, China
Preface

We are delighted to introduce the proceedings of the 13th edition of the 2020 European Alliance for Innovation (EAI) International Conference on Mobile Multimedia Communications (MOBIMEDIA). This conference has brought researchers, developers and practitioners around the world who are leveraging and developing multimedia coding, mobile communications and networking fields. The MOBIMEDIA 2020 was intended to provide a unique international forum for researchers from industry and academia to study new technologies, applications and standards.

The technical program of MOBIMEDIA 2020 consisted of 103 full papers. The conference tracks were: Track 1 - Basic Theories, Key Technologies and Artificial Intelligence for Next-Generations Wireless Communications; Track 2 - Intelligent Technologies for Subspace Learning and Clustering of High-dimensional Data; Track 3 – Main Track; and Track 4 – Late Track.

Aside from the high quality technical paper presentations, the technical program also featured a keynote speech and two special sections. The keynote speech was Prof. Wei Zhang from The University of New South Wales, Australia. The two special sections organized were Basic Theories, Key Technologies and Artificial Intelligence for Next-Generations Wireless Communications (BTKT-AI) and Intelligent Technologies for Subspace Learning and Clustering of High-dimensional Data (ITSLC). The BTKT-AI section aimed to address the key basic theories and technical ideas for the problems that the next generation of intelligent mobile communication will face. The ITSLC section aimed to gain insights into key challenges, and intelligent technologies for subspace learning and clustering are adopted to deal with high-dimensional data.

Coordination with the steering chairs, Imrich Chlamtac and Honggang Wang were essential for the success of the conference. We sincerely appreciate their constant support and guidance. It was also a great pleasure to work with such an excellent organizing committee team for their hard work in organizing and supporting the conference. In particular, the Technical Program Committee, Dr. Yun Lin, Dr. Lin Ma and Dr. Tianyi Zhou who have completed the peer-review process of technical papers and made a high-quality technical program. We are also grateful to Conference Managers, Lukas Skolek for his support and all the authors who submitted their papers to the MOBIMEDIA 2020 conference and sections.

We strongly believe that MOBIMEDIA conference provides a good forum for all researcher, developers and practitioners to discuss all science and technology aspects that are relevant to multimedia coding, mobile communications and networking fields. We also expect that the future MOBIMEDIA conference will be as successful and stimulating, as indicated by the contributions presented in this volume.

Dr. Yun Lin
Contents

Basic Theories, Key Technologies and Artificial Intelligence for Next-Generations Wireless Communications

Heterogeneous wireless sensor network routing protocol for an adaptive gray wolf optimizer
Chao Cheng, QingShan Han, Guoli Cheng, Shuang Zhai

Automatic Modulation Classification Using Hybrid Convolutional Neural Network
Xiaofei Zhang, Li Anping, Chai Lu, Ma Xiaoying, Wei Meiyong

Optimization research of radio monitoring network for the Beijing 2022 Olympic Winter Games
Xiaoying Ma, Meiyong Wei, Xiaofei Zhang, Xu Yang

Feature Fusion Convolutional Network Based Automatic Modulation Classification
Jiajie Mai, Shanchuan Ying, Nanxin Wang, Sai Huang

Automatic Modulation Classification Using Dense Memory Fusion Network
Anping Li, Juanjuan Huang, Xu Yang, Xiaofei Zhang, Meiyong Wei

A Calculation Method of vCPU Occupancy Rate of Virtual Machine Forwarding Process
Hua Lu, Qinshu Chen, Jidong Zhang, Xuefei Duan

A large capacity programmable packet forwarding device
Qinshu Chen, Hua Lu, Fusheng Zhu, Hui Li

Spectrum-efficient D2D-based puncturing algorithm for eMBB and URLLC coexisting networks
Yuhong Zhu, Kun Jiang, Linlin Zhao, Xuefen Chi, Shuang Wang

Research of Frequency Allocation Based on Improved Genetic Algorithm
Bo Li, Yunyi Zhai, Fugang Sun, Kun He, Richeng Guo, Cuican Wang

Transfer Learning Based Screen Defect Classification
Li Yilei, Li Chengyuan, Zhang Yifan, Chang Shuo, Zhang Fan, wang zixuan

Improved Loss Function for Defect Detection of Mobile Phone Screen
wang zixuan, Li Chengyuan, Zhang Yifan, Zhang Fan, Chang Shuo, Li Yilei

Intelligent Technologies for Subspace Learning and Clustering of High-dimensional Data

Detecting Urban Functional Area Based on Clustering on Base Station Data
Yuan Cheng, Yahong Wang, Ronghua Chi

A Clustering Analysis Method Based on Wilcoxon-Mann-Whitney Testing
Yuan Cheng, Weinan Jia, Ronghua Chi

An Uncertain Trajectory Modelling Method Based on Kernel Density Estimation
Yuan Cheng, Ronghua Chi, Yahong Wang
A Similarity Between Uncertain Data Measurement Method Based on stochastic simulation
Yuan Cheng, Ronghua Chi, Dapeng Lang

Discriminative Graph Learning via Low-Rank Constraint for Semi-Supervised Learning
Wang Ying, Ao Li

Interaction representation-based subspace learning for domain adaptation
Jia jia Chen, AO LI

Joint Cross-view Heterogeneous Discriminative Subspace Learning via Low-rank Representation
Yu Ding, Ao Li

Low Rank Sample Reconstruction-Based Semi-supervised Feature Subspace Learning
Ruoqi An, AO LI

Semi-supervised Corrupted Face Classification via Graph Learning
ZHONG YISHENG, AO LI

An Intelligent Fault Diagnosis Model Based on FastDTW for Railway Turnout
WEN-Jiang JI, Yuan ZUO, XING-Hong HEI, rong Fei

Security & Safety

Light Deep Learning based Edge Safety Surveillance
Yimo Lou, Wengang Cao, Zhimin He, Guan Gui

A Practical Low-Cost Security Solution for Log Management and File Integrity Monitoring
Lei Chen, Ming Yang, Hayden Wimmer, Brittany Wilbert

Analysis the Reliability Capability of Cognitive Frequency Hopping Communication System
Chenxi Li, Peihan Qi, Danyang Wang, Xiaoyu Zhou

Dynamic Machine Learning Algorithm for AODV Routing Attacks Detection
Md Raqibull Hasan, Yanxiao Zhao, Guodong Wang, Yu Luo, Lina Pu

Association rule mining of network security monitoring data based on time series
Fang XUE, Rong LIU

Design of encryption and decryption system of heterogeneous database based on Data Mining
Yan-hui BAI, Xi-jie MA

Blockchain-Based Vehicular Collaborative Computing
Qibin Zheng, Rongqing Zhang, Shengjie Zhao

Blockchain-Based Collaborative Decision-Making in Vehicular Networks
Dezhen Wang, Rongqing Zhang, Shengjie Zhao

Data transmission deployment and reliability analysis in enterprise informatization construction
Jun Li, XiaoLing Liu
Signal-level Honeypot: A Covert Communication and Interference Collection System
Cheng Chang, Xingsheng Zhu, Yue Gu, Zhijun Deng, Sheng Liu

Research on data transmission encryption algorithm of wireless sensor network in cloud storage
Xian-chun ZHOU, Qi-xiang WU

Communication networks & Coding Analysis

Analysis of Transmission Delay of Wireless Sensor Network Based on GNSS Signal
Rong LIU, Fang XUE

A resilient multi-path routing for the SDN based space-based network
Feihu Dong, Xi Chen, Yasheng Zhang, Zhongyuan Jiang, Ci He

Passive Radar based on the Specific Sequence of Wireless Network
Qian Zhao, Hui Zhao

Data transmission pattern recognition method for communication network with terminal band constraint
Xin Yuan CHEN, Jun Li

Fast Topology Inference of Wireless Networks Based on Hawkes Process
Yehui Song, Jiachen Sun, Guoru Ding

Deep N-ary Error Correcting Output Codes
Hao Zhang, Joey Tianyi Zhou, Tianying Wang, Ivor W. Tsang, Rick Siow Mong Goh

A Machine Learning Method for Optimizing Partition of Prediction Block in Coding Unit in H.265/HEVC
Wenchan Jiang, Ming Yang, Ying Xie, Zhigang Li

Study of Synthetic Airspeed Algorithm Based on Machine Learning for Lift Coefficient Curve Fitting
Dianzhong Chen, Yue Xu, Lei Wang

Does More Bandwidth Really Not Matter (Much)?
Seraj Al Mahmud Mostafa, Mike Wittie, Utkarsh Goel

Research on Non-stationary Blind Separation Method for Unfixed Signal Based on Extended Joint Diagonalization
Jun Li, Xin-Yuan Chen

FPGA Design and Optimization Implementation of GPS Positioning Algorithm
Wei He, Zhi Jiang, Zengshan Tian, Mu Zhou, Shuai Lu

Heading estimation algorithm for complex environment based on GPS single baseline
Zengshan Tian, Shuai Lu, Wei He, Mu Zhou, Zhi Jiang

Electromagnetic & Media Access Control

Radio Frequency Fingerprinting Driven Drone Identification Based on Complex-valued CNN
Hao Gu, Yu Wang, Guan Gui, Sheng Hong, Hao Huang, Jie Yang, Miao Liu, Jinlong Sun, Yun Lin
Clustering-Based Edge Compression Method with Application to Electromagnetic Object Recognition
Ziyan Yan, qibin zheng, yun lin, Qingjiang Shi

Efficient Resource Allocation For UAV Swarm Communication Systems With Awareness of Environmental Interference
Yinfei Qian, Lingyun Zhou, Ziyan Yan, Qingjiang Shi

Research on statistical method of channel information transmission risk probability in large-scale wireless communication system
feng ma, Hong-e WU

Radio Frequency Fingerprint Identification Based on Multi-Intervals Differential Constellation Trace Figures
yang yang, Aiqun Hu, jiaobao yu

Study on the optimization model of rural spatial main function area distribution based on 5g Technology
wen wen, Yan-dong Li

Research on 5G millimeter wave phased array antenna for broadband communication
Jihao Li, Zhenfeng Yang, Yibing Li, Xiaohang Sun, Ji Lv, Wei Zhong, Junfeng Dong

D2D & IoT

Research on Access Rate Optimization Algorithm in D2D Communication System
Yanliang Ge, Qiannan Zhao, Guanghua Zhang, weidang lu

Research on the performance of multi-user detection algorithm based on serial and threshold in the IoT
Jingqiu Ren, jianpeng Zhang, Zonglin Gu, Guanghua Zhang, weidang lu

Support Precise Latency for Network Based AR/VR Applications with New IP
Lijun Dong, Lin Han, Richard Li

A Novel Sample-Enhanced Dataset based on MFF for Large-Angle Face Recognition
He Wang, Yan Wang, Jie Liu, Guisheng Ying

A Three-party Repeated Game Model for Data Privacy in Mobile Edge Crowdsensing of IoT
Mingfeng Zhao, Lei Chen, Jinbo Xiong, Youliang Tian

The design of wireless communication and positioning system of electric vehicle based on CAN bus
Xi-jie MA, Tian XU, Yu-chao Li, Chen-yue WU, Yanhui Bai

Remote monitoring system for embedded network equipment faults based on Internet of Things
weixue liu

Embedded intelligent vacuum monitoring method of high voltage circuit breaker based on Internet of things
weixue liu
Multimedia platform & analysis

Research on multimedia digital real-time signal transmission based on probability and statistics
Man Li, Yu-min Yu, Feng Ma

Multimedia self-learning behavior monitoring data mining system based on Web
Chun-Sheng Deng, Hai-Bo Cheng, Jian-Ming Lin, Ju-Rong Yang, Yan-Chen ZHU

Research on data transmission method of new media marketing analysis based on Web4.0
Qi-xiang Wu, Lin He, Xian-chun Zhou

Designing and Testing of a Semi-Physical QUIC Video Streaming Simulation Platform
Jun Zhao, Siqie Zhang, Ben Liu, Hongjie Yu, Mingwu Yao, Fuzheng Yang

Signal processing method of campus mutual aid platform based on mobile terminal
Yi Huang, Jie Gao

Joint scheduling of audio teaching resources in NFV service chain deployment
Wei-dong Xu, Bing-yan Yu

New Energy & Smart City

A NewOOK Wake-Up Receiver With Sampling Technique
Tao Ma, Zhiyuan Tao, Jicheng Zhang

Research on node importance evaluation of railway logistics distribution network based on Rough Set
Xiu-miao Liu

Multi objective optimization model of railway transportation logistics network based on improved ant colony algorithm
Xiu-miao LIU

Hierarchical planning control model of smart city based on urban micro community
Wen Wen, Yan-dong Li, Shu Cheng
Research on Real-time Mining of New Energy Vehicle Fault Diagnosis Data under Cloud Computing
jiangwei, Yan XU, Chun-hua KONG

Numerical simulation and analysis of heat dissipation characteristics of battery box of electric vehicle based on coupling model
jiangwei, Chun-hua KONG, Yan XU

A Constellation Diagram Design Method with Minimum Average Energy: An Idea from Structural Chemistry
Yan Li

**Vision & Images Analysis**

Location of feature points in 3D reconstruction of multi vision color image based on principal component analysis
Huan-ping FENG, Li-wei ZHANG

Edge tracking method of damaged mural images based on deep learning
Li-wei ZHANG, Huan-ping FENG

A Novel 3D Reconstruction Algorithm of CT Images Based on Improved Marching Cubes Algorithm
Dongxue Su, Lin Ma

Lightweight Deep Learning Model for Invoice Image Classification
run wang, Chonghu Cheng, Yang Meng, Yingyi Sun, Jie Yang, Guan Gui

Research on image sensitive information recognition based on machine learning
Kaiyong Li, Duo-lu Mao

High speed linear array CCD image data acquisition system based on machine learning
Kaiyong Li

Automatic detection method of boundary data of gemstone bearing image based on Fourier transform
caihong li, Xiao-Jun ZHU

Signal Modulation Recognition Method based on Time-frequency Image
Yuqian Li, Cheng Chang, Chengzhuo Shi, Sen Wang

**Systems & Applications**

Design of early warning system for reservoir flood level based on probability analysis
Wen-bing JIANG, Li LIU

The design of high speed multiplex data transmission system based on single chip microcomputer
Rong-rong CUI, Yan-yan Ren

The design of data acquisition system of program automatic shelling based on ARM
Yan-yan Ren, Rong-rong CUI
Design of distributed storage system of digital media information based on Metadata
Xiao-Jun ZHU, caihong li

Data storage system of personalized multimedia network teaching resources based on Hypertext
Fei-fei WANG, Yan LIANG

Abnormal Data Mining Method in Environmental Monitoring Data of Animal Husbandry Farm
Xiao-hua XU, changxi chen

Design of multi-agent based decision system for building spatial planning
lin LIU, Jia Jie YU

Research on power control system of inspection robot based on wavelet transform
Hong-mei GUAN, Yue LIU

Research on Fault Location Method for Wind Turbine Line in Power System
Yue LIU, Hong-mei GUAN

Design of medical database information query system based on Android
Yan-Chen ZHU, Jian-Qiang Du, Zhi-Biao Li, Guo-Yue Zhong, Chun-Sheng Deng

Measurement method of enterprise information asymmetry based on probability analysis
Li LIU, Wen-bing JIANG

Case studies & Prediction & Educational Application

Predicting Sales of Cross-chain Discrete Manufacturing Products Based on LSTM
Zhicheng Wang, Ye Tao

Prediction model of spatial distribution pattern of building based on Neural Network
Jia Jie YU, lin LIU

Sentiment Prediction for User Comments on Home Appliances Products
Ye Tao, Cao Shi, Canhui Xu, Ruichun Hou, Zhifang Xu

Design of distance learning process monitoring system based on 5G communication network
Bing-yan YU, Wei-Dong XU

Design of Education Management System Based on Wireless Self-organizing Network
XiaoLing Liu, Jun LI

Research on the management of multimedia online education based on the virtual architecture of network function
Li Li TANG, Li TANG

Access control method of multimedia distance education resource database based on Web Platform
Yan LIANG, Fei-fei WANG
Design of Online Education Resource Scheduling System Based on Multi-level SDN Architecture

Li TANG, Li Li TANG
Heterogeneous wireless sensor network routing protocol for an adaptive gray wolf optimizer

1st Chao Cheng1; 2nd Qingshan Han2; 3rd Guoli Cheng2; 4th Shuang Zhai1
1School of Computer Science and Engineering, Changchun University of Technology, Changchun, China;
2National Engineering Laboratory, CRRC Changchun Railway Vehicles Co., Ltd., Changchun, China;
1st chengchao@ccut.edu.cn; 2nd hany0301@163.com; 3rd chengguoli@cccar.com.cn; 4th zhaishuang@ccut.edu.cn

Abstract: Wireless sensor network (WSN), plays an increasingly important role in information collection. In this paper, firstly, in order to adapt to the actual conditions, the communication process of the node's energy is limited, and a three-stage energy heterogeneous network model is designed. Secondly, for the convergence node frequent task forwarding and complex cluster first-round energy consumption, by combining the optimal number of cluster heads with the gray wolf optimization algorithm, a new fitness function is designed that integrates the remaining energy of the nodes and the distance from the nodes to the base station. In addition, an improved iterative factor is introduced to enhance the ability of local search in cluster head selection, so as to improve the accuracy of cluster head search. Finally, the simulation results show that the proposed method extends the lifetime of the network 50%, reduces the process of energy consumption, and improves the throughput of network data 30%.

Keywords: Energy efficient; gray wolf optimizer (GWO); balanced cluster structure; wireless sensor network

1 Introduction

With the development of wireless communication technology and sensor technology, the self-organizing wireless sensor network with micro and low-power nodes plays an increasingly important role in information acquisition under the special circumstances [1]. The WSN nodes, which is randomly monitored around the clock, is widely used in the fields of environmental monitoring, military reconnaissance and medical data collection [2-4]. Because the nodes in WSN are powered by a limited capacity micro-battery, it is a hot issue to find a WSN routing protocol with high energy utilization and strong fault tolerance to reduce the energy loss of the node, thus prolonging the network life cycle [5-7].

The WSN routing model can be divided into isomorphic networks and heterogeneous networks according to the consistency of sensor type and communication radius distance [8,9]. One of the most representative routing protocol is the Low Energy Adaptive Clustering (Leach) [10] proposed by Heinzelman in 2000. The algorithm takes the dynamic "wheel" as the working period, and randomly selects the cluster heads (CHs) by setting the threshold and the probability combination, but the probability of the lower energy node being selected as the CHs is the same as the probability of the normal node. In order to improve the simple probability to determine the energy consumption, many scholars have designed quite a few
new protocols to improve the performance of LEACH [11-12] by reducing the energy consumption of the CHs. In [13], the LEACH-EA algorithm, which uses enhanced network communication to reduce energy consumption, improves the survivability of the first node in the literature. However, the way in which cluster heads are selected based on probability does not change. In [14], an energy balance hierarchical routing algorithm (EBHRA) is proposed, which uses the residual energy of the node as the CHs election weight. However, the method of considering only the remaining energy of the node is too single. The possibility of CHs increases the energy loss of long-distance communication networks. A multi-factor cluster head selection strategy for residual energy and node density decision factors is proposed in the literature [15], which improves the decision factor of cluster head, but improves the network survival cycle by about 10% to 60%. In order to improve the traditional node clustering method, Tapswini Samant et al. proposed to reduce the number of data transmissions and reduce the network energy consumption low energy threshold sensitivity protocol (TEEN) by setting the node sensing threshold [16]. Since cooperative communication requires multiple nodes, the TEEN protocol is mainly used for cluster formation. On this basis, the TEEN-vector quantization (TEEN-V) protocol proposed by D.E. Boubiche et al., which is used for cluster communication in cooperative networks to improve the survival period of the network as much as possible [17]. However, since the ideal homogeneous network structure does not exist under realistic conditions, the homogeneous network model has great limitations in the process of production practice.

Therefore, it is important to design a heterogeneous WSN network with high reliability, integrity of data transmission, and enhanced heterogeneity of nodes [18,19]. Smaragdakis et al. proposed the earliest heterogeneous stable election protocol (SEP) to extend the death period of the first node in the network [20], but did not consider the residual energy and node location information of the node. A multi-hop routing communication protocol (MCR) is proposed in the literature [21], which changes the election mode of CHs by reducing network stability, but the energy level is too simple to set the level to achieve the complexity of calculating heterogeneous. Young et al. proposed a distributed energy efficient clustering (DEEC) network heterogeneous model, which selects the CHs by the ratio of the residual energy of the surviving node in the network to the average energy of the entire node [22]. If only from the perspective of energy consumption, when the CHs are far away from the base station BS node, there is still a chance to be elected cluster heads to undoubtedly increase the network energy burden; In [23-24], an enhanced DEEC algorithm (E-DEEC) for enhanced heterogeneous LEACH (EHE-LEACH) and three nodes was proposed. Therefore, the importance of routing node density and super node to enhance the stability and heterogeneity of the network life cycle is emphasized. In [25], Ya liang et al. combined heterogeneous data with tensor multi-clustering method (TMC) to provide a future research value for evaluating the cluster performance of data. The main contributions of this paper are as follows:

- According to the energy model in the wireless sensor network, the optimal number of cluster heads is selected. Combined with the K-MEANS algorithm to complete the process of starting clustering, reduce the energy loss in the process of cluster establishment.
- Adopting the Gray-wolf Algorithm decision-making model of sub-region, combining the energy heterogeneity and computing heterogeneous characteristics of heterogeneous wireless sensor networks to select the CHs candidate nodes that are most suitable for the network.
- Construct a information evaluation model that combines the distance from node to the base station and the node's own energy, introduces the adaptive factor of local
optimization of weight renewal, enhances the local optimization mechanism of the algorithm, and finds the most suitable cluster head.

The rest of the paper is organized as follows. In Section II mainly introduces the energy model in the wireless sensor network, which is the basis of the design of this paper. In Section III proposes an adaptive gray wolf algorithm routing decision model, and the Section IV is to verify the simulation results. Finally, the conclusion are drawn in Section V.

2 Network and energy model

According to the heterogeneous characteristics of sensor nodes, heterogeneous WSNs can be classified three types: energy heterogeneity, heterogeneous communication capability and heterogeneous computing power. Energy heterogeneous means that network node configurations have different initial energy; Heterogeneous communication capability is the difference between node transmission rate, communication link and communication protocol. The computing capability is heterogeneous in the node processing processing capability, sensing capability and storage space difference. In heterogeneous WSNs, whether the communication capability is heterogeneous or the computing power is heterogeneous, as the network operation will lead to node energy heterogeneity, energy isomerism is the basis for studying heterogeneous WSN.

The energy consumption model of the first-order wireless communication mode used in this paper is shown in Figure 1. The energy of the nodes in the model is limited, and the radio signals consume the same initial energy in all directions [26].

![Energy dissipation model](image)

According to the energy loss model, the energy consumed when transmitting 1-bit information is shown as [27]:

\[
E_{TX}(l,d) = E_{TX-elec}(l) + E_{TX-mp}(l,d) = \begin{cases} l * E_{elec} + l * \epsilon_{mp} * d^4 & (d \geq d_0) \\ l * E_{elec} + l * \epsilon_{mp} * d^2 & (d < d_0) \end{cases}
\]

where \( E_{TX-elec} \) and \( E_{TX-mp} \) is the energy consumed by the transmitter when transmitting or receiving 1 bit of data, \( \epsilon_{mp} \) is the power amplification factor of the multipath attenuation model, and \( \epsilon_{fs} \) is the power amplification factor of the free space model. When the transmission distance \( d \geq d_0 \):

\[
d_0 = \frac{\epsilon_{fs}}{\sqrt{\epsilon_{mp}}}
\]
3 Routing decision based on adaptive grey wolf algorithm (AD-GWO)

The three most suitable sensor nodes are selected by designing a combination of node residual energy and node-to-base station distance characteristics. Give the three nodes the strongest guiding effect, guiding the evolution direction of the whole population, thus approaching the best cluster head position. In order to reduce the time and space loss of the iterative process, an adaptive iterative factor selection method is proposed. In order to reduce the energy loss of nodes joining the cluster network in the random clustering process, selecting the appropriate CHs is an indispensable process for establishing a heterogeneous network. The specific analysis is as follows:

3.1 Energy-based optimal cluster head number selection decision

According to the physical protocol and the running process, the main energy consumed by the CHs is mainly divided into the energy consumed in the data acquisition phase, the energy in the data capacity and the processing phase. The free decay model is mainly used in this paper. Assuming that N random nodes are randomly deployed in an M*M region, the average number of non-cluster head members in each cluster is N/K-1 [28], and the energy consumption of the CHs is as shown:

\[ E_{CH} = l * E_{elec} * \left( \frac{N}{K} - 1 \right) + l * E_{fs} * \frac{N}{K} + (l * E_{elec} + l * e_p * d_{toBS}^2) \]  

The non-CHs consumes energy as shown in equation (4)(5):

\[ E_{NCH} = l * E_{elec} + l * e_p * d_{toch}^2 \]

\[ E[d_{toch}^2] = \iint (x^2 + y^2) \rho(x, y) dx dy = \frac{M^2}{2\pi K} \]  

Then the total energy consumed in a cluster.

\[ E_{ele} = E_{CH} + \sum E_{NCH} = E_{CH} + \left( \frac{N}{K} - 1 \right) * E_{NCH} \]

Then the total energy consumed in the K clusters is:

\[ E_r = K * E_{ele} = l * \left[ 2 * E_{ele} * N + N * EDA + k * e_p * d_{toBS}^4 + N * e_p * M^2 \right] \]

If the energy consumption in each of the above equations is regarded as a function of K [29], then:

\[ K = \sqrt{\frac{e_p * N * M^2}{2\pi e_{mp} * d_{toBS}}} \]

In the above formula, the simulation results in Chapter Section IV show that when the base station is located at (50, 50), d_{toBS} is about 141m, and according to the minimum of 50m in the literature, the value of K in this paper is in [4,7]. According to the simulation results, the following table can be obtained:

The (CHs) receives information by all CNs members in the cluster, and transmits collects information to the base station(BS) through data fusion. However, the distance d_{ch} (CHs to the BS) is not same, and the process of energy consumption of the CHs determines the robustness of entire network. If Number_{CH} < K (K is the number of CHs actually needed), the
CHs cannot cover all the monitoring areas; and the energy loss during data transmission will increase accordingly. Therefore, it is especially important to choose the appropriate CHs.

Table 1. The survival time corresponding to different K values

<table>
<thead>
<tr>
<th>K value</th>
<th>Number</th>
<th>Initial node death</th>
<th>All nodes die</th>
</tr>
</thead>
<tbody>
<tr>
<td>K=5</td>
<td>1</td>
<td>1261</td>
<td>2033</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1286</td>
<td>2053</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1047</td>
<td>2059</td>
</tr>
<tr>
<td>K=6</td>
<td>1</td>
<td>1060</td>
<td>2030</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>968</td>
<td>1998</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>928</td>
<td>2013</td>
</tr>
<tr>
<td>K=7</td>
<td>1</td>
<td>998</td>
<td>2018</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>993</td>
<td>2016</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1085</td>
<td>2014</td>
</tr>
</tbody>
</table>

Different from the traditional WSN network establishment process, this paper first compares the distance between each node and BS. If the basic communication requirements are met, direct communication is performed. Otherwise, the distance from each node to the CHs will enter the cluster establishment phase. This approach reduces the energy consumption during the setup phase of the clusters complex clusters. The detailed pseudo code is shown in Algorithm 1:

Algorithm 1 Clustering process

Input: Randomly generate N node coordinates \( N_i=(x_i, y_i) \) Max range of communication \( d_0 \)
Output: Cluster head node coordinates
1: \( K_i(i, j) = \emptyset \)
2: while \( d_{ij} > d_0 \) do
3:     for \( i = 1; i \leq K; i \) do
4:         for each centroid do
5:             Calculate the distance between the centroid and the data point
6:         end for
7:     end for
8:     for each cluster do
9:     end for
10:     Find the centroid closest to them
11:     end for
12:     The current node belongs to the cluster
13:     The current node belongs to the cluster
14: end while
15: Nodehead then communicate directly with
16: return \( K_i(i, j) \);

3.2 Adaptive Grey Wolf Optimization Algorithm (Ad-GWO) Model

As a typical cluster intelligent optimization algorithm, GWO [30] simulates the predation behavior of the grey wolf population, and compares the wolf group tracking process with the prey as the optimization process, so as to achieve the optimal solution. Applying GWO to the cluster head optimization problem of WSNs, the position of the wolf group represents the position of the sensor. Compared with the position where the prey is hunt, the position of the CHs is also the position of the optimal solution expressed in this paper. In the gray wolf algorithm, the \( \alpha \) wolf, the \( \beta \) wolf, and the \( \epsilon \) wolf represent the first dominant node the second dominant node, and the third dominant node in the wireless sensor network. They are at the
top of the population, and have the best decision-making ability compared to the common
wolves. Other wolves must obey the instructions of the highest priority first dominant node
compared to the three highest priority sensor nodes. According to the existing algorithm, the
GWO algorithm is mainly divided into two main processes of prey and hunting [31-33]. The
distance between the prey and the wolves needs to be ascertained in the enveloping process as
follows:

\[ \vec{D} = \left| C \vec{X}_p^t - \vec{X}_t^t \right| \quad (9) \]

where \( \vec{X}_p^t \) is the position of the prey when iterating to \( t^{th} \), \( \vec{X}_t^t \) is the position of the gray wolf
at this moment, \( C \) is a constant factor, so the following formula.

\[ C = 2 \ast r_1, r_1 \in [0, 2] \quad (10) \]

Then according to the surrounding prey information, we can calculate the specific
direction of the wolves in the next step, then the next round of wolves is:

\[ \vec{X}_{t+1} = \vec{X}_t - S \ast \vec{D} \quad (11) \]

in which \( S \) is the convergence factor expressed by the following formula:

\[ S = 2 \ast s \ast r_2 - s, r_2 \in [0, 2] \quad (12) \]

The hunting process in the traditional algorithm will directly determine the final position
of the three wolves with the highest priority, and the expression will be brought into the
expression of Eq. 11 as:

\[ \vec{X}_{a+1} = \vec{X}_a - S \ast D_a \quad (13) \]

\[ \vec{X}_{b+1} = \vec{X}_b - S \ast D_b \quad (14) \]

\[ \vec{X}_{c+1} = \vec{X}_c - S \ast D_c \quad (15) \]

Then the best cluster head position is available:

\[ \vec{X}_{p+1} = \frac{\vec{X}_{a+1} + \vec{X}_{b+1} + \vec{X}_{c+1}}{3} \quad (16) \]

3.2.1 Ray wolf fitness function model based on energy and distance

According to the energy consumption model established in Section 2, When the distance
between the node and the base station is greater, the energy consumed by the node will also be
greater, and whether the current node has the opportunity to become a cluster head is based on
the residual energy of the current node. Selecting nodes with high residual energy helps to
improve the survival time of the network and prevent the occurrence of energy holes. Therefore,
this paper designs an adaptive value function based on the residual energy of the
node and the communication distance between the node and the BS as follows:

\[ F = \begin{cases} \gamma \ast \frac{E_i}{E_{avg}} + \lambda \ast \frac{D_i}{D_{avg}} & E_i \neq 0 \\ 0 & E_i \leq 0 \end{cases} \quad (17) \]
In the above formula, both $\gamma$ and $\lambda$ are influence factors, ignoring other losses in the transmission process. It is assumed that the cluster head is only affected by both distance and energy, and the sum of the influence factors is 1. $E_i$ is the current node energy, $E_r$ is the total energy in the current cluster, $D_i$ is the distance from the current node to the base station, and $D_{avg}$ is the average distance from the node in the current cluster to the base station. At the same time, in order to emphasize the position of the first dominant node, the second dominant node and the third dominant node of the wireless sensor network, the position of the cluster head node in the whole heterogeneous network is determined, and the optimal weighting factor is re-set, and the improved most The location of the cluster head is as follows:

$$E_i = \sum_{i=0}^{n} E_i \quad (E_i \geq 0)$$

$$F_a = \frac{F_a}{F_a + F^p + F^\delta}$$
$$F^p = \frac{F^p}{F_a + F^p + F^\delta}$$
$$F^\delta = \frac{F^\delta}{F_a + F^p + F^\delta}$$

3.2.2 An improved convergence factor adaptive adjustment strategy

According to the design principle of the algorithm, in the early iteration, the convergence speed of the function is faster. As the number of iterations increases, the iteration speed of the algorithm gradually decreases. It can be seen from the above evidence that the iteration factor $s$ plays a key role in the degree of convergence of the algorithm. In this paper, the nonlinear adjustment strategy of a cosine function is used to extend the original $s \in [0, 2]$ interval to the traditional GWO algorithm based on the nonlinear reduction of the cosine function on $[0, \pi / 2]$:

$$s = 2 \cos\left(\frac{t}{t_{max}} \cdot \frac{\pi}{2}\right)$$

Where $t_{max}$ is the maximum number of iterations and $t$ is the current number of iterations. The fitness value of each wireless sensor network node in each cluster can be found from the adaptation value in Eq.15. Comparing each adaptation value $F_i$ with the average fitness value, if the self-adaptation value is higher than the average fitness value, then the individual should be adjusted to the direction of the prey in the hunt. Otherwise, the influence range of the control parameters should be expanded to enhance the global search interval as shown below.

$$s = \begin{cases} 
2 \cos\left(\frac{t}{t_{max}} \cdot \frac{\pi}{2}\right) & F_i \geq F_{Avg} \\
-2t + 2 & F_i < F_{Avg}
\end{cases}$$

The detailed pseudo code for Ad-GWO is as follows:
4 Simulation results

In order to evaluate the algorithm of this paper and the performance indicators of existing algorithms, this paper proposes a variety of simulation evaluation methods. In the simulation area of 100m*100m, 100 nodes are randomly thrown around the BS node located in the central area, and a three-level energy heterogeneous network is constructed as shown in Figure 2; Among them, 40% of the energy is deployed as the ordinary node, and the E0 level node is represented by ‘o’; A three-level energy node with 20% energy of $1/2E_0$ is deployed, denoted by ‘*’, the remaining nodes are normal secondary nodes, denoted by ‘#’, and the BS node is denoted by ‘X’. In the formulas 1 to 8, the optimal K value is derived. In this paper, the K-means algorithm is used to cluster the unordered network nodes, and the clustering strategy is adopted to help determine the cluster head node and the base station. Minimum distance between, reduce network energy consumption, and the final clustering result is represented by Tyson polygon as shown in Figure 3.

![Fig. 2. Energy dissipation model.](image-url)
In the test to simulate the effect of the AD-GWO algorithm, the initial experimental parameters determine the performance of the entire experimental results. The specific parameters are shown in the Table 2 below. Since the unreasonable energy allocation and consumption protocol accelerates the process of node death, the more energy the node stores, it means that the benefits of maximizing network survival directly reflect the survival value of the agreement.

Table 2. Key parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Font size and style</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network Size</td>
<td>{100^2}</td>
</tr>
<tr>
<td>Number of Sencer nodes</td>
<td>100</td>
</tr>
<tr>
<td>Proportion of CHs</td>
<td>Table 1</td>
</tr>
<tr>
<td>Data Aggregation Energy Cost</td>
<td>(E_{DA}=50) nJ/bit</td>
</tr>
<tr>
<td>Transmitter/Receive</td>
<td>(E_{d/re}=50) nJ/bit</td>
</tr>
<tr>
<td>Packet Size</td>
<td>4000 bits</td>
</tr>
<tr>
<td>Transmitter Amplifier</td>
<td>0.0013 pJ/(bit\cdot m^4)</td>
</tr>
<tr>
<td>Transmitter Amplifier</td>
<td>0.0013 pJ/(bit\cdot m^4)</td>
</tr>
<tr>
<td>(\lambda)</td>
<td>0.7</td>
</tr>
</tbody>
</table>

4.1 Residual energy balance ratio(REB)

The remaining energy balance ratio(REB) is expressed as the ratio of the remaining energy of the current node to the total energy of the node, since the initial energy priorities are different, the energy loss values of the various nodes are different at different locations. Figure 4 shows that the AD-GWO algorithm proposed in this paper has four dead nodes in 1000 iterations, and the REB ratio of most nodes exceeds 50%\(E_0\). Figure 5 is the energy balance ratio of the improved heterogeneous LEACH (HLEACH-e) algorithm under the same number of iterations. In this mode, there are 59 dead nodes, and most of the REB ratio are between [40%,50%]\(E_0\). In Figure 6, the number of dead nodes based on the improved heterogeneous
SPE algorithm is about 11, and the REB ratio is between [20%,60%] \( E_0 \); In Figure 7, the energy nodes of the heterogeneous Fuzzy C-means clustering DEEC(FCM-DEEC) die are about 11, but most of the energy nodes of the remaining nodes are distributed between [20%,40%]\( E_0 \); And a considerable number of nodes are dying, so it can be clearly seen from the above data that the proposed adaptive gray wolf algorithm has significant effects on solving the energy balance problem in heterogeneous wireless sensor networks.

**Fig. 4.** AD-HGWO 1000 generation energy balance ratio.

**Fig. 5.** HLEACH-e 1000 generation energy balance ratio.
4.2 Node life cycle

In the heterogeneous wireless sensor network, due to the large difference of the three levels of energy, the improved LEACH algorithm highlights the serious first node death based on the probability disadvantage; Although the improved FCM-HDEEC algorithm makes up for this deficiency, the problem of cluster head selection does not essentially solve the inherent defects in the algorithm process; The SPE algorithm itself adopts two
nodes with different initial energy, and designs different cluster head election thresholds, which further increases the probability that the advanced node becomes the cluster head, and improves the death time of the first node; However, this approach does not take into account the negative impact of the cluster head node and base station distance. The probability of death of the node in a certain period of time is further increased, and the lifting effect is not obtained. The simulation results are shown in Figure 8:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Initial Node Died</th>
<th>Half Node Died</th>
<th>All Node Died</th>
</tr>
</thead>
<tbody>
<tr>
<td>HLEACH-e</td>
<td>386</td>
<td>936</td>
<td>2131</td>
</tr>
<tr>
<td>SEP</td>
<td>813</td>
<td>1086</td>
<td>1352</td>
</tr>
<tr>
<td>FCM-DEEC</td>
<td>820</td>
<td>1515</td>
<td>2805</td>
</tr>
<tr>
<td>AD-HGWO</td>
<td>500</td>
<td>1995</td>
<td>3972</td>
</tr>
</tbody>
</table>

Table 3. The survival time corresponding to different K values

![Figure 8](image)

**Fig.8. Stability period with respect to number of rounds**

In Table 3, the initial node death algebra, 50% node death algebra, and all node death algebras of various algorithms are therefore, this paper has profound significance and influence in improving the network life cycle.

4.2 Number of data packets received by the base station

The more data received by the base station, the longer the survival time of the network, which means that the amount of information monitored by the network will be larger in complex environments. This indicator directly reflects the effectiveness of the proposed algorithm. As shown in figure 9, the data in this paper is about 36679, which is 65%, 49%, and 50% improved by HLEACH-e algorithm, FCM-DEEC algorithm, and SPE algorithm respectively.
The average node residual energy is the ratio between the total energy of the current node and the number of current nodes. When the average node residual energy means the higher the survival value of the network, the greater the use value of the network. In the algorithm proposed in this paper, the slope is the smallest compared to the other three algorithms, which means that the energy loss in the same network living space is slower. It is about 33% higher than the HLEACH-e algorithm and FCM-DEEC algorithm, and nearly 50% higher than the SPE algorithm. The method proposed in this paper saves the total energy consumption of the network and improves the energy utilization of a single node.
5 Conclusion

This paper proposes an adaptive gray wolf algorithm for the three-level energy heterogeneous wireless sensor network, which improves the network life cycle and improves the network residual energy usage rate. Firstly, a reasonable clustering structure is set by the energy consumption model of the wireless sensor network. The K-means algorithm is used to ensure that the nodes in each cluster are in a reasonable range, and the irrelevant energy loss of the cluster members in selecting the unreasonable cluster head is avoided. Secondly, the fitness function is established by combining the residual energy of the network node with the distance from the node to the base station. Make full use of the relationship between node energy and position in the information transfer process, combined with the logical composition of the gray wolf algorithm, select the three most adaptable nodes in each cluster, and iteratively select the cluster head nodes that best reflect the current cluster structure. Complete the information of the clustered transfer sensor; finally, simulate the real network environment by using various evaluation criteria. The experimental results show that the model has a good network life cycle and has better adaptability than the traditional homogeneous and heterogeneous network models. In addition, although the proposed algorithm improves the life cycle of heterogeneous WSN networks, it increases the computational complexity of the algorithm. The base station and the network node of the algorithm are not mobile, and the delay and jitter in the information transmission process are not considered. Therefore, future work can further consider the scope of adaptation of the algorithm.

References


Automatic Modulation Classification Using Hybrid Convolutional Neural Network

1st Xiaofei Zhang¹, 2nd Anping Li¹, 3rd Lu Chai², 4th Xiaoying Ma¹, 5th Meiyiing Wei¹
{zhangxf@srrc.org.cn, lianping@srrc.org.cn¹, chaibyr@bupt.edu.cn², maxiaoying@srrc.org.cn¹, weimeiyiing@srrc.org.cn¹}

The state radio monitoring center¹,
Key Laboratory of Universal Wireless Communications, Ministry of Education,
Beijing University of Posts and Telecommunications, Beijing, P.R.China, 100876²

Abstract. Automatic modulation classification (AMC) plays an essential role in signal demodulation and interference identification. In this paper, we propose a novel AMC method using the Hybrid Convolutional Neural Network (HCNN), which combines with two different convolutional neural networks (CNNs) jointly using various signal features. In the former CNN, spectral correlation features (SCFs) are generated as network input, to classify FSK and BPSK. In the latter CNN, the Attention-based Densely Convolutional Neural Network (AD-CNN), which is trained using regular constellation images (RCs), is proposed to identify the modulation formats that are hardly recognized by the former CNN, such as QPSK, 16-QAM and 64-QAM. The simulation results demonstrate that HCNN displays superior classification performance than existing AMC methods with lower computational complexity.

Keywords: AMC, Deep convolutional neural network, Hybrid convolutional neural networks, Spectral correlation features, Regular constellation images.

1 Introduction

With the rapid development of wireless communications, new paradigms of telecommunications emerge currently [1]. However, the shortage of wireless spectrum is becoming an outstanding problem. To monitor spectrum resource tension and increase spectrum efficiency, the radio spectral sensing is introduced in commercial and civil applications [2]. With the task of ensuring the proper functioning of radio spectral sensing, automatic modulation classification (AMC) is utilized essentially as the link between signal receiving and demodulation to identify the signal format corrupted by noise [3].

Generally, AMC can be divided into two methodologies, i.e., likelihood-based methods (LB) and feature-based methods (FB) [4]. LB methods have a better integrated performance of precision. However, LB methods need prior knowledge of statistical information of the signals, which increases the computational complexity in practice [5]. Comparatively, FB methods deal with modulation classification straightforwardly without prior information [6]. Various features are widely utilized in FB methods including high-order statistics [7], cyclostationary spectrum, etc. Kim et al. implemented the maximum of cyclostationary spectrum over frequency to reduce the computational complexity [8]. However, cyclostationary-based AMC lacks the ability to
deal with the classification between M-ary phase shift keying (M-PSK) and M-ary quadrature amplitude modulation (M-QAM). Huang et al. proposed grid constellation matrix and fully convolutional network for classification through a short training process. However, it is limited to classify Frequency-shift keying (FSK) using this method [9]. On this basis, this paper focuses on improving classification accuracy of FB methods.

The convolutional neural network with strong representative ability can be applied in AMC for higher classification accuracy [10]. Based on this, the authors proposed a low complexity deep neural network (DNN) to identify the signal modulation formats [11], but it hardly extracts high-dimensional features since the lack of the convolutional layers. In [12], convolutional neural network is utilized in AMC to extract features from constellation images. However, CNN with constellation images have limitations when the classification is conducted among FSK and M-QAM.

In this paper, a novel AMC method named Hybrid Convolutional Neural Network (HCNN), is proposed involving Deep Convolutional Neural Network (DCNN) and Attention-based Densely Convolutional Neural Network (AD-CNN). HCNN can effectively classify seven widely-used modulation formats with high classification accuracy and low computational complexity. Several contributions of this paper are listed as follows.

1. Considering the various of signal features, the signal spectral correlation features (SCFs) and regular constellation images (RCs) are represented as the inputs of HCNN respectively. With SCFs and RCs, signal high-dimensional features are considered sufficiently.
2. Deeper network structure and dropout layers, which can extract high-dimensional features directly and avoid overfitting, are exploited to obtain better performance in DCNN.
3. AD-CNN combines the densely convolutional neural network and the attention block. It enhances the ability of feature extraction ulteriorly and reduces the network parameters simultaneously.
4. Concretely, DCNN provides the classification ability among 2FSK, 4FSK, MSK and BPSK. Meanwhile, AD-CNN is utilized to improve the classification accuracy among QPSK, 16-QAM and 64-QAM. Simulation results show that higher classification accuracy is achieved comparing to existing AMC methods with lower computational complexity.

The remainder of this paper is organized as follows. Section II provides an overview of the signal model and the proposed SCFs and RCs. In Section III, we explain the network structure of HCNN. Section V exhibits simulation results and conclusion of this paper is given in Section VI.

2 Signal model and features extraction

2.1 Signal model

Assume that radio frequency signals are received by the single-antenna receiver. The signal transformed to baseband can be formatted as follows,

\[ y(n) = h e^{j(2\pi f_c n + \omega_t)} x(n) + w(n), n = 1, \ldots, N \]  (1)
where \( h \) denotes Rayleigh fading channel coefficients, \( f_0 \) and \( \theta_0 \) denote the frequency and phase offset, respectively. \( N \) denotes the signal sample length. \( x(n) \) denotes the \( n \)-th complex sample in received segment signal. \( w(n) \) denotes the additive white gaussian noise (AWGN) with mean zero and variance \( \sigma^2 \). Meanwhile, the received signal-to-noise ratio (SNR) is defined as \( h^2/\sigma^2 \).

### 2.2 Spectral correlation features

Cyclostationary signal \( x(t) \) is a kind of random signal whose statistical characteristic parameters change periodically, i.e.,

\[
R_x(t, \tau) = E\{x(t)x^*(t + \tau)\}
\]  

(2)

where its autocorrelation coefficient \( R_x(t, \tau) \) and mean value \( E\{x(t)\} \) both change with \( t \) in a period \( T \) for any value of \( \tau \). Compute the Fourier series of \( R_x(t, \tau) \), then we get the cyclic autocorrelation coefficient. Considering the signal cyclic ergodicity, can be formatted as follows.

\[
R_x^\alpha(\tau) = \lim_{T \to \infty} \int_{-T/2}^{T/2} x(t + \frac{\tau}{2})x^*(t - \frac{\tau}{2})e^{-j2\pi \alpha t} dt
\]

(3)

To derive a proper signal processing and surpass the influence of AWGN, we obtain the signal cyclic spectrum, which is the Fourier series of \( R_x^\alpha(\tau) \) and can be defined as follows,

\[
S^\alpha_P(t, f) = \frac{1}{T}X_T(t, f + \frac{\alpha}{2})X_T^*(t, f - \frac{\alpha}{2})
\]

(4)

where \( T \) denotes the signal segment length, \( \alpha \) denotes the cyclic frequency, and \( X_T \) denotes finite time Fourier transform.

In a real scenario, signal samples are always be divided into \( K \). Then we average the cyclic spectrum of each segment, which can be formatted as follows.

\[
S^\alpha_P(t, f) = \frac{1}{K} \sum_{k=1}^{K} S^\alpha_P(t, f)
\]

(5)

Considering various modulation formats have various signal cyclic spectrums, signal cyclic spectrum can be exploited for AMC. By normalizing the 3-dimensional cyclic spectrum along the \( Z \) axis, the 3-dimensional cyclic spectrum can be normalized to the 2-dimensional cyclic spectrum in the \( X-Y \) plane. Therefore, we extract the SCFs from signal cyclic spectrum, which are related to the carrier frequency \( f \) and symbol rate \( R_s \). Some examples of SCFs under different modulation formats are presented in Figure 1. The figure shows that SCFs among 2FSK, 4FSK, MSK, BPSK have distinct patterns, while QPSK, 16-QAM, 64-QAM represented in the same form as SCFs.
Generally, SCFs have two advantages. Firstly, different SCFs are generated by different modulation formats so as to classify modulation. Secondly, SCFs have the robustness to AWGN, which could be exploited for modulation classification appropriately at low SNR.

![Spectral Correlation Features](image1.png)

**Fig. 1.** Spectral Correlation Features for seven modulation formats.

### 2.3 Constellation Features

To further classify modulation formats of QPSK, 16-QAM, 64-QAM, received complex signal symbols are transformed into RCs as the input of the network. We plot the real and imaginary parts of the signals by transforming the signal samples into 2-dimensional scatter diagrams as RCs. Practically, RCs are widely used to learn effective representations from signals.

A generation diagram is conducted in **Figure 2** to illustrate RCs of three widely-used modulation formats (i.e. QPSK, 16-QAM, 64-QAM) at different SNRs. For different modulation formats under the same SNR, RCs are used to represent signal minutiae features. Generally, RCs provide a relatively new approach to extend its representation discrimination.

![Regular Constellation Images](image2.png)

**Fig. 2.** Regular Constellation Images for three modulation formats under different SNRs.
3 HCNN based classification method

3.1 Structure of HCNN

Our proposed HCNN consists of two CNNs, so as to classify seven modulation formats of BPSK, QPSK, 2FSK, 4FSK, MSK, 16-QAM, 64-QAM. As shown in Figure 3, the former CNN, DCNN, is designed to classify modulation formats except QPSK, 16-QAM, 64-QAM, since the characterization limitation of SCFs. The latter CNN, AD-CNN, trained using signal RCs, is designed to classify QPSK, 16-QAM, 64-QAM.

![Figure 3](image-url)

Fig. 3. The structure of HCNN.

3.2 Structure of DCNN

DCNN includes three convolutional blocks, one fusion block and three fully-connected blocks. Each convolutional block consists of four convolutional layers followed by batch normalization. In first three convolutional layers, convolutional kernel size decrease to 3×3 and stride step set as 1 in order to realize sparse connectivity between layers. Moreover, ReLU activation function is utilized to increase sparsity and nonlinearity. Specially, the fourth layer, whose kernel size is 2 × 2 and stride step is 2, replaces max-pooling layer to prevent feature loss in down-sampling. Besides, the channels are incremental among convolutional blocks, intending to get sufficient specific receptive fields. The fusion block unifies the form of feature representations for the fully-connected blocks by convolutional layer and fusion layer. In fully-connected blocks, there are 3 fully connected layers whose channel numbers are 256, 64, 4, respectively, reducing output channels and improving computational efficiency. The dropout is set as 0.6 to solve overfitting problem [14] and increase robustness to model mismatches. Additionally, the output channel number of the last hidden layer is limited to 2, which indicates that the dimension of the deep feature is reduced to 2 and beneficial to two-dimensional visualization. Finally, with Softmax [15], the output layer could sum all neurons to one and map multiple scalars to probability distributions.
3.3 Structure of AD-CNN

AD-CNN uses RCs to solve the classification among QPSK, 16-QAM and 64-QAM. Considering the similarity of RCs between 16-QAM and 64-QAM in low SNRs, AD-CNN is involved essentially to optimize the CNN structure so as to increase classification accuracy. As shown in Figure 5, AD-CNN consists of four parts, attention blocks, convolutional layers, max-pooling layers and fully-connected blocks.

Initially, attention block is composed of dense convolutional block and attention layer. For dense convolutional block, it has direct connections among nonadjacent layers, which is inspired by ResNet [16] and shown in Figure 6. To improve the feature utilization, each layer in the dense convolutional block receives the composition of preceding layers’ output, meanwhile, as the input for subsequent layers. Incorporating both accuracy and efficiency, dense convolutional block brings two advantages. Firstly, it solves the problem of network gradient disappearance.
and explosion. Secondly, it reduces the parameters that high-dimensional features can be extracted with low computational complexity.

Assume the network includes $L$ layers, on which adopt a non-linear transformation $H_l(\cdot)$. The output of the $l-th$ layer can be denoted as $x_l$. Therefore, the input of $l-th$ layer, which contains the features from all the previous layers, can be denoted as follows,

$$x_l = H_l([x_0, x_1, \ldots, x_{l-1}])$$  \hspace{1cm} (6)

where $x_0, x_1, \ldots, x_l$ refer to the connection of the function diagram generated by layers.

Besides, attention layer is added after every two $1 \times 1$ convolution layer and $3 \times 3$ convolution layer in dense convolutional block, which is pictured in Figure 7. For attention layer, it could recalibrate the feature channel weights by calculating the interdependence among channels and improve significant performance of the most advanced architecture. As shown in Figure 8, there are four steps for the attention layers to recalibrate the features. Firstly, in order to construct the interdependence among channels, we convert the feature map $U \in \mathbb{R}^{H	imes W \times C}$ on the spatial dimension $H \times W$ to the channel descriptor. Next, the feature map $U$ is shaped to $1 \times 1 \times C$ by a global average pooling layer. Thirdly, the gating mechanism composed of full
connection layers and sigmoid activation function is used to learn the nonlinear interaction among multiple channels. The feature map is shaped to $1 \times 1 \times C$ and then decoded back to $1 \times 1 \times C$ as the reweighed vector $V_i$ with the gating mechanism. Lastly, the output of attention block $Y$ can be formatted as follows,

$$Y = \sum U_i \cdot V_i$$  \hspace{1cm} (7)

where $U_i$ is the $U$ of each channel and multiplied with reweighed factor $V_i$ in vector $V$.

Generally, several optimizations of AD-CNN are listed as follows.

(1) In AD-CNN, feature reuse and recalibration are considered simultaneously. On the one hand, dense convolutional block achieves feature reuse and reduces the number of parameters remarkably. On the other hand, attention layer linked with convolution could reweight calibration for each feature channel. It helps to extract effective features and improve the classification accuracy.

(2) It is creative to make some optimization in the structure of AD-CNN so as to have an combination of classification accuracy and computational complexity. For instance, max-pooling layers in dense convolutional block are removed to protect low-level features. $1 \times 1$ convolutional layers are involved to protect global features and reduce feature redundancy.

4 Simulations

In this section, simulations are operated to illustrate the superiority of HCNN. The signal candidate modulation format set is $M = \{2\text{FSK}, 4\text{FSK}, \text{MSK}, \text{BPSK}, \text{QPSK}, 16 - \text{QAM}, 64 - \text{QAM}\}$. To reduce the computational complexity and improve the generalization ability, the normalized pixel size of SCFs and RCs are down sampled to $80 \times 80$. SCFs and RCs are both obtained by received signals with different pre-defined sample lengths, i.e., 1024, 2048, 4096, 9600. In the training stage, 40000 SCFs and RCs are involved respectively for each modulation at a given SNR as training data. In the classification stage, 10000 SCFs and RCs in different assumed conditions are used to calculate the classification accuracy $P_{ce}$.

In Figure 9, we compare the performance of HCNN under different settings of sample lengths. For same sample length, the HCNN performs consistently within the range of SNR, proving the practicability of HCNN. Moreover, better classification performance by employing more signal samples, proving the asymptotic behavior of HCNN.
Figure 9 is an overall methods comparison among HCNN proposed, HCNN with both DCNN, CNN with RCs [6] and CNN with SCFs [8] under different SNRs, and the sample length is set as 4096. The result verifies that proposed HCNN outperforms the others for the entire SNR range. Specifically, HCNN proposed yields 3 dB gains over HCNN with both DCNN at 95% $P_{cc}$, illustrating the superiority of the HCNN. However, the accuracy is only up to 0.63 when using a single kind of signal feature, indicating the wide range of classification modulation formats of HCNN. The reasons are from two aspects. Initially, AD-CNN in HCNN can extract more discriminative features from input features. Additionally, it is illustrated that single-feature limits the types of classification modulation. More importantly, the per epoch training time of proposed HCNN and HCNN with both DCNN are 90 seconds (s), 710s, respectively, using the same dataset and equipment (a GTX1080 GPU). It shows the lower computational complexity of HCNN.
5 Conclusion

In this paper, a novel AMC method named HCNN is proposed. First of all, the HCNN combines with DCNN and AD-CNN to make classification decisions of seven modulation formats with high accuracy. DCNN uses deep convolutional neural network and replaces the pooling operation by convolutional layer to reserve discriminative features. Especially, AD-CNN is designed to consider feature reuse and recalibration are simultaneously, in order to improve performance and reduce computational complexity. Besides, SCFs and RCs are conducted respectively to expand the modulation scopes of classification. Simulation results verify the accuracy superiority and low computational complexity of HCNN.

Acknowledgments.

This work was supported in part by the National Key Research and Development Program of China under Grant (2018YFF0301202, 2019YFB1804400), in part by the National Natural Science Foundation of China under Grant (61801052, 61227801), and in part by the Beijing Natural Science Foundation under Grants (4202046).

References


Optimization research of radio monitoring network for the Beijing 2022 Olympic Winter Games

Meiying Wei, Xiaofei Zhang, Xu Yang, Xiaoying Ma*
{weimeiyiing, zhangxf, yangxu, maxiaoying}@srrc.org.cn

The State Radio Monitoring Center, Beijing, P.R.China, 100037

Abstract. The 24th Olympic Winter Games will be held at Beijing and Zhangjiakou in 2022. The radio security is one of the most essential work to make the Olympic Games and ceremonies go smoothly. The planning of the radio monitoring network is the basis of the radio security. This article introduced a coverage capability assess method for the radio monitoring network. A optimization scheme was brought out considering the assessment results, the local geography and construction, using the Yanqing division as an example. With this scheme, the coverage capability of the monitoring network in Yanqing division will be optimised elaborately.

Keywords: radio security, radio monitoring, Olympic Winter Games , network optimise

1 Introduction

The Beijing 2022 Olympic Winter Games are planning to use 26 venues, including 13 competition venues and 13 non-competition venues. These venues are distributed in three regions: Beijing, Yanqing and Zhangjiakou[1]. As shown in figure 1, the venues are very dispersed and most terrains are mountainous and complex. This situation will bring great challenges to radio security work for the Beijing 2022 Olympic Winter Games.
In recent years, the radio monitoring networks in Beijing and Hebei province are developed rapidly. Their design concepts and construction scale are both in the leading level in our country. But in Yanqing and Zhangjiakou area, the coverage of radio monitoring stations is too small to satisfy the requirements for radio security for the Olympic Games. Thus, how to optimise the radio monitoring network is the most important work for the responsible departments.

Firstly, this article introduced a method to do the radio monitoring capacity evaluation with the ITU-R P.1546 model and typical radio stations used in the Olympic Games. Secondly, it evaluated the coverage capability of the existing radio monitoring stations for Yanqing competition division. Finally, it brought a scheme to optimise the monitor network in Yanqing using multi-type monitoring stations including fixed stations and portable/mobile stations.
stations to improve the monitoring capability, connecting with the regional geographical environment and construction condition. Through the study of this paper, the author hopes to provide theoretical basis and data support for the site selection and layout adjustment of radio monitoring network for Beijing 2022 Olympic Winter Games.

2 The coverage capability assess method for the radio monitoring network

The coverage capability evaluation of the radio monitoring network needs to define the following factors: the suitable propagation model for the radio waves, the transmitting parameters of radio stations and the receiving parameters of the monitoring systems.

The main transmitting parameters including frequency, power, signal bandwidth and antenna height. These parameters mainly decided by the service the stations designed for. As to do the evaluation work suitable to the Beijing 2022 Olympic Winter Games, typical service radio stations should be selected according to the requirements of the Games. The receiving parameters of monitoring systems mainly include the location of the monitoring system, the height of the receiving antenna, the system receiving sensitivity.

2.1 The radio propagation model in ITU-R P.1546-4 Proposal

In 2001, the international telecommunication union (ITU) put forward the ITU-R p.1546 recommendation, named "the method of face to face prediction of ground service points in the frequency range of 30MHz to 3000MHz". In this proposal, the prediction method for UHF/VHF band field intensity are given on the basis of the measured data at three transmitting frequency: 100 MHz, 600 MHz and 2000 MHz. Several probability curves of wave propagation are given when the effective transmission power is 1 kiloWatts in different propagation paths, different transmitting/receiving antenna heights and different time and places. The method and correction formula for predicting the field intensity under different transmission frequency and power are given.

This method is suitable for broadcasting, mobile and some fixed services in the UHF/VHF band at the distance between 1~1000km. The proposal is also the most commonly used model for electromagnetic compatibility analysis and international frequency coordination [2]. This paper used this propagation model to simulate and evaluate the regional coverage capability of the radio monitoring station.

2.2 Selections of typical radio stations

This paper selected six kinds of typical radio service stations used in the Olympic Games for evaluation: FM broadcasting, radio interphone relay (150MHz), digital trunked base station (350MHz), dedicated interphone (400MHz), digital TV broadcasting and digital trunked base station (800MHz) [3]. The transmitting parameters which are get from practical applications can be seen in the following table:
Table 1. Transmitting parameters for typical radio service stations

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Frequency</th>
<th>Power</th>
<th>Antenna Gain</th>
<th>Bandwidth</th>
<th>Antenna height</th>
</tr>
</thead>
<tbody>
<tr>
<td>FM broadcasting station</td>
<td>100MHz</td>
<td>10kW</td>
<td>10dBi</td>
<td>200 kHz</td>
<td>70m</td>
</tr>
<tr>
<td>Radio interphone relay</td>
<td>150MHz</td>
<td>25W</td>
<td>5dBi</td>
<td>12.5kHz</td>
<td>50m</td>
</tr>
<tr>
<td>Digital trunked base station</td>
<td>350MHz</td>
<td>25W</td>
<td>5dBi</td>
<td>12.5kHz</td>
<td>50m</td>
</tr>
<tr>
<td>Dedicated interphone</td>
<td>410MHz</td>
<td>3W</td>
<td>0dBi</td>
<td>12.5kHz</td>
<td>1.5m</td>
</tr>
<tr>
<td>Digital TV broadcasting station</td>
<td>700MHz</td>
<td>1kW</td>
<td>10dBi</td>
<td>8MHz</td>
<td>70m</td>
</tr>
<tr>
<td>Digital trunked base station</td>
<td>860MHz</td>
<td>25W</td>
<td>5dBi</td>
<td>12.5kHz</td>
<td>50m</td>
</tr>
</tbody>
</table>

2.3 Parameters of monitoring receivers

The location and height of the receiving antennas can be get from the existing monitoring network. The antenna height is set to be 10m according to experience. The system receiving sensitivity can be calculated as following:

\[
RS = -174 + NF + 10\log B(\text{Hz}) + S/N + Lf - Ga
\]

Where:
- \( RS \) is the system receiving sensitivity(dBm).
- \( NF \) is the noise factor of the receiver.
- \( B \) is the processing bandwidth of the receiver(Hz).
- \( S/N \) is the ratio of signal power to noise power.
- \( Lf \) is the feeder loss of the receiving system(dB).
- \( Ga \) is the antenna gain of the receiving system(dB).

According to empirical values and the existing monitoring network, these parameters are set as following: \( NF = 12\text{dB} \), \( B = 10\text{kHz} \), \( S/N = 10 \), \( Lf = 5\text{dB} \), \( Ga = 0\text{dB} \), so we can get that \( RS = -107 \text{ dBm} \).

3. The Coverage capability evaluation of the monitoring stations

The coverage capability of the monitoring stations in Yanqing division is evaluated in this part as an example. Nearby the Yanqing division, there are three radio monitoring stations: Yanqing station, Changping station and Mangshan station. Their system parameters and the assessment method mentioned in part 2 are used in the assessment. The evaluation results can be seen in the following figures:
FM broadcasting (10kW)
Aim area: 7.9642 sq.km.
Covered area: 7.9642 sq.km.
Coverage rate: 100%

150MHz Radio interphone relay (25W)
Aim area: 7.9642 sq.km.
Covered area: 0.5950 sq.km
Coverage rate: 7.4709%

350MHz Digital trunked base station (25W)
Aim area: 7.9642 sq.km.
Covered area: 0.0000 sq.km.
Coverage rate: 0.0000%

400MHz Dedicated interphone (3W)
Aim area: 7.9642 sq.km.
Covered area: 0.0000 sq.km.
Coverage rate: 0.0000%
As the results showed above, the radio monitoring network in Yanqing can only be used to monitor FM broadcasting stations which have large transmit power and low frequency, and also the coverage rate is only 49.2433%. Its capability for monitoring the other five kinds of radio services is basically blank. It can draw a conclusion that the monitoring capability is fall far short of what is required to do the radio security work in major regional Yanqing for the Olympic Games.
4 Optimization scheme for the radio monitoring network

4.1 Layout optimization of fixed or large radio monitoring stations

In order to maximize the use of existing monitoring resources, this paper brought the scheme which appropriately increased the monitoring station facilities, with the complete integration of the old and new systems, to protect radio safety for the Beijing Olympic Winter Games.

First, two virtual fixed radio monitoring stations located at the top of Haituo Mountain and Yanqing Olympic village respectively were added for coverage capacity evaluation. The distribution diagram of the two new virtual fixed radio monitoring stations is shown in Figure 3.

![Fig. 3. Distribution map of new virtual fixed radio monitoring stations in Yanqing.](image)

The evaluate results comparison before and after the addition of the two new virtual fixed radio monitoring stations can be seen in the following figures:
FM radio broadcast (10kW) - before
Area: 7.9642 square kilometers
Covered area: 7.9642 square kilometers
Covered volume ratio: 100%

FM radio broadcast (10kW) - after
Area: 7.9642 square kilometers
Covered area: 7.9642 square kilometers
Covered volume ratio: 100%

150MHz Radio interphone relay (25W) - before
Area: 7.9642 square kilometers
Covered area: 0.5950 square kilometers
Covered volume ratio: 7.4709%

150MHz Radio interphone relay (25W) - after
Area: 7.9642 square kilometers
Covered area: 7.9642 square kilometers
Covered volume ratio: 100%
350MHz Digital trunked base station (25W) - before
Area: 7.9642 square kilometers
Covered area: 0.0000 square kilometers
Covered volume ratio: 0.0000

350MHz digital trunked base station (25W) - after
Area: 7.9642 square kilometers
Covered area: 7.9642 square kilometers
Covered volume ratio: 100%

400MHz Dedicated interphone (3W) - before
Area: 7.9642 square kilometers
Covered area: 0.0000 square kilometers
Covered volume ratio: 0.0000

400MHz Dedicated interphone (3W) - after
Area: 7.9642 square kilometers
Covered area: 3.2862 square kilometers
Covered volume ratio: 41.2629%
Fig. 3. Comparison of coverage simulation results before and after adding these virtual fixed monitoring stations in Yanqing.

Table 3. Coverage rate of the newly added virtual radio monitoring stations and comparison between before and after the addition.

<table>
<thead>
<tr>
<th>Parameter Business</th>
<th>Frequency</th>
<th>Power</th>
<th>Haituo Mountain station</th>
<th>Yanqing Olympic village station</th>
<th>Overall coverage before</th>
<th>Overall coverage after</th>
</tr>
</thead>
<tbody>
<tr>
<td>FM broadcasting station</td>
<td>100MHz</td>
<td>10kW</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
The simulation results showing that after the addition of the two new stations, except for the dedicated interphone, the coverage for other typical services are 100%. Compared with before, the overall coverage has been greatly improved. It can be seen that the addition of two new virtual fixed stations in the Yanqing Division can basically guarantee that the radio monitoring coverage of key protection areas is able to reach 100% for several high-power typical service stations such as FM, digital trunked communication, and digital television, but the coverage for low power stations is still poor.

### 4.2 Layout optimization of portable/small radio monitoring station

In order to increase the coverage of low power stations and realize the multi-station joint coverage for important protection areas, it is necessary to add portable/small monitoring stations near the National Alpine Ski Centre and the National Sliding Centre.

Considering the large area of the open field and the complex terrain of the mountains, it is planned to set up a station at the high point near the end of the giant slalom of the National Alpine Ski Centre for covering the Haituo mountain and both sides of the valley. The other station can be located in the National Sliding Centre to monitor the electromagnetic environment around. The distribution diagram of the two new virtual radio monitoring stations is shown in Figure 4.
Fig. 4. Distribution map of the two new virtual radio monitoring stations in Yanqing.

After adding the two virtual portable/small stations, the monitoring coverage of 400MHz dedicated intercom service was recalculated, and the result is compared with the coverage only adding virtual fixed stations, as shown in Figure 5.

<table>
<thead>
<tr>
<th>400MHz Dedicated interphone (3W) with adding two virtual fixed stations.</th>
<th>400MHz Dedicated interphone (3W) with adding the virtual fixed stations and small stations.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area: 7.9642 square kilometers</td>
<td>Area: 7.9642 square kilometers</td>
</tr>
<tr>
<td>Covered area: 3.2862 square kilometers</td>
<td>Covered area: 7.9642 square kilometers</td>
</tr>
<tr>
<td>Covered volume ratio: 41.2629%</td>
<td>Covered volume ratio: 100%</td>
</tr>
</tbody>
</table>

Fig. 5. Comparison of coverage simulation results after adding two virtual small monitoring stations in Yanqing.

It can be seen that through adding the 2 fixed radio monitoring stations and 2 portable/small stations, the radio monitoring network in Yanqing competition area can basically achieve 100% coverage for the typical services in theory, also for low power equipments like 400MHz dedicated interphones. With this layout, the coverage capability can meet the security requirements for various radio stations in the 2022 Beijing Olympic Winter Games.

5 Conclusion

In this paper, through calculating the coverage of the radio monitoring network in Yanqing competition area, the layout optimization suggestions of the new adding radio monitoring stations are proposed based on the coverage evaluation results, so that the monitoring coverage effect of the competition area is greatly improved, and the monitoring coverage rate can reach to 100%. The full monitoring coverage of typical service stations provides theoretical and data support for the radio management department to optimize the construction of the regional radio monitoring network, and lays a certain foundation for improving the refined technical supervision ability of the radio safety guarantee of the 2022 Olympic and Paralympic Winter Games.
Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant (2018YFF0301202).

References

[1] 北京2022年冬奥会及冬残奥会场馆和交通规划，2017
Feature Fusion Convolutional Network Based
Automatic Modulation Classification

Jiajie Mai, Shanchuan Ying, Nanxin Wang, Sai Huang
{maijiajie@bupt.edu.cn, yingshanchuan@bupt.edu.cn, nancingwang@bupt.edu.cn, huangsai@bupt.edu.cn}
International School, Beijing University of Posts and Telecommunications, Beijing, China
Key Laboratory of Universal Wireless Communications, Ministry of Education

Abstract. Automatic modulation classification (AMC) technology, which utilizes to classify different kinds of signals using various expert features, typically the constellation map and the cyclic spectrum density graph, plays a significant role in spectrum monitoring and radio communication. Aiming at the commonly used features-based (FB) approach in practice, we propose a novel AMC model that jointly utilizes the constellation map and cyclic spectrum density graph as the signal features in this paper. In order to provide a solution for the model, we propose a feature-based and supervised network in Single in Single out (SISO) system to identify seven kinds of signals, which is called Feature Fusion Convolutional Network (FFCN). Simulation results are provided to show that FFCN can classify different types of signals with high probability. Typically when signal to noise ratio (SNR) is low, the success rate of the proposed network is 2% higher on average, which proves that our network has better performance.

Keywords: automatic modulation classification, residual neural network, constellation, cyclic spectrum density graph

1 Introduction

Along with the nowadays rapid development of the wireless communications, mountainous technologies like software radio and the fifth generation (5G) wireless systems arise, and radio communication is playing an increasingly essential role in the advanced area [1]. Meanwhile, spectrum monitoring, which ensures the normal operation of wireless services and strengthens the spectrum management, gets harder and the public have paid more and more attention on it [2] [3]. In the whole process of modulation format categories, automatic modulation classification (AMC) is vitally needed due to its strong ability to improve the performance of classifying signal, typically for signals corrupted by different kinds of noise and fading [4]. Not only used in modulation classification, but AMC also takes responsibility for identifying the interference signal, filtering unhealthy and dangerous information in many security applications for decades, which acts as the bridge between receive signal and signal demodulation.

Generally, there are two basic method of the AMC algorithm: likelihood-based (LB) and features-based (FB) approach [5]. Swami et al. achieved AMC using the approach of LB [6] and Ghasemzadeh et al. employed FB to recognize the modulations [7]. The LB method is used to calculate the likelihood equation of the received signal in different modulation modes.
When the channel and noise are fixed, the algorithm can maximize the probability of correct classification so it is regarded as the best classifier in the scenario. Despite of its excellent performance, it may cause the high computational complexity and poor performance when handling small-scale or sparse data. Compared to the likelihood-based approaches, the FB method extracts the signal features involving cumulant, wavelet feature, cyclostationary characteristic, etc., to classify the format with lower computational complexity. Although the performance of FB is worse than LB in many scenario, FB reduces the computational complexity and achieve the low time cost and high robustness. As a result, FB is widely implemented in practical applications.

There are many contributions on AMC using classifiers belonging to the supervised machine learning. The authors of [8] utilized Support Vector Machine (SVM) to classify the signal, while the authors of [9] employed k-NearestNeighbor (KNN). Rather than using classifiers, Kim et al. used Deep Neural Network to achieve the goal [10]. The contributions above demonstrates that the machine learning has been used widely in AMC [11] [12]. Residual neural network (ResNet) [15] is a sort of supervised machine learning and can identify patterns in data according to diverse inputs such as examples, direct experience, or instruction. In this article, we employ ResNet18 which is a ResNet whose number of layers with weights is 18. After unceasing training, ResNet is able to make better decisions in the future based on the learning experience.

Regarding to the approaches of processing received discrete data by antennas, the constellation has advantages of recognizing Quadrature Amplitude Modulation (QAM) signal while the cyclic spectrum density has better performance when recognizing Frequency Shift Keying (FSK) signal. Huang et al. employed the constellation maps [13] and Ma et al. take advantages of the cyclic spectrum density maps to process the received data [14], respectively. However, we can not know which modulation the received signal belongs to in advance, so it is necessary to find a network that has relatively great performance whatever the modulation type of the received signal is. Therefore, we need to take into account their own advantages of the two approaches of processing data and to combine them together.

In order to provide a optimal solution for automatic modulation classification, we propose a Feature Fusion Convolutional Network (FFCN) based on modified ResNet, which can apply what has been learned in the past to newly coming data using labeled examples. The main contribution of our work are summarized as follows.

1) In contrast to traditional methods, we propose a novel AMC model based on the discrete signal processed by both constellation map and cyclic spectrum density graph. We normalize the heterogenous data from the two maps into the same specification and combine it in the same dimension, which allows estimating modulation format by taking advantages of the two approaches.

2) According to the form of input, we modify Resnet18 in the proposed FFCN. Depending on the efficiency and maximum a posteriori, cross entropy is selected as our loss function to promote the classification ability of ResNet18 to the maximum extent.

3) Extensive simulations are carried out and it is proven that the proposed network is able to automatically learn more discriminative features for recognition performance.
improvement due to the advantage of powerful capability in extracting features of matrix-shaped data.

The rest of the paper is organized as follows. Section II presents the detailed introduction of the two ways to process signal data and the system model. The network structure and data combining methods are showed in Section III. Section IV indicates the simulation results and Section V draws a conclusion of this paper.

2 SYSTEM AND SIGNAL MODEL

There are two parts in this section. In the first part, the system model is outlined to show the whole process of AMC. In the latter part, we introduce the additive Gaussian noise (AWGN) channel and seven sorts of signal which would be classified.

2.1 System Model

The overall structure of signal transmission and reception is shown in Fig. 1. Firstly, different signals with different modulation format are generated in the source and transmitted to the receiver via the antenna after modulation. And in actual situation of spectrum monitoring radio detection, it usually does not require a large number of devices, so we assume that AMC is applied to a single receive, that is, the cooperation between different devices in different environments is not considered then the receiver can complete the task independently. Then, in the part of the receiver, after gaining the signal from a specified channel by the antenna, AMC technology is utilized to recognize the signal’s modulation format through two main step that are feature extraction and classification respectively. Eventually, an estimated modulation format will be given to the demodulator to obtain the exact information contained.

![Fig. 1. System model](image-url)
2.2 Signal Model

In this paper, there are seven kinds of signals would be classified: Binary Phase Shift Keying (BPSK), Quadrature Phase Shift Keying (QPSK), Binary Frequency Shift Keying (BFSK), Quadrature Frequency Shift Keying (QFSK), 16QAM, 32QAM and 64QAM. Initially, we provide the expression of baseband signals and analyze the noise. The received baseband signal can be given as follows.

\[ r(n) = h e^{j\phi(n)} x_s(n) + w(n) \]  

(1)
where \( n = 1, 2, ..., N \). And \( h \) is the channel parameter that follows Rayleigh distribution and is invariant during the identification process. \( f_0 \) and \( \hat{\phi} \) denote the frequency and phase offset. \( x(n) \) represents the \( n \)th complex symbol with unit average power, while \( w(n) \) represents the noise. As far as the noise concerned, we choose the additive Gaussian noise. The reason of it is that we prefer synthetic noise to real noise, which brings the advantage that it’s easy to analyze the problem and design the algorithm and AWGN is a simple and good approximation simulation to deal with complex situations without knowing the actual noise distribution. The based function of Gaussian Distribution is given as follows.

\[
f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)
\]  
(2)
where \(\mu\) represents the mean value, and \(\sigma\) represents the standard deviation.

The seven involved signals can be divided into three sorts: Multiple Phase Shift Keying (MPSK), Multiple Frequency Shift Keying (MFSK) and Multiple Quadrature Amplitude Modulation (MQAM). We then discuss the expressions of the three sorts of signals in discrete forms after they are received by antennas.

1) The transmission signal of MPSK modulation can be expressed in complex numbers, and the expression is:

\[
\chi_j(n) = A \sum_{m} \exp\{j(2\pi m + \phi)\} P(m - n T)
\]  

(3)
In the expression above, $A$ represents the amplitude having a determined value. $P(n)$ represents the baseband pulse which has a duration of $T_s$, whose waveform depends on the selection of modeling filters. And $f_c$ is the center frequency. Moreover, the value of $\phi$ which represents the corresponding phase of baseband signal should satisfy:

$$
\phi \in \left\{ \frac{2\pi \times 1}{M}, \ldots, \frac{2\pi \times M}{M} \right\}
$$

(4)
where $M$ is the mode of the modulation.

2) MFSK: The transmission signal of MFSK modulation can be expressed in complex numbers, and the expression is:

$$x_c(i) = \sum_{\mathcal{S}} \exp \left( j(2\pi f_s i) \right) F(i - mI_c)$$  (5)
And the frequency of the sub-carriers, $f_i$, should satisfy:

$$f_i \in \{f_1, \ldots, f_M\}$$

(6)

In the expressions above, $A$ and $M$ also represent the amplitude having a determined value and the mode of the modulation respectively. The rectangular pulse, denoted as $P(n)$, has a duration of $T_s$ and $f_{space}$ represents the frequency space, whose expression is $f_{space} = f_{0} - f_{..} = 1, \ldots, M-1$. Furthermore, the center frequency, denoted as $f_c$, has the following expression.

$$f_c = \frac{1}{M} \sum_{i=1}^{M} f_i$$

(7)

3) MQAM: The signals can be expressed as:

$$S_{M^4M} = a_{in} g_{T}(n) \cos(\varphi) - a_{out} g_{T}(n) \sin(\varphi)$$

(8)

In the expression above, $a_{in}$ and $a_{out}$ are the in-phase and orthogonal component of the amplitude respectively, both of whose values are integers ranging from $1$ to $\sqrt{M}$. Additionally, $g_{T}(n)$ is the pulse of symbols formed by baseband filters.

### 3 PROPOSED METHOD FOR MODULATION CLASSIFICATION

In Section 3.1, we introduce two sorts of data, constellation and cyclic spectrum density, which act as the heterogeneous input of ResNet18, and we also deliver how to formulate and normalize them into the specific matrix forms. Then, the structure of ResNet18 is given in Section 3.2. Eventually in Section 3.3, a brief introduction will be given about the chosen loss function, which is employed to optimize the whole performance.

#### 3.1 Heterogeneous Input Data

**Constellation.** In the field of digital communication, digital signals are often represented on a complex plane to visually represent signals and the relationship between signals. The constellation has two axes. The horizontal X-axis is related to the in-phase carrier and the vertical Y-axis is related to the orthogonal carrier. After receiving diverse signals, we restore them to raw IQ sequences and organize which into the constellation map responding to Real and imaginary axis.

Fig. 2 shows the normalized constellation images for the input data. It demonstrates that the difference between BFSK and QFSK is scarcely possible to identify no matter at any signal to noise ratio (SNR), which is why we select the cyclic spectral density to make up this defect.
Cyclic Spectrum Theory. If the mean and autocorrelation function of random signal $x(t)$ is a periodic function of time, cycle is $T$, then $x(t)$ is called generalized cycle stationary signal. Its average $m_x(t)$ and autocorrelation function $R_x(t,\tau)$ should respectively meet the function below.

$$
\begin{align*}
m_x(t) &= m_x(t+T) \\
R_x(t,\tau) &= R_x(t+T,\tau)
\end{align*}
$$

(9)

Expand the periodic function $R_x(t,\tau)$ with Fourier series, and we have:

$$
R_x(t,\tau) = \sum_{n=-\infty}^{\infty} \tilde{R}_n e^{j\frac{2\pi}{T}nt} = \sum_{n=-\infty}^{\infty} \tilde{R}_n e^{j\frac{2\pi}{T}nt} \tag{10}
$$

where $\tilde{\omega} = \frac{2\pi}{T}$ and the Fourier coefficient $\tilde{R}_n$ is:
The coefficient \( R_\alpha^\alpha (\tau) \) represents the cyclic autocorrelation intensity of the frequency \( \alpha \), which is also a function of \( \tau \) referred to as the cyclic (self) correlation function. The frequency \( \alpha \) is called the cyclic frequency of the signal \( x(t) \), which can be multiple in cyclostationary signals (including zero cycle frequency and non-zero cycle frequency). Among them, the zero cycle frequency corresponds to the stationary part of the signal, and only the non-zero cycle frequency can indicate the signal's cyclic stability. If \( \alpha = 0 \), then \( R_0^0 (\tau) = R_0 (\tau) \) is the autocorrelation function of the stationary signal.

Additionally, the Fourier transform of \( R_\alpha^\alpha (\tau) \) is known as the Cyclic spectral density (or spectral correlation density function). The coefficient \( R_\alpha^\alpha (\tau) \) is known as the Cyclic spectral density (or spectral correlation density function).

\[
R_\alpha^\alpha (\tau) = \frac{1}{T} \int_{-T/2}^{T/2} x(t) x^*(t-\tau) e^{-j2\pi \alpha \tau} d\tau
\]

\[
= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t) x^*(t-\tau) e^{-j2\pi \alpha \tau} d\tau
\]

\[
= \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} x(t) x^*(t-\tau) e^{-j2\pi \alpha \tau} d\tau
\]

Fig. 3 shows the normalized cyclic spectrum density maps of all types of signals at different SNR that the height of the 3D model represents the magnitude of the density value. It indicates that the difference between BPSK, BFSK, QPSK, QFSK is obvious and easy to recognize whatever the SNR is. However, in the aspect of QAM signals we can hardly distinguish whether the signal is QPSK, 16QAM, 32QAM or 64QAM whatever the SNR is, which is the disadvantage using cyclic spectrum density to classify the QAM signals and it has been proven that the features provided by the constellation can successfully complement.

**Process of Formulation and Normalization.** After obtaining the images of the two formats, rational process of formulation and normalization is needed to implement the complementation of the heterogeneous data. Firstly, we formulate the two kinds of images into \( 50 \times 50 \) matrix in disparate approach. In terms of the constellation, we fill out each grid of the matrix with the number of IQ points involved in every single grid, after which we normalize the matrix through the function given below.

\[
X_{\text{Nor}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}
\]

where \( X_{\text{Nor}} \) is the normalized number of IQ points. \( X_{\text{max}} \) and \( X_{\text{min}} \) represent the maximum and minimum number of IQ points respectively, and then \( X \) is the number of IQ points in a certain
grid. Consequently, we obtain one of the heterogeneous input of ResNet18.

![Cyclic spectrum density map of seven types of modulations versus SNR](image)

Fig. 3. Cyclic spectrum density map of seven types of modulations versus SNR

In terms of Cyclic Spectrum Density, the data has been normalized through been divided by the maximum value when the map is generated. We calculate the each average value of $50 \times 50$ grids based on the number of sample points in a certain grid, after which we acquire the other heterogeneous input.

### 3.2 Structure of Modified ResNet18

In this model, we prefer ResNet18 to be the classifier. The architecture of the ResNet18 model is shown in Fig. 4, which consists of an input layer on the left and 17 convolution layers followed by an average pool using softmax. Moreover, detailed introduction of the residual blocks are displayed in Table 1. The sample matrix of heterogeneous data is first sent to the input layer, and after various non-linear operation of the hidden layers, estimated modulation formats will be given in the final stages.

![Structure of modified ResNet18](image)

Fig. 4. Structure of modified ResNet18

<table>
<thead>
<tr>
<th>DETAILS OF THE RESNET18 STRUCTURE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid Concatulation Matrix</td>
</tr>
<tr>
<td>Cyclic Spectrum Density Matrix</td>
</tr>
<tr>
<td>Conv 3x3, 64 Channels, 1 stride</td>
</tr>
<tr>
<td>Conv 3x3, 64 Channels, 1 stride</td>
</tr>
<tr>
<td>Conv 3x3, 64 Channels, 1 stride</td>
</tr>
<tr>
<td>Conv 3x3, 512 Channels, 1 stride</td>
</tr>
<tr>
<td>Average Pool, 1024, 64, softmax</td>
</tr>
<tr>
<td>FLOP L1=10^6</td>
</tr>
</tbody>
</table>
To make the network more practical and perform well, we modify it on two points. The first and foremost one is the modification before the second convolution layer. In the initial design of ResNet18, the first step in network operation is to pass the feature through a $7 \times 7$ convolution layer and a $3 \times 3$ maximum pooling layer. The significance of this block is to reduce the scale of the input feature, which is normally conducive to the process of feature extraction. If it is for large-sized data such as high-resolution pictures, this structure can definitely achieve good outcomes. But since the quality and length of the signal received in a single time, the number of sample points generated is usually too small to build such a large matrix. Therefore, we changed the original block to a convolution layer with stride equals to 1, just like the convolution network followed. In this way, not only can we prevent loss of information due to downsizing, we can also implement the process of feature extraction, which is crucial to the classification. The next modification is selection of loss function, which we will introduce in the next section.

Due to the sufficiently good performance of ResNet18, we do not make changes to its main framework. The consecutive convolution layers aim at extracting features, which play a crucial role in the process of feature extraction. And the link between two convolution layers is the rectified linear unit (ReLU) activation function. The most creative and incredible structure is definitely the residual block, which is implemented by the shortcut connection and we can operate an element-wise addition between the input and output of this block. This operation does not append extra parameters and calculations to the network, but greatly improve the performance of this model in terms of training effect and speed.

### 3.3 Loss Function

To enhance the recognition of this model, we use cross entropy function as our loss function, which is the derivative of relative entropy. Relative entropy, also known as KL divergence, can be used to measure the difference between two distribution if we have two separate probability distributions $p(x)$ and $q(x)$ for the same random variable $x$. In the context of machine learning, $\frac{\mathbb{E}_x \{ \log p(x) \}}{\log \mathbb{E}_x \{ p(x) \}}$ is often called the information gain achieved if $p$ is used instead of $q$. The function of KL divergence is given below.
The first part of the last equation happens to be the entropy of $p$ and the latter part of the equation is the cross entropy. In machine learning, we need to evaluate the difference between label and predictions. KL divergence is definitely a suitable choose. Additionally, due to the invariance of the first part, $-\mathcal{H}(p(x))$, in the optimization process, we only need pay attention to the cross entropy.

4 NUMERICAL RESULTS

In this section, simulations are conducted to verify the performances of our proposed FFCN. The modulation set is $M = \{\text{BPSK, QPSK, BFSK, QFSK, 16QAM, 32QAM, 64QAM}\}$, and the calculation of the average probability of correct classification $P_{c}$ is showed below.

\[
P_{c} = \sum_{i=1}^{n} P(\hat{H}_i = H_i | H_i) P(H_i)
\]

where $\hat{H}_i \in M$. In this equation above, $S$ is the total number of candidate modulations and $P_{c}$ is the prior probability, and it is assumed that the prior probability of each modulation is identical. Furthermore, $P(\hat{H}_i = H_i | H_i)$ represents the probability that the modulation is recognized as $H_i$ under the hypothesis $\hat{H}_i$.

Fig. 5 shows the probability of correct identification for seven modulation models when SNR increases from -5 to 10dB with FFCN. The seven modulation modes include BPSK, QPSK, BFSK, QFSK, 16QAM, 32QAM and 64QAM. From the chart we can see that no matter which sorts of signals, the success rate of them continuously increase and remain stable at approximately 97% eventually at a SNR of 10. Moreover, FFCN shows its outstanding performance in the aspect of BPSK and QPSK, with the figure of over 95% when the SNR is around 5. To sum up, we can draw a conclusion that the FFCN performs well in classifying all the seven kinds of signals.

Fig. 6 shows the probability of correct identification for 3 kinds of features with different dimensions or data-processing approach when SNR increases from -5 to 10dB with FFCN. In the proposed model, the two channel of heterogeneous data take a role of the input. From the graph its clear that the novel type of input outperforms both the constellation and the cyclic spectrum density throughout the range of SNR. More specifically, the success rate of FFCN is
over 90% when SNR is increased to 2 eventually. Additionally, the performance of cyclic spectrum density is better than constellation when SNR is low, while it is inverse when SNR is high. It is proven that to employ the combination of heterogenous data is feasible and has better performance than to adopt single data.

Fig. 5. Performance comparison of different kinds of modulation versus SNR

Fig. 6. Performance comparison of different kinds of inputs for seven types of modulations versus SNR

Fig. 7 shows the probability of correct identification for 4 diverse network (classifier) when SNR increases from -5 to 10dB with the same input of FFCN. From the illustration, when SNR is high, it can be observed that the success rate of all the four classifiers are similar because of the low noise level. In the condition of SNR 0 and 5, the figure of FFCN is outstanding at 95% and 88% respectively, which demonstrate the advantages of ResNet18. However, in the condition of low SNR (-5), the performance of ResNet18 appears slightly mediocre with the figure of around 63%, followed by SVM and RF with almost the same
success rate (60%) and KNN owns the best performance.

![Fig. 7. The probability of correct classification for seven types of modulations versus SNR with different classifiers](image)

### 5 CONCLUSIONS

This paper proposed a novel AMC model in SISO system using supervised learning and FB approach, which was based on the discrete heterogeneous data processed by both constellation map and cyclic spectrum density graph. In contrast to the traditional modulation classification method, we utilized the modified ResNet18 to optimize the unsatisfactory performance of the FB approach that accomplishes the prediction according to the original input matrices after normalization and combination. Due to the potent ability of ResNet18 in the aspect of classification, more discriminative features could be detected and the probability of correct identification could be improved. After extensive simulation, it had been proven that with the suitable operation between the data processed by two methods, the FFCN showed an outstanding performance compared to several popular classifier used today like KNN, SVM and RF, especially at low SNR.

### 6 ACKNOWLEDGEMENT

This work was supported in part by the National Key Research and Development Program of China under Grant (2018YFF0301202, 2019YFB1804400), in part by the National Natural Science Foundation of China under Grant (61801052), and in part by the Beijing Natural Science Foundation under Grants (4202046).
References


Automatic Modulation Classification Using Dense Memory Fusion Network

Anping Li¹, Juanjuan Huang², Xiaofei Zhang¹, Xu Yang¹*, Meiying Wei¹
{lapaning@srrc.org.cn¹, jjhuang@bupt.edu.cn², zhangxf@srrc.org.cn¹,
yangxu@srrc.org.cn¹, weimeiying@srrc.org.cn¹}

The State Radio Monitoring Center, Beijing, P.R.China, 100037¹.
Key Laboratory of Universal Wireless Communications, Ministry of Education,
Beijing University of Posts and Telecommunications, Beijing, P.R.China, 100876².

Abstract. Automatic modulation classification (AMC), as a key technology of cognitive radio (CR), aims to identify the modulation format of the received signal. In this paper, we propose a novel dense memory fusion neural network (DMFN) based AMC method where grid constellation matrix (GCM) extracted from the received signals with low computational complexity are utilized as the input of DMFN. In DMFN, densnet with densely connected structures is designed to extract high representative feature of GCMs, the unit of long short-term memory (LSTM) and fully connected layer are used to make classification decisions. Extensive simulations demonstrate that DMFN yields significant performance gain and takes higher robustness comparing with other methods. In addition, DMFN based AMC scheme achieves 90% classification accuracy at 4dB when the symbol length is set as 512, which illustrates its outstanding performance.

Keywords: Cognitive Radio, Automatic Modulation Classification, Deep Learning, Dense Memory Fusion Neural Network.

1 Introduction

With the rapid growth of mobile devices, the radio frequency (RF) spectrum resource become more scarce[1][2][3]. The development of radio technology is largely constrained by the shortage of RF spectrum resource. To alleviate the spectrum shortage, cognitive radio (CR)[4]. Cognitive Radio proposes a dynamic spectrum access method, which can dynamically access the idle band to avoid interference by perceptually the state of the target band[5]. It shows that there is a need for improved spectrum sensing and signal identification algorithms to enable sensors and radios to detect and identify spectrum users and interference in the widest possible range, thereby improving the signal-to-noise ratio (SNR). In order to ensure the normal usage of the spectrum in CR, automatic modulation classification (AMC), which servers as an intermediate step between signal detection and demodulation, is a promising key technology to identify the modulation formats of the target signals corrupted by noise and interference. AMC is widely used in military and civil communication fields, such as software radio, cognitive radio, spectrum detection and management, signal detection, adaptive modulation transmission, threat analysis, interference identification, signal authentication, electronic reconnaissance, non-cooperative communication, etc.[6]. It plays an important role in intelligent signal analysis.
In general, AMC methods can be roughly divided into two different types, which are the likelihood-based (LB) and feature-based (FB) approaches [7][8]. LB method take AMC as a hypothesis testing problem, which can get the optimal solution under Bayesian sense by minimizing the probability of misclassification. However, when the receiving signal appears parameters of an unknown probability distribution, the complexity of the method increases and the robustness decreases, which greatly limits its application in practice [9][10]. Though FB methods are sub-optimal methods, it is easier to apply in practical implementations than the former method because of its low computational complexity and less prior knowledge required. So this paper is going to improve the performance of FB methods to get more classification accuracy gain.

The FB-based AMC method can be roughly divided into two steps. The first step is to extract statistical features from the received signal. The second step is to use the extracted feature parameters to design a classifier for classification. In recent years, many researchers use some novel features to improve the performance of the FB method. Smith et al. in [11] used the high-order cumulant feature as the input of a fully connected neural network to implement the AMC method, however, the computational complexity of higher-order cumulants is high. The extracted characteristic feature include amplitude-phase information, and signal constellation representation. Plenty of novel neural networks are utilized to process those features, such as Alexnet [12], Resnet [13]. However, both of them can not satisfy the requirement of high performance at low SNR. Some other researchers pay attention to change the loss function to constrained neural network to converges. Sai Huang et al. use contrastive loss [14], compressive loss [15] to enhance intra-class aggregation and inter-class separability, which may make it harder to converge for networks. Moreover, part of his effort was devoted to identifying overlapped sources by using multi-gene genetic programming with structural risk minimization rinciple [16].

In this paper, we propose a novel dense memory fusion neural network (DMFN) based AMC method. Firstly, we utilize grid constellation matrix projection to generated GCMs, which contains the constellation information of the received signal and discards some redundant information. Secondly, a dual model fusion network structure is designed to learn the deep information of GCMs and improve AMC performance. Thirdly, simulations are conducted to verify the performance and robustness of proposed DMFN based AMC method.

2 Problem statement

2.1 Signal model

Assume that the radio frequency signal is received by the signal antenna system and transformed to baseband signal. Therefore, a general expression for the complex envelope of the received baseband signal is given by

\[ y(n) = z(n) + \varepsilon(n), \]

\[ z(n) = As^{i\pi/2n^\pi/\varepsilon^2} \sum_{l=\infty}^{\infty} x(l) h(nT - lT + \varepsilon^2 T), \]  

(1a)

(1b)
where $A$ and $x(t)$ represent the unknown amplitude and symbol sequence, respectively. $\hat{h}(\cdot)$ denotes the residual channel effects and is invariant during the AMC process. $f_0$ is the frequency offset, and $\theta$ represents the phase jitter. $T$ and $\varepsilon$ are the symbol spacing and timing errors, respectively.

$\varepsilon(\tau)$ is the additive white Gaussian noise. Suppose that the average power of each symbol is normalized, the signal-to-noise ratio (SNR) is formulated by

$$\text{SNR} = \frac{\mathbb{E}[\{x(t)\}]^2}{\mathbb{E}[\{\varepsilon(\tau)\}]}$$

(2)

### 2.2 System model

Fig. 1 presents the modulation classification Architecture of this paper, which can be roughly divided into two modules, i.e., feature extraction module and LSTM-based classification module. Firstly, the raw IQ signal received by the antenna receiver should go through the feature extraction module, which can extract GCM features different from the raw IQ feature and the traditional constellation. Secondly, we utilize the GCM as the input of the classification, which is composed of DenseNet and LSTM. Finally, we get the hypothesis $H_k$ of this AMC issue, in which $k$ is set as 5.

![Fig. 1. Dense memory fusion neural network based AMC scheme.](image)

### 3 DMFN based modulation classification

#### 3.1 Grid constellation matrix

Considering that the received signal can be separated by two parts: the imaginary part and the real part [17], we often project the raw "IQ" signal into cartesian coordinate system through mapping the I component to the X-axis, and the Q component to the Y-axis. I, Q components denote the imaginary and real part of the received complex signal, respectively. Regular constellation images (RC) are generated by mapping the I, Q data into scattering points on the complex plane.

In order to facilitate the network to learn effective representations from the signal, the received signal are transformed into GCM as the input of the network. Firstly, we pre-draw the $M \times N$ grid on the RC. Secondly, we calculate the number of sample points that fall in each
grid, and finally, after all the sample points are calculated, we normalize the $M \times N$ grid, resulting in a grayscale matrix image that represents the density of the constellation map.

In this paper, the predetermined size of the grid $M \times N$ are set to 40, both. Some example of GCM for five modulation formats versus SNRs are given in Fig. 2. The highlight pixels in the image indicate that signal symbols are clustered in these area.

![Fig. 2. GCMs for five modulation formats with different SNRs.](image)

### 3.2 Structure of DMFN

In order to explore deep representations and recognize the signal format. DMFN, a data-driven automatic modulation classification structure is designed. As is shown in Fig., the DMFN mainly consist of two modules, the feature extraction module and the classification module. The feature extraction module extracts the deep representations from the raw GCM images. The classification module utilizes the deep high-dimensional representations to perform the final classification decision. As the GCM is quite different with the pictures of actual physical environment. Therefore, the DMFN is different with the deep learning networks using in computer vision field. Firstly, We set the convolutional kernel size as $3 \times 3$ to actualize sparse connectivity. Secondly, we utilize the LSTM to transform the high-dimensional feature into one-dimensional feature using as the input of the last fully connected layer.

In the feature extraction module, the first $3 \times 3$ convolution is utilized to pre-process the raw GCM features and obtain the preliminary representations. The max-pooling performs down-sampling to reduce the size of the preliminary representations. Next, in order to incorporate efficiency and accuracy, we adopt four dense blocks to extract the underlying high-dimensional features. The four dense blocks have the similar structure. Take dense block 1 as example, it contain six units, one convolution and one pooling. The six unit make up the densely connections, each unit connects with all previous units to reuse the feature and strengthen feature propagation. The convolution after six units is used as the transformation to
connect the dense blocks. Besides, rectified linear unit (Relu) activation function in the convolution is able to enhance the nonlinearity of the DMFN. The last part of dense block 1 is pooling, which is utilized to enhance the sparsity of the feature generated from the former convolution and reduce the number of network parameters.

After the feature extraction, we get the high-dimensional feature which can represent the raw signal format. However, the high-dimensional feature is unable to perform the finally classification directly. Hence, we use the classification module to reduce the dimension and improve the computation efficiency. The classification module consists of LSTM layer and fully connected layer, we first use the LSTM layer to transform the high-dimensional feature into one-dimensional feature. Then, fully connected layer output the unified units whose number is equal to the modulation formats. the value of the unified units are utilized to conduct the finally classification.

![Fig. 3. The structure of DMFN.](image)

### 3.3 Loss function

The most widely used classification loss function in multi-classification problem, cross-entropy loss, is presented as follows:

\[
J_c(w) = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i | x_i; w)
\]

\[
= -\frac{1}{N} \sum_{i=1}^{N} \log \left( \sum_{j=1}^{\pi} e^{\theta_j x_i} \right) \delta(y_i, \pi_j)
\]

\[
= -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{\pi} \log \left( \frac{e^{\theta_j x_i}}{\sum_{j=1}^{\pi} e^{\theta_j x_i}} \right) \delta(y_i, \pi_j)
\]

Where \( \delta(\cdot) \) is the indicator function, and \( x_i \in \mathbb{R}^d \) denotes the deep feature of the \( i \)-th sample, belonging to the \( y_i \)-th class. The embedding feature dimension \( d \) is set to 32 in this paper.
$W_j \in \mathbb{R}^d$ denotes the $j$-th column of the weight $W \in \mathbb{R}^{d \times m}$ and $b_j \in \mathbb{R}^k$ is the bias term. The batch size and the class number are $N$ and $K$, respectively.

In order to prevent the network weights from taking extremely large values and over-fitting, we add a regularization term to the cross-entropy loss function. So the loss function is given by:

$$J(w) = J_1(w) + J_2(w) = J_1(w) + \alpha \frac{1}{N} \sum_{i} \sum_{j} \|W^{(i)}_j\|_2$$

(4)

Where $\| \|$ denotes the Euclidean norm. $W^{(i)}_j \in \mathbb{R}^d$ denotes the $j$-th column of the weight $W \in \mathbb{R}^{d \times m}$ of the $i$-th sample and $\alpha$ is the coefficient of the regularization term, which is specified by the user.

We update the weight $w$ utilizing mini-batch stochastic gradient descent and error back propagation algorithm until the loss converges to a constant, and then get the optimal parameter $w^*$ of DMFN, which is given by

$$w^* = \arg \min_w J(w)$$

(5)

4 Simulation

In this section, extensive simulations are conducted to illustrate the superiority and offset robustness of the proposed DMFN. An open source dataset platform, RadioML (https://github.com/radioML), is used to generate the modulated signal for fair comparison. We utilize the RadioML platform to generate the modulation signals and simulate each modulation format as SNR ranges from -6 dB to 14 dB with a step of 2 dB. The candidate modulation set is $M = \{$BPSK, QPSK, 8PSK, 16QAM, 64QAM$\}$. Ten thousand samples for each modulation format at certain SNR are used to train and verify the performance of proposed AMC scheme, and one thousand samples to test. In the training process, the Monte Carlo trails are used to calculate the probability of correctly classifying $P^c$, which is given as follows

$$P^c = \sum_{k=1}^{K} P(H = H_k | H_{y}, H_{z}) P(H_{z}), H_{y} \in M$$

(6)
Fig. 4 shows the classification performance of DMFN versus SNR with different symbol lengths. It is obvious that better classification performance is achieved by using more signal symbols. The classification accuracy reaches 94.94% when the SNR is 6dB and the symbol length is 128.
Fig. 5 illustrates that the performance of DMFN is better than the other two AMC methods, i.e. Resnet[18], Alexnet[19], under the same symbol length and SNR. Those two deep learning neural network structures are selected to conduct comparison. We slightly modified the number of neurons in the final fully connected layer from the original 1000 to the 5, which is the number of modulation format. As is shown in Fig. 5, given the same symbol length, for Resnet34 and Alexnet, the former is better than the latter overall. However, Both of them are slightly insufficient compared with DMFN. The performance gain yielded by DMFN deep learning model can be explained by the combination of densenet and LSTM structure.
Fig. 6. The robustness of DMFN and the other AMC methods.

Fig. 6 compares the robustness of proposed DMFN, Resnet34, Alexnet. The robustness of DMFN based AMC scheme versus normalized frequency offset with symbol length $N = 512$ is depicted. The symbol length and SNR are fixed at 512 and 6dB, respectively. The range of frequency offset is set from 0 to $5 \times 10^{-4}$ with a step of $10^{-4}$. It can be observed that the frequency offset severely degrades the classification performance and DMFN performs best of all, proving its robustness to frequency offset.

5 Conclusions

In this paper, we proposed a GCM based AMC method named DMFN. Firstly, GCM are utilized as the initial input of the network. Secondly, a densnet construction is conducted to extract distinguishable feature from GCMs and a LSTM based classifier is designed to make classification decisions. Finally, extensive simulations are designed to verify the superiority and robustness of DMFN compared with other existing methods.

Acknowledgement. This work was supported in part by the National Key Research and Development Program of China under Grant (2018YFF0301202, 2019YFB1804400), in part by the National Natural Science Foundation of China under Grant (61801052, 61227801), and in part by the Beijing Natural Science Foundation under Grants (4202046).
References

A Calculation Method of vCPU Occupancy Rate of Virtual Machine Forwarding Process

1st Hua Lu¹, 2nd Qinshu Chen¹, 3rd Jidong Zhang¹, 4th Xuefei Duan¹
{lhl_zqh@aliyun.com¹, chenqinshu@gdcni.cn¹, zhangjidong@gdcni.cn¹,
duanxuefei@gdcni.cn¹}

 Guangdong Communications & Networks Institute¹

Abstract. Network Function Virtualization (NFV) is a technology that implements network functions through virtualization on x86 servers. In order to improve the forwarding performance of a VM (virtual machine), usually the most effective method is to use multiple cores and exclusive methods of the vCPUs in the VM to monopolize the physical CPU resources for multiple forwarding processes to use independently. The forwarding process discards the interrupt-based asynchronous signal sending mechanism to avoid the impact of interrupt switching on the forwarding overhead. Instead, it uses a while 1 dead-loop to poll the packet receiving queue. Once there is a packet in the packet receiving queue, the packet is immediately forwarded. Because the vCPU bound to the forwarding process works in an infinite loop polling mode, the vCPU occupancy display is 100% regardless of whether it is in the no-load or full-load phase. Because the VM cannot obtain the real load of each forwarding vCPU in real time and expand it in time, it will cause VM to lose a lot of packets due to overload operation, which will affect the quality of the service carried by the VM. This article studies how to measure the real utilization of vCPU in real time. The results show that the real vCPU occupancy can be accurately calculated using this solution.

Keywords: DPDK, SR-IOV, OVS-DPDK, NFV, vCPU, VM, Occupancy Rate, Forwarding Process.

1 Introduction

Network Functions Virtualisation aims to transform the way that network operators architect networks by evolving standard IT virtualisation technology to consolidate many network equipment types onto industry standard high volume servers, switches and storage, which could be located in Datacentres, Network Nodes and in the end user premises. It involves the implementation of network functions in software that can run on a range of industry standard server hardware, and that can be moved to, or instantiated in, various locations in the network as required, without the need for installation of new equipment [1].

With the continuous advancement of operator network reconstruction, innovative technology architectures represented by SDN/NFV are accelerating the network towards full cloudification, and the transformation of the network into a data centered cloud-based network has become the focus of industry attention, such as foreign AT&T Domain 2.0, China Telecom's CTNet2025, China Mobile's Novonet, and China Unicom's CO reconstruction. At present, the general consensus reached in the industry is that 5G networks will be based on SDN/NFV technology and cloud computing technology to implement network virtualization.
and cloud deployment. The "China Telecom 5G technology White Paper" released by China Telecom puts forward the logic architecture of 5G target network, referred to as "three clouds" network architecture. The "three clouds" architecture meets the characteristics of flexible, intelligent, integrated and open 5G network in the future, and is also built on the technology foundation of SDN/NFV and network cloud[2]. With the maturity of virtualization technology, considering the issues of bandwidth and latency, the concept of edge computing has begun to be introduced [3]. With edge computing, a large number of computing tasks can be processed directly near the source of the data generation, which greatly eases the transmission pressure on the network. At the same time, tasks are processed at the edge, which reduces the delay caused by data transmission speed and bandwidth limitations, and users get faster response time. Edge computing is generally a micro data center composed of several servers. Under the condition of limited resources, the dynamic elastic scalability of NFV virtualization technology is used to achieve the efficient use of edge computing resources.

The problem of insufficient software forwarding performance in virtualization has been greatly improved after the rise of DPDK technology. In this article, we start with the analysis of virtualization technology and introduce several typical software forwarding models. In order to improve the software forwarding performance, the DPDK needs to bind the core to exclusive vCPU resources for the while loop packet processing. The vCPU utilization counted by the general operating system = (1-percentage of time occupied by the idle process). Since the vCPU bound by the DPDK runs in an endless loop, the vCPU will not enter the idle process, so the vCPU utilization is always 100%.

This brings up a question: What is the actual vCPU utilization rate, and does it need to be expanded? Virtualization has the ability to flexibly scale in/out, so how to accurately calculate the actual vCPU occupancy rate at the current moment makes it possible to implement automatic elastic scaling of virtual network elements based on a preset threshold of vCPU occupancy. In this paper, the method of real-time quantitative measurement of vCPU's real occupancy is researched and described, in order to achieve real-time accurate display of the vCPU under the situation of dead loop running vCPU real vCPU occupancy.

Chapter 2 of this article first introduced virtualization-related technologies, and described several typical general server-based forwarding models. Finally, in Chapter 3, the vCPU load indeterminacy when the VM implements soft forwarding is studied in depth.

2 Virtual machine forwarding

2.1 Virtualization technology

Virtualization currently uses Intel's x86 architecture. Intel's virtualization technology (VT) mainly includes VT-x technology for CPU processors, VT-d technology for I/O chipset and VT-c technology for network interface cards. These are hardware-assisted virtualization technologies that provide strong underlying technical support for the widespread development of NFV.

Virtualized network elements in the CT industry generally use the open source Linux operating system, which uses "network interface card reception-> network interface card interruption-> kernel reception-> send messages to user processes-> switch to user mode-> user process processing packet" packet processing mode, which involves multiple memory copies, kernel mode/user mode switching, and process scheduling, which is very inefficient.
What is more serious is that the VM runs on the Host machine, which means that the VM user application layer receives a packet and needs to pass through the kernel processing of the Host OS kernel and the VM OS kernel twice. For virtualized network elements with large forwarding traffic, the performance loss is unacceptable. The DPDK (Data Plane Development Kit) technology was born [6]. DPDK runs in user mode and uses the PMD (Poll Mode Driver). PMD consists of APIs provided by specific drivers in user space for setting up devices and their corresponding queues. Abandoning the interrupt-based asynchronous signaling mechanism brings great cost savings to the architecture. Avoiding interruption performance bottlenecks is one of the keys for DPDK to improve packet processing speed. At the same time combined with zero-copy, memory huge pages, NUMA, SR-IOV and other technologies, greatly improved software forwarding capabilities [4][8].

2.2 Virtualized forwarding model

In the pure server case, software forwarding mainly has the following three forms:

1) VM forwarding based on OVS-DPDK

OVS-DPDK [7] runs in Host OS user mode, DPDK on a VM runs in Guest OS user mode, and VF is a virtual network interface card assigned to the VM. As shown in Fig 1: Packet interaction for two VMs in the same Host, it is forwarded through the OVS-DPDK of the Host, see the curve marked by the label a in the figure; For packet interaction between two VMs across the Host, the OVS-DPDK bypass kernel receives packets directly from the physical NIC. OVS-DPDK sends packets to the corresponding VM through table matching, and the VM user mode DPDK bypasses the VM OS kernel to directly receive packets. See curve marked with label b in the figure. This forwarding model is very flexible, and because the Host and Guest kernel processing is bypassed, forwarding performance is also guaranteed to some extent, but because OVS-DPDK also needs to occupy a certain amount of CPU resources, it will affect the VM's resource occupation allocation to a certain extent. At the same time, OVS-DPDK serves as a unified output for all VMs on the Host, and the forwarding performance of all VMs cannot exceed the forwarding performance of OVS-DPDK, which is a bottleneck point in forwarding performance.

![Fig. 1. VM forwarding model based on OVS-DPDK](image-url)
2) Virtual forwarding based on SR-IOV

The VM uses SR-IOV technology to directly take packets from the physical network interface card for processing, completely bypassing the kernel mode and user mode processing on the Host, and its performance is almost equivalent to the processing of receiving and sending packets on the Host. A physical network interface card supporting VT-c technology can virtualize multiple vfs. Each vf corresponds to an independent packet sending and receiving memory space on the physical network interface card. These vf are allocated to the corresponding VMs respectively. One correspondence makes it possible for the VM to read and write packets directly from the physical NIC memory. This SR-IOV method, combined with DPDK, completely bypasses the processing of the forwarded packets by the Host OS and Guest OS kernels, and the software forwarding performance is greatly improved [5]. The VM forwarding model of SR-IOV+DPDK is also the current choice to pursue the ultimate pure software forwarding. As shown in Fig2: Packet interaction for two VMs in the same Host, skip the Host OS and forward it directly by the switching chip on the physical network interface card, see the curve marked by the label a in the figure; For packet interaction between two VMs across the Host, the VM directly writes the packet to the physical NIC memory and sends it to the external switch, and the other VM reads the packet directly from the physical NIC memory of the Host, and combines the packet processing with the user mode DPDK, see the curve marked with label b in the figure. This forwarding model is the optimal model of VM pure software forwarding, but it is less flexible than the above model 1 due to the requirement of hardware characteristics support for the physical network interface card.

![Fig. 2. SR-IOV-based VM forwarding model](image)

3) VM forwarding based on Intelligent Ethernet Card

This model is derived from model 1, which sinks ovs-dpdk into the Intelligent Ethernet Card to release the occupation of OVS-DPDK on the Host CPU resources in model 1. The powerful throughput capacity of the special forwarding chip in the Intelligent Ethernet Card will not become the forwarding performance bottleneck of the VM belonging to the Host. In addition to OVS-DPDK, some services in the VM can also be sunk, such as tunnel encapsulation and decapsulation, fragmentation and reassembly, load balancing, and so on.
According to the actual service deployed by the VM, the corresponding service sink processing is performed to valuable CPU resources are released for use by VM. As shown in Fig 3: Packet interaction for two VMs in the same Host, skip the Host OS and forward it directly by the switching chip on the Intelligent Ethernet Card, see the curve marked by the label a in the figure; For packet interaction between two VMs across the Host, the VM directly writes the packet to the Intelligent Ethernet Card memory and sends it to the external switch. The other VM reads the packet directly from the Host's physical network interface card memory and combines the packet processing with the user mode DPDK, see the curve marked by label b in the figure; Some services functions sink the Intelligent Ethernet Card, for this type of forwarded packet, it can be forwarded directly after being processed by the Intelligent Ethernet Card without being sent to the CPU for processing, which can save valuable CPU resources, see the curve marked by label c in the figure.

Among the above three virtual forwarding models, the VM dpdk and Host ovs-dpdk in model 1 and the VM dpdk in models 2 and 3 will all involve vCPU binding operations to improve software forwarding capabilities.

2.3 Soft forward CPU load uncertainty

Network Function Virtualization is a technology that implements network functions through virtualization on x86 servers. In order to improve the forwarding performance of a VM, usually the most effective method is to use multiple cores and exclusive methods of the vCPUs in the VM to monopolize the physical CPU resources for multiple forwarding processes to use independently. The forwarding process discards the interrupt-based asynchronous signal sending mechanism to avoid the impact of interrupt switching on the
forwarding overhead. Instead, it uses a while 1 dead-loop to poll the packet receiving queue. Once there is a packet in the packet receiving queue, the packet is immediately forwarded. As shown in Fig.4, in the VM for network function virtualization, there are multiple vCPUs, and the forwarding process is bound to these vCPUs in an exclusive way of binding cores, and an endless loop polling method is used for efficient forwarding service processing.

![Fig. 4. Multi-vCPU VM](image)

Because the vCPU bound to the forwarding process works in an infinite loop polling mode, the vCPU occupancy display is 100% regardless of whether it is in the no-load or full-load phase.

When a network-capable VM is deployed, it estimates the processing throughput required by the VM based on information such as the current business scenario and the number of users, and deploys redundant capabilities. As business scenarios change and the number of users increases, the redundant processing capabilities of the originally deployed VMs will become less and less redundant, or even exceed the original processing VM’s forwarding processing capabilities, causing packet processing to be discarded in a timely manner.

Currently, the maximum throughput of VMs with network functions in typical scenarios is generally provided. In the daily operation and maintenance process, various statistical (VM throughput, packet loss, and other statistical information) of VMs that rely on human resources to detect network functions are used. Plan for subsequent expansion. This method relying on human guarantees is inefficient, and is subject to human factors, which has greater risks. Also, different types of x86 CPUs have different processing capabilities, which increases the difference in the maximum throughput value of the VM originally given.

## 3 Technical solutions and implementation

In order to improve the operation and maintenance efficiency in the virtualization scenario and reduce the risks caused by differences in physical equipment and human factors, this paper proposes a calculation method for the real CPU occupation rate of the vCPU bound by the endless loop operation of the VM forwarding process of network functions, which can conveniently and intuitively display the real CPU usage of each vCPU of the VM to which the network function virtualization belongs, and make a VM capacity expansion plan in advance. The technical scheme is as follows:
First, we need to set the weight value according to the different packets that the VM needs to process. The packet weight value is obtained based on the CPU resources consumed by one packet processing. The size of the weight value is determined by the implementation complexity of different services, so the weight value remains unchanged after the service function goes online.

Secondly, it is necessary to obtain the crystal clock of the physical CPU of the general server as the reference clock, and based on the reference clock, simulate the processing of forwarded packets to obtain the processing time of the baseline service (IPV4) packet, according to the reference clock and the processing time of the baseline service packet, get the total number of packets that can be processed in a vCPU usage calculation cycle.

Finally, based on the statistical values of different service packets processed by the VM in a calculation cycle, and combining different weight values of different service packets, the number of limit service packets can be calculated. According to the formula:

\[
\text{CPU usage} = \frac{\text{number of baseline service packets}}{\text{total number of packets}} \times 100\%
\]

We can get the real vCPU usage in the current calculation cycle.

It should be noted that hardware resources such as the CPU computing unit and CACHE are shared resources. Due to the different load of the bearer service in different time periods, the use of shared hardware resources is different. This has an impact on the accuracy of the originally calculated baseline service packet processing time. To reduce the impact of shared hardware resource usage on the accuracy of baseline service packet processing time, the baseline service packet processing time needs to be updated regularly. Therefore, it is necessary to set a system timer, periodically simulate the processing of forwarded packets, and periodically obtain the processing time of the baseline service (IPV4) packets to use more accurate values in the calculation of the next cycle.

In order to facilitate the understanding of this technical solution, this section describes the specific implementation of the technical solution in combination with specific service functions. The definition symbols used in the text are explained here:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fwdi</td>
<td>Identifier of the forwarding process i, used to identify multiple forwarding processes. The value of i is greater than 0 and less than the number of vCPUs occupied by the virtual machine (VM)</td>
</tr>
<tr>
<td>Tcpu</td>
<td>Crystal clock of server physical CPU</td>
</tr>
<tr>
<td>Tstd,i</td>
<td>Baseline packet processing time. Here, the IPV4 packet is used as the baseline packet to obtain the processing time. i is the number of forwarding processes</td>
</tr>
<tr>
<td>Tload</td>
<td>One vCPU usage calculation cycle</td>
</tr>
<tr>
<td>Tcyc</td>
<td>Running timer period</td>
</tr>
<tr>
<td>Numstd,i</td>
<td>The total number of baseline packets that can be processed in a vCPU usage calculation period. i is the number of forwarding processes</td>
</tr>
<tr>
<td>(\beta_{v4})</td>
<td>IPV4 service weight value. Since IPV4 is a baseline service, the weight value is 1</td>
</tr>
</tbody>
</table>

Table 1. Symbol used in text
The calculation method for the real occupancy rate of vCPUs in an infinite loop includes the acquisition of the physical CPU crystal clock, the acquisition of baseline packet processing time, the acquisition of the total number of baseline packets that can be processed by the vCPU occupancy calculation cycle, the acquisition of the weight relationship between various services and the baseline packets, obtain statistics of various service packets processed during the vCPU occupancy calculation cycle. As shown in Fig 5:

Fig. 5. Calculation of initial state occupancy

The specific implementation steps are as follows:

After the VM is powered on, the occupancy calculation module obtains the crystal clock \( T_{cpu} \) of the physical CPU of the general server where the VM is located as the reference clock of the forwarding process Fwd.

The occupancy calculation module uses \( T_{cpu} \) as the reference clock and records the start time \( T_1 \). At the same time, it instructs the simulator to trigger the IPV4 baseline packet to the forwarding process Fwd, for processing. After the processing of the forwarding process is complete, record the time \( T_2 \) to obtain the baseline packet processing time \( T_{std,i} = T_2 - T_1 \).

Divide a vCPU occupancy calculation period by dividing the processing time of a single baseline packet to obtain the number of baseline packets that can be processed during the calculation period, that is, \( \text{Num}_{\text{std,i}} = \frac{T_{std}}{T_{std,i}} \).

The weight presetting module presets a weight ratio relationship between processing of each service packet and processing time of the baseline packet.
The forwarding module counts the number of packets processed by each service.

The occupancy calculation module obtains the weight value of each service packet and the number of processing each service packet in each calculation cycle. According to the formula $\text{Counter}_{std,i} = \sum \text{Counter}(k) \cdot \beta(k)$ (where $k$ is Different services, such as: ipsec, nat, etc.), get the number of baselined packets.

Divide the number of baselined packets $\text{Counter}_{std,i}$ by the total number of baseline packets $\text{Num}_{total,i}$ that a vCPU occupancy can process in a cycle to get the true vCPU occupancy, ie: $\text{Load}_i = \frac{\text{Counter}_{std,i}}{\text{Num}_{total,i}} \cdot 100\%$.

Considering that in a virtualized environment, the server's physical CPU, CACHE, and other hardware resources are shared and used, the processing time of baseline packets may change under different traffic processing situations, so the timer device needs to be enabled during the running state. Calculate the baseline packet processing time periodically to obtain a more accurate total number of baseline packets that can be processed in the vCPU usage calculation period. As shown in Fig 6:

![Calculation of operating occupancy](image)

In the running state, the crystal clock $T_{cpu}$ of the server's physical CPU is used as the reference clock, the system timer is started, and the timer period is set to $T_{cyc}$. The timer expires, triggers a message, and dynamically adjusts the processing time of the baseline packet. By periodically adjusting the baseline value, it can reflect more real processing power according to the usage of physical resources, making the calculated vCPU usage more accurate.

Through the above measurement method, a more accurate vCPU occupancy situation can be obtained in real time. Compared with the 100% displayed by the operating system vCPU monitoring tool, the vCPU occupancy rate described in this article can dynamically and accurately display the real-time vCPU occupancy situation. Fig 7 is a graph of the vCPU occupancy curve when the vCPU is running different services with the same throughput, using the operating system's own monitoring tools and using this solution's measurement method.
The general server configuration used in this article is as follows: Intel XEN Gold 5118 processors (12 cores each, 2.3GHz) * 2, 128G DDR4 Memory and an Intel X520 10GbE PCIe dual port network interface card. The software uses DPDK 18.11 version, combined with software code to achieve service functions of IPV4, NAT and IPSEC. Use Spirent TestCenter tester to directly connect with server's X520 network interface card to send and receive packets. The figure shows the vCPU usage of different services (IPV4, NAT and IPSEC) at the same throughput.

![Comparison of vCPU utilization](image)

**Fig. 7.** Comparison of vCPU utilization

## 4 Conclusion

This article proposes a method for calculating the real vCPU occupancy when the 100% occupancy of the VM's forwarding process is displayed, and the baseline service packet processing time is obtained by timing the simulated packet processing. According to different service statistics and corresponding the weight value normalizes the service flow model during the calculation period of the occupancy rate, so as to obtain the real vCPU usage. In addition, considering the impact of hardware resource sharing usage, a method of periodically updating the processing time value of the baseline service packet is adopted to achieve the purpose of truly reflecting the vCPU usage. Combining alarms, statistics, and graphical displays allows operations and maintenance personnel to easily and intuitively detect changes in forwarding throughput and schedule capacity expansion in a timely manner.

**Acknowledgement.** This work was supported by the National Key Research and Development Program of China under Grant (2019YFB1804400).

**References**


A large capacity programmable packet forwarding device

1st Qinshu Chen1, 2nd Hua Lu*, 3rd Fusheng Zhu1, 4th Hui Li2
{chenqinshu@gdcni.cn, luhua@gdcni.cn, zhufusheng@gdcni.cn}

Guangdong Communications&Networks Institute

Abstract. In the 5G/B5G era, the need of various new services for network bandwidth, delay and cost is far different, so various emerging services may have their own connective protocols according to their own special applications, which requires the core exchange equipment in the network to support programmability. However, neither using network processor to realize packets forwarding nor using fixed pipeline to realize packets forwarding can not meet the needs of 5G/B5G new applications. In this paper, a new architecture of switching chip is proposed. By using programmable packets parser, programmable packet process unit and programmable packet editor, the requirements of 5G/B5G new applications’ rapid deployment, large capacity, low delay and low energy consumption are met.

Keywords: programmable packet parser, programmable packet processor, programmable packet editor

1 Introduction

In 5G/B5G era, with the rise of new services such as automatic driving, telemedicine and virtual reality, various emerging services may define their own communication protocols according to their own special applications due to the distinct demands of various new services for network bandwidth, delay and cost, requiring the core equipment in the network to support programmability. The programmability of services depends on that of switching chip, making it necessary to have programmable packet parser, programmable packet process and programmable editing function in the switching chip of the switching equipment. At present, the implementation of switching chip either uses network processor to make packets forwarded or uses fixed pipeline to make packets forwarded. Although making packets forwarded by using network processor supports software programming according to the needs of services, it leads to large search delay and inability to support large bandwidth, can not reach the level of T bps switching capacity and not meet the requirements of 5G/B5G high-capacity & low delay due to the multi-core time-sharing reuse of the same piece of RAM; while fixed pipeline forwarding can support the level of T bps switching capacity, but once the design is completed, it cannot be modified through software programming to satisfy the needs of 5G/B5G new applications’ rapid deployments due to
the need to design in accordance with the networks protocol in advance. These con-
tradictions become more and more serious with the development of 5G/B5G mobile
Internet and Internet of things. In this paper, we study a new programmable switching
architecture and realize programmable packet parser, programmable packet pro-
cessing unit and programmable packet editer to solve the above problems.

2 Programmable packet process flow

![Diagram of Programmable Switch Chip Block Diagram]

The structure of programmable switch chip is shown in 0, which is composed of pro-
grammable packet parser module, several packet processing units and programmable
packet editer module. Firstly, the packets enter the programmable packet parser mod-
ule for analysis, obtaining the structure of the packet, putting the parse results (mainly
packet type, destination MAC address, source MAC address, VLAN, ETH type, DIP,
SIP, IP protocol number etc. The keys carried by the packet in this paper can be user-
defined keys, not limited to the keys described above) into the meta-
data of the packet
and send them to the subsequent module PPU (packet process unit) for processing.
The structure of each PPU is the same, as shown in 0. Each PPU takes the corre-
sponding keys from the meta-data and those configured in the lookup table to match.
If matched, the contents of the corresponding entries in the table are taken out and put
into the meta-data of the packet, and then sent to the next PPU for processing. After
processed by PPU’s, the meta-data of the packet is sent to the packet editer module for
generating new packet. Finally, forward the packet to the destination port indicated by
the meta-data.

3 Programmable packet parser

The main function of the programmable packet parser is to analyze the packet accord-
ing to the packet parse database configured by users and move the key fields of the
packet to the meta-data. The parse flow is to determine the next node (i.e. the next
state) of the packet according to the current node state in the parse tree and the match-
ing keys extracted from the packet, then jump to a new node in the packet parse tree,
and then to determine the next node (i.e. the next state) in the parse tree according to
the state of the new node and the next keys to be matched, and so on to the end node of the packet parse tree.

To realize the parse process described above, a parse database is required which is composed of the matching rule table and the action table. The matching rule table in the parse database needs to support the masked matching function, as can be realized by the ternary content addressable memory. The matching rule table needs to store the keywords to be compared and the current status, while the former need to support masked matching as its length may be less than the width of the matching rule table and the bits not concerned can be masked. The action table stores the next state, the address of the key to be extracted from the packet next time and the operation instructions that are needed to move the keys of the packet header to the meta-data.

The packet parser first obtains one initial state and four initial offset addresses according to the input port register, obtaining four corresponding keys from the packet according to the latter. After splicing these keys with the initial state looks up to the matching rule table. If hitting the entry of the matching rule table, it reads the action table according to the address of the entry, obtaining the next state and the four offset addresses from the action table in the database. Each key offset can extract one byte key from the packet header, while data offset and data length, corresponding to each other one by one, are given in the packet parse database. Data length is two bits, indicating how length can be extracted by data offset which can be one byte, two bytes, four bytes and six bytes. Once matching can get up to eight valid data offsets and data lengths, that is to say, up to eight fields can be extracted from the packet and put on the meta-data. According to the key offset obtained from the matched entry of the packet parse database, the keyword corresponding to four bytes is extracted from the packet, after that the next state is obtained from this table, with the corresponding search result obtained from the packet parse database again, that is, the next state and the offset addresses of four matched keys. At the same time, the packet parse database provides up to eight offset addresses and data lengths. According to the offset addresses and data lengths, up to eight data are extracted from the packet and stored on the meta-data. Then the process described above is repeated to match the packet parse database again until the next state is 255. Since the packet parse database can be configured by software, the packet parser can analyze any packet by configuring packet parse database according to the format of the packet.
4 Programmable packet process unit

![Programmable packet process unit](image)

The whole programmable packet process unit is composed of the lookup table condition decision module, lookup key generation & hash calculation module, lookup table read/write control module and lookup table result process module, as shown in 0.

4.1 Lookup condition decision

The conditional decision module of lookup table consists of \( m \) lookup table decision makers, each of which is composed of \( i \) \( j \)-bit comparators that is able to perform the judgment of greater than, less than, equal and not equal; each comparator has a control register which stores the immediate number or the address of the match key in the packet meta-data and the flag of the match key is an immediate or not. If it is an immediate number, the value of the compared number is stored; if not, it’s the match key address of the packet meta-data. Simultaneously, the mask of the match key is also included. Before the comparison operation, the match key first needs to do and operation with the mask register by bit, then performing the comparison operation. In the end, the judgment results of \( i \) comparators and \( k \) key fields from the meta-data according to the address indicated by the lookup table adjudicator are combined to search the TCAM. The \( m \) lookup table decision makers can be used independently, each of which corresponds to a lookup table; on the other hand, multiple lookup table decision makers can be used jointly to search a lookup table as a decision condition, at that time the number of lookup tables supported by the PPU becomes smaller yet the supported lookup table decision conditions are more complex than the single lookup table decision maker. The flag of whether the lookup table inside the PPU or outside the PPU and the number of the lookup table to indicate which lookup table to look up are given before it is read.
4.2 Key generation and hash calculation

Each lookup table contains \( p \) configuration registers, which are used to indicate how to generate keywords to search the lookup table. Each register contains a byte of the corresponding key in the position of the packet meta-data and the bit length of the byte’s valid bit. In terms of the register, each byte of the match key is extracted from the meta-data, which extract the bytes that are combined into a temporary lookup table key according to the register’s order. Finally, the invalid bits inside the temporary lookup table key are wiped to generate the lookup table key. According to the configuration, the corresponding hash function is selected to calculate the hash index. In this paper, the double hash lookup is supported. Thus, for each lookup key, two different hash functions are used to calculate its hash index simultaneously.

4.3 Lookup table read/write control

PPU lookup table read/write control is to flexibly change the size of each lookup table by configuring registers according to the work scenario of the switch, so that the packet forwarding chip can meet the needs of various services in various scenarios as well as make the best use of RAM resources. This module is divided into the read/write control of the lookup table inside the PPU and that outside the PPU, but the two of lookup table read-write control modules are almost the same.

The hash table attribute register includes the depth of the hash table, the byte width of the hash table, whether the hash table supports single hash or double hash, whether the hash table carries the statistical counter pointer or not and the base address of the statistical counter pointer, whether the hash table carries the flag of the instruction pointer or not and the base address of the instruction pointer, etc. At the same time, each RAM inside the PPU read/write control module also contains the attribute register describing the RAM block. The register includes the logical table number of the RAM, the RAM block’s location in the row and column of the lookup table and the hash function number support by the RAM block. In order to reduce the hash conflict, multiple entries are usually stored in a hash index.

When the hash table read/write control logic inside the PPU receives the lookup table request, it is decoded into the logical table number of the hash table in accordance with the request source. Each RAM block in the PPU checks its RAM attribute register according to the request and the address of lookup table, then checking the logic number of the request lookup table and the number of configured in RAM block attribute register, judging the address of the lookup table in the RAM block of the lookup table. In the meantime, the RAM address is generated according to lookup table address and the RAM is read, after which the data is put into the result register according to the logical table column number in the attribute register of RAM block.

The hash_mux_result is divided in light of the table byte width configuration in the attribute register of the hash table, finally obtaining multiple hash_result data. The
hash table supports a hash index to contain 8 entries in general, causing hash_mux_result to split 8 hash_result data of the same byte width in the end. The width of partitions and the number of data to be partitioned are all configured by software. Then the match key, whose width is configured by the hash table attribute register, is taken out from the hash_result as well as compared with the hash table. If match, the final result of the hash table is taken out according to the property register configuration of the hash table, including the instruction pointer offset and the statistics count offset. Afterwards, the instruction pointer offset and the instruction pointer base address configured in the attribute register of the table are added to send to the lookup table result processing module as the instruction pointer. At last, the offset of the statistics count and the base address of the statistics count in the attribute register of the lookup table are added to generate the statistics count pointer, after which initiate the request of searching the statistics count lookup table.

4.4 Lookup table result process

The lookup table result process module is mainly based on the processing results of the previous PPU or the packet parser module and the results of this lookup table to merge and generate a new meta-data, transferring to the next PPU or programmable packet editor module. The lookup table process module generates an instruction database, each entry of which contains u ALU (arithmetic logical unit) instructions. Each instruction can perform the following actions:

- Move the lookup table result data to the meta-data
- Move the data on the meta-data to other locations of the meta-data
- Perform logical operation, including operation and, or, not, exclusive-OR
- Perform arithmetic operation, add and subtract, but not multiply or divide
- Perform shift operations including left shift and right shift
- Set the immediate number to the meta-data

5 Programmable packet editor

The programmable packet editor module uses the results of the PPU process to extract the meta-data from the key field and merge them into a new packet header, then taking out the payload from the packet buffer, splicing the new packet header and the payload to a new packet. The principle of the programmable packet editor is that the sequence of protocol header key fields of a packet is sequential and fixed, for example, a UDP packet must be Ethernet L2 header + IPv4 header + UDP header + payload; a TCP packet must be Ethernet L2 header + IPv4 header + TCP header + payload, while the packet of Ethernet L2 header + UDP header + IPv4 header + payload will not occur. Therefore, we only need to build a most complete protocol header as well as define the location where each protocol header key fields exist in the meta-data as a packet editing database. During packet editing, we can take out each byte from the meta-data according to the location indicated by the packet editing database and then splice it to form a new packet header.
In order to better describe the above packet editing flow, this paper presents the L2 packet editing process. The first line of 0 is the meta-data content sent by the previous PPU, while the second line describes the address of each byte on the meta-data, which does not exist in reality, just for the convenience of description. On the meta-data, the 0 to 5 bytes store the SA (MAC source address) of the packet which is 48 bits and 6 bytes in total. The highest byte of SA is S5 which is stored in the position 5 of the meta-data, with the second highest byte S4 stored in the position 4 of that, the other bytes are stored as shown in 0, which is not described in this paper. In 0, the orange field is SA field, 6 bytes; the green field is DA field, 6 bytes; the yellow field is Ethernet type field, 2 bytes; the blue field is VLAN tag field, 4 bytes, including 16 bits VLAN tag protocol field, 3 bits priority COS field, 1 bit CF field and 12 bits VLAN ID field.

The packet editing database stores the position information of the corresponding field of the packet in the meta-data, as shown in 0. The first is the DA field of the packet, which has 6 bytes in total. There lies the location of the packet in the meta-data, with the next one of the SA field in the meta-data, the third one of the VLAN field in the meta-data and the last one of the Ethernet type field in the meta-data. During packet editing, the corresponding fields are taken out from the meta-data according to the address stored in the packet editing database, spliced into packets. For example, when editing a packet, the highest byte of DA is taken from the packet editing database at the address 45 of the meta-data, using which to take the corresponding content D5 from the meta-data, placed at the position of the first byte of the packet. Then the position 44 of the next byte of DA is taken from the packet editing database, from which the meta-data takes the content D4, spliced after the D5. Similar is the other fields until a new packet header is assembled finally as in 0. Afterwards, a new data packet composed of the original payload is spliced with a new packet header, forwarded to the output port indicated by the meta-data.
6 Summary

In this paper, a large capacity programmable packet forwarding scheme is proposed to solve the problem that the current network processor cannot meet the requirements of low latency and large capacity exchange as well as to avoid the problem that the fixed pipeline forwarding cannot meet that of the rapid deployment of new services. The forwarding behavior can be changed according to the needs of the services by software programming, which can meet the needs of 5G/B5G new services rapid deployment, large capacity, low latency and low energy consumption.

Acknowledgement. This work was supported by the National Key Research and Development Program of China under Grant (2019YFB1804400).

References

4. O. Flauzac, C. Gonzalez, and F. Nolot SDN Based Architecture for Clustered WSN. Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 9th International Conference on, Blumenau, 2015, pp. 342-347.
Abstract. In this paper, we propose a spectrum-efficient device-to-device (D2D)-based puncturing algorithm to multiplex enhanced mobile broadband (eMBB) and ultra-reliable low-latency communication (URLLC) traffic. The users with URLLC traffic communicate directly in D2D mode by reusing the resource blocks (RBs) allocated to cellular user with eMBB traffic. First, we derive the achievable transmission rate of cellular link that suffers from time-varying interference caused by different D2D links in two-timescale scheduling slot structure. Additionally, we calculate the achievable transmission rate of D2D link considering the short-packet nature of URLLC traffic. Then, the delay and reliability quality of service (QoS) of URLLCs are studied based on the effective capacity theory. Finally, the puncturing algorithm is formulated as a throughput maximization problem subject to the QoS constraints of eMBB and URLLC traffic. The transmitting power of eMBB users and URLLC users, as well as the resource reusing pattern are jointly optimized.

Keywords: eMBB (enhanced mobile broadband); URLLC (ultra-reliable low-latency communication); D2D (device-to-device) based puncturing algorithm; effective capacity.

1 Introduction

Enhanced mobile broadband (eMBB) and ultra-reliable low-latency communication (URLLC) are two typical types of services in the fifth generation (5G). eMBB is supposed to support 0.1-1 Gbit/s user experience rates, and URLLCs require the end-to-end delay below 1 ms with the probability of at least 99.999% [1]. Though the URLLC traffic is sporadic, it should be supported by large bandwidth as similar as the eMBB traffic. Due to scarcity of bandwidth resource, multiplexing eMBB and URLLC traffic, which refers to transmitting the diverse traffic on the same bandwidth resources, may be a feasible approach.

For eMBB and URLLC multiplexing scenarios, 3GPP RAN WG1 developed a two-timescale scheduling slot structure. The transmission time interval (TTI) of eMBB is one millisecond, while the TTI of URLLC traffic is shorter than one millisecond [2]. In the two-timescale scheduling slot structure, the puncturing-type multiplexing schemes were proposed, where the URLLC traffic prefers to be scheduled immediately once it arrives even if the eMBB traffic is transmitting for guaranteeing the delay quality of service (QoS) of URLLC traffic. In the case of a single URLLC user, some researchers focused on mitigating the rate loss of the punctured eMBB traffic through joint scheduling [3] or rotating constellations [4], rather than satisfying the QoS of URLLC traffic. In [5], a multi-user-punctured scheduler was developed for achieving a balance between the spectral efficiency of eMBB and average latency of URLLC. However, guaranteeing the average delay is not enough for URLLC traffic, the delay and reliability QoS of which can be represented by the statistical delay QoS including the delay bound and delay bound violation probability.

Device-to-device (D2D) technique is a viable way to reduce the end-to-end delay of URLLC. In the existing works, both the cellular links and the D2D links were utilized to carry URLLC traffic [6-8]. Only the spatial diversity gain of D2D was exploited to boost reliability of URLLCs. In the conventional D2D communications, resource allocation schemes have been studied to maximize spatial multiplexing. However, they are not suitable for the case of eMBB and URLLC multiplexing. In the two-timescale scheduling slot structure, since the TTI of URLLC is smaller than the one of eMBB, the resource of a cellular link carrying eMBB traffic may be selectively multiplexed by several different D2D links with URLLC traffic. Therefore, the interference suffered by one transmission of a cellular link is time-varying and hard to be evaluated.

In this paper, the spatial multiplexing gain of D2D technology is exploited to multiplex the URLLC traffic and the eMBB traffic. To guarantee the end-to-end delay QoS of URLLC, the users with URLLC
traffic employ the D2D mode to communicate directly via reusing the resource blocks (RBs) allocated to
the cellular users with eMBB traffic. The statistical delay QoS in effective capacity theory is adopted to
characterize the reliability and delay QoS of URLLCs. Based on the effective capacity theory, we propose
a novel D2D-based puncturing algorithm with the aim of achieving high system throughput while
satisfying the heterogeneous QoS of eMBB and URLLC traffic. The main contributions of this paper are
as follows.
(1) In the D2D-based multiplexing scenario, we calculate the average power of time-varying interference
caused by multiple D2D transmissions and derive the achievable transmission rate of the cellular
link. Then we derive the achievable transmission rate of D2D link considering short-packet
transmission feature of URLLCs.
(2) We derive the effective capability of D2D link with URLLC traffic. Based on the effective capacity
theory, we quantify the constraints on the transmission rate of D2D link considering the statistical
delay QoS requirement of the random and sporadic URLLC traffic.
(3) The D2D-based puncturing algorithm is formulated as a throughput maximization problem subject to
the rate requirement of eMBB traffic and the statistical delay QoS requirement of URLLC traffic.
The problem is solved by the particle swarm optimization (PSO) algorithm.

2 System model

In this paper, $M$ cellular links carrying eMBB traffic and $N$ D2D links carrying URLLC traffic are
considered to coexist in a single cell of wireless network. $M = \{1, 2, \ldots, M\}$ and $N = \{1, 2, \ldots, N\}$
denote the sets of cellular users and D2D links respectively. The frequency division duplex (FDD) is
utilized. In this paper, the D2D links reuse the UL resources of cellular users. The BS is assumed to know
the channel statement information (CSI) of cellular users and D2D links. The channel gain between the
cellular user $i$ and the BS is denoted by $g_{i}^{d}$. The channel gain between the transmitter of D2D link $j$ and
the receiver of D2D link $j$ is denoted by $g_{j}^{d}$. The channel gain between the cellular user $i$ and the receiver of D2D link $j$ is
denoted by $g_{i}^{d}$. The channel gain between the transmitter of D2D link $j$ and
the receiver of D2D link $j$ is denoted by $g_{j}^{d}$. The pathloss model for D2D
links is $F = 148 + 40 \log_{10} \left( d / 1000 \right)$ and that for cellular links is
Rayleigh fading model for small-scale part is also considered.

Fig. 1 The communications scenario

In this paper, the TTI of eMBB traffic is 1 ms, while the TTI of URLLC traffic is $\omega t_{s}$, and $\omega t_{s} < 1$ ms.
$\omega$ denotes the number of OFDM symbols occupied by each D2D link. $t_{s}$ denotes the duration of one
OFDM symbol. An example of the system model is depicted in the Fig.1, where the D2D link 1 and D2D
link 2 multiplex the UL resources of eMBB user 1. The transmissions of these two D2D links occupy the
same subcarriers but different symbols.

3 D2D-based puncturing algorithm for eMBB And URLLC multiplexing scenario

3.1 Analysis on the achievable transmission rates of eMBB and URLLC Traffic
In this paper, $\rho_{i,j}$ is defined to reflect the resources reusing relationship between the cellular link and the D2D link. If the RBs of the cellular user $i$ are reused by the D2D link $j$, $\rho_{i,j} = 1$; otherwise, $\rho_{i,j} = 0$. In our scheme, we regulate that each D2D link shares at most one cellular user’s RBs, i.e., $\sum_{j \in M} \rho_{i,j} \leq 1, \forall i \in \mathbb{N}$.

And $\lambda$ is defined as the maximum number of D2D links that are allowed to multiplex on the RBs allocated to one cellular user. We assume that the data in one transmission block of eMBB traffic is jointly coded and decoded. Hence, the average power of interference caused by multiple D2D links is considered in deriving the transmission rate of cellular communication. The average interference is given as

$$N_i^{\text{total}} = \sigma^2 \sum_j \rho_{i,j} g_j^k e^{\gamma_j} / \tau_j, \lambda \cdot \sigma \leq 14$$

(1)

$\lambda \cdot \sigma \leq 14$ holds for the normal cyclic prefix (CP) condition that one TTI of eMBB traffic contains 14 OFDM symbols.

According to the average power of the time-varying interference, the SINR of the cellular user $i$ is given as

$$\gamma_i^c = \frac{P_c g_i^k e^{\gamma_i}}{N_0 + N_i^{\text{total}}} = \frac{P_c g_i^k e^{\gamma_i}}{N_0 + \sigma^2 \sum_j \rho_{i,j} g_j^k e^{\gamma_j} / \tau_j}, \lambda \cdot \sigma \leq 14$$

(2)

where $N_0$ denotes noise power of wireless channel. $P_c$ denotes the transmission power of cellular users.

Since $\sum_j \rho_{i,j} \leq 1$, the SINR of D2D link $j$ is given as

$$\gamma_j^d = \frac{\sum_j \rho_{i,j} P_d g_j^e e^{\gamma_j}}{N_0 + \sum_j \rho_{i,j} P_d g_j^e e^{\gamma_j}}$$

(3)

According to Shannon formula, the achievable transmission rate of cellular users $i$ can be calculated as

$$R_i^c = B_0 \log_2 (1 + \gamma_i^c)$$

(4)

where $B_0$ denotes the bandwidth resource that is allocated to the cellular user $i$.

Because the short data packets with finite block-length channel codes are transmitted in URLLCs, the block error rate (BLER) of the short packet transmission is inevitable. For a given block-length $m$ and BLER $\phi$, the achievable transmission rate of D2D link $j$ is approximated by [10]

$$R_j^d = B_0 \left[ \log_2 \left( 1 + \gamma_j^d \right) - \frac{\left( 2 + \gamma_j^d \right) \gamma_j^d \left( \log_2 e \right) - 1}{\left( 1 + \gamma_j^d \right)^m f_2^{-1}(\phi)} \right]$$

(5)

where $f_2^{-1}(x)$ denotes the inverse function of the Gaussian Q-function. According to our multiplexing scheme, the achievable sum-rate of the system is given as

$$R_s = \sum_i \left[ R_i^c + \sum_j \rho_{i,j} R_j^d \right] - B_0 \left[ \sum_j \log_2 \left( 1 + \gamma_j^d \right) + \sum_j \rho_{i,j} \log_2 \left( 1 + \gamma_j^d \right) - \frac{\left( 2 + \gamma_j^d \right) \gamma_j^d \left( \log_2 e \right) - 1}{\left( 1 + \gamma_j^d \right)^m f_2^{-1}(\phi)} \right]$$

(6)

3.2 Statistical delay QoS analysis of URLLCs
In this paper, the arrival process and the service process of dynamic queuing system are assumed to be stationary ergodic stochastic processes. Let \( Q_j(\infty) \) denotes the stationary queue length of the transmitter of D2D link \( j \). \( Q_j^{\text{max}} \) denotes the threshold of queue length. On the basis of large deviation principle (LDP), the relationship holds as follow [11]

\[
-\lim_{Q_j^{\text{max}} \to \infty} \frac{1}{Q_j^{\text{max}}} \log \left( \Pr \left[ Q_j(\infty) > Q_j^{\text{max}} \right] \right) = \theta_j
\]

(7)

where \( \theta_j \) is QoS exponent of the URLLC traffic carried by the D2D link \( j \). The larger the \( \theta_j \) is, the more stringent the QoS requirement is.

The statistical delay QoS requirement of the URLLC traffic carried by the D2D link \( j \) is defined as \( \langle D_j^{\text{max}}, \varepsilon_j \rangle \), where \( D_j^{\text{max}} \) and \( \varepsilon_j \) denote the delay bound and the maximum allowable delay violation probability, respectively. Let \( D_j(\infty) \) denote the stationary delay of the packets transmitting on the D2D link \( j \). The statistical delay QoS requirement holds for \( \Pr \left\{ D_j(\infty) > D_j^{\text{max}} \right\} \leq \varepsilon_j \). From [12], the approximated function of delay violation probability is given as

\[
\Pr \left\{ D_j(\infty) > D_j^{\text{max}} \right\} \approx \delta_j \exp \left\{ -\delta_j E_j^\varepsilon(\theta_j) D_j^{\text{max}} \right\}
\]

(8)

\( \delta_j \) denotes the probability of the event that the buffer of the transmitter of D2D link \( j \) is non-empty, and \( \delta_j \leq 1 \). Then

\[
\Pr \left\{ D_j(\infty) > D_j^{\text{max}} \right\} \leq \exp \left\{ -\delta_j E_j^\varepsilon(\theta_j) D_j^{\text{max}} \right\} \leq \varepsilon_j
\]

(9)

\( E_j^\varepsilon(\theta_j) \) denotes the effective bandwidth of URLLC traffic carried by the D2D link \( j \). When the average arrival rate is \( \lambda_j \), the effective bandwidth for a Poisson process is

\[
E_j^\varepsilon(\theta_j) = \frac{\ln(1/\varepsilon_j)}{D_j^{\text{max}} \ln \left[ 1 + \frac{\ln(1/\varepsilon_j)}{\alpha_j D_j^{\text{max}}} \right]}
\]

(10)

The effective capacity characterizes the maximum constant arrival rate that a system could support so as to satisfy the statistical delay QoS requirement [13]. The transmissions of D2D links are regarded as service process in this paper. The effective capacity of D2D link \( j \) is given as

\[
E^\varepsilon_j(\theta_j) = \frac{1}{\theta_j} \ln \left[ E \left( e^{-\delta_j^\varepsilon(\theta_j)} \right) \right]
\]

(11)

For URLLCs, the delay bound is much smaller than the coherence time of wireless channel in typical scenarios except high mobility scenarios where the velocity is larger than 100 km/h. Hence, the channel is referred to as quasi-static fading channel, the random fading coefficients of which keep constant in the duration of transmitting one data packet of URLLC traffic. Then, the effective capacity of D2D link \( j \) carrying URLLC traffic is derived as

\[
E_j^\varepsilon(\theta_j) = G_j \left[ \log(1 + \gamma_j) - \sqrt{\frac{(2 - \gamma_j) \gamma_j \log(1 + \gamma_j) f_j^\varepsilon(\theta_j)}{(1 + \gamma_j)^2 m}} \right]
\]

(12)

According to [14], for statistical independent arrival and service processes, the statistical delay QoS requirement characterized by \( \theta_j \) can be guaranteed, if and only if the effective capacity is greater than effective bandwidth, i.e.,

\[
E_j^\varepsilon(\theta_j) \geq E_j^\varepsilon(\theta_j)
\]

(13)

3.3 Problem Formulation

In this paper, we formulate the spectrum-efficient D2D-based puncturing algorithm as a system throughput maximization problem with multiple constraints including the statistical delay QoS constraints of URLLCs, which is formulated as
Constraints in (14.a) restrict that each D2D link only reuse at most one cellular user’s bandwidth resource. Constraints in (14.b) allow each cellular user’s RBs to be multiplexed by at most $\lambda$ D2D link(s). Constraints in (14.c) limit the transmitting power of cellular users and D2D links. Constraints in (14.d) guarantee the SINR of cellular link which reflect to the rate constraints of eMBB traffic. Lastly, the constraints in (14.e) hold when the statistical delay QoS requirement of URLLC traffic is satisfied. In this paper, we employ the PSO algorithm with penalty function to cope with the optimization problem.

4. Simulation Results

In our simulations, $M (=10)$ cellular users with eMBB traffic and $2N (=10-40)$ URLLC terminals (communicating via D2D links) are uniformly and randomly distributed in a cell (with radius of 50 m and 100 m respectively). The maximal transmit power $P_{d}^{\text{max}}$ of D2D transmitter is set to be 0.1 w. The required minimal SINR of cellular user is 0.2. The delay and reliability QoS requirement of URLLC traffic is 0.8 ms and 1-10^{-5} respectively. The numerical results are averaged over 500 times.

![Fig.2 The throughput versus the maximum transmitting power of eMBB users.](image)

Fig. 2 shows the system throughput of our D2D-based puncturing algorithm versus the allowable maximum transmitting power of cellular users. It shows that the system throughput is unaffected by the maximum transmitting power in case of more (i.e., $N > 5$) D2D links existing. When $N = 5$, the system throughput increases lightly along with the maximum transmitting power of cellular users. On the one hand, a larger transmitting power of cellular users with eMBB traffic can elevate the transmission rate of cellular link. On the other hand, a larger transmitting power also causes more interference to D2D links with URLLC traffic. When the number of D2D links is small, the larger transmitting power is taken by our algorithm for the cellular user to improve their transmission rates. When there are more D2D links, a moderate transmitting power is obtained by our algorithm for reducing interference and satisfying the statistical delay QoS of URLLCs.
Fig. 3 The throughput versus the maximum distance of D2D links.

Fig. 3 shows the system throughput of our D2D-based puncturing algorithm in the scenarios with different maximum communication distances of the D2D links. From Fig. 3, we can find that increasing the maximum communication distance of the D2D link makes the system throughput decrease. Further, as Fig. 3 shown, the decreasing rate of the throughput curve for the case of 5 D2D links is obviously lower than others. In case of existing 5 D2D links, the D2D transmissions cause less interference than other cases. Hence, for the case of $N = 5$, increasing the maximum distance for the D2D link does not result in a dramatic reduction in system throughput. Fig. 4 depicts the system throughput versus the average arrival packet rate of URLLC traffic in each D2D link. It shows that packet arrival rate has little effect on the system throughput when the arrival rate is smaller than 0.1. It is indicated that our puncturing algorithm is robust to the variation of the average arrival rates.

5. Conclusions

In this paper, we proposed a spectrum-efficient D2D-based puncturing algorithm for eMBB and URLLC traffic co-existence networks, where D2D links with URLLC traffic reused the bandwidth resources allocated to the cellular link carrying eMBB traffic. Our novel proposed puncturing algorithm relied on the effective capability theory to provision the heterogeneous QoS guarantees of eMBB and URLLC traffic, and achieved high throughput of system at the same time.

Acknowledgement

This work was supported by Jilin Provincial Science and Technology Department Key Scientific and Technological Project (No.20190302031GX), Changchun Scientific and Technological Innovation Double Ten Project (No.18SS010), National Natural Science Foundation of China (No.61671219), Jilin Province Development and Reform Commission Project (No.2017C046-3), and National Natural Science Foundation of China (No.61801191).
References

[2] Chairman’s notes 3GPP: 3GPP TSG RAN WG1 Meeting 88bis, Available at http://www.3gpp.org/ftp/TSG RAN/WG1 RL1/TSGR1 88b/Report/; April 2017
Research of Frequency Allocation Based on Improved Genetic Algorithm

Bo Li1,*, Yunyi Zhai2 and Fugang Sun1, Kun He1, Richeng Guo1, Cuican Wang1,  
{libo1669@163.com1, zhaiyy@bupt.edu.cn2, gullit_sfg@163.com3}

1Systems Engineering Research Institute, Beijing 100094, China.  
2Key Laboratory of Universal Wireless Communications, Ministry of Education, 
Beijing University of Posts and Telecommunications, Beijing 100876, China.

Abstract. With the development of science and technology, more and more attention has been paid to wireless spectrum allocation technology. In this paper, we propose a improved genetic algorithm which is suitable for spectrum allocation on open space platforms. It uses the method of sequential allocation and solves the problem of channel multiplexing by a new designed interference model under the situation of tight spectrum resources. The simulation results show that the improved genetic algorithm can effectively reduce the total interference of signal and improve the convergence speed of reallocation after the number of devices has changed.

Keywords: Spectrum Allocation; Genetic Algorithm; Sequential Allocation; Interference Model; Convergence Speed

1 Introduction

Wireless spectrum, the foundation of network speed and quality, is an important but nonrenewable resource. Different signals working on the same frequency may produce mass mutual interference. In addition, the signal working frequency is concentrated, and the spectrum utilization is low [1]. With the further development of technology, more wireless devices will be put into use. It will certainly cause congestion of spectrum resources and inefficient use of frequencies. Thus, more and more attention has been paid to spectrum allocation techniques that can effectively reduce interference between devices and improve spectrum utilization.

Spectrum allocation technology is a soft strategy. It is the process to assign operation frequency for radio equipment, which is the ultimate manifestation of frequency management, in order to make sure that the radio equipments operate normally without any mutual interference [2].

Traditional distribution methods include exhaustive search algorithm and order distribution method. However, in recent years, researchers have proposed many intelligent algorithms that simulate natural ecosystem mechanisms which is suitable for spectrum allocation problems. These methods can effectively increase the speed of algorithm. There are Simulate Anneal (SA) [3], Ant Colony Optimization (ACO) [4], Genetic Algorithm (GA) [5], and etc. So far, several techniques have proposed some spectrum allocation schemes based on genetic algorithms:
A GA-Based effective model for channel allocation is proposed by Yasmina El Morabit et al [6]. In [6], they have proposed GA in cognitive radio network to obtain the optimum radio configurations. They focus on spectrum allocation with the quality of service requirements is specified in the inputs. The objectives are combined to one multi-objective fitness function using weighted sum approach so that each objective can be represented by a rank which represents the importance of each objective.

A radio frequency allocation method based on improved GA is proposed by Changsheng Yin et al [7]. In this paper, they design an improved GA to enhance efficiency and real-time of frequency assignment. They combine GA with greedy algorithm when generating the initial population. Meanwhile, a hybridization method was introduced to improve the roulette selection efficiency. These improve the convergence speed of the algorithm.

However, the existing improved GA algorithm does not take into account the problem of channel multiplexing when spectrum resources are limit. To solve this problem, we provide a practical new solution to the frequency allocation problem by improving genetic algorithms. In this algorithms, we considered the power path loss of the signal and the interference tolerance of the devices when designing the fitness function. It can reduce the total interference to the devices in the final allocation result. In addition, the existing improved GA algorithm does not take into account the problem of spectrum reallocation in the actual open space spectrum allocation scenario. In this paper, devices are assigned in groups and we used the sequential allocation method. It can greatly reduce the convergence speed of reallocation.

The rest of this paper is organized as follows. Section 2 outlines the system model of spectrum allocation on open space platforms. Section 3 describes the specific method and principle of the allocation algorithm. Section 4 illustrates the simulation cases and result. Finally, Section 5 concludes the paper.

2 System Model

Fig. 1 illustrates the system model of spectrum allocation on open space platforms. There are multiple platforms in an open space and each platform has multiple devices. All of which are zoned on a same frequency band. If devices on the same platform occupy the same channel, it will cause a lot of mutual interference. To avoid this situation, we directly prohibit devices from same platform occupying the same channel. In addition, comparing with the distance between platforms, the size of platform can be ignored. Thus we assume that each platform is a single point and all the devices are located at the point in the modeling process. Mutual interference of devices on different platforms is the main factor and spectrum allocation is based on minimum total interference of platform. In order to simulate the frequency characteristics of different devices, each device has a different operating frequency band range.

The following parameters are set in this paper. The available frequency band is \((f_L, f_H)\), \(f_L\) as the starting frequency point, \(f_H\) as the ending frequency point. Every \(\Delta f\) is set to a channel and there is \(M\) free channels. A total of \(T\) platforms are distributed randomly on the circular area with a radius of \(R\), with \(N\) devices on each platform.

In this paper, two different scenarios are designed, that is sufficient spectrum scenarios and insufficient spectrum scenarios, by changing the above parameters.
2.1 Scenario 1: Sufficient spectrum resources

In this scenario, we simulate a situation with sufficient channels. The number of channels required by the devices are less than the number of free channels, so that different devices are able to occupy different channels. In theory, we can find an allocation scheme without mutual interference between devices. We get the final allocation scheme through the algorithm and calculate the interference and convergence speed.

After the initial allocation, keep the position of platforms and the number free channels unchanged. Add a new platform with a random position and the past allocation scheme no longer meet new requirement. The spectrum needs to be reallocated and then calculate the interference and convergence speed.

2.2 Scenario 2: Insufficient spectrum resources

In this scenario, we simulate a situation with insufficient channels. The number of channels required by the devices are more than the number of free channels, so that different devices are not able to occupy different channels. In theory, some channels will be occupied by multiple devices simultaneously. Thus, an allocation scheme that allows efficient channel multiplexing is vital. We get the final allocation scheme through the algorithm and calculate the interference and convergence speed.

After the initial allocation, keep the position of platforms and the number free channels unchanged. Add a new platform with a random position and the past allocation scheme no longer meet new requirement. The spectrum needs to be reallocated and then calculate the interference and convergence speed.
3 Improved Genetic Algorithm

3.1 Genetic Algorithm

The genetic algorithm (GA), a search algorithm based on the mechanics of natural selection and genetics, combines a strategy of "survival of the fittest" with a random exchange of information, but structured[6].

![GA flow chart](image)

**Fig. 2.** GA flow chart.

**Fig. 2** illustrates basic flows of the GA. Brief explanations of each step involved of the GA is as follow [8].

Population initialization: Randomly generate a chromosome of length i based on the available matrix defined by the input parameters. An initial population includes j chromosomes.

Fitness: It represents the evaluation of fitness of each chromosomes.

Selection: Selecting the superior chromosome from the population and eliminating the inferior chromosome is called selection.

Crossover: Crossover refers to the operation of replacing and reorganizing part of the structure of two parent chromosomes to generate new chromosomes.

Mutation: Mutation is a random change in some genes on the chromosomes.

3.2 Interference Model

Co-channel interference is the most basic form of interference. It refers to the wanted signal and other unwanted interference occupying same channel. This interference signal may originate from devices of other platform or intentionally applied interference signals.

The past allocation method only distinguished whether interference exist when analyzing co-channel interference. It makes the allocation scheme suitable for scenarios with sufficient spectrum resources to avoid co-channel interference. However, this kind of method does not take into account the problem of channel multiplexing in spectrum tight scenarios. In the process of spectrum allocation with channel multiplexing, it is necessary to keep the total system interference to a minimum. Thus, more complex evaluation criteria for interference values are demanded.

In this paper, we use the total power of the interference signal of each channel as the criterion for evaluating the interference. Assume that there is a device A on a platform and
device B on another platform, and they occupy the same channel. Device A may interferes with B. The signal power of A is \( P_A \) dBm. During signal transmission, the power of the signal will be attenuated. When the signal reaches another platform, the signal power decays to \( P_A' \).

We regard \( P_A' \) as the interference of device A to B. The formula for path loss is:

\[
L = 32.44 + 20 \times \log(D) + 20 \times \log(M)
\]

(1)

Where in (1), \( L \) represents the signal path loss, and its unit is dB. \( D \) represents the distance of signal traveling in space, which can be simplified to the distance between two platforms in our scenarios, and its unit is km. \( M \) represents the signal operating frequency, and its unit is MHz.

According to the system model setting above, there is seventy devices on seven platforms and a device may occupy plural channels. Therefore, a channel may be occupied by more than two devices. The interference experienced by one device is the superposition of the power of other devices. The formula for power superposition is:

\[
I = 10 \times \log \left( \sum_{i=1}^{N} 10^{I_i/10} \right)
\]

(2)

Where in (2), \( N \) represents the number of interfering devices. \( I \) represents the total interference, and its unit is dBm. \( I_i \) represents interference from each device, and its unit is dBm.

We use interference matrix to represents the interference of each device on a platform. The form of the matrix is shown in (3).

\[
\begin{pmatrix}
I_{11} & \cdots & I_{1M} \\
\vdots & \ddots & \vdots \\
I_{N1} & \cdots & I_{NM}
\end{pmatrix}
\]

(3)

Where in (3), \( M \) represents the number of channels. \( N \) represents the number of devices. \( I_{i,j} \) represents total interference of a device in one channel.

In addition, the actual communication devices has a certain anti-interference ability. Small amplitude interference will not affect the transmission of wanted signals. We set the interference threshold in order to simulate the anti-interference ability of devices. When the total interference power is less than \( I_v \), we can account that the interference will not affect the operation of devices. This kind of interference will not affect channel multiplexing. Thus, it can be removed from the interference matrix when calculating reward.
3.3 Sequential Spectrum Allocation

Scenarios for spectrum allocation on open space platforms are more complicated than the ordinary allocation. All devices are naturally classified by each platform. Differences between platforms are issues that need to be analyzed. Besides, each platform is not connected to the network at the same time, but in a certain order. Therefore, we carried out sequential spectrum allocation in order to reflect the differences of each platform.

The ordinary GA is limited by population and length of chromosome. It has a problem with early convergence, especially in complex scenarios. In these scenarios, the ordinary GA often fails to get the global optimal solution. Sequential spectrum allocation can solve this kind of problem. It splits a long chromosome into segments to reduce the length, improving gene richness and ability to find needed genes. Therefore, sequential spectrum allocation can effectively improve the ability to find global optimal solutions in complex scenarios and reduce the amount of calculations per generation.

However, sequential spectrum allocation brings a new problem with the increasing number of generations. We set a new parameter $T_n$ to reduce generations. In one GA process, if fitness value has not changed for $T_n$ generations, this GA process ends and the next GA process starts. Besides, if the fitness value meet the upper limit $\text{Maximum fitness}$, this GA process ends too. The methods can greatly reduce the number of useless generations, especially in the first few GA processes.

In addition, sequential spectrum allocation can greatly reduce the convergence speed of reallocation. Assume that the first allocation has ended and there is a platform that changes its position. Traditional GA requires a new spectrum allocation for all platforms. However, new allocation scheme in this paper has completely different reallocation methods. Due to the independence of the device allocation process on each platform, it allows use to directly read the previous allocation results and reallocate spectrum for the new platform based on the past results.

The basic steps of the sequential spectrum allocation are:

Algorithm 1 Sequential Spectrum Allocation;

**Input:** $TT$, $\text{Maximum fitness}$, $T_n$

**Output:** Allocation scheme matrix $A$ and fitness of the best gene of each generation $\text{Max} - \text{Reward}$

1. $T = 150$;
2. randomly initialize chromosomes;
3. for $i = 1 : TT$ do
4. $T_s = 0; t = 0; i = i + 1$;
5. while ($t < T$) do
3.4 Algorithm Convergence Criterion

Improved Genetic Algorithm is a kind of optimization problem. The allocation scheme is to find the optimal solution under the current conditions. We use the sum of the interference from each channel as the fitness function in the genetic algorithm and define a standard named Max-Reward to evaluate it.

\[
\text{Max-Reward} = \max \left\{ \text{Maximum fitness} - \sum_{m=1}^{\mathcal{M}} \sum_{n=1}^{\mathcal{N}} \left( I_{m,n}(i) \cdot a_{m,n}(i) \right) \right\}
\]

where \( I_{m,n}(i) \) is the allocation schemes under the constraints of \( T_{m,n} \), \( \mathcal{M} \), \( \mathcal{N} \) and other parameters. The product of \( I_{m,n}(i) \) and \( a_{m,n}(i) \) represents total interference to the devices. During the improved Genetic Algorithm, our purpose of the algorithm is to find an allocation scheme that maximizes Max-Reward.

Since the first generation \( t = 1 \), we define one selection, crossover, and mutation as one generation. The algorithm converges when all platforms are allocated and Max-Reward has not changed multiple generations in a row. Convergence speed is the number of generations required. We think that the less generations the algorithm needs, the faster the convergence speed.
4 Simulation

4.1 Scenario 1: Sufficient spectrum resources

In this scenario, we set a simulation where the spectrum resources are sufficient. This scenario is described above.

The unchanged parameters are shown in Table 1.

First, we make an initial allocation with four platforms that need no more than forty channels. The comparison result is shown in the Fig. 3.

Table 1. Parameters of GA and improved GA in Scenario 1.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>GA</th>
<th>improved GA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 Scenario Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area radius/R(km)</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Spectrum range(MHz)</td>
<td>850-1000</td>
<td>850-1000</td>
</tr>
<tr>
<td>Channel bandwidth/Δf (MHz)</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Number of free channels/M</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Number of platforms/TT</td>
<td>4/5</td>
<td>4/5</td>
</tr>
<tr>
<td>Number of devices per platform</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td><strong>2 Algorithm Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>Generation/t</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>Maximum fitness</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>Crossover rate</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Mutation rate</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Interference threshold(dBm)</td>
<td>-60</td>
<td>-60</td>
</tr>
</tbody>
</table>
As the result illustrates, both GA and improved GA reach maximum fitness and figure out the global optimal solution. However, the convergence speed of improved GA is much faster than ordinary GA. Therefore, the improved GA can significantly reduce the number of iterations in the situation where the spectrum resources is sufficient.
Then, add a new platform and reallocate the spectrum. There are fifty devices and fifty free channels. The improved GA is able to analyze the past allocation scheme and only allocate new platforms based on the past solutions. However, the ordinary GA needs to reallocate all the five platforms. The comparison result is shown in the Fig. 4.

As the result illustrates, the improved GA reaches maximum fitness and figures out the global optimal solution but the ordinary GA fails. Besides, the convergence speed of improved GA is faster than the ordinary GA.

The improved GA significantly reduced length of chromosomes, which gives every excellent gene more chance to inherit. However, ordinary GA, limited by length of chromosomes, has converged before finding the optimal solution. In addition, the improved also significantly reduced number of platforms that need to reallocate, which improved convergence speed.

4.1 Scenario 2: Insufficient spectrum resources

In this scenario, we set a simulation where the spectrum resources are insufficient. This scenario is described above.

The unchanged parameters are shown in Table 2.

First, we make an initial allocation with six platforms. In this situation, interference is inevitable. The purpose of the allocation is to reduce interference and multiplex channels. The comparison result is shown in the Fig. 5.
Table 2. Parameters of GA and improved GA in Scenario 2.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>GA</th>
<th>improved GA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1 Scenario Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Area radius/R(km)</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Spectrum range(MHz)</td>
<td>850-1000</td>
<td>850-1000</td>
</tr>
<tr>
<td>Channel bandwidth/ $\Delta \nu$ (MHz)</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Number of free channels</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Number of platforms</td>
<td>6/7</td>
<td>6/7</td>
</tr>
<tr>
<td><strong>2 Algorithm Parameters</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Generation/t</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>$\tau_n$</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>Maximum fitness</td>
<td>2000</td>
<td>2000</td>
</tr>
<tr>
<td>Probability of crossover</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Probability of mutation</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Interference threshold(dBm)</td>
<td>-60</td>
<td>-60</td>
</tr>
</tbody>
</table>

Fig. 5. The max-reward of each generation in the initial allocation.

As the result illustrates, neither ordinary GA nor improved GA reach maximum fitness. This complex scenario increases generations of the improved GA. In return, the improved GA provides an ability on greatly reducing interference when multiplex channels.

Then, add a new platform and reallocate the spectrum. Spectrum resources become more scarced. The comparison result is shown in the Fig. 6.
As the result illustrates, both ordinary GA and improved GA scheme are suffer from a huge interference. However, the improved GA has better max-reward and convergence speed.

The ordinary GA does not take space factor into account but the improved GA does. Therefore, especially in the complex scenario where the spectrum resources are insufficient, improved GA can effectively reduce interference between co-channel signals and increase the speed of reallocation.

5 Conclusion

In this paper, a new spectrum allocation on open space platforms based on improved Genetic Algorithm has been proposed. The major differences are sequential spectrum allocation and interference model using in the improved Genetic Algorithm. Finally, the simulation shows that the algorithm is more efficient to reduce co-channel interference and increase convergence speed.
Acknowledgement

This work was supported in part by the National Key Research and Development Program of China under Grant (2019YFB1804400), in part by the National Natural Science Foundation of China under Grant (61801052), and in part by the Beijing Natural Science Foundation under Grants (4202046).

References

Transfer Learning Based Screen Defect Classification

Yilei Li, Chengyuan Li, Yifan Zhang, Shuo Chang, Zixuan Wang and Fan Zhang
{ liyilei@bupt.edu.cn, lcyuan@bupt.edu.cn, zhangyf@bupt.edu.cn, changshuo@bupt.edu.cn,
  heiyouwangzi@bupt.edu.cn, zhangfan2015@bupt.edu.cn }

Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunication, Beijing, P.R.China, 100876.
School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing, P.R.China, 100085.

Abstract. For the screen defect classification, human inspectors and traditional machine learning algorithms are inefficient and inaccurate. Convolutional neural network (CNN) driven by data are feasible solutions. However, real training images are limited in the industrial scenario, which causes overfitting. Hence, in this paper, a novel learning based method is proposed for defect classification, which is based on the CNN. To alleviate the problem of limited images, two strategies are introduced into model learning. For the training data, a data generation module is implemented to enlarge the training dataset. For the CNN model learning, transfer learning is applied in the whole training process. To verify the proposed method, various experiments are carried out. The results indicate that our screen defect classification model achieves superior performance.

Keywords: Defect Classification, Transfer Learning, CNN, Generation Module.

1 Introduction

In the field of industrial manufacturing, defect classification for liquid crystal display (LCD) screens is a hotspot. In general, there are different kinds of defects appearing on LCD screens [1], such as point defects, mura defects, leakage defects, etc. Some of the examples are shown in Figure 1.
To determine whether an LCD screen is defective or not, human inspectors are employed on the production line. But there are three drawbacks for employing human inspectors: The standards of defects are difficult to unify; The efficiency of human inspectors is low; The cost of employment is high.

Since using manual classification of screens has many disadvantages, machine learning methods are proposed [1-5]. Fundamentally, general method is mainly divided into two steps. First, the features of samples are extracted with common feature extraction algorithms, for example, HOG [4], Harr [2], and Gabor [5], etc. Second, a classification model is applied to return a final result. Generally, SVM is employed as the classification model [6][7]. However, the features extracted are not specifically designed for defect classification.

Nowadays, deep learning methods can extract features directly. However, deep learning models are data driven. In the industrial scene, the dataset is limited. Therefore, in this paper, we propose a novel method — Transfer Learning based mobile screen Defect Classification (TLDC) to tackle this issue. The framework of TLDC is depicted in Figure 2. With exemplar images as input, the generation module can produce many defective samples. Then, we use a data pre-processing module to exclude the backgrounds of the LCD screens and get regions of interest (ROI). Finally, the processed samples are sent to the transfer learning module to retrain a new binary classifier, where two classes are defective images and no-defective images.

Overall, TLDC is designed for industrial screen defects classification. TLDC adopts two strategies: one is the automatic generation strategy, the other is transfer learning. With the help of the generation module and transfer learning, our model is able to avoid overfitting due to insufficient data.

The main contributions of this paper are summarized as follows. A generation module is proposed. It can imitate real samples to generate similar defective samples and enlarge dataset. A pre-processing module is introduced to remove the influence of the black background.
Transfer learning is applied in our scheme when images are insufficient. Various comparative experiments are carried out to verify the effectiveness of TLDC.

The rest of the paper is organized as follows. Section 2 reviews the related work including traditional machine learning methods, deep learning models and transfer learning. The proposed TLDC approach is presented concretely in Section 3. The experimental results are given in Section 4, and finally some conclusions are drawn in Section 5.

![Fig. 2. The framework of TLDC.](image)

## 2 Related Work

### 2.1 Traditional Machine Learning Methods

Harr is a traditional machine learning method for visual object classification [22], but the impact of uneven illumination in industrial manufacturing cannot be well balanced [2]. Navneet Dalal et al. proposed HOG descriptors, which can significantly outperform existing feature sets for human classification [4]. However, HOG descriptors are extremely sensitive to noise, which affects the accuracy of industrial classification [8]. Different from the HOG and Harr operator, the Gabor filter has better performance in extracting the information of the frequency domain. But it performs unstably when abrupt changes happen [5]. After feature extraction, a classifier is commonly employed to classify features, such as SVM [3]. However, SVM performs inferiorly when large-scale samples are trained.

Even though the traditional machine learning methods can accomplish defects classification in place of humans, they still has two drawbacks. First, the process of feature extraction
and classification is relatively independent. Second, its performance is unsatisfactory when the dataset is large.

2.2 Deep Learning Models

Deep learning methods train a multi-layer neural network model usually using massive labelled data to obtain feature maps. Specifically, deep learning methods can obtain features directly in various scenarios, for example, natural language processing (NLP) [9], medical image recognition [10], and industrial detection [11], etc. CNN is a classical deep learning model widely utilized in image classification [21] due to its superior performance and robustness [20]. Particularly, CNN consists of convolutional layer, pooling layer, and fully connected layer (FC) [12].

2.3 Transfer Learning

In 2013, Nitish Srivastava [13] et al. proposed a way of improving classification performance for classes which have few training examples. The key idea is to discover classes which are similar and transfer knowledge among them. Later, some researchers discovered that initializing a network with transferred features from almost any number of layers can produce a boost to generalization [14]. On the whole, transfer learning is employed for promoting the model performance in multiple tasks, such as sign language [15], face sketch recognition [16], and so on.

3 Model Design

The objective of our work is to judge whether an LCD screen is defective or not. Therefore, TLDC method is proposed for this purpose. The framework of the proposed transfer learning based method is shown in Figure 2, and the details of generation module, ROI module and transfer learning strategy are elaborated as follows.

3.1 Generation Module

The generation module is able to produce three types of screen defects: point, mura and line defects. Specifically, the generation principle of point and mura defects is similar. Practically, we select a point randomly in the ROI region as the center and divide the neighborhood of this center into five levels. Figure 3 shows the definition of the neighborhood of the point defect and mura defect.

For any point in this neighborhood, its pixel value is modified according to the following steps:
First, the RGB pixel value of the center point \((x, y)\) is calculated. Second, in the \(i\)th \((i\) ranges from 0 to 4) level neighborhood, the pixel value in this area increases or declines by a given offset on the basis of the calculated value in the first step.

For point defects, another step is added. The image needs to be operated with linear filtering [17]. In addition, the generation of line defects varies from other types. A chosen point as the starting point of the line extends for a length in a random direction, and the pixel value of the generated line is a fixed value.

![Fig. 3. Neighborhoods of generated defects.](image)

![Fig. 4. Comparisons of generated defects and real defects.](image)
Due to the data augmentation by generation module, the limited training dataset is enlarged. Figure 4 displays the contrast between the generated defective images and the real ones.

### 3.2 ROI Module

In order to remove the black background, we find the corner spots of the screen by the Hough transform, and then extract the ROI area of the origin image by the perspective transformation. The flowchart of the ROI module is demonstrated in Figure 5.

![Fig. 5. The flowchart of getting ROI.](image)

Next, each step in Figure 5 is introduced separately. Hough transform is a commonly applied feature extraction technique in image processing [18]. Hough transform calculates the local maximum value of the cumulative result in $\theta$-$\rho$ parameter space. The principle of the Hough transform is to convert the linear equation in the Cartesian coordinate system to the linear equation in the polar coordinate system. The detailed process of the Hough transform is listed as follows. Accumulator $H$ is initialized to all zeros. For each edge point $(x, y)$ in the image, when $\theta$ ranges from $-90^\circ$ to $180^\circ$, $\rho$ and $H(\theta, \rho)$ are calculated by the following equations:

\[
\rho = x \sin \theta + y \cos \theta. \tag{1}
\]
\[
H(\theta, \rho) = H(\theta, \rho) + 1. \tag{2}
\]

where $\rho$ is the distance from the straight line to the origin, and $\theta$ is the angle between the straight line and the x-axis.

The value of $(\theta, \rho)$ is found when $H(\theta, \rho)$ is a local maximum. The detected line in the image is given by equation (1).

Because the LCD screens in our task are rectangular, we match four boundaries by Hough transform, and four intersections of boundary lines are the corners of the image. According to the corner information, the ROI area of the screen is obtained. Finally, the screen is projected into a fixed-size rectangular area through perspective transformation to obtain a ROI area of uniform size.

### 3.3 Transfer Learning Strategy

The transfer learning strategy has the ability to apply knowledge learned in the source domain to the target domain. In this paper, the source domain is a new dataset mixed with
generated defective images and real ones, and the target domain is a dataset which makes up of real samples. Mixed images are the input of the VGG-Net [19] to obtain a pre-trained model.

Then we transfer the feature information to the classification model with fewer samples. Specifically, the convolutional and pooling layers are transferred to the classification model. After the transfer of the convolutional and pooling layers, the classification layer needs to be reconstructed. This paper mainly retrains the FC layer of the model transferred from the source domain. Since the screen classification task is a binary classification, the parameters of the last FC layer are modified to be two. And the last FC layer is retrained. Finally, the output is passed through the SoftMax classifier as the final classification output layer. Figure 6 illustrates the process of transfer learning in this paper.

\[ l = - [y \log(p) + (1 - y) \log(1 - p)] \]  

(3)

Where \( y \) is the label of the sample, the positive class is 1 and the negative class is 0, \( p \) is the probability of the positive class.

4 Experiments

4.1 Dataset
The images used in this paper are taken by industrial cameras mounted on the production line. Each image is annotated by experienced workers. In our experiment, 168 defective images and 1032 no-defective images are collected, which is presented in Table 1.

Table 1. Dataset Configuration.

<table>
<thead>
<tr>
<th>Images</th>
<th>Training Set</th>
<th>Valication Set</th>
<th>Test Set</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defective Images</td>
<td>860</td>
<td>86</td>
<td>86</td>
<td>1032</td>
</tr>
<tr>
<td>No-defective Images</td>
<td>140</td>
<td>14</td>
<td>14</td>
<td>168</td>
</tr>
<tr>
<td>Total Numbers of Set</td>
<td>1400</td>
<td>100</td>
<td>100</td>
<td>1200</td>
</tr>
</tbody>
</table>

4.2 Experimental Setup

This experiment is implemented on a Dell R7300 server with an NVidia 1080Ti GPU. All the codes in this experiment are based on the Pytorch and python3. For the algorithm optimization, the learning rate is 0.01, and the epoch is 100. TLDC is updated by stochastic gradient descent (SGD) [23].

4.3 Results and discussion

Pretraining and no-pretraining. In order to verify whether the VGG-Net based model needs pretraining, we set three models for comparison. Their accuracy on the test set is shown in Table 2. The first model in Table 2 refers to the VGG-Net with primitive parameters. The second model is pretrained with ImageNet dataset [12]. The last one is pretrained with mixed dataset. The mixed dataset consists of generated and real defective images. The pretrained VGG-Net with mixed dataset outperforms the VGG-Net about 14% on the test set. However, VGG-Net pretrained on ImageNet dataset [12] performs worse than others. In conclusion, model pretrained on a mixed dataset performs the best.

Table 2. Pretrain and no-pretrain.

<table>
<thead>
<tr>
<th></th>
<th>VGG</th>
<th>VGG(ILSVRC’14)</th>
<th>VGG(Mixed Dataset)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy(test)</td>
<td>83%</td>
<td>73%</td>
<td>97%</td>
</tr>
</tbody>
</table>

The mixing ratio. The comparison results of different mixing ratios are presented in Figure 7. It can be seen that when the mixing ratio of real to generated images is 1:15, model performs best. Note that, while the proportion of real samples is too high or too low, the accuracy of the model declines. Hence, we can conclude that dataset with appropriate mixing rate can help models gain improvement.
Fig. 7. Accuracy of different mixing ratios. From left to right, the mixing ratios of the real images to the generated ones are 0:1, 1:20, 1:15, 1:10.

In summary, selecting a mixture of real and generated samples of appropriate proportions for transfer learning can help model achieve high accuracy in industrial tasks.

5 Conclusion

In this paper, a new data generation strategy and transfer learning strategy have been introduced for defect classification. Compared with other machine learning methods, our method can solve the issue of limited dataset and learn from domain knowledge. In view of proposed methods, our classifier achieves optimal performance. Although this paper focused on screen defects, the methods mentioned in this paper can be carried out in other similar tasks. In future work, the problem of industrial classification and complex scenes will be further explored.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant (61801052), in part by the National Key Research and Development Program of China under Grant (2019YFB1804400, 2018YFF0301202), and in part by the Beijing Natural Science Foundation under Grants (4202046).

References


Improved Loss Function for Defect Detection of Mobile Phone Screen

Zixuan Wang, Chengyuan Li, Yifan Zhang, Shuo Chang, Yilei Li and Fan Zhang
{beiyouwangzi@bupt.edu.cn, lcyuan@bupt.edu.cn, zhangyf@bupt.edu.cn, changshuo@bupt.edu.cn, livilei@bupt.edu.cn, zhangfan2015@bupt.edu.cn}

Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunication, Beijing, P.R.China, 100876.
School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing, P.R.China, 100085.

Abstract. Due to the excellent feature learning and representation capabilities of deep learning, the method based on deep learning for mobile phone screen defect detection is gradually being applied to industrial detection. Nowadays, the cross-entropy loss commonly used in deep learning only focuses on the differences between different classes with less intra-class differences. It leads to poor discriminative ability of the model when the similarity between training samples is high. The contrastive loss reduces the intra-class variations and can distinguish between similar objects from different classes. Given the above analysis, we propose a Siamese network for mobile phone screen defect detection (SMSDD) using combined contrastive loss and cross-entropy loss, thereby enhancing the discriminative ability of model. Numerical results show that SMSDD achieves comparable performance.

Keywords: Defect Detection, Deep Learning, Cross-entropy Loss, Contrastive Loss, Siamese Network.

1 Introduction

Nowadays, mobile phones have become an indispensable part of daily life. The transparent glass screen of the mobile phone is the main window of human-machine interaction. If there are defects on the phone screen, it will affect the user experience directly. Therefore, to satisfy the high-quality and high-resolution requirements of the phone screen, screen defect detection is necessary.

Till now, the methods of mobile phone screen detection mainly rely on manual detection or traditional image processing. Traditional manual detection has shortcomings such as low speed, low accuracy and inconsistent detection standards, etc. Therefore, it can not adapt to the current industrial detection requirements [1]. Although methods based on image processing are faster, more reliable, and more accurate than manual detection, the feature extraction capability is unstable. Due to influences of illumination, camera noise, screen texture, etc., these methods may result in inaccurate detection results [2].

With the deepening of the research on Convolutional Neural Network (CNN), the method based on CNN has achieved superior results on various computer vision tasks [3-5],[8]. And methods based on CNN for mobile phone screen defect detection have become a good solution
The loss function is an important part for CNN. On one hand, it controls the ultimate goal of network optimization. On the other hand, it determines the learning direction of the network. The cross-entropy loss [9] is used to classify the images to a certain class generally. It mainly penalizes the inter-class loss, and there is no analysis of intra-class differences. In the screen defect detection, the area containing the defect accounts for a little proportion of the screen, mainly less than 0.1%. This leads to the similarity of some samples in defective and non-defective classes is high. Meanwhile, as shown in Figure 1, the defects have large differences in shape and small appearance similarity. So, such variations within the defects could overwhelm the variations within class differences and make screen defect detection challenging. This requires the model to consider the intra-class compactness while ensuring the inter-class separability. For the purpose, the contrastive loss [10] which mainly concentrates on reducing intra-class differences can meet the demand in the screen defect detection. So, we add the contrastive loss as a part of loss function of the network.

Fig. 1. Six different types of screen defects. They are various in area, shape, pixel value, etc.

In this paper, the mobile phone screen defect detection is transformed into binary classification, which only contains no-defective images and defective images. A Siamese network is proposed for mobile phone screen defect detection (SMSDD), which is supervised by the improved loss function that combines the cross-entropy loss with contrastive loss. Extensive experiments are conducted to prove the effectiveness of the proposed method on Mobile-Screen-Defect dataset.

The rest of this paper is organized as follows. Section 2 reviews the related work. In Section 3, details of the proposed SMSDD and the improved loss function are presented. In Section 4, the implementation details in model training are presented. The numerical results are provided in Section 5. Finally, Section 6 concludes the paper.
2 Related Work

2.1 Traditional Image Processing For Screen Defect Detection

Kim [11] et al. proposed a method that detected defects by reducing the threshold level of gray unevenness, which can detect most types of defects other than line defects. Lee and Yoo [12] et al. used the background subtraction method, with a two-dimensional curve to estimate the image background, to detect mura. With Fourier transform to eliminate the background interference of defects, Tsai and Tseng [14] et al. utilized a simple method of threshold processing to complete the image segmentation. In short, traditional image processing methods have high time complexity and weak robustness.

2.2 Convolutional Neural Network For Classification

With the rapid development of deep learning, many works for large-scale image classification have been proposed [13]. Krizhevsky [3] et al. introduced deep CNN into image classification for the first time. Simonyan and Zisserman proposed a very deep CNN in [15], and the performance was significantly improved by increasing the depth and using small convolutional filters of size 3×3. The loss function commonly used in the classification system is the cross-entropy loss [9], which mainly penalizes the inter-class loss. However, defective and no-defective mobile phone screen samples have similar features. The variations within class differences are not enough obvious, where the result of classification is easy to be disturbed. So, only using the cross-entropy loss cannot classify samples accurately. Therefore, it is not a wise choice to use a general image classification system for the mobile phone screen defect detection.

Fig. 2. The framework of SMSDD.
2.3 Contrastive Loss For CNN

The contrastive loss [10] is mainly used in Siamese framework. It can effectively reduce the intra-class variations by pulling feature vectors from the same instance together. Based on the contrastive loss, Yi Sun [16] et al. increased the inter-personal variations by drawing features extracted from different identities apart, while reduced the intra-personal variations by pulling features extracted from the same identity together in face recognition task. S. Chopra [17] et al. enhanced the robustness of the network to the nonlinear geometric transformation of the data by controlling the "semantic" distance between the paired data. Defect samples of mobile phone screen usually have great differences in area, pixel value, etc. Intra-class differences have great impact on classification [18] and we need to consider the intra-class compactness. In this case, the contrastive loss is helpful to enhance the classification ability of the model for the mobile phone screen defect detection.

3 The Proposed SMSDD Scheme

In this section, details of the proposed SMSDD are firstly described. Then, a brief introduction about contrastive loss and cross-entropy loss is given. Finally, We visualize the effects of the improved loss function.

3.1 Framework

The framework of SMSDD is depicted in Figure 2. The proposed network consists of two branches, and they share the same structure and weight. As shown in Figure 2, the backbone of Siamese Network is VGG16 [13] without fully connected (FC) layer. Samples \( a \) and \( b \) stand for input pair-data to update the Siamese network and the size is 230×110×3. The class label is considered to be positive if there are no defects in the screen:

\[
 y = \begin{cases} 
 0, & \text{no - defective} \\ 
 1, & \text{defective} 
\end{cases} 
\]  

(1)

The pair label is determined by whether \( a \) and \( b \) belong to the same class:

\[
 Y = \begin{cases} 
 1, & \text{the same class} \\ 
 -1, & \text{other} 
\end{cases} 
\]  

(2)

The loss function of the network is shown on the right side of Figure 2. For fully connected layer, it has two kinds of outputs, corresponding to the defective or no-defective image. The output of fully connected layer is used to calculate the cross-entropy loss. Under the supervision of the pair label, the outputs of Siamese Network are used to calculate the contrastive loss. Finally, the sum of the cross-entropy loss and contrastive loss is used to propagate gradients for updating the network parameters.

3.2 Loss Function
**Cross-entropy Loss.** Cross entropy loss is mainly used for classification tasks. In the screen defect detection task, there are two classes: no-defective and defective images. The corresponding cross-entropy loss is:

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N} [y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log (1 - \hat{y}^{(i)})]$$  \hspace{1cm} (3)

where $\hat{y}$ is the score of a single exemplar-candidate pair and $y$ is the class label. $N$ represents the total number of samples in the training set.

The cross-entropy loss mainly focuses on increasing the margin among candidates from different classes. Namely, there is no consideration about the intra-class variations, where the prediction result of the model is greatly reduced.

**Contrastive Loss.** We want the feature representations of defects are close enough for positive pairs, and far away at least by $m$ for negative pairs, where $m$ is a hyper parameter, which means that we only consider paired data belonging to different classes with Euclidean distance less than $m$. Therefore, we employ the contrastive loss:

$$L_2 = \frac{1}{2N} \sum_{k=1}^{N} YD_w^2 + (1 - Y)\max (m - D_w, 0)^2$$  \hspace{1cm} (4)

Where

$$D_w(X_1, X_2) = \|X_1 - X_2\|_2 = \left(\sum_{i=1}^{P} (X_{1i} - X_{2i})^2\right)^{\frac{1}{2}}$$  \hspace{1cm} (5)

$D_w$ is the Euclidean distance between the sample feature $X_1$ and $X_2$, where $P$ is the eigenvector dimension. The variable $Y$ stands for the pair label. When $Y = 1$, the two candidate samples ($X_1, X_2$) are from the same class, and the loss function in equation (4) becomes equation (6):

$$L_S = \frac{1}{2N} \sum_{k=1}^{N} YD_w^2$$  \hspace{1cm} (6)

So, if the Euclidean distance between the two samples in the feature space is large, the loss will be large. This means that the current model does not meet the requirement of intra-class compactness and needs to continue training. Otherwise, when $Y = 0$, the loss function is:

$$L_D = (1 - Y)\max (m - D_w)^2$$  \hspace{1cm} (7)

With the decrease of Euclidean distance $D_w$, the loss will be larger, which meets the requirement of inter-class separability.

**3.3 Combined Loss**

By integrating the contrastive loss into cross-entropy loss, a new loss function to supervise the SNMSDD is:

$$L = L_1 + \beta L_2$$  \hspace{1cm} (8)
In equation (8), the hyperparameter $\beta$ is a weight to balance the effect between the cross-entropy loss and contrastive loss.

![Figure 3](image.png)

**Fig. 3.** Visualization of effects of the original and improved loss function.

Figure 3 visualizes effects of the original and improved loss function. For the sample C (black), although it belongs to class B (red), it has some same characteristics as class A (green). If we use the original loss function, its position in feature space is close to class A, which makes it difficult to classify. In Figure 3(b), the improved loss function is used to control the distance between A and B in a certain range, which draws the sample C away from class A, making the model easy to distinguish similar samples.

4 Implementation Details

The proposed algorithm SMSDD is implemented in python on a Dell R7300 server with an NVidia 1080Ti GPU. The experiment is performed on Linux with a processor of Intel (R) Xeon (R) CPU E5-2620 v3 @ 2.40GHz and 21G RAM. The learning rate for model training is 0.01 in all epochs.

5 Experiments

5.1 Dataset and Preprocessing

**Dataset.** Screen images are taken by industrial cameras mounted on the production line as experimental data, and each image will be read and marked by experienced workers. We collected 1200 images, including 168 images of defective products and 1032 images of no-defective products.

**Preprocessing.** Due to the extensive black background in collected images, most of black border areas will be extracted during the feature extraction process, and the features of the mobile phone screen will be ignored. To avoid the influence of the background in the training,
the Hough transform [19] is used to find the edge points and the perspective transform is used to correct the region of interest (RoI) of screens. After the preprocessing, 80% of images are used as the training set for model training, 10% are used as the validation set, and the remaining 10% are used as the test set.

![Original Image](Image1.png) ![Final ROI](Image2.png)

**Fig. 4.** The process of getting RoI.

**Table 1.** The loss function of different models.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Loss Function</th>
<th>Hyper-parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMSDD-A</td>
<td>Cross-entropy Loss</td>
<td>$\beta = 0$</td>
</tr>
<tr>
<td>SMSDD-B</td>
<td>Cross-entropy Loss + $\beta$ Contrastive Loss</td>
<td>$\beta = 0.1$</td>
</tr>
<tr>
<td>SMSDD-C</td>
<td>Cross-entropy Loss + $\beta$ Contrastive Loss</td>
<td>$\beta = 0.01$</td>
</tr>
<tr>
<td>SMSDD-D</td>
<td>Cross-entropy Loss + $\beta$ Contrastive Loss</td>
<td>$\beta = 0.001$</td>
</tr>
<tr>
<td>SMSDD-E</td>
<td>Cross-entropy Loss + $\beta$ Contrastive Loss</td>
<td>$\beta = 0.0001$</td>
</tr>
</tbody>
</table>

**5.2 Evaluation**

To evaluate the performance of the proposed model, five contrastive experiments are carried out, as shown in Table 1.

**Loss Evaluation.** The comparison of training loss of different models are shown in **Figure 5**. SMSDD-A only use the cross-entropy loss to update the network parameters. Because it only penalizes the inter-class loss and has no consideration about the intra-class variations, the training result of SMSDD-A is inferior. Finally, the loss only converges to 0.23.

When $\beta = 0.1$, the cross-entropy loss plays a small role in model updating. This makes SMSDD-B unable to increase the margin among candidates from different classes correctly. According to the experimental result, SMSDD-B is the worst. Due to the appropriate values of $\beta$, SMSDD-C and SMSDD-D have a quick convergence speed and the final loss is low. They also perform well in validation and test sets, which can meet industry standards. Because $\beta$ in SMSDD-E is small, the contrastive loss has less effect on the training process. Hence, the result of SMSDD-E is worse than the results of SMSDD-C and SMSDD-D, but better than the result of SMSDD-A.

**Accuracy Evaluation.** The results of accuracy based on different loss functions are in Table 2. Without the bells and whistles, our proposed SMSDD outperforms the original algorithm on validation and test sets in accuracy. Comparing to the model using only cross-entropy loss
(SMSDD-A), the SMSDD-C has made gains about 11% on validation set and 10% on test set in accuracy as the best choice of hyper-parameter $\beta$, respectively. This is due to the benefit of the contrastive loss, which is added to the loss function for updating the network parameters.

![Comparison of different models](image)

**Fig. 5.** Comparison of training loss of different models.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Validation Set(%)</th>
<th>Test Set(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMSDD-A</td>
<td>85</td>
<td>83</td>
</tr>
<tr>
<td>SMSDD-B</td>
<td>79</td>
<td>74</td>
</tr>
<tr>
<td>SMSDD-C</td>
<td>96</td>
<td>93</td>
</tr>
<tr>
<td>SMSDD-D</td>
<td>95</td>
<td>91</td>
</tr>
<tr>
<td>SMSDD-E</td>
<td>91</td>
<td>88</td>
</tr>
</tbody>
</table>

**Table 2.** Accuracy of different models on validation and test sets.

6 Conclusion

In this paper, we introduce deep learning into mobile phone screen defect detection. Besides, a new loss function that combines the cross-entropy loss with the contrastive loss is proposed. The new loss considers the intra-class compactness and inter-class separability. Without the bells and whistles, the proposed algorithm outperforms the method supervised by
cross-entropy loss. In the future work, we plan to introduce the new loss function into multi-classification tasks.

**Acknowledgments.** This work was supported in part by the National Natural Science Foundation of China under Grant (61801052), in part by the National Key Research and Development Program of China under Grant (2019YFB1804400, 2018YFF0301202), and in part by the Beijing Natural Science Foundation under Grants (4202046).

**References**


Detecting Urban Functional Area Based on Clustering on Base Station Data

Yuan Cheng¹, Yahong Wang², and Ronghua Chi³
{changuang7@sina.com¹, wyh811619@126.com², chironghua@126.com³}

College of Computer Science and Technology, Harbin University of Science and Technology¹,
School of Applied Science and Civil Engineering, Beijing Institute of Technology, Zhuhai²
College of Computer and Information Engineering, Heilongjiang University of Science and Technology³

Abstract. Detecting urban functional area can contribute to rational planning of urban functions, which is conducive to urban construction. Because the call data correlates with the users' locations and their behavior characteristics, we use the uncertain clustering method to find the aggregate characteristics of the users' calling behaviors, based on the data produced by base station providing the communication services, and then analyze the related land functions. Finally, we will verify the effectiveness of the proposed method through experimental results.

Keywords: Urban functional area, uncertain clustering, base station data, calling behaviors.

1 Introduction

Big data technology and machine learning methods have provided strong technical support for the development of smart city, which mainly focuses on mining knowledge from various types of data (such as taxi trajectory, call detail records, etc.) to assist in city construction. Detecting urban functional areas from user call detail records or travel behaviors contributes to urban planning. The obtained results are more effective, accurate, and more conducive to discovering the relationship between land functions and human activities, compared with traditional methods, such as field investigations or remote sensing technologies. That is because call record data or trajectory data could reflect human activities in a certain place. Some existing studies choose to judge the function of user calling locations through
discovering the characteristics of the users’ calling behaviors based on the users’ call record data[1-4]. While some studies choose to judge the function of the human movement area by analyzing the characteristics of the users’ moving behaviors based on their movement trajectory data [5-8]. However, because that the data used for analysis are mostly related to personal behavior, it may involve user privacy protection issues. The data mining methods used to discover human behavior characteristics mainly include two types, one is classification method [9-11], and the other is clustering method [12-14]. Discovering urban functional areas is mainly to explore the relationship between human behavior characteristics and regional functions, so it is not necessary to conduct analyzing based on detailed data representing individual behaviors. The data can also be appropriate as long as it could be used to analyze the aggregate characteristics of human behaviors related to land functions. Apart from that, the classification method is not suitable for the case that there is no labeled data.

Therefore, for the purpose of privacy protection, we will propose a method to analyze the aggregation characteristics of the human calling behaviors based on the statistical data about the number of persons served by the base stations in the city, and then to explore the related urban functional areas.

2 Data Processing

We will use the data collected from the base stations of Harbin providing the services of communication. Each data record represents the number of service persons monitored by each base station in the city at the collection time point. Its specific information include the geographic location of the base station, the collection time, and the number of people served, and the base station ID. There are about 3,096 base stations and about 130 million records in the dataset. The main attributes of the original data and some data examples are shown in Table 1.

<table>
<thead>
<tr>
<th>Number of users in DCH</th>
<th>Number of users in DSCH</th>
<th>Collection time</th>
<th>The wireless network controller</th>
<th>Affiliated region</th>
<th>Longitude</th>
<th>Latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>18</td>
<td>2013/2/4 1:00</td>
<td>HRBRNC04</td>
<td>Nangang</td>
<td>126.6833</td>
<td>45.75097</td>
</tr>
<tr>
<td>24</td>
<td>20</td>
<td>2013/2/4 1:01</td>
<td>HRBRNC04</td>
<td>Nangang</td>
<td>126.6833</td>
<td>45.75097</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>29</td>
<td>20</td>
<td>2013/2/4 2:00</td>
<td>HRBRNC04</td>
<td>Nangang</td>
<td>126.6833</td>
<td>45.75097</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>

Table 1. Examples of base station data.
However, when using the data in Table 1, there may exist some objective conditions that will result in some uncertainties in the data set: the users are likely to move from one base station to another base station’s coverage area near the collection time point, and it is also possible that the number of serving persons is zero in the areas with fewer people at a collection time point. The simple data preprocessing techniques such as dimensionality reduction and denoising, may ignore the actual significance of these data. Although the data is uncertain at a certain collection time point, in a coarser-grained data set, such as the sample data produced by a certain base station within one hour, the data are more likely to have the significance of reflecting the service mode of the base station. That can be explained by the law of large numbers: the uncertainty displayed by the average behavior of multiple individuals in a system will gradually decrease as the total number of individuals continues to increase.

Therefore, we will make our analysis with coarse-grained data, which is constructed by the original data with form of Table 1 in hours. Then each data record represents the information about the number of users served by the base station in one day, and is constituted by the 24-hour data. Its form is shown in Table 2. Each attribute is composed of samples collected from 1-hour data in the original dataset. That is, the data of users served by one base station which is observed at each collection time (i.e. every minute) in the original data set is used as a sample in one attribute of the coarse-grained data set, for the purpose of reflecting the distribution of users in the coverage area of each base station within each hour.

Table 2. The coarse-grained data after processing.

<table>
<thead>
<tr>
<th>Collecting date</th>
<th>The wireless network controller</th>
<th>Affiliated region</th>
<th>Longitude</th>
<th>Latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013/2/4</td>
<td>HRBRNCO04</td>
<td>Nan-gang</td>
<td>126.6093</td>
<td>45.7203</td>
</tr>
<tr>
<td>2013/2/8</td>
<td>HRBRNCO04</td>
<td>Nan-gang</td>
<td>126.6833</td>
<td>45.7509</td>
</tr>
</tbody>
</table>

The data reflect the distribution of users served by each base station in each time period. we will propose an uncertain clustering method on this data set to analyze the aggregated...
features of calling behaviors, and then detect the functions of urban areas based on the calling features.

3 Uncertain Clustering on Base Station Statistical Data

We can treat each data described in Table 2 as an uncertain object, and it represents the information about one base station providing services in one day. The reason to call it as an uncertain object is that it consists of some samples which are collected to represent the information about one base station providing services in each hour. Therefore, in order to analyze the uncertain objects efficiently, firstly we use the method proposed in the previous research result [15] to build an uncertain data model, by which we can describe the distribution of the number of users covered by one base station in each time period. Then we propose an clustering method to aggregate data represented by the uncertain data model.

The data model is constructed for each base station data after processing according to the method proposed in [15] is shown as equation (1).

\[
G(x_j) \simeq \sum_{\alpha} \sum_{\beta} C_{\alpha}^\beta e^{\frac{1}{\alpha} \left( \frac{x_j - c_{\alpha}}{h} \right)^\beta} \left( \frac{x_j - c_{\alpha}}{h} \right)^\beta
\]

\[
= \sum_{\alpha} \sum_{\beta} C_{\alpha}^\beta e^{\frac{1}{\alpha} \left( \frac{x_j - c_{\alpha}}{h} \right)^\beta} \left( \frac{x_j - c_{\alpha}}{h} \right)^\beta
\]

where \( C_{\alpha}^\beta \) is the parameters of the data model construction method. Their specific explanations can be found in [15]. The dataset described as Table 2 could reduce the uncertainties existing probably in the original statistical data, and this modelling method could ensure the high computational efficiency of model construction.

A clustering method is needed for mining the aggregating features of uncertain data,. Aiming at the characteristics of uncertain data, the basic idea of traditional clustering
algorithms can be extended to the field of uncertainty. Considering that K-Means is suitable for processing numerical attributes and has the characteristics of linear time complexity for processing large data sets, we will use the idea of K-Means to cluster uncertain data objects. When facing uncertain data objects, we need to process the uncertainties in data objects and in the clusters involved in the algorithm.

The uncertain data can be represented as equation (1). And the model describing the distribution of mean value of cluster $i$ consisting of $n_i$ uncertain objects can be represented as equation (2), where $K$ is the number of clusters. That is, similar to the uncertain object, the cluster mean is obtained from the distribution characteristics that represent the uncertainty of the objects in the cluster, and can be obtained by calculating the average of the probability density functions of all the objects in the cluster.

$$\text{pdf}(\text{mean}_i) = \frac{1}{n_i} \sum_{j=1}^{n_i} G(x_j), 1 \leq i \leq K$$  \hspace{1cm} (2)

Then we can define the similarity between uncertain objects represented by polynomials as the form of equation (1). If the coefficients of the two polynomials are expressed as vectors $v_1 = (v_{11}, v_{12}, \ldots, v_{1m})$, and $v_2 = (v_{21}, v_{22}, \ldots, v_{2m})$ respectively, the similarity is defined in equation (3), i.e. it can be obtained by calculating the cosine similarity between two objects, whose data model are $g(x_i)$ and $g(x_j)$, where $m$ is the larger number of terms in the polynomials of uncertain data model. Once we describe the cluster mean as equation (2), we can also use equation (3) to calculate the similarity between an uncertain object and a cluster.

$$p_i = \text{sim}(g(x_i), g(x_j)) = \frac{v_1 \cdot v_2}{\|v_1\| \cdot \|v_2\|}$$  \hspace{1cm} (3)

Based on the data model of the uncertain object, the similarity measurement between uncertain objects, and the definition of the cluster mean, we propose the method of clustering on uncertain data based on K-Means as follows.

**IUK-Means(D, $G(x_i)$, $K$)**

**Input:** uncertain dataset $D = \{x_1, x_2, \ldots, x_M\}$, uncertain data model $G(x_i), 1 \leq i \leq M$, number of clusters $K$;

**Output:** clustering result $C = \{C_1, \ldots, C_K\}$;

1) select $K$ objects randomly from $D = \{x_1, x_2, \ldots, x_M\}$ as the initial cluster center;

2) repeat

3) for $i = 1$ to $M$ do
4) Calculate the similaritys between object $x_i$ and the mean of each cluster, that is, $sim = \{sim_1, sim_2, \cdots, sim_K\}$ based on equation (3);

5) Assign the object $x_i$ to the most similar cluster (i.e. $sim_j = \max(sim_1, sim_2, \cdots, sim_K)$) to the cluster $C_j$;

6) end for

7) Update the data model of $K$ clusters to describe their distribution features, $pdf(\text{mean}_j)$, $(1 \leq j \leq K)$ according to equation (2);

8) Calculate the objective function $E = \sum_{j=1}^{K} \sum_{o \in C_j} \text{sim}(o, \text{mean}_j)^2$;

9) until the objective function $E$ reaches convergence.

The time complexity of this method based on the idea of K-Means for clustering uncertain objects is $O(tKdM)$, where $t$ is the number of iterations, and $d$ is the dimension of the uncertain object.

4 Experimental Results

We apply the above method on the data as shown in Table 2. 18 clusters are generated, and different clusters have different features. The features are shown in Figure 1. We can summarize the functional areas of these clusters in Table 3 based on the features of the calling behaviors.

<table>
<thead>
<tr>
<th>ID</th>
<th>Cluster</th>
<th>Functional area</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Cluster 1</td>
<td>Mixed area</td>
</tr>
<tr>
<td>2</td>
<td>Cluster 3</td>
<td>Enterprise area</td>
</tr>
<tr>
<td>3</td>
<td>Cluster 7</td>
<td>Tourist area</td>
</tr>
<tr>
<td></td>
<td>Cluster 9</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Cluster 10</td>
<td>Residential area</td>
</tr>
<tr>
<td>5</td>
<td>Cluster 15</td>
<td>Tourist area</td>
</tr>
<tr>
<td>6</td>
<td>Cluster 18</td>
<td>Mixed area</td>
</tr>
</tbody>
</table>
In the 18 clusters, the clusters in the 7th group of Table 3 seem to have similar features according to Figure 1. It seems that there are only a few people at a certain time point, and there are no people to be served at the other time. That may be areas with fewer people, which are likely to be related to local job characteristics. People covered by stations in Cluster 4 and 12 are approximate zero, so we cannot infer their corresponding functions. Stations in Cluster1 and 18 serve people with similar behavior characteristics. The difference is that the number of people covered by base stations starts to decrease after 23 o'clock in Cluster 1, while after 16 o'clock in Cluster 18. Because it remains a relatively stable number of people all over the day, it can be inferred as a mixture of office area and residential area. The person number in Cluster 3 starts to increase from 8 a.m., reaches a peak during 10 a.m. and 11 a.m., and starts to decrease significantly from 4 p.m. to 5 p.m.. Such areas are generally enterprise or institution areas with obvious rules. Cluster 7 and Cluster 9 are characterized by an increase starting from 8 a.m. and decrease from 8 p.m.. Their peaks appear at 3 p.m.. and no one appears in other time periods. Similarly, The person number of Cluster 15 starts to increase from 8 a.m. with the peak appearing at 8 p.m., and basically approximates to zero after 23 o'clock. The characteristics indicate that these base stations in the three clusters are in tourist areas, and the best time to visit the scenic spots are at the time near the peak. The number of people starts to increase after 15 o'clock and remains stable 24 o'clock, which are the features like residential areas, that is, the people covered by the base stations start to increase since the time of after work.
Fig. 1. Features of clusters generated on data described as Table 2.

Obviously, we could detect the urban functional areas according to the behavior characteristics of base stations serving people without user privacy data. If we can gather information provided by more base stations in a city, we could obtain more detailed division of functional areas.

Conclusion

In order to solve the problems about privacy leakage when analyzing the functions of the urban areas based on the detailed data of individual behaviors, this study uses the statistical data from Harbin base stations, and proposes a method to discover the land functions based on uncertain clustering. This study uses the data monitored at each gathering time as sample points to form coarser-grained data sets, builds uncertain data model and performs uncertain
clustering on this basis. The method could not only reduce the impact of the possible uncertainties existing in original data, but also detect land functions through the related aggregated features of calling behaviors based on clustering results. The experimental results indicate the effectiveness of the proposed method.

**Acknowledgments.** This work is sponsored by the Natural Science Foundation of Heilongjiang Province under grant number F2017015, and Training Program For Young Innovators in Heilongjiang General Institutes of Higher Education under grant number UNPYSCT-2017079.

**References**


A Clustering Analysis Method Based on Wilcoxon-Mann-Whitney Testing

Yuan Cheng1, Weinan Jia1 and Ronghua Chi2
{changuang7@sina.com1, jiawei@163.com1, chihong@126.com2}

College of Computer Science and Technology, Harbin University of Science and Technology1,
College of Computer and Information Engineering, Heilongjiang University of Science and Technology2

Abstract. As the core step of clustering analysis, the results of distance measurements will influence the clustering accuracy. The existing measurements are mostly based on the information about cluster features. However, the cluster features may be not sufficient enough and would result in losing data information about clusters containing a number of objects. To improve the measurement accuracy, we make full use of the distribution characteristics of objects in clusters, so we use the descriptive statistics and the Wilcoxon-Mann-Whitney rank sum test in nonparametric statistics to measure distances during clustering. Furthermore, a two-stage clustering is proposed to improve the performance of clustering analysis, from the aspects of avoiding assuming the number of clusters preliminarily, discovering clusters of arbitrary shapes and improving clustering accuracy. The experiments on multiple datasets compared with other clustering algorithms illustrate the accuracy and efficiency of the proposed clustering algorithm.

Keywords: Clustering analysis, distance measurement, nonparametric statistics, Wilcoxon-Mann-Whitney rank sum test.

1 Introduction

As a basic technology of data mining, clustering analysis is significant in discovering the characteristics of data aggregation [1]. Clustering analysis is an unsupervised method, whether the partitioning methods, hierarchical methods, density-based methods, or grid-based methods, their implicit clustering ideas are similar, that is based on the distances between objects, through the iterations to ensure the clustering quality. K-Means [2] adjusts the clusters the
objects belonging to in each iteration; DBSCAN [3] aggregates the objects that could directly density reachable from the core object in each iteration to generate new clusters; so their difference is the way to divide the objects into clusters during the clustering process. In this unsupervised analysis process, the main basis of assigning an object to a cluster is the distance measurement, including distance between objects, distance between the object and the cluster, and distance between clusters. K-Means divides objects into clusters based on the distances between objects and clusters; the judgment of directly density-reachable in DBSCAN is also based on the distances between objects; the clusters mergence is decided by the distances between clusters in agglomerative hierarchical clustering. It is obvious that the accuracy of distance measurement is the important basis for effective clustering.

The existing distance measurement between objects could be divided into multiple methods according to the attribute types and the application scenes, such as Euclidean distance, Manhattan distance, Minkowski distance, Jaccard coefficient, cosine measure and so on [4-6]. And the distance involving clusters are mostly measured based on the information reflecting cluster features. For instance, K-Means and K-medoids [7] choose the mean value or a representative object as the feature of a cluster; they assign an object into the cluster whose feature is closest to it; DBSCAN and OPTIS [8] consider the core object as the cluster feature, and decide whether to assign an object into a cluster according to whether it is density-reachable to the core object [9].

Actually, the objects in each cluster are the main factors truly reflecting cluster features, therefore the distance between clusters could be calculated through the distance between objects in the clusters, such as the minimum distance, the maximum distance, and the average distance. However the larger number of objects may affect the efficiency of distance measurement. Therefore, to improve the computation speed and scalability, Birch [10] uses zero moment, first moment and second moment to generate a three-dimensional vector, which is represented as the cluster feature to summarize cluster information and to compute the distance between clusters for hierarchical clustering. But it is not sufficient to describe the clustering information by using the representative objects or the statistics. The clusters contain a number of objects, and the existing cluster features represent the aggregation features of clusters. They would loss information reflecting the data characteristics of clusters to a certain extent. And distance measurement would be with a certain deviation, and thus affect the accuracy of clustering results.

Obviously, the distance measurement is a core step in clustering. The effective information extraction to representing the data features of clusters is the key to ensure the metric accuracy. And it is also the key to ensure the accuracy of clustering. Therefore,
researchers are trying to extract effective information about cluster features. [11] extracted some adjacent objects of centroids to summarize the cluster information. It used a group of representative objects, but the adjacent objects are not enough to reflect the general data features. [12] defined a coreset to measure distances with the idea of Birch. Although they choose a number of objects as the representative information of clusters, it is also insufficient so as to result in the information loss. The distribution of data in the clusters could reflect the general cluster data features. [13] has obtained the distribution features of clusters based on probability density function. However, this method needs to presuppose the data distribution. It is difficult to make a clearer assumption about the data distribution, due to the little knowledge about the overall information. And the incorrect assumption would result in the inaccuracy of distance measure and even the clustering results.

Nonparametric statistical methods [14,15] can be used to estimate the distribution structures based on the data information directly, rather than based on a hypothesis about the specific form of the overall distribution. The Wilcoxon-Mann-Whitney (W-M-W) rank sum test method [16,17] is one of nonparametric statistical methods. It is always used to judge whether any two sets come from a same population. In the clustering process, if two sets represented by two clusters are considered from the same population, they could be grouped into one cluster. Therefore, through this method we could reserve the original cluster information features, analyze the dissimilarity between clusters directly based on the distribution features of their data, and then determine whether to merge them into one cluster, without the hypothesis about the overall distribution form.

This paper will propose a new distance measurement method based on W-M-W rank sum test to resolve the above problems, and then propose an improved hierarchical clustering method to increase the clustering effectiveness. This method will need minimal requirements for domain knowledge to determine input parameters, discover clusters with arbitrary shape and improve clustering accuracy. Experiments on multiple datasets are used to verify the validity of the proposed algorithm. Finally, this paper is concluded.

2 Distance measurement based on Nonparametric Statistics

In data mining, especially in cluster analysis, distance measurement is the core of data analysis, its accuracy will directly affect the validity of data analysis results. There are multiple metric methods according to different data types, such as Euclidean distance, Manhattan distance, Minkowski distance, Jaccard coefficient, cosine measure and so on. While in clustering analysis, we will divide objects into some cluster, or group two similar sets
into one cluster during clustering processes. These operations are based on the distance measurements, which include distances between objects and clusters and distances between clusters. Obviously, these distances are about clusters. To ensure the objectivity and accuracy of measurements, it is necessary to consider the distribution features of the objects in the clusters with little loss of data information. With this purpose, this paper will measure distances based on the distribution characteristics of data in the clusters.

2.1 Distance between objects and clusters based on distribution characteristics

In the traditional clustering methods, the distances between objects and clusters are often transformed into the distances between objects and cluster features. The features include the cluster mean values, representative objects of clusters, coresets of clusters and so on. As mentioned above, these single descriptive features would lose the data information of clusters to some extent. And the distribution characteristics of objects in one cluster could be seen as its general cluster features. The difference between clusters is about the difference between their distribution characteristics of objects. While if we want to divide an object into some cluster, the distance between it and a cluster is the basis of assignment. Then the distribution characteristics of objects in this cluster would have an impact on distance measurement and should be considered.

The distribution characteristics of objects in a cluster are necessary to be considered when measuring distances. Obviously, the mean value as one statistical feature of a set could not completely reflect its distribution, while the representative also could not reflect the distribution characteristics of all objects in the cluster. Therefore these measure methods could not calculate differences between objects and clusters objectively. Sometimes the use of one statistic is insufficient to represent the general characteristics of data in a set after all.

Therefore, for a more accurate and objective result, we need to divide an object into some cluster with the consideration of distribution characteristics of data in this cluster. The methods that can describe the distribution features of a set include probability distribution functions and descriptive statistics. It is more time consuming to compute the probability distribution function of objects for each cluster. While multiple descriptive statistics could describe the statistical characteristics of a set from different perspectives; so it could be used to represent its distribution characteristics. Therefore, this paper will measure the distance between objects and clusters based on some statistical features of a set.

If an object belongs to a cluster, it has similar characteristics to other objects in this cluster. That is, the distribution characteristics of this cluster will not change significantly after
the object is assigned into it. This paper will consider different descriptive statistics of a cluster when measuring distances between an objects and the cluster; and analyze that whether these statistics have changed significantly after the object is divided into it. We will determine the right cluster with the smallest change of the statistics and also below a threshold.

In descriptive statistical analysis, the statistics such as mean, variance, and quantile can be used to measure the average values, central tendency and location information of data in a set respectively. These statistics describe the data information about position and dispersion of a set and actually represent the distribution characteristics of a data set.

We begin with one-dimensional data to discuss the method of determining the relationship between an object and a cluster with the above descriptive statistics. Then we extend this method to multi-dimensional data. We will match the object and descriptive statistical features of the cluster in each dimension and analyze the differences between the object and the cluster in an effective way.

Let \( o_i \) is the one-dimensional object to be assigned, the existing clusters are \( C = \{C_1, C_2, \ldots, C_n\} \). The distribution feature of cluster \( C_i \) can be described by a triple \( DF_i = \langle \mu_i, \sigma_i, m_i \rangle \), where \( \mu_i, \sigma_i, m_i \) are the mean value, variance and median, which represent the average value, dispersion and the center position of cluster \( C_i \). If the object is divided into \( C_i \), its distribution feature would be \( DF_i' = \langle \mu_i', \sigma_i', m_i' \rangle \). And the variation of distribution feature could be calculated as equation (1).

\[
\Delta_i = |\mu_i' - \mu_i| + |\sigma_i' - \sigma_i| + |m_i' - m_i| \quad (1)
\]

If \( o_i \) belongs to the cluster \( C_{\omega} \), its impact on the distribution feature of \( C_{\omega} \) should be relatively small, i.e. the value of \( \Delta_{\omega} \) should be the smallest and within a certain threshold.

For instance, there are three clusters, \( C_1 = \{4.7, 5.1, 4.8, 5.4, 5.5, 4.4, 5\} \), \( C_2 = \{5.9, 5.2, 6, 5.5, 5.8, 6.1, 5.7\} \), and \( C_3 = \{5.8, 6.3, 6.1, 7.1, 5.6, 6.7, 6.5\} \). The triples representing their distribution features are \( DF_1 = \langle 4.99, 0.39, 5 \rangle \), \( DF_2 = \langle 5.74, 0.31, 5.8 \rangle \) and \( DF_3 = \langle 6.3, 0.52, 6.3 \rangle \) respectively. The object to be divided is \( o_i = 5.7 \). The threshold of variation about distribution feature is \( \delta = 0.1 \).

We could obtain the triples \( DF_1' = \langle 5.07, 0.44, 5.05 \rangle \), \( DF_2' = \langle 5.74, 0.29, 5.75 \rangle \), \( DF_3' = \langle 6.22, 0.53, 6.2 \rangle \), if \( o_i \) is divided into \( C_1 \), \( C_2 \), and \( C_3 \) respectively. Their variations on the distribution feature are \( \Delta_1 = 0.18 \), \( \Delta_2 = 0.07 \), and \( \Delta_3 = 0.19 \) respectively, where \( \Delta_2 \) has the smallest value and \( \Delta_2 < \delta \). It can be concluded that \( o_i \) is more likely to come from the same distribution with data in \( C_2 \). Then \( o_i \) can be divided into \( C_2 \).

If we extend the above method to multi-dimensional data, we need to determine the relation between the object and the cluster distribution feature in each dimension as described above. Then integrate the analysis results on each dimension to determine the cluster having
the smallest variation about distribution feature after the object is added into it. This cluster is more similar to the object than others.

Let \( o_2 \) be the \( d \)-dimensional object to be assigned, and \( C = \{C_1, C_2, \ldots, C_n\} \) be the existing clusters. The distribution feature of cluster \( C_i (1 \leq i \leq n) \) can be described by a \( d \)-dimensional triple as equation (2).

\[
DF_i = \{DF_{i1}, \ldots, DF_{id}\} = \{\mu_{i1}, \sigma_{i1}, m_{i1} >, \ldots, \mu_{id}, \sigma_{id}, m_{id} >\} \quad (2)
\]

During the analysis, we could calculate the distribution feature in the \( k \)-th dimension of every cluster: \( DF_{ik} = \{\mu_{ik}, \sigma_{ik}, m_{ik} >\} (1 \leq i \leq n) \), when \( o_2 \) is assumed to be divided into each cluster. In addition, the variation about distribution feature in the \( k \)-th dimension could be also calculated as equation (3).

\[
\Delta_{ik} = |\mu_{ik} - \mu_{ik}^{'}, \sigma_{ik} - \sigma_{ik}^{'}, m_{ik} - m_{ik}^'| \quad (3)
\]

Then the variation in all dimensions is \( \Delta_i = \sum_{j=1}^{d} \Delta_{ij} \). Let \( C_m \) be the cluster that \( o_2 \) is most likely to be assigned. Its variation value \( \Delta_m \) should be the smallest and within a certain threshold.

We can compute the variations about distribution feature for each cluster with the assumption of the object is grouped into every cluster. The cluster having the minimum variation value and less than the threshold is the one most matching the object in statistical characteristics. If all the variation values are greater than the threshold, the object is more likely to be an outlier.

Then we take data shown in Figure 1 for instance to specify the method of assigning objects into clusters based on distribution features. The 4-dimensional object is \( o_2 = (5.7, 4.4, 1.5, 0.4) \). There are three clusters: \( C_1, C_2 \) and \( C_3 \). Their distribution features represented by 4-dimensional triples are as equation (4).

\[
DF_1 = \{4.99, 0.39, 5, 3.29, 0.24, 3.3, 1.46, 0.15, 1.4, 0.27, 0.13, 0.2\} \\
DF_2 = \{5.74, 0.31, 5.8, 2.83, 0.29, 2.9, 4.3, 0.34, 4.2, 1.37, 0.25, 1.4\} \\
DF_3 = \{6.3, 0.52, 6.3, 3.04, 0.24, 3.3, 5.37, 0.48, 5.1, 2.06, 0.22, 2\} \quad (4)
\]

The threshold of variation about distribution features is \( \delta = 0.8 \). Then we can obtain the new triples as shown in equation (5).

\[
DF_1^' = \{5.07, 0.44, 5.05, 3.425, 0.45, 3.35, 1.46, 0.14, 1.45, 0.29, 0.12, 0.2\} \\
DF_2^' = \{5.74, 0.29, 5.75, 3.025, 0.62, 2.95, 4.15, 0.25, 1.25, 0.41, 1.35\} \\
DF_3^' = \{6.22, 0.53, 6.2, 3.21, 0.53, 3.1, 4.89, 1.44, 5.1, 1.85, 0.62, 2\} \quad (5)
\]

If we assume the object is divided into these clusters respectively. The variations with the former are \( \Delta_1 = 0.665, \Delta_2 = 2.075, \Delta_3 = 2.8 \), respectively. Obviously, \( \Delta_1 \) is minimum and less than the threshold \( \delta \). Taking into account all the four dimensions, the object is more likely to be from the same distribution with data in cluster \( C_1 \). So it could be divided into \( C_1 \). This result
is different from the above 1-dimension analysis, since object $o_2$ is described by the four dimensions, and its assignment is based on distribution features on all dimensions, rather than one dimension.

<table>
<thead>
<tr>
<th>$C_1$</th>
<th>$C_2$</th>
<th>$C_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4.7, 3.2, 1.3, 0.2), (5.1, 3.3, 1.7, 0.5), (4.8, 3.1, 1.6, 0.2), (5.4, 3.4, 1.5, 0.4), (5.5, 3.5, 1.3, 0.2), (4.4, 2.9, 1.4, 0.2), (5, 3.6, 1.4, 0.2)</td>
<td>(4.7, 3.2, 1.3, 0.2), (5.1, 3.3, 1.7, 0.5), (4.8, 3.1, 1.6, 0.2), (5.4, 3.4, 1.5, 0.4), (5.5, 3.5, 1.3, 0.2), (4.4, 2.9, 1.4, 0.2), (5, 3.6, 1.4, 0.2)</td>
<td>(4.7, 3.2, 1.3, 0.2), (5.1, 3.3, 1.7, 0.5), (4.8, 3.1, 1.6, 0.2), (5.4, 3.4, 1.5, 0.4), (5.5, 3.5, 1.3, 0.2), (4.4, 2.9, 1.4, 0.2), (5, 3.6, 1.4, 0.2)</td>
</tr>
</tbody>
</table>

Fig. 1. 4-dimensional clusters to be assigned

### 2.2 Distance between clusters based on rank sum test

The main purpose of measuring distance between clusters is to merge similar clusters into one cluster. The similarity in unsupervised data analysis is based on distance measurement, while from a statistical perspective, the objects in two clusters are similar could be considered that they are more likely from the same distribution. The W-M-W rank sum test method is one of nonparametric statistics method. It can test whether two samples are from the same population without too much samples and a pre-hypothesis about data distribution.

That is, it could give a conclusion objectively. Therefore, based on the W-M-W rank sum test method, we will determine whether to merge two clusters through testing whether the objects in these two clusters are coming from the same population. If from the same population, they can be merged into one cluster; otherwise, the two clusters still exist as two separate clusters.

For any two clusters $C_1$ and $C_2$, their number of objects are $n_{c1}$ and $n_{c2}$ respectively. The number of upper limit of objects used in the rank sum test is $n_s$. When $n_{c1}, n_{c2} \leq n_s$, all objects in these clusters could involve in the rank sum test to determine whether they are from the same distribution; otherwise, we should take $n_s$ samples randomly from the two clusters respectively for the test.

Then we will take one dimensional objects as example and describe the method of deciding whether two sets need to be merged through W-M-W rank sum test. If the objects are multi dimensional data, we need to analyze in each dimension as the method; its basic idea is
that objects in two clusters are from the same population indicate these two groups of data are from the same distribution in each dimension.

Let \( C_1' = \{ x_1, x_2, \ldots, x_m \} \) and \( C_2' = \{ y_1, y_2, \ldots, y_n \} \) be the sample sets from cluster \( C_1 \) and \( C_2 \) involved in the test, \( m \) and \( n \) be their object number respectively. On the basis of W-M-W rank sum test method, we want to test whether two sets are from the same population by using sample information without the hypothesis of data distribution. Then we will conduct a hypothesis test with the sample data. If it is validated, the null hypothesis will be accepted; otherwise the null hypothesis will be rejected. Even though there is a hypothesis in this method, it is used to make a relatively objective conclusion based on data information, rather than to be a basis for subsequent analysis.

We will make the null hypothesis that the sets \( x_1, x_2, \ldots, x_m \sim F(x - \mu_1) \) and \( y_1, y_2, \ldots, y_n \sim F(x - \mu_2) \) have the similar distribution, without regarding to data symmetry. Then the problem about merging two sets could be transformed into the problem to be test: \( H_0: \mu_1 = \mu_2 \leftrightarrow H_1: \mu_1 \neq \mu_2 \). This is a bilateral test problem. The null hypothesis is that the two sets have no significant difference, come from the same distribution, and can be merged. The alternative hypothesis is that the two sets having significant differences are from different distributions and could not be merged. During the analysis we need to mix \( x_1, x_2, \ldots, x_m \) and \( y_1, y_2, \ldots, y_n \) together, and make these \( (m+n) \) numbers in ascending order. The rank of a sample is its position in this ordering sequence. In this mixed ordering sequence, let \( X \) be the sum of ranks (rank sum) of objects from \( C_1' \), while \( Y \) be the rank sum of the objects from \( C_2' \).

We use the statistics \( \min \{ W_{XY}, W_{YX} \} \) for this validation problem, where \( W_{XY} \) and \( W_{YX} \) are shown as equation (6). \( W_{XY} \) is the number of samples from \( C_2' \) whose values are greater than the values from \( C_1' \), while \( W_{YX} \) is on the contrary.

\[
W_{XY} = mn + \frac{n(n+1)}{2} - W_X \\
W_{YX} = mn + \frac{m(m+1)}{2} - W_Y
\]

If two sample sets have the same distribution, the ranks of the samples should be randomly mixed. While if they have different distributions, one of the rank sums should be greater than the other. Therefore, the rank is used to calculate the statistics, and this method can be used to analyze whether two sets are from the same population without the sample distribution.

In addition, \( Z = \frac{W_{XY} - mn/2}{\sqrt{mn(m+n+1)/12}} \sim N(0,1) \). Then we can calculate the values of \( p \) with the corresponding \( m \) and \( n \). This \( p \)-value is minimum significance level to reject null hypothesis according to the test statistics calculated by the samples\(^{[18,19]} \). Then for a given significance level \( \alpha \), we can obtain the analysis result of hypothesis testing through comparing \( p \) and \( \alpha \). If
$p > \alpha$, the null hypothesis is accepted, which indicates that there is no significant difference between data in these two clusters and they could be merged; While if $p \leq \alpha$, the null hypothesis is rejected, that is, data in the two clusters are more likely to come from different distributions and they could not be merged.

**Figure 2** describes the specific steps of determining whether two one-dimensional sets have significant differences based on the above validation method.

The time complexity of the process is $O(n^2)$, where $n$ is the threshold for the number of objects in one cluster involving the rank sum test method. Even if the cluster has a large number of objects, $n$ samples could be drawn randomly to constitute the data set to be tested for further analysis. The feasibility of this sampling method is based on the W-M-W rank sum test method, which is still feasible even with a small sample. Although not all of the objects are used to be analyzed, the random samples of objects will reflect the distribution characteristic to some extent. In addition, the test is based on nonparametric statistical method; it would take full advantage of sample data information, rather than analyze based on a hypothesis about data distribution. It tests whether two clusters are from the same distribution according to data itself. That is, it analyzes the similarity between clusters from the statistical
test perspective. Its objectivity will ensure the accuracy of the measuring results. Although traditional distance metric are also based on data information and calculate distance between objects in the clusters. However their values of distances are not the final results of the measurement. They will be used to analyze whether two clusters are similar and are needed to be merged through the comparisons about the distance values. Therefore whether two clusters are similar is a relatively comparative result. Therefore, the distance measurement method proposed in this paper has certain advantages in the accuracy and efficiency.

Multi-dimensional data need to be analyzed on each dimension as above. Once there is a significant difference to be tested in one dimension, it indicates that data are from different populations on this dimension. It is difficult to illustrate the objects in two clusters have similar features, because they are already different in one dimension. Then it can be determined that objects in two clusters have significant differences; and there is no need to merge these clusters. The time complexity of this process is \( O(dn^3) \).

Taking \( C_1 \) and \( C_2 \) in Figure 1 and another cluster \( C^* \) as an example, we will illustrate the method determining whether two clusters need to be merged based on W-M-W rank sum test method. Figure 3 describes its analysis process.

Fig. 3. Example of measuring distances between clusters based on W-M-W rank sum test

Clusters \( C_1 \) and \( C_2 \) in Figure 3 are needed to be tested for each of four dimensions. Each value of \( p \) is less than the significance level \( \alpha \). It illustrates that these two clusters have a
significant difference in all of four dimensions. Then it can be determined that $C_1$ and $C_2$ are from two different populations and they cannot be merged. While in the test cluster $C_1$ and $C^*$ in four dimensions, the $p$-values are all greater than the significant level $\alpha$. That is, there is no significant difference between $C_1$ and $C^*$, and they can be merged into one cluster.

Obviously, our proposed method would obtain a more objective result than the traditional distance metrics, because it directly determines whether to merge two clusters based on the distribution characteristics of data, rather than based on the comparison of distance values. In fact, these data are from Iris dataset, and data in $C_1$ and $C^*$ are from the same class, while data in $C_1$ and $C_2$ are from different categories. It illustrates the accuracy and validity of our method.

3 A data distribution feature oriented hierarchical clustering analysis method

Combined with the distance measurement method proposed above, this paper will propose a two-steps hierarchical clustering algorithm, so as to avoid assuming the number of clusters preliminarily, discover clusters of arbitrary shapes and improve clustering accuracy. The above distance measurement methods are the point of proposing such a clustering algorithm. In this hierarchical clustering algorithm, the distance metrics proposed in Section 2 are used to divide objects to the proper clusters and determine whether to merge clusters.

Firstly, in the first step, the idea of K-Means is used to generate a number of clusters as the initial clusters through dividing objects in the original data set. The generated number of clusters ‘$k$’ will be set to a larger value. Then the more similar objects would be divided into the same cluster. Then the idea of hierarchical clustering will be used to merge similar clusters in these initial clusters. The number of initial clusters is set to be a larger one, there will exist similar ones about distribution features among these initial clusters. Therefore in the second step we will merge the similar ones into one cluster so as to divide their objects into the same cluster. During this process, we will determine whether two clusters are similar and need to be merged. This operation will continue until all clusters are tested to have significant difference between each other, when data in different clusters are likely to come from different populations. Then the clustering process could stop. This two-steps hierarchical clustering algorithm is described as follows.

$NPSC(D, k, \delta, n_\delta, \alpha)$

Input: $D = \{x_1, x_2, \ldots, x_n\}$, dataset;
$k$, the number of the generated initial clusters;
\(\delta\), the threshold of variations about distribution features;
\(n_\delta\), the threshold of the number of objects in one cluster processed by rank sum;
\(\alpha\), the significance level;
Output: \(C = \{C_1, \ldots, C_K\}\), the clustering result;

Steps:

(1) Generate initial clusters
   1) choose \(k\) objects from dataset \(D\) to be the initial cluster centers, then obtain
   \(C' = \{C_1, \ldots, C_k\}\);
   2) repeat
   3) for \(i := 1\) to \(n\) do
   4) \(C_\omega = ocd(x_i, C, \delta)\);
   5) divide object \(x_i\) into cluster \(C_\omega\);
   6) end for
   7) update the distribution features of \(k\) clusters as equation (7):
   \[
   DF_j = \{DF_{j_1}, \ldots, DF_{j_k}\} = \{<\mu_{j_1}, \sigma_{j_1}, m_{j_1}>, \ldots, <\mu_{j_k}, \sigma_{j_k}, m_{j_k}>\} 1 \leq j \leq k
   \]
   8) compute the objective function: \(E = \sum_{j=1}^{k} DF_j\);
   9) until the objective function \(E\) converges.

(2) Merge similar ones in initial clusters \(\{C_1, C_2, \ldots, C_k\}\)
   1) Let \(K' = k\);
   2) repeat
   3) for \(i := 1\) to \(K'\) do
   4) for \(j := i + 1\) to \(K'\) do
   5) \(mb = ccd(C_i, C_j, n_\beta, \alpha)\);
   6) if \((mb = 1)\) then
   7) \(C_i, C_j\) have significant difference, do not merge them;
   8) else if \((mb = 0)\) then
   9) \(C_i, C_j\) do not have significant difference, merge them into one cluster;
   10) end for
   11) Let \(K'\) be the number of clusters after merging operations;
   12) until there exist significant differences between any two clusters.

The time complexity of obtaining initial cluster is \(O(tkdn)\), where \(t\) is the iterations, \(k\) is the number of initial clusters, \(d\) is the dimensions of data, and \(n\) is number of objects. Based on the above analysis, the time complexity of the merging step is \(O(t'k^2dn_\beta)\), where \(t'\) is the iterations for the merge step. Therefore, this proposed two-steps hierarchical clustering...
algorithm based on nonparametric statistics has the time complexity of $O(t_1^2 + k^2 n_0^2)$, where $k, n_0 \ll n$. Obviously, the proposed algorithm could be effective. The final number of clusters is generated based on the distribution features of data, rather than a pre-assumed value. In addition, the accuracy of distance metric proposed in Section 2 could ensure the accuracy of results generated by the proposed unsupervised clustering algorithm.

4 Experiments

We will choose three two-dimensional data sets and several UCI data sets, to verify the validity of the proposed clustering algorithm based on W-M-W rank sum test method. And we will compare with the following clustering algorithms: K-Means, DBSCAN, Birch, UPGMA [11], and Fast [12] about run time and accuracy of clustering results. The results will illustrate the effectiveness and practicality of our proposed algorithm.

For the data set having marked categories, we will use external indices Purity and Entropy [20] to clustering to evaluate the accuracy of clustering results. Let $C = \{C_1, \ldots, C_{K'}\}$ be the clustering result, and $P = \{P_1, \ldots, P_l\}$ represent the given categories of data, where $K'$ is the number of generated clusters, and $l$ is the number of original categories. Then Purity and Entropy can be calculated as:

Purity: $Purity = \sum_{i=1}^{K'} \max_j (n_i^j)$,

Entropy: $Entropy = \sum_{i=1}^{K'} \frac{1}{n_i} \left(-\frac{1}{\log n_i} \sum_{j=1}^{l} \frac{n_j^i}{n_i} \log \frac{n_j^i}{n_i}\right)$,

where $N$ is the number of objects in the dataset, $n_i^j$ is the number of objects divided into the $i$-th cluster which belong to the $j$-th category in the original dataset, $n_i$ is the number of objects divided into $i$-th cluster. The higher the purity is, the more accurate the clustering result is; and the lower the entropy is, the more accurate the clustering result is. Ideally, $Entropy = 0.0$ and $Purity = 1.0$.

4.1 Two-dimensional datasets

We will choose three two-dimensional graphic data sets: Aggregation, Spiral and Flame to verify the proposed method could discover clusters of arbitrary shapes. These datasets contain the similar spatial data within the same category (clusters), not only simple spherical clusters. They could also be visualized. Therefore these datasets could be used to validate the capacity of discovering clusters of arbitrary shapes. Figure 4 shows the visualized clustering results of three two-dimensional datasets obtained by our proposed algorithm NPSC. It can be
seen that NPSC could identify the clusters of data, that is, it could discover clusters of arbitrary shape better.

This is mainly due to the distance measurement method used in the proposed algorithm. It determines the similarities between clusters on the basis of distribution features of data, rather than simply based on the traditional distance metrics. This method could merge similar clusters according to the characteristics of data based on the nonparametric statistical hypothesis test method without the hypothesis of data distributions. And it is used in the second step of the proposed clustering algorithm. A number of closely similar clusters will be generated in the first step of the clustering process. Then clusters having similar distribution features discovered based on our proposed distance measurement method. And the similar clusters will be merged into one clusters. These characteristics make the proposed clustering method more suitable to discover non-spherical clusters.

![Fig. 4. Clustering results of two-dimensional datasets obtained by NPSC](image)

### 4.2 UCI datasets

Then we do clustering on the UCI datasets shown in Table 1 compared with other clustering algorithms, to verify the effectiveness and accuracy of our proposed algorithm.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Object number</th>
<th>Attribute number</th>
<th>Category number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abalone</td>
<td>4177</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>Ecoli</td>
<td>336</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Iirs</td>
<td>150</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Letter</td>
<td>20000</td>
<td>16</td>
<td>26</td>
</tr>
<tr>
<td>Yeast</td>
<td>1484</td>
<td>8</td>
<td>10</td>
</tr>
</tbody>
</table>
Figure 5 describes the comparison about the accuracy of clustering results based on evaluation indices Purity and Entropy. Obviously, NSPC obtains relatively greater Purity values and lower of Entropy values than other algorithms. It indicates that NSPC could obtain more accurate clustering results.

This is mainly due to the proposed distance measurement method, because the unsupervised clustering analysis determines the generation of clusters based on the results of distance measurement. The proposed method does not assign objects into clusters based on the relatively comparative distances as K-Means or Birch. It also does not dependent on the parameter about neighborhood radius to determine the density of clusters as DBSCAN. NSPC could not entirely depend on the numerical distance measurement results as traditional methods do. It is based on nonparametric statistical hypothesis testing method, and determines whether clusters are similar according to the distribution features of data. UPGMA and Fast have improved the problems in traditional clustering methods. However UPGMA still extracts cluster features based on neighboring objects, that is, it also depends on the distances to some extent. While NSPC draws samples in each cluster randomly during the analysis of similarity between clusters, these samples would reflect the distribution features of clusters to some extent. Fast uses the probability density function to obtain the distribution features of clusters. However this method needs to make assumptions about the distributions of data, and this assumption are more likely not to match the real data distributions. Therefore its results obtain lower Purity value and greater Entropy value compared with NSPC. It illustrates that our proposed distance metric helps to get more accurate clustering results.
**Figure 6** shows the run time comparative results between these clustering algorithms on the UCI datasets. K-Means has high efficiency due to its linear time complexity. NSPC has the run time close to Birch which is also a hierarchical clustering method. And it has a relatively high efficiency compared to the other algorithms. Because UPGMA needs to obtain neighboring objects and then calculate clustering features. Fast needs to calculate probability density distribution functions for clusters. These would take some time.

![Bar chart showing run time comparison](image)

**Fig. 6.** The comparison about run time of clustering on UCI datasets.

Obviously, our proposed algorithm could not only obtain relatively accurate clustering results, but also have high efficiency. Firstly, it is due to the use of our proposed distance metric based on the distribution features of data during clustering. Secondly, it relies on our proposed two-steps clustering process. The former could help to ensure the accuracy and effectiveness of clustering. Because we use several descriptive statistics to represent distribution features of data in a cluster when measuring distances between an object and a cluster. It analyzes the distribution feature variations once the object is divided into a cluster. So it could get a more objective similarity result between the object and a cluster. We use W-M-W rank sum test method to measure distances between clusters. It could avoid the inaccuracy problems when determining whether to merge clusters according to a less objective comparison value in the traditional metrics. It could also ensure the efficiency of clustering process through not using all objects in the clusters. The latter uses a two-steps process, so that the number of generated clusters would not depend on a pre-assumed value. It could determine when the clustering process is terminated through the analysis of data distribution based on the nonparametric statistical hypothesis test. The final number of clusters does not be relative with the initial number set by the parameter.
Conclusion

This study aims at the purpose of distance measurement in unsupervised clustering: to generate new and accurate clusters. So a distance metric is proposed based on descriptive statistics and nonparametric statistical methods. In addition, a two-steps hierarchical clustering algorithm is also proposed. The distance measurement method based on nonparametric statistics could take full advantage of the distribution features of data. It could obtain clusters in a more straightforward and more objective way compared with the traditional distance metrics. The hierarchical clustering algorithm could avoid the pre-assumed initial number of clusters with its two-steps characteristics; the final number of clusters does not be relative with the initial number set by the input parameter. It could also discover clusters of arbitrary shapes and obtain more accurate results due to the distance metrics: it determines the similarities between clusters on the basis of distribution features of data. Therefore the proposed distance measurement method could provide a stronger support for unsupervised clustering analysis.

Acknowledgments. This work is sponsored by the Natural Science Foundation of Heilongjiang Province under grant number F2017015, and Training Program For Young Innovators in Heilongjiang General Institutes of Higher Education under grant number UNPYSCT-2017079.

References


An Uncertain Trajectory Modelling Method Based on Kernel Density Estimation

Yuan Cheng¹, Ronghua Chi² and Yahong Wang³
{changuang7@sina.com¹, chironghua@126.com², wyh811619@126.com³}

College of Computer Science and Technology, Harbin University of Science and Technology¹,
College of Computer and Information Engineering, Heilongjiang University of Science and Technology²
School of Applied Science and Civil Engineering, Beijing Institute of Technology, Zhuhai³

Abstract. The accurate analysis of trajectories is of great significance for route selection, traffic status analysis, and urban traffic planning and so on. Existing researches lack effective methods for dealing with possible uncertainties in trajectories caused by objective environment and subjective intention etc. This work studies the method of constructing an uncertain model for the trajectories with the same starting point and end point based on kernel density estimation, to discover the distribution characteristics of the trajectories between two points in historical data, and to lay the foundation for trajectory prediction. Finally, the validity of the proposed method is verified on the real trajectory dataset.

Keywords: Uncertainties, kernel density estimation, modelling method, distribution characteristics.

1 Introduction

With the advent of lots of mobile equipments, the information about human location can be acquired. The trajectory historical data can be used to predict human mobility trajectory. The accurate and efficient prediction results could provide the required information and help to users, and improve the quality of location-oriented applications and services. Research on trajectory prediction include methods based on individual trajectory information entropy [1], trajectory frequent pattern mining [2-4], complex information-based behavior pattern mining [5,6], and hybrid methods [7]. However, human being mobility behaviors are extremely complicated, and there may exist many uncertain factors, such as objective environment and subjective intention, which will affect the accuracy of prediction. Although the mobility...
behaviors are complex and changeable, in general, human behavior is more purposeful and the end point of a trajectory is relatively clear. So we can predict the end point of a trajectory.

Generally, the prediction method of the trajectory end point needs to match the trajectory to be predicted with the historical trajectories, and select the end point of the most similar trajectory to be the prediction result. The matching methods of trajectories include calculating the similarity between trajectories based on distance [8], Markov model based on state transition probability [9-11], and Bayesian mixed model [12-14]. To judge the end point, almost all of them need to judge the matching degree between the trajectory to be predicted and the historical trajectory data. However, human mobility behaviors have strong uncertainties. A user's current trajectory may not be the same as any historical trajectory, but it may have the same start point and end point with a historical trajectory. It is difficult to calculate the matching degree between such trajectories containing uncertainties used the existing methods.

A trajectory is made up of several location information that users have passed. If a user's mobility trajectory is considered to be uncertain, the trajectories having the same start and end points constitute an uncertain trajectory dataset representing all passing routes between these two points. That is, an uncertain trajectory is a sample set containing all possible routes between two points. Then we could obtain the behavior distribution characteristics of human from all trajectories between two points in the uncertain dataset. This distribution characteristics of human behaviors are specifically embodied as the distribution characteristics of location information in the trajectory. If it is considered that the location points in the trajectories obey a certain density distribution, it is more likely to select the location having a higher probability density in a uncertain dataset, when analyzing the current trajectory of a user.

The kernel density estimation method could be used to construct a probability density function that is more according with the actual distribution without assuming the distribution type previously. Therefore, we will use its ideas, and propose a trajectory modeling method based on kernel density estimation. The constructed model represents the distribution characteristics of locations in the trajectories of two points, and lays a foundation for the end point prediction of a trajectory with uncertain information.

2 Uncertain Trajectory Modelling Method

The ideal trajectory information contain the start point, end point, necessary path information of objects, and have less noise or redundancies. This moving mode is simple and
easy to calculate similarities between trajectories. However, in some real mobile scenarios, due to the complexity of people and society, human mobility behaviors have strong uncertainties, so that there may exist different trajectories when moving to the same target as the end point. For example, someone may choose a relatively long route to avoid traffic congestion, or choose a casual route for interests or special purposes. It is clear that on the one hand, it is difficult to accurately determine the details of the trajectories (e.g. the next accurate moving position), on the other hand, it is difficult to construct a model for a trajectory based on a single or simple hypothetical theoretical model. The existence of complexity and uncertainty make a challenge for trajectory modeling and analysis.

We do not simply ignore the uncertainty in the trajectory, but treat all the trajectories between two points in historical data as samples to build an uncertainty-oriented trajectory model between two points. So the uncertainties reflect all the possible routes from a start point to an end point, and can be represented as the distribution features of trajectories used probability density function. Based on the model describing the distribution of trajectories between points, and the individual's current movement route, we can analyze his most likely movement direction. The specific constructing trajectory model method is described as follows.

First, dataset $D$ represents a historical trajectory set, in which each one records the location information of an individual collected at different time points in a mobility behavior, that is, each record consists of the object's location information series. Sequence composition. Then the trajectory set $D = \{T^1, T^2, \ldots, T^n\}$, where $T^i (1 \leq i \leq n)$ is the $i$-th trajectory sequence. It can be denoted as $T^i = \{u_i^1, u_i^2, \ldots, u_i^n\} (1 \leq i \leq n)$, i.e. an ordered set of positions collected at $d$ time points of the $i$-th trajectory, in which $u_i^j = (x_i^j, y_i^j, t_i^j), 1 \leq j \leq d$ is the position information collected at the $j$ time point, representing that two-dimensional position information $(x_i^j, y_i^j)$ is collected at the sampling time $t_i^j$.

That is, the trajectory dataset consists of several trajectory sequences like $T^i$, which consists of position information (trajectory points) acquired at different time. The trajectory dataset can also be represented as $D = \{(u_1^1, u_1^2, \ldots), (u_2^1, u_2^2, \ldots), \ldots, (u_n^1, u_n^2, \ldots)\}$, then a location set $U$ can be acquired: $U = \{u_1^1, u_1^2, \ldots, u_n^1, u_n^2, \ldots\}$ through extracting location information from $D$. Actually, some locations in $U$ may be the same or nearby with each other geographically. Similar to the setting of a bus stop, the locations close to a certain station belong to the station in the reachable range. If you aim at a destination close to a certain station, you can get off at that station in advance. In fact, locations with similar coordinates can be treated as the same in trajectory prediction and planning application. Similarly, we can
obtain the set $V$ containing trajectory points after processing the locations having similar coordinates in $U$ as the same points. The integrated trajectory point set $V = \{v_1, v_2, \ldots, v_k\}$, apparently $|V| \leq |U|$, is helpful for reconstructing historical trajectories and analyzing their characteristics between the same origin and destination.

Then we can also reconstruct the trajectory sequence $T_i^*(1 \leq i \leq n)$ in $D$ as $T_i^* = \{v_1', v_2', \ldots, v_n'\}$, and there may exist several identical trajectory points in different trajectories. And the reconstructed trajectory dataset $D^* = \{T_1^*; T_2^*; \ldots; T_n^*\}$. We can extract trajectories from the trajectory sequences having the same origin and destination in $D^*$. These trajectories may contain different trajectory points, which reflect different behaviors of different users. And this type of trajectories with the same start point $s$ and end point $e$ as well as passing through different trajectory point sets is called an uncertain trajectory $UT$ between $s$ and $e$. Its formal description is a quadruple $UT = (s, e, ns, ts)$, consisting of the start point $s$, the end point $e$, the trajectory point set $ns$, and the trajectory set $ts$ it passing by, where $s, e \in V$, $ns \subset V$, $ts$ is the set of all trajectories passing through $s$ and $e$ in $D^*$, which can be described as:

$$ts_{s,e} = T_{s,e}^1 \cup T_{s,e}^2 \cdots \cup T_{s,e}^m.$$

Each data in $D^*$ represents an actual user's mobility trajectory sequence $T_i^*$, in which there may exist at least one trajectory between any two points. For example, $T_{v_i}^v$ is the reachable trajectory between $v_i$ and $v_j$ in $T_i^*$. If we take $\forall v_i, v_j \in V$ as the start point and the end point of the uncertain trajectory respectively, and assume that each trajectory sequence in $D^*$ includes the routes between these two trajectory points $v_i$ and $v_j$, we can obtain the uncertain trajectory $UT_{i,j} = \{\{v_i, v_j, ts_{i,j}\}\} \{v_i, v_j \in V\}$ between $v_i$ and $v_j$, where $ts_{i,j} = T_{i,j}^1 \cup T_{i,j}^2 \cup \cdots \cup T_{i,j}^m$ represents all the routes that users are likely to choose when passing these two points in historical data.

Then, through analyzing the trajectory dataset $D^*$, we can get an uncertain trajectory dataset $UTD = \{UT_{i,j}\}$, where each element represents an uncertain trajectory sequence $UT_{i,j} = \{\{v_i, v_j, ts_{i,j}\}\} \{v_i, v_j \in V, i \neq j\}$. It is clear that there is not only one reachable sub route between $v_i$ and $v_j$ in $ts_{i,j}$, but it covers all trajectories that users have passed by between two points in the collected historical data. Therefore, the routes between any two trajectory points could reflect the behaviors of different users in the mobility process, and could provide abundant information resources for the subsequent trajectory prediction.

Although it is difficult to analyze the purpose of individual when selecting their moving route based on the trajectory dataset, it is possible to analyze the distribution characteristics of
mobility behaviors between two points through the trajectories contained in $ts_{i,j}$. The trajectory points passed by the trajectories in $ts_{i,j}$ are extremely complicated and uncertain, which may have randomness due to some objective environment or subjective intentions in the mobility behaviors. Therefore, the density distribution of the trajectories generally do not obey some simple hypothetical distribution, such as normal distribution, or power law distribution, etc. The kernel density estimation method can be used to obtain the probability density function consistent with the actual distribution characteristics without assuming data distribution previously. We will use the kernel density estimation method to obtain the distribution characteristics of the trajectory points in an uncertain trajectory.

The trajectories in the uncertain trajectory $ts$ can be regarded as the samples representing the reachable paths between two points. When analyzing an uncertain trajectory, we will extract the samples about the reachable paths to get the uncertain trajectory between two points from the dataset, and use kernel density estimation to obtain the probability density function that can reflect the distribution characteristics of mobility behaviors in the uncertain trajectory. In addition, the trajectory points and their numbers may differ from each other in different trajectories, so it is difficult to directly analyze the distribution density of each uncertain trajectory. While all trajectories in $ts$ are composed of trajectory points, so when analyzing the characteristics of behaviors between two points, we can refine the analysis objects to the trajectory points, then we can construct the probability density function to describe the distribution characteristics of trajectory points in the historical data.

As defined as above, $UT = (s,e,ns,ts)$ is the uncertain trajectory between the start point $s$, and the end point $e$, $s,e \in V$. $ns_{s,e}$ is the set of trajectory points contained in all trajectories in $ts$, and $ns_{s,e} = \{v_{1}^{e},v_{2}^{e},...,v_{m}^{e}\}$, where $m$ is the number of trajectory points. According to the kernel density estimation, we can obtain the probability density function of $ns_{s,e}$: $f^{e} = \frac{1}{m_{v}} \sum_{i=1}^{m} K\left(\frac{x-v_{i}^{e}}{h}\right)$, where $x$ represents the d-dimensional independent variable. Since the probability density function is constructed for the trajectory points, the independent variable in this study is a 2-dimensional variable. $h^{d}$ represents the optimal bandwidth with $d$ dimensions. Then the density uncertainty trajectory between $v_{s}$ and $v_{e}$ is $UTf = (s,e,f)$.

In order to extract the information that is conducive to predicting and analyzing the user's trajectory behaviors, it is an important basis to extract the uncertain trajectories and construct their models between any two points. For an uncertain trajectory, the construction of its model includes not only the start point, end point of the trajectory, and multiple trajectories generated by the users between two points, but also the distribution characteristics of the trajectory.
points covered in these routes. That is, the uncertainty of a trajectory can be reflected as the multiple optional routes generated by different users, and its description can be a probability density function reflecting the distribution characteristics of the trajectory points contained in the possible routes. Actually, the model constructed no longer pays attention to the trajectory details difficult to determine, but pays more attention to the target of the trajectory and the possible trajectory patterns.

3 Experiments

We will verify the effectiveness of proposed method on GPS trajectory data Geolife collected by Microsoft Research Asia [15], a publicly available dataset commonly used in the field of trajectory research. This data is recorded by Map Life, including GPS coordinates of 182 users in 5 years, including latitude, longitude, altitude, and time. The dataset contains 17,621 trajectories, 24874410 location information, with a total distance of 1292951 kilometers and a total duration of 50176 hours.

In order to extract reasonable trajectory data, the start point and end point of a trajectory need to be reasonably determined, and the continuous position points between two points construct a reasonable trajectory. In addition to the original trajectory data in Geolife, any trajectory in the dataset may contain multiple sub-trajectories. Although these sub-trajectories do not belong to the actual complete trajectory in the original data, they are likely to be the basis for analyzing the mobility behaviors of other users. It is necessary to extract the information of these sub-trajectories and treat them as samples of uncertain trajectories. Considering that the users may break the original trajectory into several parts through some actions, such as resting and eating midway. These breaks can be seen as trajectory points and be used to extract sub-trajectories. Therefore, to obtain more abundant trajectory data, firstly we calculate the frequency of the retention time in the trajectories, and we can get frequency distribution of retention time generated by different users in different trajectories as shown in Figure 1.

Figure 1 shows that the retention time and its corresponding frequency tend to be linear in the coordinate system. The general statistical period in this trajectory dataset is 1 second or 5 seconds, and more than 5 seconds can be regarded as a break. When the retention time is longer than 10 seconds, its frequency decreases significantly. Therefore, the dwell points with more than 10 seconds are considered as a stagnation point. If the stagnation point belongs to a trajectory point in a trajectory, it is used to obtain the sub-trajectory of the original data. That
is, the The start and end points of an uncertain trajectory are defined according to the stagnation points. We obtain the trajectories passing through any two stagnation points from the original dataset, and treat them as the samples in an uncertain trajectory between these two stagnation points. Then we can construct its uncertain trajectory model: calculate the probability density function of the trajectory points contained in these trajectory samples, and analyze the distribution characteristics of the included trajectory points.

We will verify the effecttiveness of the proposed modeling method through the prediction of target of a trajectory. Given a user's trajectory, we predict his destination, according to the current trajectory points passed by already, through matching with the trajectories extracted from the dataset. Figure 2 depicts the results about prediction accuracy on the Geolife trajectory dataset. It can be seen that the prediction accuracy can reach 70% with 40% input data. As the size of the input trajectory data to be predicted continues to increase, the prediction accuracy continues to improve. Moreover, the variance of the prediction results is small, indicating that the proposed method not only has high prediction accuracy but also has good stability.

**Conclusion**

To reduce the impact of uncertainties on trajectory mining, this study uses the kernel density estimation method is to construct models for the uncertain trajectories and represent the distribution characteristics of their uncertainties through probability density functions. Kernel density estimation method could obtain more objective distribution characteristics of uncertain trajectories conforming to the actual distribution of data. The proposed method fully
considers the uncertainties reflecting mobility behaviors and any possible location in the historical trajectories. Therefore, it can obtain more accurate modeling results and lay the foundation for trajectory prediction. Finally, experiments on the real dataset also verify the effectiveness and reliability of the proposed method.

Fig.2. Prediction accuracy result

Acknowledgments. This work is sponsored by the Natural Science Foundation of Heilongjiang Province under grant number F2017015, and Training Program For Young Innovators in Heilongjiang General Institutes of Higher Education under grant number UNPYSCT-2017079.

References


A Similarity Between Uncertain Data Measurement

Method Based on stochastic simulation

Yuan Cheng¹, Ronghua Chi² and Dapeng Lang³
{changuang7@sina.com¹, chironghua@126.com², langdapeng.heu@gmail.com³}

College of Computer Science and Technology, Harbin University of Science and Technology¹,
College of Computer and Information Engineering, Heilongjiang University of Science and Technology²,
College of Computer Science and Technology, Harbin Engineering University³

Abstract. The distance measurement between uncertain data is an important basis for accurate clustering. Taking full advantage of the uncertainty characteristics of the object will help to represent the uncertain data more accurately and calculate its distance. Based on the probability distribution function to represent the characteristics of uncertainty distribution, this paper studies a method for measuring distance between uncertain objects based on stochastic simulation. The effectiveness of the proposed method is verified by experiments.

Keywords: uncertain data, distance measurement, probability density function, stochastic simulation.

1 Introduction

Under the background of the rapid development of the Internet+ and big data, data acquisition methods are more abundant, which may make data uncertain, including the uncertainties of data values caused by inaccurate raw data or the use of coarse-grained datasets, i.e. the uncertainties of attribute values. The existence of uncertainty increases the complexity of data, so that traditional data mining methods are difficult to directly use for mining knowledge.

In the clustering for uncertain data, due to the uncertainties in data, the distance measurement between data objects may be disturbed to also be uncertain, which may further
affect the clustering results. Therefore, related researches mostly construct a data model capturing the uncertainty distribution of object, measure distance between uncertain data, and finally based on a traditional clustering algorithm, such as K-Means, DBSCAN, or OPTICS, form clustering algorithms, such as UK-Means [1], FDBSCAN [2], FOPTICS [3], etc., to cluster uncertain data [4-6]. These algorithms usually represent an uncertainty object as an area of uncertainty, and use the probability density function (Probability Density Function) to build a model that conforms to the actual probability distribution of uncertain data as much as possible. That is, the data model could describe the uncertainty characteristics of an uncertain object. When measuring the distance between uncertain objects, because of the existence of uncertainty, some studies using the expected distance method [1-4], which is calculated based on the sample points representing the uncertain area, and is linearly integrated into one scalar value. Its calculation efficiency is greatly affected by the number of sample points. Therefore, in order to improve the calculation efficiency, [1,6] reduce the amount of calculation when calculating the expected distance based on the pruning strategy, and improve the calculation efficiency of the expected distance by defining the cluster boundary. [7] calculates the expected distance based on information theory; and [8] improves the calculation efficiency of the expected distance based on the center boundary. The difference between uncertain objects is measured using the expected distance, and the distance described by probability can be linearly integrated into a scalar value. Another research idea is to calculate the similarity between objects based on uncertain data model. The uncertainty model represents the distribution characteristics of uncertainty, so the distance between uncertain objects represented by this distribution function can be obtained by comparing the differences between the distribution functions [9,10]. KL divergence is a common method for measuring the difference of probability distribution. However, a major problem of KL divergence is that it does not have symmetry, while the distance between objects in data mining needs to be symmetric. Therefore, the difference between the uncertainty distributions based on the KL divergence is not the most appropriate. We will propose a method to measure the distance between uncertain objects to solve these problems.

2 Similarity Measurement Based on Stochastic Simulation

2.1 Distance measurement between uncertain objects
The uncertain data model uses probability density function to represent the distribution characteristics of uncertainty. The difference between functions is mainly reflected as the difference of the area formed by the definition domain and the function curve. The stochastic simulation is an effective method to calculate such area. There are two probability density functions representing the uncertainties of two objects A and B in Figure 1. The difference between the two functions in the definition domain \([a, b]\) is shown as the shaded part in Figure 1, that is, the difference in area of non-intersecting parts formed by two curves and x axis in the definition domain. Therefore, we calculate the distance between uncertain data represented by the probability model based on stochastic simulation method.

Through the stochastic simulation method, we perform a large number of random samplings on the definition domain of the probability density function, and perform the integral operation on the difference of the corresponding function values to obtain an approximate difference with higher accuracy. For uncertain objects \(x_i, x_j\) and \(y_i\), the probability density functions describing their uncertainty distribution characteristics are \(f_{x_i}\) and \(f_{y_i}\). Then we generate \(I\) random numbers: \(\tilde{z} \in I_{x_i} \cup I_{x_j} (1 \leq j \leq I)\), in the union of their definition domains. After the random numbers are generated, we can obtain the integral of the difference between the probability density function in the specific definition domain, which represents the difference between the probability of the values of uncertain objects. The difference is the distance between uncertain data objects, which is shown in equation (1).

\[
d'(x_i, y_i) = \langle z - c \rangle \int \left| f_{x_i}(z) - f_{y_i}(z) \right| dz
\]

(1)

The above method is described as follows:

\[\text{DBU}(\mathcal{E}_x, \mathcal{E}_y, I_{x_i}, I_{x_j}, I)\]

Input: \(\mathcal{E}_x, \mathcal{E}_y\): the probability density functions of uncertain objects \(x_i, x_j\) and \(y_i\);
\(I_{x_i}, I_{x_j}\): the definition domains of \(x_i, x_j\) and \(y_i\);
\(I\): the number of generated random numbers;
Output: \(d'\): the distance between \(x_i\) and \(y_i\).

Steps:
1) Define \([a, b] = I_{x_i} \cup I_{x_j}\);
2) Let \(d' = 0\);
3) for \(z\) from 1 to \(I\):

\[\text{DBU}(\mathcal{E}_x, \mathcal{E}_y, I_{x_i}, I_{x_j}, I)\]

Input: \(\mathcal{E}_x, \mathcal{E}_y\): the probability density functions of uncertain objects \(x_i, x_j\) and \(y_i\);
\(I_{x_i}, I_{x_j}\): the definition domains of \(x_i, x_j\) and \(y_i\);
\(I\): the number of generated random numbers;
Output: \(d'\): the distance between \(x_i\) and \(y_i\).

Steps:
1) Define \([a, b] = I_{x_i} \cup I_{x_j}\);
2) Let \(d' = 0\);
3) for \(z\) from 1 to \(I\):
4) generate $\varepsilon$ random numbers in $I_{\alpha} \cup I_{\beta}$;
5) $d^i = (\beta - \alpha) \times (\tilde{a}_i - \tilde{b}_i)$;
6) end for
7) return $d^i$.

Fig. 1. Distance measurement between uncertain objects

2.2 Distance measurement on clusters

In different clustering processes, in addition to measuring the distance between objects, dividing the clusters that objects belong to also involves measuring the distances about clusters, including: ① the distance between uncertain data object and cluster, ② the distance between the clusters.

In the uncertainty-oriented clustering method, the objects in the clusters are described by the uncertain data model. It is difficult to measure distance directly based on the set of uncertain models represented by the probability density function. Therefore, we can firstly extract the cluster characteristics based on uncertain data model, and then use the above DBU method to measure the distance about clusters represented by the data model.

Figure 2 shows a set of probability density functions of objects within a cluster. Firstly, based on the stochastic simulation method, $\tilde{l}$ random numbers are generated in definition domain, and the values of each random number on each probability density function are obtained to the set $\tilde{e}_j(1 \leq j \leq l)$. If they are regarded as the samples, we can get its uncertain data model $\tilde{E}_i$ based on [11] for cluster $\tilde{e}_i$. Therefore, the clusters consisting of uncertain data
can be expressed as a form of uncertain data model based on stochastic simulation, which is used to represent the characteristic of the cluster.

The distance between uncertain data object and the cluster can be converted into a distance between the uncertain data object and an uncertain model representing the cluster characteristics. While the distance between clusters can be transformed into the distance between two uncertain data models representing cluster characteristics. Then according to the above DBU method, the distance between the probability density functions, i.e. uncertain data model, could be calculated based on the random simulation.

3 Experiments

This paper will compare the effectiveness of distance measurement methods on 5 UCR datasets: Beef, Coffee, Fish, OliveOil, Trace [12], with the other two methods: expected distance [1], and KL divergence based method [10]. The accuracy of the clustering results could indicate the effectiveness of the proposed distance measurement method. We use Entropy and Purity [13] to measure the clustering accuracy. The smaller the value of the former is, the better the clustering is. The latter is the opposite. With the same clustering method, the more accurate the clustering is, the better the distance measurement is.

There is no uncertainties in the original UCR datasets, we need to construct uncertainties based on the method mentioned in [14] first. The uncertainties can be described by the samples representing the possible values, so we choose Gaussian distribution to generate \( z \) samples for each object \( o \) in dataset \( D \). There is a parameter \( \sigma \in [0, \varepsilon] \) that is the unified standard deviation when using Gaussian distribution, where \( \varepsilon \) represents the uncertainty. Generate a center point \( \xi = \sigma \), and then generate the other \( z - 1 \) samples based on Gaussian
distribution with a $\sigma$ value and $\mu = \mu$ for any object $o$ in $D$. If we choose different parameters, we could obtain different probability density functions for uncertainties of each object $o$ in $D$. Then we can get uncertain objects with different uncertainty distribution characteristics.

Therefore, we will use the above method to generate the uncertainties for the UCR datasets, and verify the effectiveness of our proposed method, comparing with the other two distance measurement methods, based on the idea of K-Means clustering algorithm.

**Figure 3** shows the comparison results of clustering accuracy. Compared with the other two distance measurement methods, it can be seen that the proposed method could obtain more accurate clustering results on the datasets. This is because the construction of uncertain data models and the distance measurement between uncertain objects both make full use of the characteristics of uncertainty.

![Figure 3](image)

**Fig.3.** Comparison of clustering accuracy on UCR datasets

**Figure 4** shows the comparison of the execution time of the clustering process when the number of sample points are gradually increased. The results indicate that on several datasets,
the DBU method performs better than the expected distance, and is close to the execution time of the method based on the KL divergence. It shows that the proposed method could not only help to obtain more accurate clustering results, but also have higher calculation efficiency.

**Conclusion**

This study is oriented to data with attribute uncertainty. To solve the problems in the calculating distance between uncertain objects, a distance measurement method based on stochastic simulation is proposed based on the uncertainty distribution characteristics represented by probability density function. The proposed method could not only ensure the accuracy of the metric due to the calculation based on uncertainty distribution characteristics, but also improve the efficiency of the distance calculation used stochastic simulation. The validity and accuracy of the proposed method are verified through experiments on UCR datasets.

**Acknowledgments.** This work is sponsored by the Natural Science Foundation of Heilongjiang Province under grant number F2017015, and Training Program For Young Inno-
vators in Heilongjiang General Institutes of Higher Education under grant number UNPYSCT-2017079.

References


Discriminative Graph Learning via Low-Rank Constraint for Semi-Supervised Learning

Ying Wang¹, Ao Li¹, Gunaglu Sun¹, Xin Liu¹
{viewwangying@163.com, dargonboy@126.com, sunguangl@hrbust.edu.cn}

School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China¹

Abstract. Graph learning is more and more widely used in discovering associations and mining relationships between data and data. Our paper proposed a discriminative graph learning for semi-supervised learning method discriminative graph learning with low rank constraints. Traditional semi-supervised learning methods exist the problem of similarity between similar species is not high. How to improve the similarity between similar species has become a problem we need to solve in this field. In order to solve this problem, this paper proposed a new semi-supervised learning means, which mainly includes two parts: 1) Representing the local structure of the research sample by low-rank representation, compressing similar content, and easily finding the focus of image classification; 2) In order to improve the similarity between classes, a model that uses similarity constraints between the data to be classified. The proposed means is used for comparative experiments of classification on multiple data sets. Through experiments show that the means proposed in this paper performs better than traditional semi-supervised classification methods in terms of accuracy of classification.

Keywords: Image classification, graph learning, low-rank constraints, semi-supervised.

1 Introduction

With the increasing application of face recognition and image classification technology in our daily lives, how to extract features more conducive to classification from high-dimensional and complex data has become the focus of today's research [1]. Recently, because supervised learning and unsupervised learning have many problems in real life and cannot practical applications, researchers have combined supervised learning with semi-supervised learning and proposed semi-supervised learning [2]. Semi-supervised learning is to learn the features of known label images to extract the features that have a higher proportion of image classification, and classify the image data of unknown labels according to the learning results[3][4][5].

Inspired by the methods proposed by the above researchers, this paper proposes a discriminative graph learning with low rank constraints for semi-supervised learning. This method puts low rank constraints and semi-supervised learning into the same framework and expresses low rank Coefficients and discriminant graph learning errors refer to semi-supervised learning as two different constraints, which improves the similarity between classes, makes semi-supervised learning more accurate, and reduces the time to mark unknown label data.

The main contributions of this paper are:

1) Put low-rank constraints and semi-supervised learning into the same method, and use low-rank representations to transform high-dimensional data into low-rank representations, so that semi-
supervised learning becomes less computational and takes less time.

2) In order to improve the accuracy of semi-supervised learning for the classification of unknown label data, a discriminative graph learning error is constructed, which improves the accuracy of semi-supervised learning.

2 Method description

2.1 Low rank constraints for semi-supervised learning

2.1.1 Low rank constraints

Since the low rank model was proposed, the low rank model has been applied to various fields such as subspace learning[6], subspace clustering[7], and image processing[8][9]. The low-rank model can represent the sample data with a low-dimensional structure. This representation can play a role in eliminating the noise of the sample data to a certain extent. Compared with the method of directly processing the dry data by using the low-rank model to eliminate the noise of the sample data, the performance of this method is more prominent and increases the robustness of semi-supervised learning. The function of this model is:

\[ \|Z\|_* + \|E\|_1 \] (1)

The above formula \(\|Z\|_*\) represents the kernel norm of the matrix and \(\|E\|_1\) represents the \(L_1\) norm of the matrix, which is often used for data containing outliers. The model \(\|Z\|_* + \|E\|_1\) is a low-rank representation model. By applying a low rank constraint to the coefficient representation matrix \(Z\), it is easier to find the local similarity of the training sample itself [10], and it is beneficial to eliminate noise interference in the sample data.

2.1.2 Discriminative graph learning error

In order to further enhance the accuracy of semi-supervised learning so that the method can maintain high performance even under noisy conditions, we propose a discriminant graph learning error constraint term. The constraint term is obtained by training the sample data and finding the matrix \(D = [d_1, \ldots, d_n] \in R^{K \times N}\) according to the training result. The matrix \(D\) is composed of 0 and 1.

When only one value in a column is 1 and the other values are 0, the image belongs to this kind. For example, suppose \(Z = [z_1, \ldots, z_4]\) exists, where \(z_1\) belong to class 1, \(z_2, z_3\) belong to class two, \(z_4\) belong to class three, then \(D\) should be defined as

\[
D = \begin{bmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 \\
0 & 1 & 1 & 0 \\
0 & 0 & 0 & 1 \\
\end{bmatrix}
\]

According to \(D\) defined above, we construct the discriminant graph learning error as

\[ \|WZ - D\|^2_F \] (2)

The purpose of constructing this constraint term is to reduce the error when classifying unknown label data.
The idea of constructing this term is that we want to reduce the error in data classification of unknown labels as much as possible, so ideally it is \( \|Z - D\|_F^2 \), but because of the error, we construct the coefficient matrix \( W \), and the term construction is complete.

Among them, \( \| \cdot \|_F^2 \) is the Frobenius specification, which is commonly used to process data containing Gaussian noise. The matrix \( W \) is the coefficient matrix.

According to the definition of the Frobenius-norm, equation (2) is equivalent to the following equation:

\[
\text{Tr} \left\{ (WZ - D)^T \wedge (WZ - D) \right\}
\]

According to the two constraints described above, the function model can be defined as

\[
\min_{Z,E,W} \|Z\|_F + \lambda \|E\|_F + \text{Tr} \left\{ (WZ - D)^T \wedge (WZ - D) \right\}
\]

\[\text{s.t.} \ X = XZ + E \]

2.2 Optimization

To solve the convergence of the objective function (4), we use the inexact ALM algorithm. In order to make our solution more convenient, the auxiliary variable \( J \) is introduced, so the objective function can be written as:

\[
\min_{Z,E,J,W} \|J\|_F + \|E\|_F + \text{Tr} \left\{ (WZ - D)^T \wedge (WZ - D) \right\}
\]

\[\text{s.t.} \ X = XZ + E, Z = J \]

According to the Lagrangian multiplier method, the Lagrangian function of equation (5) is:

\[
\min_{Z,E,J,W} \|J\|_F + \|E\|_F + \text{Tr} \left\{ (WZ - D)^T \wedge (WZ - D) \right\} + \mu \left( \|X - XZ - E\|_F^2 + \text{Tr}(S^T(Z - J)) + \frac{\mu}{2} \|Z - J\|_F^2 \right)
\]

\[\text{s.t.} \ X = XZ + E, Z = J \]

In order to solve the above formula, the ALM algorithm that has been proposed is used to solve the unknown variable in iterations and find the minimum value. The approach we take is to solve only one unknown variable during each solution and treat the other unknown variables as constants.

We first solve the low rank representation \( J \). When solving \( J \), we consider other unknown variables as fixed values. After \( K \) iterations, equation (6) can be written as:

\[
\min_J \|J\|_F + \text{Tr}(S^T(Z - J)) + \frac{\mu}{2} \|Z - J\|_F^2
\]

The above formula can be rewritten as:

\[
\min_J \frac{1}{\mu} \|J\|_F + \frac{1}{2} \|J - (Z + R/\mu)\|_F^2
\]

We use the (SVT) method in the literature [11] to solve problem (8).

To initialize \( W \), we use a multiple ridge regression model [12]. The model consists of the quadratic loss and \( L_2 \) norm regularization. So \( W \) can be updated as:
\[
W = \arg \min_{W} \|WX - Q\|^2 + \lambda_2 \|W\|^2_2
\]  
(9)

Ignore the variable (5) that has nothing to do with \(Z\), the original formula can be rewritten as:

\[
\min_{Z,E,J,W} \left\{ (WZ - D)^\top \wedge (WZ - D) \right\} + \text{Tr}(R^\top (X - XZ - E)) \\
+ \frac{\mu}{2} \|X - XZ - E\|^2_F + \text{Tr}(S^\top (Z - J)) + \frac{\mu}{2} \|Z - J\|^2_F
\]  
(10)

Find the partial derivative of the above formula about \(Z\), and set the partial derivative equal to 0 to solve \(Z\).

Similarly, in order to solve \(E\), we omit variables not related to \(E\), then we can rewrite equation (5) as:

\[
E = \min_{E} \frac{\lambda_1}{\mu} \|E\| + \frac{1}{2} \|E - (X - AZ + Y/\mu)\|_F^2
\]  
(11)

The solution of the above formula is proposed in [12]. When we let \(\Psi = X - AZ + Y/\mu\), then then i-th first column in \(E_{k+1}\) can be expressed as:

\[
E^{k+1}(i,:) = \begin{cases} 
\|\Psi\| - \frac{\lambda_1}{\mu} \Psi, & \text{if } \frac{\lambda_1}{\mu} < \|\Psi\| \\
0, & \text{otherwise}
\end{cases}
\]  
(12)

According to the previous analysis, we summarize the previous solution method as the following algorithm:

**Algorithm 1:**

Input: data set \(X\), parameter \(\lambda_1\), \(Z = J = 0\), \(E = 0\), \(W = 0\), \(\mu = 0.1\), \(\mu_{\text{max}} = 10^{10}\)

\[S = 0, R = 0\]

Output: \(Z_k, W_k, E_k\)

1. While not convergence do
2. Updated \(J_{k+1}\) using (8).
3. Updated \(W_{k+1}\) using (9).
4. Updated \(Z_{k+1}\) using (10).
5. Updated \(E_{k+1}\) using (12).
6. Updated \(R_{k+1}\) and \(S_{k+1}\)

\[
R_{k+1} = R_k + \mu_k (X - XZ_{k+1} - E_{k+1})
\]
\[
S_{k+1} = S_k + \mu_k (Z_{k+1} + J_{k+1})
\]
7. \( k = k + 1 \)
8. End while

3. Experiments

In the experiment, we compare the method proposed in this paper with several existing semi-supervised learning methods such as FME, RLSR, RRPC, and SDA. So as to have a more accurate understanding of the performance of ours.

First, we compared our method with several means in Extended YaleB[13]. When we set the training sample to 10, the comparison data is as follows:

Table 3.1 Comparison of accuracy in Extended YaleB

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FME</td>
<td>77.56</td>
</tr>
<tr>
<td>RLSR</td>
<td>83.47</td>
</tr>
<tr>
<td>RRPC</td>
<td>82.81</td>
</tr>
<tr>
<td>SDA</td>
<td>74.67</td>
</tr>
<tr>
<td>Ours</td>
<td>88.49</td>
</tr>
</tbody>
</table>

We compared our method with several other methods in AR Face[14].

Table 3.2 Comparison of accuracy in Extended AR Face

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FME</td>
<td>85.69</td>
</tr>
<tr>
<td>RLSR</td>
<td>84.43</td>
</tr>
<tr>
<td>RRPC</td>
<td>85.41</td>
</tr>
<tr>
<td>SDA</td>
<td>74.67</td>
</tr>
<tr>
<td>Ours</td>
<td>91.74</td>
</tr>
</tbody>
</table>

We compared our method with several other methods in COIL20.

Table 3.3 Comparison of accuracy in Extended COIL20

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FME</td>
<td>69.74</td>
</tr>
<tr>
<td>RLSR</td>
<td>75.26</td>
</tr>
<tr>
<td>RRPC</td>
<td>65.84</td>
</tr>
<tr>
<td>SDA</td>
<td>72.15</td>
</tr>
<tr>
<td>Ours</td>
<td>76.12</td>
</tr>
</tbody>
</table>

From the above experimental results, we can see that our proposed method has higher accuracy than other semi-supervised learning methods.

4. Conclusion

This paper proposes a discriminant graph learning with low rank constraint for semi-supervised learning to classify data of unknown labels. This method designs two constraint terms based on low rank representation, discriminant graph learning error and iterative solution method. ALM effectively solves the convergence problem of the objective function, and compares it with the existing semi-supervised learning methods on three different data sets. Through experiments we learned that the method proposed in this paper is better than some current semi-supervised learning methods.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China under Grant 61501147, in part by the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province under Grant UNPYSCT-2018203, in part by the Natural
Science Foundation of Heilongjiang Province under Grant YQ2019F011, in part by the Fundamental Research Foundation for University of Heilongjiang Province under Grant LGYC2018JQ013, and in part by the Postdoctoral Foundation of Heilongjiang Province under Grant LBH-Q19112.

References


Interaction representation-based subspace learning for domain adaptation

Jiajia Chen 1, Ao Li 2, Deyun Chen 3, Kezheng Lin 4
1. School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China
2. School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China
Email: {544953065@qq.com(Jiajia Chen), dargonboy@126.com (Ao Li)}

Abstract
Generally, Transfer Learning (TL) or Domain Adaptation (DA) are used to solve domain inconsistency, but the conventional domain adaptation methods mostly only consider local information and ignore global information, and only consider one-way data and ignore the possibility of two-way data. Therefore, in this paper, we proposed an interactive representation-based framework for domain adaptation. In the novel framework, two low-rank based interactive representation models are built on both of source and target domains, which can be used to better align distribution discrepancy. Then, a distance constraint is designed to model the subspace relationship between source and target domain. Finally, the label-based regression is jointly used to earn extra discrimination for classification. Experiments on a number of public databases demonstrate that our method has competitive performance among comparison methods.

Keywords: Subspace learning, Domain adaptation, Interaction representation

1 Introduction
In most machine learning methods, training set and test set are generally considered to have similar distribution. However, in the real world, the training and test set are affected by light, Angle, background, Fig.1 show a few of pictures of different distributions, Therefore, the question of domain inconsistencies arises, that is, training sets and test sets cannot have similar distributions. However, domain inconsistency will lead to the reduction of classification accuracy when classifying images[2]. Usually, the solution to the problem of domain inconsistency is Transfer Learning or Domain Adaptation[3].

Supervised[4], semi supervised[5] and unsupervised[6] are three DA strategies. Domain adaptation can be used to solve the question of data coming from two different domains but related to each other. The method of domain adaption is to explore the invariant structure in the two domains, through which a subspace can be found to connect the source and target domain[12]. In this paper, we think that the domain adaptive method only considers the unidirectional alignment of data sets and ignores the possibility of bidirectional alignment. Therefore, we advance a subspace learning method based on interactive representation. The main contents of this method are as follows:

(1) Inspired by the alignment of two domains, we propose an interactive representation model based on two low rank. In this model, we make the two domains align with each other. In addition to making the target domain use the linear representation of the source domain, we can
also make the source domain use the linear representation of the target domain.

(2) A distance constraint is added to the model, which makes the two domains approach each other infinitely, and further achieves the effect of mutual representation.

(3) In order to obtain extra discrimination information, we explore a great of label information in source domain to strengthen the robustness of the model.

![Examples of different distributions in dataset CMU](image)

Fig. 1. Examples of different distributions in dataset CMU

## 2 Related Works

We will introduce two topics related to research work, namely low rank and subspace learning.

### 2.1 Subspace Learning

In the field of pattern recognition and image classification, Subspace learning has been used. It is expected to find a subspace in which all required data are kept. Subspace level method can align the features of both domains. For example, PCA will project data along the direction of the maximum variance. PCA pursues to maximize the internal information of data after dimensionality reduction. It does not consider classification information. LDA seeks the most effective recognition direction by minimizing the ratio between in-class and inter-class scatterers. What LDA pursues is that data points after dimension reduction can be distinguished as easily as possible. LDA pursues that the data points after dimensionality reduction can be distinguished as easily as possible, that is, the data has better separability in the low dimensional space, so that the data samples after dimensionality reduction have not only the largest distance between classes, but also the smallest intra class variance in this new dimensional space. SA[1] mainly searches for a linear mapping by aligning the subspaces covered by vector features, which can find a domain invariant feature space. The ULRA method proposes a special logarithmic determinant function to approximate the rank function. It reduces the contribution of large singular value to kernel norm while keeping the contribution of small singular value close to zero[15].

### 2.2 Low-Rank

During the old days, low-rank learning has become a study hotspot and has been widely used in many applications[7], such as image clustering, discriminant subspace learning and outlier test. When data is affected by noise, Wright[8] proposed a robust principal component analysis method for recovering the low-rank property of data structure. RPCA is a typical low order method of study. The principle of the observation matrix can be broken down into low rank matrix and the sum of sparse matrix. The purpose of RPCA is to recover from low rank matrix to the observation matrix. It can correspond to a rank minimum optimization problem, in the shadow removal and background modeling has a good effect in actual application. When images, high-dimensional data and some unbalanced data are encountered, Peng proposes a classification method based on Discriminant ridge regression (DRM). DRM does not need to find binary similarity, but estimates
the representation of a new example as a soft similarity vector by minimizing the fitting error in
the regression framework, so that the model has a continuous neural network supervision label[14].

Low rank said it would RPCA from a single subspace extends to more subspace, but most of
the low rank representation may encounter the problem of insufficient samples. If only to make a
study of domain data information for the final model will be no good robustness. Therefore, the
potential of low rank representation method was used to solve this problem are put forward[9][10].

Recently, a few studies have proposed the introduction of low rank constraints in transfer
learning[11]. The learning method of low-order transfer subspace is to constrain the Shared
subspace of two domains, while the adaptive method of low-rank domains is to use low-rank
representations to reduce the difference in domain distribution[13]. Our method is different from
the method above, which only adopts one-way domain alignment in subspace learning, while our
proposed IRSL method is the interactive representation of two domains, not only the training set
alignment test set, but also the test set alignment training set.

3 IRSL: Interaction representation-based subspace learning

3.1 Mathematical Notation

We define the characters required below. The identifier of the source domain is
\( S = \{X_s, y_s\} \), The identifier of the source domain is \( T = \{X_t, y_t\} \), where \( X_s \in \mathbb{R}^{D \times n_s} \) and
\( X_t \in \mathbb{R}^{D \times n_t} \) are samples, \( y_s \) and \( y_t \) are labels. In the source domain, \( n_s \) represents the amount
of samples and \( P_s \in \mathbb{R}^{D \times d} \) represents the projection. In the target domain, where \( P_t \in \mathbb{R}^{D \times d} \)
represents the projection, \( n_t \) represent the amount of samples. \( d \) represents the dimension of the
primitive samples. \( d \) is the dimension of the subspace. \( Z \in \mathbb{R}^{n_t \times m} \) represents the reconstruction
matrix.

3.2 Formulation

In this section, the IRSP method can solve the domain inconsistency problem by learning a
subspace. In IRSP, we introduce a subspace learning model based on interactive representation.

First of all, we can learn \( P_t \) from the invariable information in the data. Then, we expect to find
an invariant subspace. In this subspace, the target domain and the source domain are infinitely
close. Therefore, the target domain and the source domain can be expressed linearly with each
other. In this way, the two domains achieve the interactive effect, and the two domains achieve a
better effect in alignment with each other. Mathematically, A low-order model can be formed, that
is, a low rank constraint is used to realize the reconstruction matrix, and its expression is as
follows:

\[
\begin{align*}
\min_{P_t, P_s, P_t} & \left\| P_t^T X_t - P_s^T X_s Z \right\|_F^2 + \lambda \| Z \|_F + \left\| P_s^T X_s - P_t^T X_t Z \right\|_F^2 + \eta \| Z \|_2 \\
\end{align*}
\]

Then, we hope to settle the question of domain inconsistency by aligning the two domains.
Therefore, we set the distance constraint on the subspace of the two domains so that the two fields are close. This distance constraint can be achieved by minimizing L2 norm:

$$\min_{P_i, r_i} \left\| P_i - P_r \right\|_2^2$$  \hspace{1cm} (2)

This formula can retain the useful information of two datasets to a great extent, meanwhile, adjust the subspaces of two domains. Therefore, a better $P_t$ can be obtained.

For the final classification model, it is not enough to use only subspace to distinguish, and extensive label information in the source domain is ignored. Therefore, we suggest labels in the two domains into the classification model to obtain additional discrimination information and improve the resolution of the model. In order to make the best of the known label information in the source domain, we define the constructed label matrix $Y$ as:

$$Y_{[i, j]} = \begin{cases} 1, & \text{if } x_j \in c_i \\ -1, & \text{otherwise} \end{cases}$$ \hspace{1cm} (3)

The purpose of introducing labels into the model is to find a discriminative $P_t$, which is similar to a subspace between domains. The formula can be expressed as follows:

$$\min_{P_i} \left\| P_t^T X - Y \right\|_2^2$$ \hspace{1cm} (4)

Where $X = [X_s, X_t]$.

We can get the final formula by Eqs. 1, 2 and 4, as follows:

$$\min_{r, P, X, Z, L} \left\| P_t^T X_i - P_t^T X_s, Z, X_s \right\|_2^2 + \lambda \left\| L \right\|_1 + \left\| P_t^T X_s - P_t^T X_s, Z_s \right\|_2^2 + \eta \left\| Z_s \right\|_1 + \beta \left\| P_t - P_r \right\|_2^2 + \frac{1}{2} \left\| P_t X - Y \right\|_2^2$$ \hspace{1cm} (5)

Where $\lambda, \eta$ and $\beta$ are trade-off paramenters to balance the constraints.

3.3 Optimization

It can be seen from formula (5) that when $Y$ is fixed, four variables will be involved. Therefore, solving formula (5) by the method of inexact augmented Lagrange multiplier method (IALM), in which two auxiliary variables L1 and L2 are involved. Formula (5) is converted into:

$$\min_{r, P, X, Z, L} \left\| P_t^T X_i - P_t^T X_s, Z, X_s \right\|_2^2 + \lambda \left\| L \right\|_1 + \left\| P_t^T X_s - P_t^T X_s, Z_s \right\|_2^2 + \eta \left\| Z_s \right\|_1 + \beta \left\| P_t - P_r \right\|_2^2 + \frac{1}{2} \left\| P_t X - Y \right\|_2^2$$ \hspace{1cm} (6)

s.t. $Z_1 = L_1, Z_2 = L_2$

Using the variable alternation strategy, we can get the following results by derivation:

$$P_t = \left(2\partial + 2X_t X_t^T - 2X_t Z_t Z_t^T X_t^T + XX^T \right)^{-1} \left(2\partial P_t + 2X_t Z_t Z_t^T P_t X_t^T - 2X_t Z_t Z_t^T P_s + XY \right)$$ \hspace{1cm} (7)

$$P_s = \left(2\partial + 2X_s Z_s Z_s^T X_s^T + 2X_s X_s^T \right)^{-1} \left(2\partial P_t - 2X_s Z_s Z_s^T P_t + 2X_s Z_s Z_s^T P_t \right)$$ \hspace{1cm} (8)

$$Z_t = \left(2P_t X_t^T P_t X_t + \mu \right)^{-1} \left(2P_t X_s^T P_t X_s + \mu \left( L_1 - \frac{Y}{\mu} \right) \right)$$ \hspace{1cm} (9)
\[
Z_2 = \left(2P_t X_t^T P_t^T X_t + \mu \right)^{-1}\left(2P_t X_t^T P_t^T X_t + \mu \left(L_2 - \frac{Y_2}{\mu} \right) \right) \tag{10}
\]

\[
L_1 = \arg \min_{L_1} \|L_1\|_2 + \frac{\mu}{2} \left\|Z_1 - L_1 + \frac{Y_1}{\mu} \right\|_2^2 \tag{11}
\]

\[
L_2 = \arg \min_{L_2} \|L_2\|_2 + \frac{\mu}{2} \left\|Z_2 - L_2 + \frac{Y_2}{\mu} \right\|_2^2 \tag{12}
\]

where \( Y_1 \) and \( Y_2 \) are Langrange multiplier, \( \mu > 0 \) is a penalty parameter. \( P_t \) can be obtained by IALM algorithm.

### 4 Experiment

**A. Data set**

4DA data set: There are four domains in 4da: Amazon (A), Webcam (W), DSLR (D) and Caletch(C). There are 10 shared categories in these four domains. We experiment with these four domains alternately as source and target domain, that is to say, we have carried out cross domain experiments of 12 tasks.

CMU PIE Face data set: PIE includes 41368 pictures taken by 68 people in 13 different postures and 21 different kinds of light. The size of these 41368 pictures is 32x32. We divide these pictures into 5 different postures, P1 is left postures, P2 is up postures, P3 is down postures, P4 is positive postures, P5 is right postures.

**B. Experimental setting**

The dimension \( d \) of the learning subspace is set to the category number \( c \) in each data set. According to a large number of experiments, the optimal accuracy can be obtained by setting the trade-off parameter \( \alpha \) to 25, as shown in Figure 2.

**C. Comparisons With Other Approaches**

In this section, we compare with SA, JDA and TSL in the above two datasets. Table 1 and Table 2 shows the results of the above method comparison. From the table, we can see that the classification performance of the proposed IRSP method is the best.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>C→A(1)</td>
<td>48.02</td>
<td>51.46</td>
<td>52.30</td>
<td>57.10</td>
</tr>
<tr>
<td>C→W(2)</td>
<td>31.86</td>
<td>41.36</td>
<td>40.34</td>
<td>57.97</td>
</tr>
<tr>
<td>C→D(3)</td>
<td>42.68</td>
<td>46.50</td>
<td>49.04</td>
<td>50.32</td>
</tr>
<tr>
<td>A→C(4)</td>
<td>34.37</td>
<td>43.90</td>
<td>43.28</td>
<td>44.70</td>
</tr>
<tr>
<td>A→W(5)</td>
<td>33.90</td>
<td>33.90</td>
<td>34.58</td>
<td>47.80</td>
</tr>
<tr>
<td>A→D(6)</td>
<td>38.85</td>
<td>33.76</td>
<td>38.85</td>
<td>47.13</td>
</tr>
<tr>
<td>W→C(7)</td>
<td>30.01</td>
<td>31.17</td>
<td>31.43</td>
<td>38.47</td>
</tr>
<tr>
<td>W→A(8)</td>
<td>32.15</td>
<td>36.33</td>
<td>34.66</td>
<td>41.34</td>
</tr>
<tr>
<td>W→D(9)</td>
<td>83.44</td>
<td>77.71</td>
<td>79.62</td>
<td>83.35</td>
</tr>
</tbody>
</table>
### 5 Conclusion

In this paper, we propose a subspace learning based on interactive representation framework to solve the question of domain inconsistency. In this framework, we align the two fields and make the two fields as close as possible. In order to achieve this better, we further constrain the distance of the model to make the two domains closer. Finally, in the source domain, we take full advantage of the known label information to obtain additional classification information, so as to increase the robustness of the model. A great of experiments show that our way is superior to the existing method.

**Acknowledgment.** This work was supported in part by the National Natural Science Foundation of China under Grant 61501147, in part by the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province under Grant UNPYSCT-2018203, in part by the Natural Science Foundation of Heilongjiang Province under Grant YQ2019F011, in part by the Fundamental Research Foundation for University of Heilongjiang Province under Grant LGYC2018JQ013, and in part by the Postdoctoral Foundation of Heilongjiang Province under Grant LBH-Q19112.

**References**


Joint Cross-view Heterogeneous Discriminative Subspace Learning via Low-rank Representation

Yu Ding¹, Ao Li¹, Kezheng Lin¹, Xin Liu¹
{xuedinger66@163.com(YuDing), dargonboy@126.com(Ao Li), link@hrbust.edu.cn(Kezheng Lin)}

¹School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China

Abstract. The cross-view data are very unforced to capture due to the fact that different viewpoints or data collected from different sensors are already very common in recent years. However, cross-view data from different views present a significant difference, that is, cross-view data from different categories but in the same view have a higher similarity than the same category but within different views. To solve this problem, we have developed a dual low-rank representation framework to unbind these interleaved structures in a learning space. In addition, we consider that each cross-view sample of the same category is from isomorphic and heterogeneous information of two interlaced structures. Hence, we propose a powerful joint cross-view heterogeneous subspace feature learning model. In addition, the subspace learned by our algorithm contains more useful information and is more adaptable to cross-view data.

Keywords: Multi-view discriminative analysis, Cross-view subspace learning, Low-rank representation.

1 Introduction

Cross-view data, which depicts the same object from multiple angles, and contains richer classification and identification information than traditional single-view data, has been a hot research topic in the past few years[1]-[4]. Because different views data often has some special information, the samples of different views may be located in independent spaces. This phenomenon makes retrieval and recognition of objects or people very difficult. Hence, the requirements for the application of cross-view data methods are increasing.

In the past decade, many cross-view methods have emerged and achieved satisfactory results. Most of these methods learn a view-specific projection, projecting all views into a common or invariant subspace, where different views are represented as common or invariant representations. The most classic method is Canonical Correlation Analysis (CCA)[5], which establishes the relationship between two different views by maximizing the correlation coefficient of the two views and projects it into a common shared subspace. However, CCA can only be used in the case of two views. In order to overcome this difficulty, Multi-view CCA (MCCA)[6] was proposed. MCCA obtains multiple view-specific transformations between multiple views by maximizing the correlation coefficients between multiple views. In [7], Multi-view Discriminant Analysis (MvDA) was proposed, which adds Fisher constraints to the multi-view projection subspace to allow different views data to be projected into a discriminative shared subspace.
These methods usually project different views data into the same common subspace[8]-[10], and we find that cross-view data presents two different but intertwined structures in the original high-dimensional space, for example, a sample from cross-view data has both a class structure representing its own semantic information and a view structure containing its view information. The heterogeneous information brought by the view structure can affect the performance of retrieval and recognition. Therefore, we need to reduce the variance of the view to decompose the two structures to obtain the cross-view discriminant subspace. In addition, the dual low-rank representation method has been widely used recently[11]-[13], which has the ability to resist noise and makes the learned subspace robust. However, they ignore the similarity information between different views, for example, different views of the same sample contain part of the same feature information and unique information for each view. Therefore, cross-view subspace learning should not only discriminate learning based on the manifold structure of data, but also use complementary information between different views.

In this paper, we propose a novel cross-view discriminative feature subspace learning model called Joint Cross-view Heterogeneous Subspace Learning (JCHSL). Our algorithm learns the low-dimensional feature subspace through the joint heterogeneous information of different views, and adds discriminant graph variance constraints to unlock the potential double structure of cross-view data. In summary, the main contributions of our algorithm are composed of the following three aspects: (1) We use a dual low-rank representation to describe the manifold structure of cross-view data in a high-dimensional subspace. By discriminating view variance constraints, complex data is projected into a low-dimensional subspace structure that is easy to identify and its class structure and view structure are decomposed into each other. (2) We have designed a joint view heterogeneous feature subspace constraint that can extract hidden view heterogeneous information from different view data.

The other chapters of this article are designed as follows. Section II introduces the basic theory of the algorithm. In section III, we illustrate our proposed model and its optimization methods. The performance of the model is demonstrated through experiments in section IV, and the conclusions can be found in section V.

2 Related works

In this section, we introduce the related works of our proposed method, which are two different representation methods.

2.1 Ridge regression

We assume that a set of data consists of n samples \( X = [X_1, X_2, \ldots, X_n] \), and the output corresponding to the sample set \( Y = [Y_1, Y_2, \ldots, Y_n] \). We need find a regression coefficient matrix matrix \( W \) to transform and reduce feature information for data \( X \) to touch semantic dictionary. Then, we use label information to construct the loss function as follows:

\[
\min_{W} \left\| Y - WX \right\|_2^2 + \lambda \left\| W \right\|_2^2
\]  

(2.1)
Ridge regression is a biased regression method dedicated to collinear data analysis, which is essentially an improved linear regression method. The linear regression model is widely used as an empirical risk function in supervised learning. Because there is a collinear relationship between cross-view data, we use ridge regression to obtain more stable regression coefficients.

2.2 Low-rank representation

Giving a set of data $X = [X_1, X_2, \ldots, X_k]$ from $k$ subspaces, which can be composed of a linear combination of dictionary and sparse noise matrix as follows:

$$\min_{Z, E} \text{rank}(Z) + \lambda \|E\|_1$$

s.t. $X = AZ + E$ \hspace{1cm} (2.2)

in which $A$ is a low-rank dictionary of the data. $Z$ is the newly learned affine matrix with low-rank constraint, and $\text{rank}(\cdot)$ represents the rank of the matrix. We assume the noisy data $E$ is sparse and handle it by $l_1$-norm constraint. $\lambda$ is the balanced parameter between linear representation and noise. However, the traditional low-rank representation only has a certain effect on ordinary data. For cross-view data, the data from the same view in the learned subspace will be close to each other, greatly reducing the accuracy of recognition. Therefore, our model uses a dual low-rank representation framework to solve this problem.

3 The proposes algorithm

In this section, we first introduce the proposed algorithm called Joint Cross-view Heterogeneous Subspace Learning. Then, an feasible solution is obtained though iterative method as well as complexity analysis.

3.1 Notations

We assume that there is a set of cross-view data $X = [X_1, X_2]$ from two views. The samples of the $i$ th view $X_i = [X_{i1}, X_{i2}, \ldots, X_{ic}]$ is consisting of $l$ classes, where $X_{ir} \in \mathbb{R}^{d \times m_r}$ denotes that $d$ is the original dimensionality of training data and $m_r$ is the number of samples from the $i$ th view and the $r$ th class $\left(m_i = \sum_r m_r, m = \sum_i m_i\right)$. We construct the class structure matrix $Z_c \in \mathbb{R}^{m \times n}$ and the view structure matrix $Z_v \in \mathbb{R}^{n \times m}$ to adapt cross-view data subspace structure. Furthermore, we use the error matrix $E \in \mathbb{R}^{p \times m}$ to fit the noisy data.

As a kind of feature subspace learning method, it is common to learn a projection matrix $P \in \mathbb{R}^{d \times p}$, where $p$ is reduced dimensionality. In addition, we design three weight matrices $W_0, W_1, W_2 \in \mathbb{R}^{p \times d}$, where $l$ is the number of classes for different categories of sample data,
to quantify cross-view feature in order to increase generalization for projection subspace.

### 3.2 Dual low-rank discriminative subspace learning

To the cross-view data, the samples from different views but within the same class has a large divergence due to two interlaced structures, which are class structure and view structure. As we discussed in section II, we adopt class structure matrix $Z_c$ and view structure matrix $Z_v$ as two affine matrices via dual low-rank representation to describe the class structure and view structure of cross-view data as follows:

$$\min_{Z_c, Z_v, E} \text{rank}(Z_c) + \text{rank}(Z_v) + \lambda \|E\|_1$$

subject to $X = A_c Z_c + A_v Z_v + E$.

In which $A_c \in \mathbb{R}^{d \times m}$ and $A_v \in \mathbb{R}^{d \times m}$ are the class structure dictionary and view structure dictionary of the samples, respectively. $\lambda$ is the trade-off parameter. $\text{rank}(\bullet)$ is the rank of the matrix, and in order to solve the minimizing rank problem, we use $\|\bullet\|_*$ to replace equivalently. With equation (3.1), the two interwoven structures, the class structure and view structure, are constructed independently. However, due to the characteristics of cross-view data, these two structures cannot be decomposed clearly by unsupervised methods.

In order to our model to achieve the desired effect, we design a supervised discriminant view variance constraint term to help class structure and view structure decomposition of the cross-view data. Then, we promote our model as follows:

$$\min_{P, Z_c, Z_v, E} \|Z_c\|_1 + \|Z_v\|_1 + \lambda \|E\|_1 + \alpha \zeta(P, Z_c, Z_v)$$

subject to $P^T X = P^T (A_c Z_c + A_v Z_v) + E, P^T P = I$,

where $\zeta(P, Z_c, Z_v)$ is the supervised regularization term and $\alpha$ is the balancing parameter.

In addition, the orthogonal constraint $P^T P = I$ is imposed to eliminate trivial solution and reduce the redundancy. More specifically, we design regularization terms based on Fisher's principle as follows:

$$\zeta(P, Z_c, Z_v) = \text{tr} \left( P^T (A_c^T L_c Z_c^T A_c - \phi A_c L_c Z_c^T A_c^T) P \right)$$

in which $L_c$ and $L_v$ are the Laplacian operator of $\mathcal{W}^c$ and $\mathcal{W}^v$. $\phi$ is the balance coefficient inside the regularization term as 1. The forms of $\mathcal{W}^c$ and $\mathcal{W}^v$ are as follows:
in which $l_i$ and $l_j$ are the labels of sample $x_i$, $x_j$, respectively. $x_i \in \Pi_i^v(x_j)$ denotes $x_i$ is $k_1$ the nearest adjacency of data $x_j$ within the same class. $x_i \in \Pi_i^v(x_j)$ means is $k_2$ the nearest adjacency of the same view data $x_j$. We design a supervised regularized view variance term in order to make within-class samples close to each other and between-class samples far away from each other, thereby unlocking two intertwined structures of cross-view data.

3.3 Joint view heterogeneous feature subspace learning

In order to obtain heterogeneous information and isomorphic information between different view data, we designed a joint view heterogeneous constraint so that the low-dimensional subspace contains these useful information. Therefore, we propose the final model by joint view heterogeneous constraints as follows:

$$
\begin{align*}
\min_{P, Z, \tilde{Z}, \tilde{F}, W_0, W_1, W_2} & \left\| (\varphi W_0 + (1-\varphi) W_1)^T P^T X - Y \right\|_F^2 + \left\| (\varphi W_0 + (1-\varphi) W_2)^T P^T X - Y \right\|_F^2 \\
& + \epsilon \left( \left\| W_1 \right\|_F^2 + \left\| W_2 \right\|_F^2 \right) + \gamma \left\| W_0 \right\|_F^2 + \left\| Z_e \right\|_F^2 + \left\| Z_v \right\|_F^2 + \lambda \left\| E \right\|_F^2 + \alpha \zeta \left( P^T Z_e + P^T Z_v \right) \\
\text{s.t.} & \ P^T X = P^T (A_e Z_e + A_v Z_v) + E, P^T P = I
\end{align*}
$$

where $Y_i = [Y^1, Y^2, \ldots, Y^m]$ is a class labels matrix from $i$ th view. The zero-mean vector $Y^j = [-1, -1, \ldots, 0, \ldots, -1] \in \mathbb{R}^d$ represents the $j$ th column of $Y_i$. We use a weight matrix $W_0$ to quantify the view-shared feature that is owned jointly by all different views. And the weight matrices $W_1, W_2$ are used to elicit the individual component. In addition, we use the Frobenius norm constraint weight matrix to make our algorithm generalizable and eliminate trivial solutions. Two parameters $\gamma$ and $\epsilon$ can adjust the values of the isomorphic weight matrix and the heterogeneous weight matrix, respectively. We also use the parameter $\varphi$ to control the proportion of isomorphic and heterogeneous information in the low-dimensional subspace.

3.4 Optimization
In order to simplify the matrix we introduce the auxiliary matrix $M$ and use the original data $X$ as its own class structure dictionary $A_c$ and view structure dictionary $A_v$. Our objective function is rewritten as follow:

$$
\min_{P,Y,Z,W_1,W_2} \left\{ \left\| \left( \varphi W_0 + (1 - \varphi) W_1 \right)^T M^T X_1 - Y_1 \right\|_F^2 + \left\| \left( \varphi W_0 + (1 - \varphi) W_2 \right)^T M^T X_2 - Y_2 \right\|_F^2 \\
+ \varepsilon \left\{ \left\| W_1 \right\|_F^2 + \left\| W_2 \right\|_F^2 \right\} + \gamma \left\| Z_c \right\|_2 + \left\| Z_v \right\|_2 + \lambda \left\| E \right\|_2 + \alpha \sigma (P,Z_c,Z_v) \right\}
$$

s.t. $P^T X = P^T X (Z_c + Z_v) + E, P^T P = I, P = M$

We choose Augmented Lagrangian Methods (ALM) to solve the optimization problem. However, our objective function is non-convex. We need use first order Taylor expansion to simplify the objective function, and then use the Alternating Direction Multiplier Method (ADMM) to solve it.

Firstly, we transform the equation (3.6) into a graceful Lagrangian form:

$$
\min_{P,Y,Z,W_1,W_2,M} \left\{ \left\| \left( \varphi W_0 + (1 - \varphi) W_1 \right)^T M^T X_1 - Y_1 \right\|_F^2 + \left\| \left( \varphi W_0 + (1 - \varphi) W_2 \right)^T M^T X_2 - Y_2 \right\|_F^2 \\
+ \varepsilon \left\{ \left\| W_1 \right\|_F^2 + \left\| W_2 \right\|_F^2 \right\} + \gamma \left\| Z_c \right\|_2 + \left\| Z_v \right\|_2 + \lambda \left\| E \right\|_2 + \alpha \sigma (P,Z_c,Z_v) \right\} + \| P^T X - P^T X (Z_c + Z_v) - E \|_F^2 \\
+ \beta \| P - M \|_F^2
$$

where $Q$ is the Lagrange multiplier and $\mu > 0$ and $\beta = 1$ are the regularization parameters. Then, the equation is rewritten to a quadratic form via merging some terms as follows:

$$
\min_{P,Y,Z,W_1,W_2,M} \left\{ \left\| \left( \varphi W_0 + (1 - \varphi) W_1 \right)^T M^T X_1 - Y_1 \right\|_F^2 + \left\| \left( \varphi W_0 + (1 - \varphi) W_2 \right)^T M^T X_2 - Y_2 \right\|_F^2 \\
+ \varepsilon \left\{ \left\| W_1 \right\|_F^2 + \left\| W_2 \right\|_F^2 \right\} + \gamma \left\| Z_c \right\|_2 + \left\| Z_v \right\|_2 + \lambda \left\| E \right\|_2 \\
+ \eta (P,Z_c,Z_v,E,Q,\mu) - \frac{1}{\mu} \| Q \|_F^2 + \beta \| P - M \|_F^2
$$

(3.8)
where $\eta(P, Z_c, Z_v, E, Q, \mu) = \alpha^2 (P, Z_c, Z_v) + \frac{\mu}{2} \|P^T X - P^T X (Z_c + Z_v) - E + \frac{Q}{\mu}\|_F^2$.

Similar to the conventional ALM, variable $P, Z_c, Z_v, M, W_0, W_1, W_2$ and $E$ cannot be addressed simultaneously, but they are solvable individually when fixing other variables. To solve each sub-problem, $\eta$ is approximated by the first order Taylor expansion. We define the right-bottom of the variable plus $t$ as the optimized solution at the $t$ th time. Then, each sub-problem at the $t+1$ th time is as follows:

Updating $Z_c$:

$$Z_{c,t+1} = \min_{Z_c} \left\{ \frac{1}{\mu} \|Z_c\|_F + \frac{1}{2} \|Z_c - Z_{c,t} + \nabla Z_c \eta\|_F^2 \right\}$$

(3.9)

where

$$\nabla Z_c \eta = 2\alpha X^T P_i P_i^T X Z_{c,t} L_c - Q_i^T P_i^T X - \mu X^T P_i \left( P_i^T X - P_i^T X (Z_{c,t} + Z_{v,t}) - E_i \right)$$

and

$$\rho = \|P_i^T X\|_2^2$$. This can be addressed by singular value thresholding effectively[14].

Updating $Z_v$:

$$Z_{v,t+1} = \min_{Z_v} \left\{ \frac{1}{\mu} \|Z_v\|_F + \frac{1}{2} \|Z_v - Z_{v,t} + \nabla Z_v \eta\|_F^2 \right\}$$

(3.10)

where

$$\nabla Z_v \eta = -2\alpha X^T P_i P_i^T X Z_{v,t} L_v - Q_i^T P_i^T X - \mu X^T P_i \left( P_i^T X - P_i^T X (Z_{c,t} + Z_{v,t}) - E_i \right)$$. 

Equation(3.9) can be addressed in the same way to Equation(3.8).

Updating $E$:

$$E_{t+1} = \min_{E} \left\{ \frac{\lambda}{\mu} \|E\|_F + \frac{1}{2} \|E - \left( P_i^T X - X (Z_{c,t+1} + Z_{v,t+1}) \right) + \frac{Q}{\mu} \|_F^2 \right\}$$

(3.11)

This can be addressed by shrinkage operator[15].

Updating $P$:

$$P_{i,t+1} = \left( 2\alpha X Z_n X^T + \mu X_n X_n^T \right)^{-1} \left( \beta M + X_n \left( E - \frac{Q}{\mu} \right)^T \right)$$

(3.12)

where we define $Z_n = Z_{c,t+1} L_{c,t+1} Z_{v,t+1}^T - Z_{v,t+1} L_{c,t+1} Z_{v,t+1}^T$ and $X_n = X - X (Z_{c,t+1} + Z_{v,t+1})$ for simplicity.
Updating $M$:

$$M_{r+1} = \min_M \left\{ \left\| \varphi W_0 + (1 - \varphi) W_1 \right\|^2_F + \left\| \varphi W_0 + (1 - \varphi) W_2 \right\|^2_F + \beta \left\| P - M \right\|^2_F \right\}$$  \hspace{1cm} (3.13)

It is difficult to solve the equation (3.13) with non-convex constraints directly on Euclidean space. We use a gradient based approach to optimize the problem on the Stiefel manifold[15].

Algorithm 1

**Input:** data matrices $X_1, X_2, Y_1, Y_2$, parameters $\alpha, \lambda, \varepsilon, \varphi, \gamma$

**Initialize:** $\theta = 10^{-6}$, $\psi = 1.1$, $\mu = 0.1$, $\mu_{\max} = 10^6$, $\beta = 1$, $t = 0$, $t_{\max} = 10^2$

**while not converged or $t \leq t_{\max}$ do**

1. Optimize $Z_{c,r+1}$ according to (3.9) by fixing other parameters;
2. Optimize $Z_{v,r+1}$ according to (3.10) by fixing other parameters;
3. Optimize $E_{r+1}$ according to (3.11) by fixing other parameters;
4. Optimize $P_{r+1}$ according to (3.12) by fixing other parameters;
5. Optimize $M_{r+1}$ according to (3.13) by fixing other parameters;
6. Optimize $W_{0,r+1}$ according to (3.14) by fixing other parameters;
7. Optimize $W_{1,r+1}$ according to (3.15) by fixing other parameters;
8. Optimize $W_{2,r+1}$ according to (3.16) by fixing other parameters;
9. $P_{r+1} \leftarrow \text{orthogonal}(P_{r+1})$;
10. Optimize the multiplier $Q_{r+1}$,
    $$Q_{r+1} = Q_{r} + \mu \left( P_{r+1} ^T \left( X - X \left( Z_{c,r+1} + Z_{v,r+1} \right) - E_{r+1} \right) \right);$$
11. Update the parameter $\mu$ by $\mu = \min(\mu_{\max}, \psi \mu)$;
12. Check the convergence conditions $\left\| P_{r+1} ^T \left( X - X \left( Z_{c,r+1} + Z_{v,r+1} \right) - E_{r+1} \right) \right\|_\infty < \theta$;
13. $t = t + 1$.

**end while**

**Output:** $P, M, Z_c, Z_v, E, W_0, W_1, W_2$

Updating $W_0$:
\[ W_{0,t+1} = \varphi \left( \varphi^2 M^T X_1X_1^T + \varphi^2 M^T X_2X_2^T + \varepsilon I \right)^{-1} \left( M^T X_1 \left( Y_1^T - (1-\varphi)X_1^T MW_1 \right) + M^T X_2 \left( Y_2^T - (1-\varphi)X_2^T MW_2 \right) \right) \]  

(3.14)

Updating \( W_1 \):

\[ W_{1,t+1} = (1-\varphi) \left( (1-\varphi)^2 M^T X_1X_1^T + \varepsilon I \right)^{-1} M^T X_1 \left( Y_1^T - \varphi X_1^T MW_0 \right) \]  

(3.15)

Updating \( W_2 \):

\[ W_{2,t+1} = (1-\varphi) \left( (1-\varphi)^2 M^T X_2X_2^T + \varepsilon I \right)^{-1} M^T X_2 \left( Y_2^T - \varphi X_2^T MW_0 \right) \]  

(3.16)

Updating \( Q \):

\[ Q_{t+1} = Q_t + \mu \left( P_{v,t}^T \left( X - X \left( Z_{z,t+1} + Z_{z,t+1} \right) \right) - E_{z,t+1} \right) \]  

(3.17)

Finally, the detail process of optimization is listed in Algorithm 1. We set the parameters \( \mu, \mu_{\text{max}}, \theta, \psi, l_{\text{max}} \) and tune the trade-off parameters \( \alpha, \lambda, \varepsilon, \varphi, \gamma \) by the experiment. And we initialize the matrices \( P, M, Z_{c}, Z_{v}, E, W_1, W_2 \) at random.

4 Experiments results

In this section, we evaluate the performance of our algorithm on four standard data sets.

4.1 Datasets introduction and experimental setting

The CMU-PIE Face database contains 68 different people, and each category of cross-view face samples have 21 different illumination conditions and 9 different poses. We crop the face images to \( 64 \times 64 \) size. We adopt 5 poses P05, P09, P14, P27, P29 and randomly divide the samples in each pose set into a test set and a training set. To 5 poses of the CMU-PIE faces, we divided each 2 poses into same group, where Case1:{P05,P09}, Case2:{P05,P14}, Case3:{P05,P27}, Case4:{P05,P29}, Case5:{P09,P14}, Case6:{P09,P27}, Case7:{P09,P29}, Case8:{P14,P27}, Case9:{P14,P29}, Case10:{P27,P29}.

The COIL-100 object database includes 100 objects, a total of 7200 images, each object obtained 72 images and these images are caught with 5 degree rotation. The object images are cropped to \( 32 \times 32 \) size and divided to two subsets as “C1” and “C2”. In addition, C1 contains
the images in two point of view $V_1 \left[ 0^\circ, 85^\circ \right]$ and $V_2 \left[ 185^\circ, 265^\circ \right]$. Similarly, $C_2$ obtains the images in $V_3 \left[ 90^\circ, 175^\circ \right]$ and $V_4 \left[ 270^\circ, 355^\circ \right]$.

The Extended YaleB face database consists of 16128 images, under 28 years old, including 64 images in different illumination conditions and 9 postures. We crop the face images to $32 \times 32$ size. We divide each person’s images into four poses $P_1, P_2, P_3, P_4$ through experiments, which are approximately positive, so that samples can be more relevant to our experiments. We further partition each two poses into one group, where $V_1\{P_1, P_2\}$ and $V_2\{P_3, P_4\}$.

4.2 Experimental results

In order to evaluate the performance of our algorithm, we have selected several classic methods for comparison, such as PCA, LDA, LPP, LatLRR, LRCS, SRRS, MvDA, RMSL. In addition, we use the k-Nearest Neighbor classifier for performance evaluation of the extracted feature information by each method. To 5 poses of the CMU-PIE faces, the results of our algorithm and comparative experiment are shown in Tables 1&2. To COIL-100 objects, we select one from $V_1$ and one from $V_2$ as training set from each set of perspectives, and another perspective as a test set. There are four experimental groups to evaluate the performance of all algorithm, as shown in Figure 1. For extended YaleB faces, our experimental setup is similar to COIL-100 objects. The recognition rates of the algorithm are shown in Figure 2.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Case1</th>
<th>Case2</th>
<th>Case3</th>
<th>Case4</th>
<th>Case5</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>48.81±0.73</td>
<td>50.89±1.00</td>
<td>50.50±1.14</td>
<td>49.07±1.04</td>
<td>48.36±1.33</td>
</tr>
<tr>
<td>LDA</td>
<td>62.48±0.78</td>
<td>66.31±1.33</td>
<td>66.76±1.67</td>
<td>62.16±1.44</td>
<td>61.27±1.35</td>
</tr>
<tr>
<td>LPP</td>
<td>62.40±1.08</td>
<td>59.25±0.02</td>
<td>60.17±0.03</td>
<td>61.97±0.14</td>
<td>65.72±0.11</td>
</tr>
<tr>
<td>LatLRR</td>
<td>65.07±1.00</td>
<td>65.36±1.75</td>
<td>66.61±1.23</td>
<td>62.47±1.8</td>
<td>63.65±3.11</td>
</tr>
<tr>
<td>LRCS</td>
<td>95.68±1.01</td>
<td>91.83±0.64</td>
<td>92.30±0.59</td>
<td>95.48±0.43</td>
<td>89.60±1.08</td>
</tr>
<tr>
<td>SRRS</td>
<td>85.31±1.02</td>
<td>82.04±2.12</td>
<td>82.33±1.03</td>
<td>85.22±1.08</td>
<td>83.37±1.12</td>
</tr>
<tr>
<td>MvDA</td>
<td>95.71±2.2</td>
<td>92.02±0.8</td>
<td>91.5±1.76</td>
<td>95.42±0.18</td>
<td>90.62±0.22</td>
</tr>
<tr>
<td>RMSL</td>
<td>97.14±0.02</td>
<td>92.97±0.01</td>
<td>93.70±0.07</td>
<td>97.26±0.17</td>
<td>91.85±0.04</td>
</tr>
<tr>
<td>Ours</td>
<td>98.39±0.03</td>
<td>93.82±0.04</td>
<td>94.50±0.05</td>
<td>98.19±0.11</td>
<td>92.47±0.12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Case6</th>
<th>Case7</th>
<th>Case8</th>
<th>Case9</th>
<th>Case10</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>48.43±0.73</td>
<td>45.51±1.62</td>
<td>55.28±1.37</td>
<td>48.64±1.76</td>
<td>49.68±0.86</td>
</tr>
<tr>
<td>LDA</td>
<td>61.50±1.49</td>
<td>56.54±1.05</td>
<td>66.96±1.32</td>
<td>61.10±1.56</td>
<td>61.83±1.48</td>
</tr>
<tr>
<td>LPP</td>
<td>66.13±0.65</td>
<td>63.34±0.05</td>
<td>59.29±0.14</td>
<td>58.10±0.01</td>
<td>63.72±0.13</td>
</tr>
<tr>
<td>LatLRR</td>
<td>63.09±0.97</td>
<td>61.04±1.39</td>
<td>66.1±1.84</td>
<td>60.78±2.01</td>
<td>60.42±1.26</td>
</tr>
<tr>
<td>LRCS</td>
<td>89.23±0.62</td>
<td>95.57±1.02</td>
<td>87.42±0.54</td>
<td>90.88±0.92</td>
<td>90.64±0.71</td>
</tr>
<tr>
<td>SRRS</td>
<td>86.17±0.44</td>
<td>82.89±0.32</td>
<td>77.45±0.64</td>
<td>81.64±0.78</td>
<td>82.18±0.83</td>
</tr>
<tr>
<td>MvDA</td>
<td>91.03±0.19</td>
<td>95.32±0.18</td>
<td>87.32±0.17</td>
<td>90.86±0.23</td>
<td>92.38±0.21</td>
</tr>
<tr>
<td>RMSL</td>
<td>92.99±0.11</td>
<td>97.55±0.08</td>
<td>88.47±0.01</td>
<td>92.02±0.12</td>
<td>92.36±0.06</td>
</tr>
<tr>
<td>Ours</td>
<td>93.95±0.08</td>
<td>98.35±0.04</td>
<td>89.48±0.07</td>
<td>92.19±0.01</td>
<td>93.89±0.15</td>
</tr>
</tbody>
</table>
By comparing Table 1&2 and Figure 1&2, we can observe that our algorithm is more accurate than other traditional methods for cross-view data.

4.3 Sensitivity analysis of proposed algorithm

In this part, we will test the influences of the parameters and feature dimensionality selected by the algorithm on the recognition results.

Our algorithm has four parameters $\lambda, \alpha, \varepsilon, \gamma, \phi$. We adjust parameters $\alpha$ and $\lambda$ at the same time. We evaluate them respectively on CMU-PIE faces Case8 and COIL-100 Case1, and the Figure 3 shows the results. In addition, we individually adjust parameters $\varepsilon, \gamma$ and $\phi$. We evaluate them on CMU-PIE faces Case8, and the Figure 4 shows the results. Because parameters $\alpha$ and $\lambda$ come from the same framework and parameters $\varepsilon, \gamma$ and $\phi$ come from the joint cross-view heterogeneous subspace learning framework, they will affect each other in their own framework. From the Figure 3&4, the recognition rates are hardly sensitive to $\lambda, \varepsilon, \gamma, \phi$. For the parameter $\alpha$, the recognition rate has small fluctuations in a narrow
range. We can obtain almost consistent classification results in a wide range. The results points out that our algorithm is stable to parameter selection.

![Figure 3](image)

**Fig. 3.** The performance of our algorithm is evaluated on the four parameters influence $[\alpha, \lambda]$ on CMU-PIE faces Case8 (a) and COIL-100 Case1 (b), where the value from -4 to 4 denotes $[10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4]$.

Afterwards, we proof the dimensionality influence of our method in CMU-PIE faces Case8 and Coil100 Case1, and the Figure 5 shows the experiment results. From the results, algorithm performance is not sensitive to the dimensions. For the CMU-PIE faces Case8, classification performance increases slightly when the dimensionality goes up. Performance reach the highest around 300. In the Extend YaleB faces Case1, classification performance cuts down enough slightly with increasing of the dimensionality.

![Figure 4](image)

**Fig. 4.** The performance of our algorithm is evaluated on the four parameters influence $[\phi, \gamma, \phi]$ on CMU-PIE faces Case8, where the value from -4 to 4 denotes $[10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1, 10^2, 10^3, 10^4]$.
Fig. 5. Classification rates of different dimensionality on CMU-PIE faces

Case 8 and Coil100 Case 1.

5 Conclusion

To solve the difficulties caused by cross-view data, we propose a joint cross-view heterogeneous subspace learning method based low-rank constraint for image feature extraction and recognition tasks. In detail, we have established a dual low-rank representation framework to unlock the potential double structure of cross-view data through supervised view variance constraints. Also, we have established a joint cross-view heterogeneous subspace framework that combines the isomorphic and heterogeneous features of images from different perspectives to preserve more useful information for image recognition. At the same time, we propose a feasible solution to ensure convergence. Through experiments on three different public cross-view datasets, compared with other methods, our method has obtained obvious advantages.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China under Grant 61501147, in part by the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province under Grant UNPYSCT-2018203, in part by the Natural Science Foundation of Heilongjiang Province under Grant YQ2019F011, in part by the Fundamental Research Foundation for University of Heilongjiang Province under Grant LGYC2018JQ013, and in part by the Postdoctoral Foundation of Heilongjiang Province under Grant LBH-Q19112.
References


Low Rank Sample Reconstruction-based Semi-supervised Feature Subspace Learning

Ruoqi An¹, Ao Li¹, Deyun Chen¹, Xin Liu¹
{anruoqi77@126.com, dargonboy@126.com, chendeyun@hrbust.edu.cn}

School of Computer Science and Technology, Harbin University of Science and Technology, Harbin, China¹

Abstract. Feature subspaces have been widely applied in image classification. However, majority conventional subspace learning models are supervised, the data containing only partial labeled samples will lead to unsatisfactory classification results in practical applications. To address this problem, in this paper, we design a graph semi-supervised feature subspace model named low rank sample reconstruction-based semi-supervised feature subspace learning, which combines graph-based semi-supervised learning and low rank sample reconstruction into a unified framework. The proposed model implementing low rank constraint on representation coefficient, exploits the graph-based label propagation algorithm to predict the unlabeled labels to obtain the labels of all samples. The experimental results demonstrate that the robustness and the classification efficiency of our model.

Keywords: Image classification; Subspace learning; Graph semi-supervised learning; Low Rank Sample Reconstruction;

1 Introduction

High-dimensional data brings difficulties to image processing, so that how to eliminate redundant data in high-dimensional data becomes a crucial problem to be solved in pattern recognition [1] and computer vision [2]. Dimension reduction can effectively improve performance and efficiency, whose purpose is to eliminate redundant information to enhance the recognition speed. Subspace learning have been extensively used to dimensionality reduction methods and achieved impressive performance. As the unsupervised subspace algorithms, Principal Component Analysis (PCA) aims to seek a subspace whose projection sample has the most massive variance [3]. Meanwhile, Linear Discriminant Analysis (LDA) is an effective method in the supervised subspace learning [4], in which it captures the features with the largest discriminative low dimensional features from high dimensional data, so that the ratio of the dispersion between samples and the dispersion within samples is the largest. Locality Sensitive Discriminant Analysis (LSDA), as a further extension of LDA, aims at preserving the discriminant structure and local geometry of the data [5].

Low-rank representation model (LRR) is introduced for learning valid features from data containing redundant information and mining potential manifold structures. In this decade, LRR has been widely applied to dictionary learning[6], transfer learning[7], [8], domain adaptation [9]. In recent years, particularly, several improved LRR-based subspace models have been proposed. For instance, a subspace clustering approach about low rank structured representation was proposed by Yao et al. [10]. In order to
make the clustering more robust, this method applied the global structure to obtain the low rank representation and generated low rank structure to extract the neighborhood construction. Other LRR-based feature subspace learning models with outstanding performance can be referred to [11], [12], [13].

Most subspace models with the impressive performance are supervised. However, in many practical problems, only a small number of labeled data are available, it caused weak discrimination. Several semi-supervised models for feature subspace have been presented in recent years. The semi-supervised discriminant analysis (SDA) was proposed by Cai et al. [14], in which the method improved the discriminant ability of the model through labeled samples and explored manifold structure through unlabeled samples. Flexible manifold embedding (FME) is a semi-supervised learning framework [15], in which semi-supervised dimension reduction is performed by using the manifold structure of unlabeled sample and information of the labeled sample. Chen et al. proposed the method named rescaled linear square regression (RLSR), which make the feature selection model solve the optimal projection matrix under global and sparse conditions [16].

Inspired by the above semi-supervised model, a semi-supervised model named low rank sample reconstruction-based semi-supervised feature subspace learning is proposed in this paper. The core idea of our model is to combine label propagation and feature subspace learning to obtain better classification results.

The key contributions of the proposed model can be summarized. 1) we proposed a novel semi-supervised feature subspace learning model, which integrates the LRR model, label propagation, and feature subspace learning. The coefficients of LRR simultaneously constrain the prediction of label propagation information and feature subspace learning. 2) Our designed optimization solution applies the Lagrange multiplier [17] to calculate the objective function and guarantee its convergence. 3) We compared different traditional semi-supervised methods to conduct the experiment by using three publicly available datasets and two different classifiers (KNN and SRC). Numerous experiments results show the effectiveness of our model.

The other section in the paper is arranged as follows. In the second section, we briefly introduce graph-based methods and sample reconstruction constraints in related work. And then, in the third section, we introduce the proposed semi-supervised learning method in detail and listed the solution process of the objective function. In the fourth section, we evaluate the proposed method through comparative experiments and parameter analysis with traditional methods. Finally, the conclusion is presented in the fifth section.

2 Related work.

2.1 Graph-based methods

Graph-based methods have excellent potential because of their diverse structures that can represent similarities between samples. In graph-based methods, given a weighted graph \( G = (X, E, W) \), where \( X = \{x_i\}_{i=1}^{n} \) represents the vertex set. And, \( W \) is the weight of the undirected edge set \( E = \{x_i, x_j\} \). Then, the graph-based algorithms will be defined as:

\[
\sum_{i} \sum_{j} ||g_i - g_j||W_{ij}
\] (1)
where \( g_i = g(x_i) \) is the loss function, which is required to be smooth the entire graph and \( W_{ij} \) is the coefficient of the weight matrix \( W \). By Gaussian kernel function, \( W \) is calculated as follows:

\[
W_{ij} = \exp\left(-\mu \|x_i - x_j\|^2\right)
\]

(2)

The labels of unlabeled samples are predicted by the connection between unlabeled samples and labeled samples. The graph-based objective function can be written as follows:

\[
\sum_{i} \|y_i - g_i\|^2 + \sum_{i} \sum_{j} |g_i - g_j|^2 W_{ij}
\]

(3)

Where \( y_i \) denotes the predicted labels. The labeled information is transmitted by the similarity between vertices. The more extensive similarity between the sample vertices, the labeled information is more easily propagated. More methods about graph-based can be reference at [18], [19], [20].

2.2 Sample reconstruction constraint

The sample reconstruction constraint based on the core of the Fisher criterion is modeled to eliminate noise interference [21]. Assuming \( X = [x_1, x_2, ..., x_n] \) is the training set, \( Z \) is representation coefficient matrix, the reconstructed constraint item obtained from sample self-representation \( X = XZ \) as follows:

\[
f(P, Z) = \text{Tr}\left(S_u(P^T XZ)\right) - \text{Tr}\left(S_u\left(P^T XZ\right)\right)
\]

\[\text{s.t. } X = XZ + E, P^TP = I_p\]

(4)

Where \( P \) is feature subspace, which is an orthogonal matrix to reduce redundant information. \( S_u(P^T XZ)\) and \( S_u\left(P^T XZ\right) \) denotes the inter-class and within-class scatter divergence matrix, \( \text{Tr}(\cdot) \) denotes the trace. Since \( Z \) is non-convex, adding a relaxation term to the constraint term can be expressed as:

\[
f(P, Z) = \|P^T XZ (I - D_u)\|_F^2 \|P^T XZ (D_u - D_v)\|_F^2 + \eta \|P^T XZ\|_F^2
\]

\[\text{s.t. } X = XZ + E, P^TP = I_p\]

(5)

Where \( I \) is an identity matrix, \( D_u \) and \( D_v \) are two constant-coefficient matrices, and \( \eta \) is regularization parameter. In detail, \( B_u(i, j) = 1/n_u \), when \( x_i \) and \( x_j \) belongs to the same class. Otherwise, \( B_u(i, j) = 0 \).

3 Our model

The proposed objective function and its detailed description are presented in this section, and we also design the optimal scheme to solve the objective function by using Lagrange multiplier method.

3.1 Proposed method
To identify the noisy information in the original samples and obtain the robust subspace, we implement low rank constraint on representation coefficient. Meanwhile, the low rank constraint is used to predict the labels of unlabeled samples during the graph-based label propagation process. The low rank constraint can effectively eliminate the noise component in the sample. Moreover, the reconstruction of samples can further explore the nearest neighbor relationship, which makes the different classes of samples far away from each other and improve the discriminability of the model. The objective function with low rank representation constraint can be obtained:

$$\min_{F,Z} \|Z\|_{*} + \lambda_1 \|E\|_1 + \lambda_2 \left(\|P^T X Z (I - D_s)\|_F^2 - \|P^T X Z (D_s - D_s)\|_F^2 + \eta \|P^T X Z\|_F^2\right)$$

subject to $X = XZ + E, P^T P = I_p$.

where $\|Z\|_*$ denotes the nuclear norm of $Z$. $\lambda_1$ and $\lambda_2$ denote two regularization parameters.

Next, by integrating graph learning and low-rank reconstructed subspace learning into the same learning framework, we extend the semi-supervised method to feature subspace learning. In the graph propagation algorithm, the similarity between the two samples is constrained by low rank representation coefficient. The final objective function we can obtain as followed:

$$\min_{F,R,Z} \frac{1}{2} \|F - B\|_F^2 + \sum_\gamma Z_{\gamma} \|F_{\gamma} - F\|_F^2 + \frac{1}{2} \|F - P^T X\|_F^2 + \|Z\|_1 + \lambda_1 \|E\|_1 + \lambda_2 \left(\|P^T X Z (I - D_s)\|_F^2 - \|P^T X Z (D_s - D_s)\|_F^2 + \eta \|P^T X Z\|_F^2\right)$$

subject to $X = XZ + E, P^T P = I_p, Z_{ij} \geq 0$.

where $B = [B_1, B_2, ..., B_m]$ denotes the matrix with all the label vectors of unlabeled samples and labeled samples. $B_i = [-1, ..., -1, 1, -1, ..., -1]^T \in \mathbb{R}^C$ is the $i$-th column of $B$. The $c$-th element of $B_i$ is 1 and the other elements is -1 when $B_i$ belongs to $c$ class. $B_j = [0, ..., 0, 0]^T \in \mathbb{R}^C$ is the label vector of $j$-th unlabeled sample. $F$ represents the learned label matrix through the similarity measure with $B$. In equation (7), the first two items learn the complete label information through label propagation, $Z_{ij}$ is used as similarity weights to constrain the relationship between two labels, we also introduce a non-negative constraint on $Z$ as the non-negative regularization parameter.

We have described the graph-based semi-supervised approach and low rank reconstructed subspace learning jointly. Next, we will introduce the optimal scheme.

3.2 Scheme

Since the minimization for all variables is not a convex problem, we design a solution based on alternating iteration. Aim at relaxing the issue of minimization, we introduce two auxiliary variables $R$ and $J$. The Eq. (7) can be rewritten as:
Next, the Lagrangian function in Eq. (8) can be obtained as the following form:

\[ \mathcal{L}(F, P, Z, E, R, J) = \frac{1}{2} \| F - U \|_F^2 + \sum_y R_y \| F_y - F_i \|_2^2 + \frac{1}{2} \| F - P^T X \|_F^2 + \lambda_s \| E \|_F^2 + \lambda_i \left( \| P^T X Z (I - D_u) \|_F^2 - \| P^T X Z (D_u - D_i) \|_F^2 + \eta \| P^T X Z \|_F^2 \right) \]

s.t. \( X = Z + E, R_y \geq 0, P^T P = I, Z = H, Z = J \)

Furthermore, Eq. (9) can be converted into the compact form:

\[ \mathcal{L}(F, P, Z, E, R, J) = \frac{1}{2} \| F - U \|_F^2 + \sum_y R_y \| F_y - F_i \|_2^2 + \frac{1}{2} \| F - P^T X \|_F^2 + \lambda_i \left( \| P^T X Z (I - D_u) \|_F^2 - \| P^T X Z (D_u - D_i) \|_F^2 + \eta \| P^T X Z \|_F^2 \right) + \lambda_s \| E \|_F^2 \]

\[ + \mu \left( \| X - F P X - E \|_F^2 + \frac{Y}{\mu} \right) + \| Z - J \|_F^2 + \| Z - R \|_F^2 + \| Y \|_F^2 \]

Finally, the minimized problem of Eq. (7) can be rewritten as:

\[ \min_{F, P, Z, E, R, J} \mathcal{L}(F, P, Z, E, R, J) \quad \text{s.t. } R_y \geq 0, P^T P = I \]

We use an iterative method to solve each variable. At the \( k \)-th iteration, variables other than \( F \) are fixed, and the objective function formula for \( F \) is:

\[ \min_{F} \frac{1}{2} \| F - P^{(k)} X \|_F^2 + \frac{1}{2} \| F - U \|_F^2 + \sum_y R_y \| F_y - F_i \|_2^2 \]

To simplify the solution, we convert Eq. (12) to the following form:

\[ \min_{F} \frac{1}{2} \| F - P^{(k)} X \|_F^2 + \frac{1}{2} \| F - U \|_F^2 + \text{Tr} \left( F L F^T \right) \]

where \( L = D - R \) presents the matrix of graph Laplacian, \( D \) denotes the diagonal matrix \((D_u = \sum R_u + \frac{\sum R_y}{2}).\) Eq. (13) can easily solved by setting the derivative to 0:

\[ F = \left( U + P^{(k)} X \right) (2I + L)^{-1} \]

Then, we update the objective function about eliminating the terms irrelevant to \( P \):

\[ \min_{P} \frac{1}{2} \| F - P^T X \|_F^2 + \lambda_i \left( \| P^T X Z (I - D_u) \|_F^2 - \| P^T X Z (D_u - D_i) \|_F^2 + \eta \| P^T X Z \|_F^2 \right) \]

Due to the orthogonal constraint is contained in Eq. (15), the derivative can be obtained firstly:
\[
\frac{\partial \mathcal{L}_p}{\partial P} = X'XP - XF' + 2\lambda XZ (I - D_{\lambda})(I - D_{\lambda})' Z' X'
\]
\[
+ 2\lambda XZ (D_{\lambda} - D_{\lambda}) (D_{\lambda} - D_{\lambda})' Z' X' P + 2\lambda \eta XZZ' X' P
\]
(16)

Up to now, Eq. (16) can be solved regarding the methods in [22]. After fixing the variables independent of \( J \) in Eq. (10), we can obtain:

Minimize \( f \left( Z^* \right) = \left\| J - \left( Z^* + \frac{Y^*}{\mu} \right) \right\|_{F} + \eta \left\| J \right\|_{F} \)
(17)

Eq. (17) is a classical rank minimization problem by using existing technology to solve [23]. Removing the terms irrelevant to \( R \), we obtain:

Minimize \( f \left( Z \right) = \left\| Z - R + \frac{Y^*}{\mu} \right\|_{F}^{2} + \sum_{\gamma} R_{\gamma} \left\| F_{\gamma}^{k+1} - F_{\gamma}^{k+1} \right\|_{2}^{2} \)
(18)

The Eq. (18) can be rewritten as:

Minimize \( f \left( Z \right) = \left\| R - \left( Z^* + \frac{Y^*}{\mu} \right) \right\|_{F}^{2} + \sum_{\gamma} R_{\gamma} \left\| S_{\gamma}^{k+1} \otimes R \right\|_{F}^{2} \)
subject to \( R_{\gamma} \geq 0 \)
(19)

where \( S_{\gamma}^{k+1} \) is a matrix consists of \( S_{\gamma}^{k+1} = \left\| F_{\gamma}^{k+1} - F_{\gamma}^{k+1} \right\|_{2}^{2} \). Eq. (19) can be transformed as follows:

Minimize \( f \left( Z \right) = \left\| R - \left( Z^* + \frac{Y^*}{\mu} \right) \right\|_{F}^{2} + \sum_{\gamma} R_{\gamma} \left\| S_{\gamma}^{k+1} \otimes R \right\|_{F}^{2} \)
subject to \( R_{\gamma} \geq 0 \)
(20)

Eq. (20) can be regarded as the non-negative weight \( l_{1} \)-norm minimization that can be solved efficiently in [24]. Then, after solving the auxiliary variables \( J \) and \( R \), the variables other than \( Z \) are fixed, and the objective function for \( Z \) is:

Minimize \( f \left( Z \right) = \left\| P' X Z (I - D_{\lambda})_{F} - P' X Z (D_{\lambda} - D_{\lambda})_{F} \right\|_{F}^{2} + \eta \left\| P' X Z \right\|_{F}^{2} \)
\[
+ \frac{\mu}{2} \left\| X - ZX - E' + \frac{Y^*}{\mu} \right\|_{F}^{2} + \left\| Z - J^{k+1} \right\|_{F}^{2} + \left\| Z - R^{k+1} + \frac{Y^*}{\mu} \right\|_{F}^{2}
\]
(21)

The solution can be obtained simply by derivation. Finally, we express the objective function about the error matrix \( E \) by fixing other variables:

Minimize \( f \left( E \right) = \left\| E \right\|_{F}^{2} + \frac{\mu}{2} \left\| E - \left( X - ZX^{k+1} + \frac{Y^*}{\mu} \right) \right\|_{F}^{2} \)
(22)

Through defining \( \Omega = X - ZX^{k+1} + \frac{Y^*}{\mu} \), the \( i \)-th column of \( E^{k+1} \) is presented as:

\[
E_{\gamma}^{k+1} = \begin{cases} 
\frac{\left\| \Phi_{\gamma} \right\|_{2}}{\left\| \Phi_{\gamma} \right\|_{2}}, & \text{if } \lambda < \left\| \Phi_{\gamma} \right\|_{2} \\
0, & \text{otherwise}
\end{cases}
\]
(23)

We describe the detailed information of our scheme in Algorithm 1.
Algorithm 1

Input: training set $X$, label $U$, $Z = R = J = 0$, $E = 0$, $Y_1 = Y_2 = Y_3 = 0$

$\mu = 0.6$, $\mu_{max} = 10^0$, $\rho = 1.1$

Output: $F$, $P$

While not convergence do

1. Update $F^{k+1}$ using (12)

2. Update $P^{k+1}$ using (15);

3. Update $J^{k+1}$ using (17);

4. Update $R^{k+1}$ using (18);

5. Update $Z^{k+1}$ using (21);

6. Update the $Y_1^{k+1}$, $Y_2^{k+1}$, $Y_3^{k+1}$ and $\mu$;

\[
Y_1^{k+1} = Y_1^k + \mu \left( X - XZ^{k+1} - E^{k+1} \right)
\]

\[
Y_2^{k+1} = Y_2^k + \mu \left( Z^{k+1} - J^{k+1} \right)
\]

\[
Y_3^{k+1} = Y_3^k + \mu \left( Z^{k+1} - H^{k+1} \right)
\]

\[
Y_4^{k+1} = Y_4^k + \mu \left( Z^{k+1} - R^{k+1} \right)
\]

$\mu = \min \left( \mu_{max}, \rho \mu \right)$;

end while

4 Experimental results and discussion

4.1 Experimental results

We choose 3 datasets to verify the method in the experiment, respectively named Extended YaleB, COIL20, and USPS in this section. According to the experimental criteria of the semi-supervised model,
half of the whole samples were labeled and the rest of samples were not labeled. The detailed description of the dataset is shown below:

*Extended YaleB*  
Extended YaleB was photographed under a variety of controlled lighting conditions, including totally over 2,000 frontal faces images. Part of its frontal face image is shown in Fig 1(a). Each image is adjusted to $32 \times 32$ pixels. For each class, we choose 32 face images for each class and half of the facial images as training samples randomly, and the others were chosen as test samples in our experiment.

*COIL20*  
The COIL20 contains 20 objects, each of which is rotated 360 degrees horizontally to take an image every 5 degrees and 72 images of each object. We cut the size of the image data to $32 \times 32$ before the experiment as shown in Fig 1(b). 10 face images are chosen as training samples for each class.

*USPS*  
USPS database is a database for digital handwriting recognition. There are 9298 images from zero to nine in the data set. For our experiment, the whole images are $16 \times 16$ gray pixel values, which have been normalized. As shown in Fig 2(c), 10 images of per digit are selected for training and the others are test data.

Two representative methods of feature subspace learning and several novel semi-supervised approaches are chosen for comparison in experiments, including PCA, LDA, FME, RLSR, and SDA. To ensure that the experiment does not lose its generality, we use sparse representation classifier (SRC) and K nearest neighbors (KNN) to verify all methods on the test set. All experiments were performed 5 times to calculate standard deviation and mean as shown in Table 1 and 2. It is noted that, in Table 1 and 2, “Unlabeled” and “Test” represent the test sequence of the data with no labeled part of the training set and the testing set in our experiment.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Extended YaleB Unlabeled</th>
<th>Test</th>
<th>COIL20 Unlabeled</th>
<th>Test</th>
<th>USPS Unlabeled</th>
<th>Test</th>
</tr>
</thead>
</table>

Table 1. The classification results of experimental datasets with KNN (%)
<table>
<thead>
<tr>
<th>Methods</th>
<th>Extended YaleB</th>
<th>COIL20</th>
<th>USPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unlabeled</td>
<td>Test</td>
<td>Unlabeled</td>
</tr>
<tr>
<td>PCA</td>
<td>---</td>
<td>80.68 ± 1.42</td>
<td>---</td>
</tr>
<tr>
<td>LDA</td>
<td>---</td>
<td>85.5 ± 1.05</td>
<td>---</td>
</tr>
<tr>
<td>FME</td>
<td>78.35 ± 0.35</td>
<td>81.54 ± 1.88</td>
<td>73.25 ± 3.29</td>
</tr>
<tr>
<td>RLSR</td>
<td>83.34 ± 1.26</td>
<td>84.55 ± 1.35</td>
<td>76.08 ± 3.57</td>
</tr>
<tr>
<td>SDA</td>
<td>84.98 ± 0.54</td>
<td>81.23 ± 1.29</td>
<td>72.17 ± 2.35</td>
</tr>
<tr>
<td>Ours</td>
<td>86.67 ± 1.53</td>
<td>84.99 ± 0.89</td>
<td>78.27 ± 1.46</td>
</tr>
</tbody>
</table>

Compared with the comparison method, the method we proposed shows better performance on all datasets. Furthermore, the experimental results under both classifiers show that our proposed approach has more robustness and effectiveness than other methods.

4.2 Discussion on parameters and convergence

In our model, there are three regularization parameters $\lambda_1$, $\lambda_2$ and $\eta$. The COIL20 dataset was selected to discuss the effects of changes in these three parameters on the model performance by KNN classifier and the classification rate curve with the changes of parameters was drawn in Fig 2. It can be indicated from the figure that as the $\lambda_1$, $\lambda_2$ and $\eta$ change within a certain range, the classification rate does not change significantly. The fluctuation of the classification rate shows the performance stability in our model.
In order to prove the convergence of the proposed method, we select Extended YaleB and used KNN classifier to draw the convergence curve with the increase of iterations. The convergence curve in our model can rapidly converge and the convergence curve will be stable within 50 iterations as shown in Fig 3.

Fig 2. Classification results versus variational (a) $\lambda_1$ (b) $\lambda_2$ (c) $\eta$

Fig 3. Objective function values versus iterative steps.
5 Conclusion

In this paper, we propose a semi-supervised subspace learning model based on low rank sample reconstruction. To obtain complete label information in incomplete labels and improve the robustness of the model, we jointly learn low rank sample reconstruction and graph-based semi-supervised learning. To calculate the objective function and ensure its convergence, we designed an optimization solution. Model optimization can be achieved during iterative processing. We conduct comparative experiments with some classical feature subspace methods and other novel semi-supervised approaches on Extended YaleB, COIL20, and USPS and perform analysis experiments on the convergence and parameters of the model. Experimental results indicate that the effectiveness and robustness performance of this model is superior to other methods.

Acknowledgment. This work was supported in part by the National Natural Science Foundation of China under Grant 61501147, in part by the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province under Grant UNPYSCT-2018203, in part by the Natural Science Foundation of Heilongjiang Province under Grant YQ2019F011, in part by the Fundamental Research Foundation for University of Heilongjiang Province under Grant LGYC2018JQ013, and in part by the Postdoctoral Foundation of Heilongjiang Province under Grant LBH-Q19112.

References


Semi-supervised Corrupted Face Classification via Graph Learning

Yisheng Zhong, Ao Li
{951208709@qq.com(Yisheng Zhong), dargonbuy@126.com(Ao Li)}
School of Computer Science and Technology, Harbin University of Science and Technology, Harbin China

Abstract— Semi-supervised learning aims to training model with both of labeled and unlabeled data by exploring the relationships among them. Graph-based semi-supervised learning is an classical representative method that learning the class indicator matrix by propagating the similarity within the well designed graph constructed by data. However, for face data, they often happen to pixel missing or occlusion, which will degrade the graph learning performance, leading awful semi-supervised classification results. To address this problem, a novel semi-supervised corrupted face classification method via graph learning is proposed, in which the dynamic graph is learned by the completion face data recovered from the low-rank subspace. In our proposed method, the robust data representation and graph learning are implemented alternatively to obtain the overall optimal solutions. Experimental results demonstrate that our proposed method outperforms comparison methods on both of classification accuracy and robustness.

Keywords: Semi-supervised face classification, Graph learning, Self-representation model, Low-rank constraint.

I. Introduction

In the face of massive high-dimensional data, how to conduct effective data analysis and processing has become a major problem in machine learning and other fields [1]. In recent years, studying map automatically is hot, which is one of the important methods for adaptive neighbor method. We construct a matrix by setting each data point as the probability that the current data can be used as the neighborhood of another data, and this probability is used as the similarity between the two data points [2], which does not require similarity measures sensitive to noise and outliers [3], so the result obtained is of high precision.

However, in the process of graph learning, there is noise or interference in the original data, so the graph obtained may be inaccurate or suboptimal, and cannot accurately describe the true relationship between the data. In order to solve this problem, Zhao Kang [4] proposed a new robust graph learning scheme based on the adaptive neighbor method, which decomposes the original data into a low-rank matrix D ("clean data") and a sparse matrix E ("Noise/Error"). They can then use the adaptive neighbor method to build graphs on clean data D. So they can remove image disturbances and learn to map at the same time. However, when there is occlusion or partial absence of data, the results of this method are not reliable or even the learning results are not available. The Laplace Score (LS) method proposed in [5] introduced the analysis of the local structure of the data based on MaxVar. But these two methods only consider the characteristics of the data itself, ignore the correlation between the characteristics, and cannot guarantee the optimal feature subset. Inspired by the self-similarity of images, Zhu [6] believe that images should not only have self-similarity in structure, but also have the ability to express themselves in terms of feature expression. They proposed an unsupervised feature selection method based on regularized self-representation by constructing a regularized self-representation (RSR) model. This method constructs a self-representation model by assuming that each feature in the high-dimensional data can be expressed as a linear combination of other features, and removes insignificant features by adding $\ell _{1,2}$ norm constraints to the feature's weight matrix W. We borrowed the method of regularized self-representation to solve the problem of large-scale image interference or missing.

In the rest of this article, we will introduce graph learning and our multi-source robust graph learning technique in the second section. Details of the algorithm are then given in Section III. The fourth part evaluates the clustering task experimentally. The fifth part discusses semi-supervised applications and compares the data recovery effect of the sixth part.
II. Conventional Graph Learning

Given a sample set $X=\{x_1,x_2,\ldots,x_n\} \in \mathbb{R}^{m \times n}$, where $x_i \in \mathbb{R}^{m \times n}$ ($i = 1,2,\ldots,n$) represents the $i$th sample.

As locality preserving projection (LPP) [7] does, we can assign a probability $S_{ij}$ for $x_j$ as the neighborhood of $x_i$. Thus, $S_{ij}$ characterizes the similarity between $x_i$ and $x_j$ in some sense. Smaller distance $\|x_i - x_j\|_2$ indicates that $x_i$ and $x_j$ are quite similar, thus the bigger value of $S_{ij}$.

To achieve graph $S$, we can solve the following problem:

$$\min_{S} \frac{1}{2} \sum_{i,j} \left( \frac{1}{2} \| x_j - x_i \|^2 + \gamma S_{ij}^3 \right)$$

subject to $S + S^T \preceq 1$, $0 \preceq S \preceq 1$.

where $\gamma$ is a tradeoff parameter. By defining the graph Laplacian matrix $L = D - (S + S^T / 2)$, where $D$ is a diagonal matrix with $d_{ii} = \sum_j \left( (s_{ij} + s_{ji}) / 2 \right)$, (1) becomes

$$\min_{S} \text{Tr} \left( XLX^T \right) + \gamma \| S \|_F^2$$

subject to $S + S^T \preceq 1$, $0 \preceq S \preceq 1$.

By optimizing the above problem, one can learn $S$ adaptively from the data. However, this method degrades performance when it comes to input data with noise. To solve this problem, Zhao Kang [4] present a principle to robustify graph learning. They assume the data to be decomposed into two parts: 1) the clean data $D$ and 2) the corruptions $E$. And they proposed robust graph construction (RGC) model can be formulated as

$$\min_{D,E,S} \| D \|_F + \alpha \| E \|_F + \beta \text{Tr} \left( DLD^T \right) + \gamma \| S \|_F^2$$

subject to $X = D + E, S1 = 1, 0 \leq S \leq 1$.

This is called manifold RPCA (M RPCA) model in [8].

III. Proposed Model

This paper proposes a new robust learning graph scheme. We use the self-representation property of data [9], that is, every data point in the space can be effectively reconstructed through linear combination of other points in the data set to construct the similarity matrix. Specifically expressed as

$$X = XZ$$

(4)

Where $Z = \{z_1, z_2, \ldots, z_3\} \in \mathbb{R}^{n \times n}$ is a similarity matrix and each $z_i$ is a linear representation of the original data point $x_i$.

We propose a new robust learning graph scheme, and we introduced theta (4)., and then (3) becomes

$$\min_{Z,X,S} \| Z \|_F + \alpha \| E \|_F + \beta \text{Tr} \left( XZLX^T \right) + \gamma \| S \|_F^2$$

subject to $X = XZ + E, S1 = 1, 0 \leq S \leq 1$.

We then used the adaptive neighbor method to build the graph on clean $XZ$. We used a joint learning method to optimize both $S$ and $XZ$, which enabled us to construct the clustering graph while recovering low-rank clean data. The finally learned graph can be transformed into a block diagonal matrix through matrix transformation:

$$S_{n \times n} = \begin{bmatrix} S_1 & 0 & \cdots & 0 \\ 0 & \ddots & \cdots & 0 \\ 0 & \cdots & 0 & S_k \end{bmatrix}$$

To summary, compared with the existing work in the literature, the main contributions of this paper are as follows.

1) Through the establishment of self-presentation model, the vacancy of graph learning method in solving data interference or lack is filled, making the adaptive map learning method more robust.
2) Through joint solution, denoising and graph learning can be carried out simultaneously.
3) A large number of experiments have been performed on face clustering, document clustering, face / target recognition, and face image shadow removal, etc., which verifies the effectiveness of the method.
IV. Numerical Scheme

To solve (5), we first introduce auxiliary variable R to facilitate the solution of Z. Then, (5) can be written as:

$$\min_{Z,R} \left\{ R \mid + \alpha E + \| R - Z \|^2_F + \beta \| T R (X Z L Z^T X^T) + \gamma \| S \|^2_F \right\}$$

subject to

$$X = X Z + E, S 1 = 1, 0 \leq S \leq 1, R = Z$$

It can be solved via alternating direction method of multipliers. Removing the equality constraints on X and R, we obtain the augmented Lagrange function as follows:

$$L(X, Z, S, Y_1, Y_2) = \| Z \|_1 + \alpha E + \| R - Z \|^2_F + \beta \| T R (R L R^T) + \gamma \| S \|^2_F + \frac{\mu}{2} \left( \| X Z + E - X + \frac{Y_1}{\mu} || X Z - R + \frac{Y_2}{\mu} \right)$$

According to the above derivation, we can get the following algorithm flow:

**Algorithm**

Input: Data matrix X, parameters $\alpha > 0, \beta > 0, \gamma, \mu > 0$.

Initialize: $R = E = 0, Y_1 = Y_2 = 0$.

While not converge do

1. Calculate $R$ by

$$\min_\beta \| R \|_1 + \mu \| Z - R \|^2_F$$

2. Calculate $Z$ by

$$\left\{ X Z + E - X + \frac{Y_1}{\mu} \| X Z - R + \frac{Y_2}{\mu} \|^2_F + \| R - Z \|^2_F \right\}$$

3. Update $E$ according to

$$\min_\beta \alpha \| E \|_1 + \mu \left( \| E - (X - X Z - \frac{Y_1}{\mu}) \|^2_F \right)$$

4. Update $S$ using the scenario in [4];

5. Update Lagrange multipliers $Y_1$ and $Y_2$ as

$$Y_1 = Y_1 + \mu \left( X Z + E - X \right)$$

$$Y_2 = Y_2 + \mu \left( Z - R \right)$$

End while.

V. Experimental results

In this part, we will evaluate the semi-supervised classification algorithm proposed in this paper on two datasets. We first obtain the graph matrix $S$ from our proposed method and then use the graph-based propagation method to perform the semi-supervised classification task [10].

**A. Datasets description**

We evaluated the effectiveness of our method in face recognition using our proposed graph learning in the frequently used YALE and JEFFE datasets. Specifically, the YALE face database has 15 people, each of whom has 11 near-frontal images taken under different lighting conditions; The JEFFE face dataset contains 10 individuals, each with seven different facial expressions, including six basic facial expressions and one neutral expression. Figure 3(a) shows some sample photos of the two cubes. In order to demonstrate the robustness of our method, we randomly added large areas of missing or interference to these two data sets, as shown in figure 3(b).

Fig.1. Sample images of (a) YALE, (b) YALE with defects.
B. Comparison Algorithms

We compare the proposed method to two existing excellent methods.
1) LGC [10]: LGC is a widely used semisupervised classification method.
2) Semisupervised Classification With Adaptive Neighbors (SCAN) [11]: This recently developed method uses adaptive neighbors approach to construct the similarity graph. Moreover, the graph construction and clustering are formulated into a unified framework to improve the performance.

C. Experimental Results

After using different percentages of samples as labels many times, we found that the accuracy of all methods improved as the number of label samples increased. The experimental results are shown in Table 1. It can be seen that our proposed method is superior to other existing technologies when the data contains large area missing or interference. This proves the robustness of the data self-representation model, because when there are missing data, similar data can help reconstruct the data while eliminating noise, so as to construct the map more accurately, which illustrates the importance of removing noise and data reconstruction.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>YALE</td>
<td>10</td>
<td>47.33±13.96</td>
<td>45.07±1.30</td>
<td>47.38±2.57</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>63.08±2.20</td>
<td>60.92±4.03</td>
<td>65.82±1.93</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>69.56±5.42</td>
<td>68.94±4.57</td>
<td>70.74±3.48</td>
</tr>
<tr>
<td>YALE with</td>
<td>10</td>
<td>39.54±23.33</td>
<td>37.58±12.40</td>
<td>46.23±17.45</td>
</tr>
<tr>
<td>defects</td>
<td>30</td>
<td>49.12±19.12</td>
<td>46.17±2.67</td>
<td>56.46±4.83</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>63.31±9.63</td>
<td>58.85±6.32</td>
<td>62.26±5.62</td>
</tr>
<tr>
<td>JAFFE</td>
<td>10</td>
<td>96.65±7.76</td>
<td>96.92±1.68</td>
<td>95.84±4.74</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>98.86±1.14</td>
<td>98.20±1.22</td>
<td>98.74±0.97</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>99.92±0.94</td>
<td>99.25±5.79</td>
<td>99.36±3.83</td>
</tr>
<tr>
<td>JAFFE with</td>
<td>10</td>
<td>93.1±3.79</td>
<td>95.93±2.43</td>
<td>95.14±5.93</td>
</tr>
<tr>
<td>defects</td>
<td>30</td>
<td>97.46±2.14</td>
<td>97.83±4.10</td>
<td>98.56±2.50</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>98.36±2.11</td>
<td>99.03±1.35</td>
<td>99.33±1.96</td>
</tr>
</tbody>
</table>

VI. Conclusion

This paper presents a model for learning robust graphs from data to serve the semi-supervised classification. It solves the problem of robustness of adaptive neighbor graph learning. We improve the conventional model by introducing the self-representation model and the idea of joint solution. Therefore, our proposed framework can not only enhance performance of semi-supervised classification, but also improve low-rank recovery from damaged data. Experiments on several public face benchmark datasets show that the robustness of our method is superior to various advanced existing graphing learning techniques.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under Grant 61501147, in part by the University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province under Grant UNPY C-2018203, in part by the Natural Science Foundation of Heilongjiang Province under Grant YQ2019F011, in part by the Fundamental Research Foundation for University of Heilongjiang Province under Grant LGY C2018Q013, and in part by the Postdoctoral Foundation of Heilongjiang Province under Grant LBH-Q19112.

References


An Intelligent Fault Diagnosis Model Based on FastDTW for Railway Turnout

1st JI Wen-Jiang, 2nd ZUO Yuan, 3rd HEI XIN-Hong, 4th FEI Rong
{wijj@xaut.edu.cn, 759666216@qq.com, heixinghong@xaut.edu.cn, annyfei@xaut.edu.cn}

School of Computer Science and Engineering, Xi’an University of Technology, Xi’an, China
State Key Laboratory of Rail Transit Engineering Informatization FSDI, Xi’an, China

Abstract. The turnout handles the direction of the train which is one of the key equipment in the railway transportation system. In this paper, by using real action current data obtained from switch machine model No.ZD7, a turnout fault diagnosis model based on the FastDTW pattern recognition algorithm was proposed. Firstly, the original current curve was segmented relate to the features of them. Then the warping path distance between the standard sample and the tested current curve was obtained according to FastDTW algorithm. Finally a dynamic optimized threshold was used to confirm whether there is a fault happened in the turnout. According to the experiment results, the proposed diagnose model without the prior knowledge of fault samples can works well both with single and double acting type turnout machines, owning to the following elements: the diagnose accuracy can be more than 96%, the time-cost can be improved more than 5 times compared with traditional DTW based algorithms.

Keywords: turnout; switch machine; fault diagnosis; FastDTW

1 Introduction

Railway turnout system is an important part of railway infrastructure and a frequent failure part in rail transit. Its health directly affects the safety of the railway system. At present, enhancing the safety of turnouts mainly depends on setting the alarm threshold of the signal curve, Periodic inspection of system logs by professional and technical personnel to determine the health status of turnouts. However, this method is inefficient, and the phenomenon of underreporting and misreporting is common, which cannot meet the requirements of railway safety guarantee.

With the continuous development of artificial intelligence and big data technologies, various intelligent fault diagnosis methods for turnouts are gradually being widely studied. In essence, fault diagnosis is a pattern recognition problem. As a powerful pattern recognition technology, artificial intelligence has been widely used in turnout fault problems. For example, Bayesian networks[1], SVM[2], HMM[3], etc. However, most of the existing research methods of turnout fault diagnosis focus on the signal data of single action turnout. There are few related researches on the fault diagnosis of double action turnout. In view of the above reasons, based on the NO.ZD7 turnout switch operating current data, this paper proposes a FastDTW-based turnout fault diagnosis method by analyzing the warp path distance between the switchpoint current and the standard template data to be diagnosed. Experimental data shows that the method proposed in this paper does not require a large amount of prior fault data, and
the algorithm accuracy rate is more than 96%, and the time performance is improved by more than 5 times compared with the traditional DTW algorithm. It is suitable for new train control systems with higher accuracy and real-time requirements.

2 Methodology

Dynamic Time Warping (DTW) algorithm is used to calculate the distance similarity of any two time series, and it is not necessary that the two time series have the same length [4]. It uses the method of dynamic programming to find the warp path, and its time complexity is $O(m \times n)$. When the time series is long, the efficiency of the algorithm is very low. FastDTW [5] algorithm accelerates DTW algorithm by limiting the scope of path search and data abstraction. It is an improved algorithm on the time complexity of DTW algorithm. The algorithm can find an approximate optimal warp path between two time series, and the time consumption increases linearly with the increase of input time series, and its time complexity is about $O(n)$. FastDTW algorithm has three main processes: coarsening, projection and refinement. Figure 1 shows the speedup of fastdtw.

(1) Coarsening: Shrink the original time series, using half of the original time series sampling points to represent the original time series, and the value of each sampling point of the shrink time series is the mean value of two adjacent sampling points of the original time series. As shown in Figure 1, the distance matrix under the 1/1 granularity is the original distance matrix, and the 1/2, 1/4 and 1/8 granularity are the matrices during the three coarsening process of the original time series.

(2) Projection: DTW algorithm is run on the distance matrix of lower resolution. As shown in Figure 1, the black line segment is the warp path found by running DTW algorithm on this granularity.

(3) Refinement: After finding the warp path on the lower resolution matrix, the mesh that the warp path passes through is mapped to the higher resolution matrix.

![Fig. 1. The acceleration process of FastDTW.](image)

The core idea of FastDTW algorithm is to speed up DTW by limiting the search scope and data abstraction, that is, to search only the grid in the projection distortion path (the deepest meshes in Figure 1). However, the optimal warp path may not be included in the projected warp path, so the FastDTW algorithm adds a radius parameter, it is allowed to search the radius meshes outside the projection distortion path mesh, that is, the shading lighter mesh in Figure 1 (radius in Figure 1 is 1). The larger the radius, the more accurate the warp path will be. If the radius parameter is set to the same length as the input time series, the FastDTW algorithm and DTW algorithm have the same efficiency.
3 Turnout fault diagnosis model based on FastDTW

3.1 Characteristic analysis of turnout action current curve

With the continuous development of railway signal system, sensors are often equipped in key equipment of the system, such as switch machine, to monitor the running signal data of the equipment[6]. When the switch fails, its abnormal action process will be fed back to the switch machine, resulting in abnormal monitoring signal data. Therefore, the current curve of switch machine monitored in the process of switch action can reflect the health status of the switch. In this paper, the action current curve of the turnout equipment under the condition of operation at the time of delivery is taken as the standard template. If the difference between the action current curve to be measured and the action current curve of the template is large, it indicates that the turnout is likely to have faults in this action, so that the fault diagnosis of the turnout can be realized.

On the other hand, with the development of China's rail transit system, there are many types of turnout equipment, such as single acting, double acting and multi acting. The action signal data of double action and multi action turnout is more complex than that of single action. Figure 2 shows the normal current curve of single action turnout and double action turnout.

![Fig. 2. Current curve of different types of turnout.](image)

3.2 Fault Diagnosis Based on FastDTW

The warp path distance between the current curve to be diagnosed and the current curve of the template can be obtained by FastDTW algorithm. The warp path distance represents the difference between the curve to be diagnosed and the template curve. The smaller the distance of the warp path, the higher the similarity between the two curves, and the larger the distance, the lower the similarity between the two curves. Therefore, a threshold for the distance of a warp path can be set to determine the health status of the turnout. In addition, the turnout equipment has different operating conditions and different service lives, which results in different optimal diagnostic thresholds for different turnouts. Therefore, the setting of the optimal threshold needs to be continuously adjusted as the equipment changes to achieve a dynamic threshold.

Due to the varying length of each switch's action conversion process, the actual current data sampling lengths of the multiple monitoring switches are different, which leads to the difference in the distance of the warp path obtained by the FastDTW algorithm for different turnouts of the same model even if no fault occurs. But relatively large, as a result, the effectiveness of the threshold is reduced and the difficulty of determining the threshold is increased. In this paper, in order to properly eliminate the influence of the difference in the...
length of the input data, the mean value of the warp path distance is processed, and the final warp path distance is calculated by FastDTW to divide the warp path distance into the curve length to be diagnosed.

For how to confirm the fault judgment threshold, the warp path distances of normal current curve, fault current curve and its template current curve of different types of turnouts are analyzed in this paper. As shown in figure 3, figure 3(a) and figure 3(b) are respectively the warp path distance distribution diagrams of the current curve of the single action type turnout and the double action type turnout and its template current curve. It is not difficult to see from the figure that the distance between the normal curve of the single action turnout and the double action turnout and its template curve is mainly concentrated between 0.0A ~ 0.1A. The distance between the fault curve and the template curve of the warp path exceeds this value, and the distribution is relatively discrete. Therefore, a threshold of warp path distance can be set to determine the health status of turnout.

4 Model performance analysis

In order to test the performance of the proposed model, based on the actual monitoring data of single-month switch machine operating current signals recorded by a railway bureau in China, experiments were performed on single action and double action turnouts. Among them, 200 pieces of data are used to determine the dynamic threshold, 120 pieces of data are used to test the proposed method (single action and double action ratio is 1:1). At the same time, a comparative experiment with DTW(ITDM-BD) and Frechet Distance [7] is added.

4.1 Model accuracy analysis

4.1.1 Determination of the dynamic threshold

In this section, we randomly trained 200 turnout current data according to the proposed model and Fresch distance-based turnout diagnosis method to obtain the optimal threshold of the training data set. It is divided into two types of threshold determination experiments: single action type and double action type. The experimental results are as follows..

Figure 4 shows the results of single and double action turnout fault diagnosis F1-Score for ITBM-BD, ITBM-BF and Frechet Distance. We choose the highest F1-score threshold as the best threshold, then the optimal thresholds of three models for single action turnout are 0.11A,
0.11A and 2.1A respectively and the best thresholds of double action turnout are 0.09A, 0.09A and 2.3A respectively.

![Graphs of F1-score comparison of three methods.](image)

**Fig. 4.** F1-score comparison of three methods.

### 4.1.2 Model accuracy test

Based on the best thresholds determined in section 4.1.1, the model proposed in this paper is used to test the accuracy of the test set composed of 120 randomly selected data sets. The test results are shown in Table 1.

<table>
<thead>
<tr>
<th>Model name</th>
<th>Turnout type</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITDM-BD</td>
<td>Single Action</td>
<td>98.3%</td>
</tr>
<tr>
<td></td>
<td>Double Action</td>
<td>100%</td>
</tr>
<tr>
<td>ITDM-BF</td>
<td>Single Action</td>
<td>96.6%</td>
</tr>
<tr>
<td></td>
<td>Double Action</td>
<td>100%</td>
</tr>
<tr>
<td>Frechet Distance</td>
<td>Single Action</td>
<td>88.7%</td>
</tr>
<tr>
<td></td>
<td>Double Action</td>
<td>84.3%</td>
</tr>
</tbody>
</table>

As can be seen from Table 1, both ITDM-BD and ITDM-BF reach a higher F1-Score. Combining the F1-Score variation ranges of the two under different thresholds in Figure 4, it is not difficult to see that the two have little difference in accuracy performance, and both can meet actual application requirements. And both models are better than Frechet Distance based diagnosis model.

### 4.2 Model time performance analysis

In this section, 150 pieces of real data are divided into 5 test sets of different sizes, and the running time efficiency of ITDM-BD and ITDM-BF is tested respectively.

<table>
<thead>
<tr>
<th>Model</th>
<th>Turnout Type</th>
<th>Input dataset size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>ITDM-BD</td>
<td>Single</td>
<td>14.11s</td>
</tr>
<tr>
<td></td>
<td>Double</td>
<td>25.42s</td>
</tr>
<tr>
<td>ITDM-BF</td>
<td>Single</td>
<td>2.09s</td>
</tr>
<tr>
<td></td>
<td>Double</td>
<td>3.60s</td>
</tr>
</tbody>
</table>
As can be seen from Table 2, because ITDM-BF uses the FastDTW algorithm to calculate the warp path distance, its speed performance is significantly better than that of ITDM-BD, and the algorithm time efficiency is improved by at least 5 times.

5 Conclusion

For the fault diagnosis of turnout, this paper takes the current data of the switch machine monitored during the turnout operation as input, and obtains the current health status of the turnout to be diagnosed through artificial intelligence algorithms such as segmentation, pattern recognition and threshold determination. Experimental results show that ITDM-BD is not significantly different from ITDM-BF in accuracy performance, but the latter is significantly better than the former in terms of time efficiency. Therefore, ITDM-BF is more suitable for new train control systems with high accuracy and high real-time performance. In addition, without a large number of fault sample data, and for single- and double-acting turnout equipment, the model proposed in this paper can still efficiently and correctly diagnose turnout fault conditions, indicating that ITDM-BF has a wider scope of application. As the sample data is too small, it is necessary to verify whether the model in this paper is satisfied with the hyperactive turnout equipment, and the determination of the specific failure type of the turnout is still a future research direction.

References


Acknowledgments

This work was partly supported by the National key research and development program of China under Grant No. 2018YFB120500, the National Natural Science Foundation of China under Grant No. 61773313, No. 61702411.
Light Deep Learning based Edge Safety Surveillance

Yimo Lou*, Wengang Cao, Zhimin He, and Guan Gui*

College of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, 210003 China
*Email: 1219012829@njupt.edu.cn, guiguan@njupt.edu.cn

Abstract. Safety is considered as the first important factor in many industries such as construction sites. Hence, artificial intelligence based safety surveillance techniques have been received strong attentions in recent years. Conventional surveillance systems for monitoring whether the workers wearing helmets are not easy to install and carry, and the largest trouble is that the system needs considerable computation, which is not that simple to satisfy the requirement of hardware. Considering the characteristic about construction sites, in this paper, we proposed a new system based on CenterNet with MobileNet-V2 as backbone. It has a video camera, a marginal device embedded with Jetson TX2 and wireless communication routers to ensure real-time transmission about live-scene about construction sites. After inspection, the light-weight network we proposed can be run in portable marginal device smoothly and stably with slight loss of average precision.

Keywords: Intelligence safety surveillance, CenterNet, MobileNet-V2, marginal devices

1 Introduction

With the development of society, to improve environment of every aspect about human beings, new buildings spring up like mushrooms. Meanwhile, accidents happen in construction sites become a serious problem that threatens the progress of whole project. Moreover, consequences caused by accidents lead to inestimable loss to both enterprises and families of staff who are injured or dead. Thus ensuring safety of staff is significant in construction activities. From the article, we found about construction accidents in China during 2014 to 2018, 3014 accidents happened, causing nearly 4000 mortalities. More than half of these accidents (52.2%) due to falling from high place, other reasons are hitting by objects (15.2%), crane injury (7.5%), collapse (7.3%), mechanical injury (5.9%) and others (11.9%). In common sense, most of them could be avoided or less influential by wearing safety equipment, such as helmets. Hence, supervision about whether workers wear helmets is a meaningful direction to protect staff’s lives.

Traditional measures to deal with this problem cannot be deployed without human, which are waste of manpower and money. Besides, supervision by humankind has flaws that may not have solution to make up. Since majority of people are just flesh and blood without super power, the most serious problem in artificial supervision is that mortal horizon is limited. There are blind spots everywhere in construction sites, not mentioned the field operation carried out vertically, arranging supervisors at every area or height is unrealistic and a total waste. If used, not only does this method cost a large amount of money, as the number of workers increasing for supervision, but also rise the probability of potential danger.

Therefore, automatic supervision about whether workers wearing helmets turns into a meaningful topic. Bolstered by rapid development of artificial intelligence (AI), we can use this method to deal with
problem above. However, advanced method via AI generally need considerable computation, which is not easy to realize by simple portable equipment that can be used at construction sites. Besides, although the network detects helmets by frame, what is shown on screen is video composed of frames, the frequency of it must be ensured, not mentioned that it is a real-time detection. What’s more, owing to instability of sites, equipment with low cost to avoid extra loss due to damage is more suitable. Transferring the whole network to a marginal device with tiny size needs to be considered either. To sum up, a light-weight network is needed.

In this paper, we propose a new method about helmet detecting with lighter construction. We use CenterNet [1] which has MobileNet-V2 [2] as backbone. Meanwhile we transfer this network into marginal devices. By putting external device, cameras, at place we want to monitor, accent color boxes will appear in screen to show whether workers are wearing helmets.

2 Related Works

Nowadays, machine learning has integrated into every aspect of our life. G. Gui, M. Liu et al. proposed a series of method based on aviation big data and machine learning [3-4]. For some research about communication on channels and module classification, H. Huang et al. [5-12] use deep learning to improve original solution to get better result, not to mention the most popular topic, 5G/6G [13].

The whole method can be transformed into an object detection problem based on deep learning. One algorithm tried the most direct method [14], C. Wojek et al. proposed an algorithm using sliding windows to detect objects from left to right then up to down by classification. In order to detect different class from different distance, it utilizes disparate length-width ratio. Unlike sliding window, Region Convolutional Neural Network (R-CNN) proposed by R. Girshick et al. [15] obtain region of interest (ROI) by region proposals. It considers each pixel as a group, and then calculate their texture and combine two regions, which are most close to each other, until all of them combined. To upgrade the accuracy, R-CNN needs a fat lot of candidate region, large part of them are overlap. Thus Fast Region Convolutional Neural Network (Fast R-CNN) [16] proposed by the same team, which uses CNN to extract features of whole image rather than extract multiple times on every image block. Upgraded version of Fast R-CNN is Faster R-CNN [17], which let the network learn which candidate region of image are by itself instead of using regular algorithm, it has higher efficiency. Above are two-stage object detection algorithm. Liu et al. proposed Single-Shot MultiBox Detector (SSD) [18], which uses deeper layers in convolutional network to detect object. Due to decrease of Spatial dimensions and resolution caused by convolutional layer, regular structure of it can detect larger object only while execute independent detection from multiple feature maps could solve this problem. You only look once (YOLO) proposed by J. Redmon et al. [19-21] detects features by using DarkNet after convolution layer. It smooths feature maps and then unites with another feature map with lower resolution.

Departed object detection algorithm, which can be used in this condition, are insufficient to satisfy the need of light-weight we want. Thus we choose CenterNet with MobileNet-V2 as our ideal solution, which we will introduce in next part.

3 The Proposed Method

The whole system is based on Internet of Things (IoT). The data acquisition layer is a camera.
connected to marginal embedded device, power provided by Jetson TX2, and transmission layer utilizes router to data acquisition layer with server. The application layer is a screen to show the result.

![Diagram of basic framework](image)

**Fig. 1.** Basic framework of the proposed edge safety surveillance.

### 3.1 CenterNet

Object detection needs to make a rectangular box containing target with minimal size. A successful detector will list a huge amount of candidate boxes and classify them, which is low effective and a total waste, subsequent disposal cannot be avoided neither. CenterNet proposed by X. Zhou et al. has a brand new thought that we only need to locate the object as a point, which is the center point of detection box. By using heat map, detector will find its center point and then regress other attribute of this object, such as size, coordinate of 3D location, direction, posture etc.

CenterNet is an end-to-end fast object detector. We use the center points of human and helmet to represent them, then the only anchor boxes shaped as size of objective are regressed directly from the peak in heat map (one anchor for one objective, cancel Non-maximum suppression (NMS), a quiet time-consuming process). Therefore, the detection and alignment of these two objective become a standard key point estimation.

![Diagram of CenterNet structure](image)

**Fig. 2.** The structure of CenterNet used in our proposed method, we take MobileNet-V2 as backbone.

Owing to adopting fully convolutional network, CenterNet could get heat map with larger resolution (output stride of 4, a fourfold reduction compared to traditional ones) without setting anchor boxes in advance. Channels of heat map has the same quantity with classifications of objects need to detect. The
net takes first 100 peaks of heat map as center points and set a threshold to get final point.

There is deformable convolution before every upsampling to make receptive field of network become more precise instead of being limited into rectangular convolutional box with size of $3 \times 3$. Meanwhile, feature maps under downsampling, which has output stride of four possesses higher resolution than normal network.

Intersection over Union (IoU) is a criterion to measuring accuracy of standard anchor based detection, such as CenterNet. There is a threshold value, when IoU bigger than that, the anchor counts as positive to any object.

![Fig. 3. Sketch map of IoU, where purple frame is the ground truth, when $\text{IoU} = 0.7$, anchor green is positive, while anchor red is negative.](image)

the improved loss function of classification in CenterNet in our method is as follows:

$$L_k = \frac{-1}{N} \sum_{xyc} \left( \begin{array}{l}
L_c (1 - \hat{Y}_{xyc})^\alpha \log(\hat{Y}_{xyc}) \\
L_c (1 - Y_{xyc})^\beta \log(1 - \hat{Y}_{xyc})
\end{array} \right)$$

if $Y_{xyc} = 1$ otherwise

Due to images in datasets are not been fully labeled, we set a new parameter $L_c$, which represents whether the class has been labeled, $L_c = 1$ means labeled already while $L_c = 0$ means unlabeled. $N$ is number of center points in image 1, $\alpha$ and $\beta$ are hyper parameters of Focal Loss, $\hat{Y}_{xyc}$ is predictive value of object.

3.2 MobileNet-V2

MobileNet proposed by M. Sandler et al. is a lightweight CNN network focused on mobile terminal or embedded device. Consider that operating in portable marginal devices, we take MobileNet-V2 as backbone and Feature Pyramid Network (FPN) [22] as neck of CenterNet. FPN can make a tradeoff between speed and accuracy to obtain more robust and semantic information.

MobileNet-V2 improves depthwise separable convolution [23], which separates the two steps to depthwise convolution and pointwise convolution. First, do bitwise multiply by channels, while keep quantity of channels the same; second, do traditional convolution with kernel, which size is $1 \times 1$, quantity of channels can be changed. Now computation reduce to
\begin{equation}
D_K \times D_K \times M \times D_F + 1 \times 1 \times N \times D_F + D_F
\end{equation}

where \( D_F \) is the size of feature map, \( D_K \) is the size of convolutional kernel, \( M \) is quantity of input channels, \( N \) is the quantity of output channel. We set kernel size to 3, reducing computation by 8 to 9 times with slight accuracy loss compared with traditional convolution.

The batch normalization layer and ReLU6 after convolution, increasing the nonlinear variation and enhancing the generalization ability of model. However, ReLU6 consumes convolutional kernel in depthwise convolution, thus MonileNet-V2 replace ReLU6 to Linear activation function. Meanwhile, the accuracy is higher than former one.

Fig. 4. The basic convolution unit structure of MobileNet-V2.

### 3.3 Other parts

Above are related algorithm used on detection, now we introduce other parts of this system. Video camera connected with marginal embedded AI device powered by NVIDIA Jetson TX2, provides real-time scene of construction sites to monitor. Jetson TX2 adopts 256-core NVIDIA Pascal GPU and 8 GB memory, providing higher computing speed and stronger inferential capability. Router provided the function of wireless long-distance data transmission. We use OpenCV to play the result on screen.

### 4 Experiment

#### 4.1 Dataset

Data collection in our project includes three kinds: 1) pedestrian pictures from pedestrian labeled pictures in coco2017 (pedestrians labeled only, more than 60000); 2) photos taken by ourselves (both pedestrians and helmets labeled, 9000); 3) helmet pictures from Internet (helmets labeled only, more than 10000).
4.2 Training

We rescale the size of images into $512 \times 512$. The learning rate was initialized at $1e^{-3}$, and was reduced by a factor of 10 at the end of 80, 120, 103 epochs.

4.3 Result

4.3.1 Different epochs and IoUs

We test our model with different epochs and IoU, try to find best combination. The results about average accuracy are as follow:

<table>
<thead>
<tr>
<th></th>
<th>IoU=0.30</th>
<th>IoU=0.50</th>
<th>IoU=0.70</th>
<th>IoU=0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helmet</td>
<td>91.44</td>
<td>77.44</td>
<td>41.34</td>
<td>29.86</td>
</tr>
<tr>
<td>Pedestrian</td>
<td>78.88</td>
<td>70.62</td>
<td>43.66</td>
<td>33.16</td>
</tr>
</tbody>
</table>
Table 2. Average precision under epoch = 55

<table>
<thead>
<tr>
<th></th>
<th>IoU=0.30</th>
<th>IoU=0.50</th>
<th>IoU=0.70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helmet</td>
<td>93.04</td>
<td>79.42</td>
<td>42.44</td>
</tr>
<tr>
<td>Pedestrian</td>
<td>79.91</td>
<td>71.46</td>
<td>45.05</td>
</tr>
</tbody>
</table>

Table 3. Average precision under epoch = 140

<table>
<thead>
<tr>
<th></th>
<th>IoU=0.30</th>
<th>IoU=0.50</th>
<th>IoU=0.70</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helmet</td>
<td>92.04</td>
<td>79.40</td>
<td>44.86</td>
</tr>
<tr>
<td>Pedestrian</td>
<td>81.42</td>
<td>73.92</td>
<td>49.80</td>
</tr>
</tbody>
</table>

We can see from the result that when epoch up to 140, the average accuracy becomes lower, and it is a total waste of time. If we set the IoU too high, we are goanna miss some information which lead accidents happen. To sum up, we choose to set IoU = 0.5 and epoch = 55.

4.3.2 Different backbones

To show the high efficiency of MobileNet-V2, we compared our method with other algorithms about CenterNet that have other network as backbone (DLA-34, DLA34-V0, ResNet-18). All the tests are conducted under TensorRT infer optimizer. (In this section, fp32 means 32-bit floating point, fp16 means 16-bit floating point)

DLA [24]: Deep Layer Aggregation (DLA), an image classification network with hierarchical skip connections, makes the model become more precise and have smaller quantity of parameters, providing a measure for generalization for deep visualization framework and effective expansion of application.

ResNet [25]: Residual Network (ResNet) is a feature-extracting network based on CNN, which can simplify the training of networks that have deeper structure than those we used previously.

Table 4. Inference Time tested on NVIDIA GeForce GTX 1080Ti

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Mode (precision)</th>
<th>Inference Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNet-V2</td>
<td>fp32</td>
<td>3.798</td>
</tr>
<tr>
<td>DLA-34</td>
<td>fp32</td>
<td>24</td>
</tr>
<tr>
<td>DLA34-V0</td>
<td>fp32</td>
<td>12.6</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>fp32</td>
<td>5.81</td>
</tr>
<tr>
<td>MobileNet-V2</td>
<td>int8</td>
<td>1.75</td>
</tr>
<tr>
<td>DLA-34</td>
<td>int8</td>
<td>19.6</td>
</tr>
<tr>
<td>DLA34-V0</td>
<td>int8</td>
<td>6.76</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>int8</td>
<td>3.63</td>
</tr>
</tbody>
</table>
Table 5. Inference Time tested on NVIDIA Jetson TX2

<table>
<thead>
<tr>
<th>Backbone</th>
<th>Mode (precision)</th>
<th>Inference Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MobileNet-V2</td>
<td>fp16</td>
<td>22</td>
</tr>
<tr>
<td>DLA-34</td>
<td>fp16</td>
<td>186</td>
</tr>
<tr>
<td>DLA34-V0</td>
<td>fp16</td>
<td>80</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>fp16</td>
<td>41</td>
</tr>
</tbody>
</table>

We can know from the table that no matter what condition, MobileNet-V2 always has the best performance, which means it has the highest frame frequency.

4.3.3 Final test on real scenes

The final test is on videos we find about construction sites, which shows the result of detection directly.

![Scene 1](image1.png) ![Scene 2](image2.png)

![Scene 3](image3.png) ![Scene 4](image4.png)

Fig. 8. Detection result on different scenes.

5 Conclusions

In this paper, we proposed a light-weight supervision system for construction sites which can be run on marginal devices. This method used MobileNet-V2 as backbone embedded into CenterNet to realize a faster and more precise detection. With Jetson TX2, the system upgrades in stability and convenience. In the future, we plan to do further research on improving average precision, because it still has slight decrease about that than those normal weight networks. Also, we are going to add new function to supervise other accident may happen in sites, to make this system more complete.
References


A Practical Low-Cost Security Solution for Log Management and File Integrity Monitoring

Lei Chen1, Ming Yang2, Hayden Wimmer3, Brittany Wilbert4
{lchen@georgiasouthern.edu1, MYang8@kennesaw.edu2, HWimmer@georgiasouthern.edu3, BMW005@shsu.edu4}

Department of Information Technology, Georgia Southern University, Statesboro, Georgia, USA1,3, Department of Information Technology, Kennesaw State University, Kennesaw, Georgia, USA2, Google, Mountain View, California, USA4

Abstract. Log management and file integrity are among the most critical security controls in protecting valuable business data assets against internal and external security attacks. Despite the importance of these controls, many businesses, especially smaller ones, do not practically implement the controls due to reasons including cost and personnel. In this research, we propose a framework, utilizing low cost or free-of-cost tools, and offer guidance for efficient log management and integrity monitoring. A detailed list of relevant hardware, software, and tools as well as their corresponding costs is provided to assist businesses in strategic security planning.

Keywords: security, log management, file integrity monitoring, low-cost, forensics.

1 Introduction

A Data Breach Investigation report by Verizon indicated that a litter over half of small businesses were affected by malware and almost three quarters were affected by hacking incidents in 2019 [1]. Among the various reasons to this situation are the two critical difficulties that small businesses are facing: cost and personnel. Compared to mid-size and large-scale businesses, small businesses typically do not have sufficient funds or personnel to plan, implement, test or evaluate the security of their systems and networks which potentially leads to critical data breaches as well as system and network malfunctions detected long after they occurred [2][3][4]. Therefore, there is an acute need for developing a framework with a set of recommended hardware, software, and tools for log management and file integrity at low cost for businesses which cannot afford large-scale security solutions.

To answer the increasing demand for low-cost yet efficient security solutions for small businesses, our research proposes a framework solution meeting their needs. While security requires preventative, detective, and responsive mechanisms in multiple layers, such as firewalls, antivirus and antimalware software, and intrusion detection and prevention systems (IDS/IPS), our interest in this study focuses on the log management and file integrity. The rationale is that, despite the fact that many intrusions (appeared as changes to files and logs) can be potentially discovered via log management and file integrity checks, they still occurred in an environment protected by firewalls and other security mechanisms. Most small businesses either have not realized its importance or have not found a sound solution meeting their budget. In the
next few sections, a brief introduction to log management and file integrity monitoring and the corresponding standards and regulations is given [5].

2 Log Management and Requirements

In a small business, audit log data is generated by hardware, operating systems, services and applications. Log management includes the collection and inspection of such log data. Such mechanisms require monitoring audit logs and creating reports as well as analysis that can be easily reviewed by lightly trained personnel. In order to meet security standards and regulations, such as the Payment Card Industry (PCI) PCI-DSS Requirement 10.6, a formal log review structure should be in place ensuring that daily monitoring of events is performed [6].

According to HIPAA News Releases & Bulletins, every year hundreds of thousands of patients’ health information was exposed due to HIPAA privacy violations [7]. In an effort to prevent and detect such data breaching, both private industry and federal authorities have established standardized processes that include log management as a core requirement. Small businesses have the responsibility to ensure that their log management procedures and operations are meeting the requirements of the latest regulations and standards which also require the capability for forensic review when necessary.

As an example, regulatory agencies and regulations such as the PCI-DSS and the SOX (Sarbanes-Oxley Act) require companies to monitor and audit log files created in their infrastructures [8][9]. In order to avoid tremendous fines and fees due to data loss, small businesses must implement sound and efficient log management and data retention. Based on the studies of regulations and industry practices, we recommend the following process for collecting log messages and establishing and documenting standard operating procedures (SOP). A framework of log management and file integrity monitoring with details of tools and costs is presented later in this paper.

Log Retention Policy requires the length of time for which logs should be kept. Depending on regulatory requirements, log retention time must be at least 30 days, with typical retention time between one to seven years. Data storage space, e.g. hard drive size allocated for logs, can be determined based on the initial storage and long-term storage. For Linux environment log collection, rsyslog is recommended, and for Windows environment tool Snare is recommended to convert Windows formatted EVT log files to Syslog, which can then be sent to central log management locations. In this way, syslog configuration can be set for all system regardless of platforms. According to PCI-DSS Requirement 10.6 [8], log messages should be reviewed on a daily basis. Such log reviewing can be done manually or automatically, with necessary alerting mechanism set in place for the latter.

3 File Integrity Monitoring (FIM) and Requirements

File integrity monitoring (FIM) is critical implementation and practice for all companies and organizations to prevent and detect unauthorized modifications to their information assets. FIM can not only detect malicious attacks from outside of the organization, but also intentional or accidental changes to data without permission due to many reasons, such as system and
software bugs and failure, human mistakes, etc. We emphasize the importance of FIM for the following reasons. Without using FIM, an organization may overlook a potential compromise that may cripple the organization leaving little or no audit trails, which are necessary for forensic investigations. Therefore, it is wiser to invest on FIM for the protection of the information assets [10][11]. Study has shown that 87% of small businesses do not have internal documented security procedures or best practices for their environment [12], where an incidence of accidental data loss due to employee’s mistake is considered an enormous vulnerability in such environment. Hence a snapshot of the environment becomes critical to ensure a healthy baseline of hardware and software infrastructure in case of incidence.

Similar to log management, regulatory agencies and regulations also require companies and organizations to include file integrity monitoring within their environment [8]. Small businesses have to provide documentation as well as implement the technical requirements necessary to implement FIM within the environment. Following these regulations and industry practices, we suggest the following FIM configuration process, which should be included in the established and documented standard SOP of an organization.

FIM tool needs to be configured to not only protect the environment but also guard FIM itself from potential tampering. The fingerprint, a file used for verifying the integrity of data compared to its earlier version utilizing cryptographic hash functions, such as the Secure Hash Algorithm 2 (SHA-2), must be backed up and stored in a secure manner. Upon the completion of FIM tools in the environment, reviews of alerts from these tools, such as the Open Source Tripwire, need to be closely monitored. If a red flag is raised, further in-depth review should be performed and escalation to relevant parties may be necessary. At any event of adding a new patch of major system update, the fingerprint must be updated to ensure the review of system and network activities is accurate based on the current state of the environment.

4 Proposed Framework for Log Management and FIM

In this section, we first introduce a number of tools for our proposed log management and file integrity monitoring framework. These tools have been tested and are used in daily practice. We also include a section of our tool analysis and list the cost for small businesses to implement the framework.

4.1 Tools

In the selection of tools, considering minimizing the cost to small businesses or businesses with a tight financial budget, several open source or free tools are identified and reviewed. These tools combined will provide sufficient functionality for log management and FIM. Our recommended tools also avoid major licenses for use in a small business environment. Some of these tools have equivalent enterprise versions with a certain cost. This will happen when a small business has expanded to a larger size and their employment number has exceeded the allowed maximal number for the free version of the tools. However, most businesses in such scale will more likely have a better financial situation and tend to be more willing to invest in the security of their systems and networks. In our proposed framework, six software tools and services are used to create a system for log management and FIM.

1) Rsyslog
Syslog is the default Linux log manager. Rsyslog extended syslog by adding features such as listening the TCP traffic, and therefore supports both local and remote log management. It has a robust configuration system that gives very fine-grained level of control over what to collect, how to collect, where to store the log files, as well as how logs are configured for future analysis.

2) Snare
Snare has the capabilities of filtering, reviewing, and sending Windows proprietary event logs. One extra bonus feature of Snare however is it can configure/convert Windows log messages into Linux syslog format and send them to a (remote) centralized location for log management. In our daily practice, we use Snare as a compliment to rsyslog for Windows platforms.

3) Logstash
The open source log filtering tool Logstash has a simple interface for easy and quick log management, which makes it suitable for small businesses. Logstash can be used along with other tools such as Elasticsearch and Kibana (later in this section) for analyzing and filtering various types of log messages. Logstash also allows to use Bash or Perl scripts to integrate plugins for enhancing alerts and notifications. Another advantage of Logstash is that it works on multiple Linux distributions, including Ubuntu.

4) Elasticsearch
Elasticsearch is used for log collection and storing logs in server database. A highly valued feature of Elasticsearch is that it works seamlessly with Logstash to store filtered log messages at a central location, and it also works perfectly with programs such as Kibana for people who prefer GUI log management interface. Elasticsearch is available in both a demonstration level and business software packet for various Linux distributions.

5) OSSEC
OSSEC is a multi-purpose security management tool that is capable of file integrity monitoring, reporting system changes, and incident remediation, such as remote firewall changes and blocks. With both an open source and enterprise versions, OSSEC is frequently updated with new versions. OSSEC supports agent on Windows platform and both server and agent for Debian and Red Hat Linux. In the proposed framework, OSSEC is only used for file integrity monitoring purpose.

6) Kibana
Kibana allows for visualization of audit log events in the environment for efficient and easy log management, including providing clear and concise event logs understandable by lightly trained personnel. Since Kibana works well with Elasticsearch and Logstash, it is recommended in our proposed framework. Plug-ins are available for integrating Kibana into Elasticsearch and Logstash, so the installation and configuration becomes simple and straightforward.

4.2 Software Configuration
The software configurations are very important in the proposed framework for a system with different tools working seamlessly. As an example, discussions below are based on an Ubuntu Linux Server 18.04 LTS functioning as the central log management server.

Rsyslog on the server was configured to allow incoming traffic using UDP port 514. A template of log configuration script was created as shown in Figure 1. Snare was configured to allow IIS (Internet Information Service) log messages to be viewed. Logstash needs to be
configured so that data from Rsyslog, OSSEC, and Snare can be filtered and then properly recorded. A portion of such configuration script is shown in Figure 2 as an example. This filter allows for syslog files that are in rsyslog format to be filtered through Logstash.

Fig. 1. A template of log configuration script for Rsyslog.

```bash
## Update rsyslog file (assuming file is in default location
/etc/syslog.conf and has not been modified)
sed -i 's/\%\$\%ModLoad\%\%ModLoad\%\g/ /etc/syslog.conf

sh -c 'echo "\rtemplate(name="Tmp\Auth") type="string"" >>
/etc/syslog.conf

sh -c 'echo "\rstring="/\r/var/log/%HOSTNAME%\%PROGRAMNAME\Log"" >>
/etc/syslog.conf

sh -c 'echo "\"" >> /etc/syslog.conf

sh -c 'echo "\rstring="/\r/var/log/%HOSTNAME%\%PROGRAMNAME\Log"" >>
/etc/syslog.conf

sh -c 'echo "\rstring="/\r/var/log/%HOSTNAME%\%PROGRAMNAME\Log"" >>
/etc/syslog.conf

sh -c 'echo "\"" >> /etc/syslog.conf
```

Fig. 2. A portion of example Logstash configuration script.

Other configurations of these tools are quite straightforward with just one note to be addressed. The OpenSSL certificates must be installed on log management server and the Kibana server in order to make Kibana work.

4.3 Cost of Implementation

The cost of implementation consists of hardware, software, and personnel. The estimated cost addressed below is for typical scenarios in a small business environment (e.g. total employee number less than 100) [13], assuming the existence of common infrastructure that a small business would have, such as broadband internet access, properly configured routers, switches and local area networks (LANs). The details of unit costs and total costs for the hardware and software of the proposed framework can be found in Table 1.
Table 1. Approximate Hardware/software Cost.

<table>
<thead>
<tr>
<th>Hardware / Software</th>
<th>Instances Implemented</th>
<th>Unit Cost</th>
<th>Total Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desktop computer</td>
<td>1</td>
<td>$900.00</td>
<td>$900.00</td>
</tr>
<tr>
<td>Snare</td>
<td>1</td>
<td>$0.00</td>
<td>$0.00</td>
</tr>
<tr>
<td>Rsyslog</td>
<td>4</td>
<td>$0.00</td>
<td>$0.00</td>
</tr>
<tr>
<td>Logstash</td>
<td>1</td>
<td>$0.00</td>
<td>$0.00</td>
</tr>
<tr>
<td>Elasticsearch</td>
<td>1</td>
<td>$0.00</td>
<td>$0.00</td>
</tr>
<tr>
<td>Kibana</td>
<td>1</td>
<td>$0.00</td>
<td>$0.00</td>
</tr>
<tr>
<td>OSSEC</td>
<td>5</td>
<td>$0.00</td>
<td>$0.00</td>
</tr>
<tr>
<td>Windows 10 Pro</td>
<td>1</td>
<td>$199.99</td>
<td>$199.99</td>
</tr>
<tr>
<td>Ubuntu Servers (VMs)</td>
<td>4</td>
<td>$0.00</td>
<td>$0.00</td>
</tr>
<tr>
<td>1 TB Hard drive</td>
<td>2</td>
<td>$70.00</td>
<td>$140.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>$1,239.99</strong></td>
</tr>
</tbody>
</table>

The estimated personnel cost is shown in Table 2. Based on this information, if a single employee is employed to work on these tasks with the installation, configuration, and maintenance of two Ubuntu servers, two Ubuntu Desktops, and one Windows workstation allow with the daily review of the activity that results, it would cost approximately $13,417.83 at an hourly pay rate of $30.21. However, a larger infrastructure would cause an increase in this cost, particularly if there are increases in the number of systems in the environment, log messages that require review and the number of incidents that occur. This also does not include any additional roles personnel may be fulfilling.
### Table 2. Approximate Annual Personnel Cost.

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Hours Per Week</th>
<th>Hourly Cost Per Week ($)</th>
<th>Total Cost per year ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware installation (Single Cost)</td>
<td>2</td>
<td>61.41</td>
<td>61.41</td>
</tr>
<tr>
<td>Software installation – Windows 10: 1 instance (Single Cost)</td>
<td>1.5</td>
<td>46.82</td>
<td>46.82</td>
</tr>
<tr>
<td>Software installation – Virtualbox (Single Cost)</td>
<td>4</td>
<td>122.82</td>
<td>122.82</td>
</tr>
<tr>
<td>Software configuration – Virtualbox (Single Cost)</td>
<td>0.5</td>
<td>15.11</td>
<td>15.11</td>
</tr>
<tr>
<td>Virtualbox configuration – Ubuntu servers – 2 workstations and 2 servers (Single Cost)</td>
<td>6</td>
<td>187.28</td>
<td>187.28</td>
</tr>
<tr>
<td>RSyslog configuration (Single Cost)</td>
<td>2</td>
<td>$30.21</td>
<td>61.41</td>
</tr>
<tr>
<td>Snare Configuration (Single Cost)</td>
<td>2</td>
<td>$30.21</td>
<td>61.41</td>
</tr>
<tr>
<td>Logstash configuration (Single Cost)</td>
<td>2</td>
<td>$30.21</td>
<td>61.41</td>
</tr>
<tr>
<td>OSSEC Configuration – Agent: 4 instances</td>
<td>1.5</td>
<td>46.82</td>
<td>46.82</td>
</tr>
<tr>
<td>OSSEC Configuration – Server: 1 instance</td>
<td>0.5</td>
<td>15.11</td>
<td>15.11</td>
</tr>
<tr>
<td>Elasticstash configuration (Single Cost)</td>
<td>2</td>
<td>$30.21</td>
<td>61.41</td>
</tr>
<tr>
<td>Kibana Configuration (Single Cost)</td>
<td>2</td>
<td>$30.21</td>
<td>61.41</td>
</tr>
<tr>
<td>Script Creation and Testing (Single Cost)</td>
<td>25</td>
<td>755.25</td>
<td>755.25</td>
</tr>
<tr>
<td>Maintenance and Updates (estimate per week)</td>
<td>1</td>
<td>30.21</td>
<td>1570.92</td>
</tr>
<tr>
<td>Daily Log Review (estimate per week)</td>
<td>5</td>
<td>151.05</td>
<td>7854.6</td>
</tr>
<tr>
<td>Incident Response (estimate per week)</td>
<td>1.5</td>
<td>46.82</td>
<td>2434.64</td>
</tr>
<tr>
<td>Totals</td>
<td>Total Hours: 58.5</td>
<td>Total Individual Cost (Per Week): $1,629.75</td>
<td>Estimated total cost/year: $13,417.83</td>
</tr>
</tbody>
</table>

### 5 Conclusion and Future Work

Implementation of a security framework within a small business is a vital component to provide protection of the intellectual property of small organizations. Although small businesses may not hire as many employees as large corporations, the number of small businesses alone with infrastructure in risk as a result of non-existent or insufficient security may allow for these businesses to be exploited without recourse and open the potential for ransomware as well as other attacks.
Furthermore, although this work is specifically focused on small businesses with relatively tight budgets, there are many other public and private industries which may organize portions of their businesses like a small business for periods of time. Implementing some of the framework requirements may allow for these departmental structures to be transitions to more formal processes and procedures within the organization.

Small businesses have unique obstacles that are not faced by larger or more established organizations. Because they are more likely to be compromised by malicious users, the cost of deploying solutions to mitigate compromise can be unaffordable, and individuals that are tasked to manage these solutions may not have the time or resources available to effectively monitor and protect against incidents. Due to this, small businesses must find alternate solutions that will fit their environment and existing infrastructure.

The Small Business Framework with the proposed Log Management and File Integrity Monitoring is to provide small businesses starting resources that can be used to monitor their fragile environments holistically while they are growing their business. The ultimate goal of the framework is to allow the opportunity for guidance to be provided to these businesses that can be expanded on in the future. Although there are other solutions which may be used to obtain a similar result, applying some of the techniques discussed in this paper can allow small business organizations to better prepare for security events that may occur within their environment.

Cloud computing allows for small businesses and other enterprises to leverage public and private infrastructure in order to potentially lessen the cost of infrastructure and security requirements for the organization. Integrating cloud log management and file integrity monitoring into the security framework would be the next work. While integration of cloud computing is important to provide a discussion to small businesses, the importance of proper contracts and Service Level Requirements (SLA) becomes a higher necessity. Without strict contracts a cloud computing organization could pose a greater risk to the security data of the organization than the organization’s technological structure. It must also be considered how the cloud computing structure, regardless of if it is created by a small business itself or a third party, is properly segmented and protected from threats.

References


Analysis the Reliability Capability of Cognitive Frequency Hopping Communication System

Chenxi Li¹, Peihan Qi¹,², Danyang Wang¹, and Xiaoyu Zhou¹

¹ State Key Laboratory of Integrated Services Networks, Xidian University, Xi’an
² Collaborative Innovation Center of Information Sensing and Understanding
henxili@stu.xidian.edu.cn

Abstract. In this paper, we propose a method that can effectively measure the reliability capability of cognitive frequency hopping (CFH) communication systems. Considering that various malicious interference devices and advanced interference technologies have appeared in the wireless communication environment, a novel method called CFH has been proposed recently to effectively resist the interference devices. This method can evaluate the occupation of frequency hopping frequency slots, and thus dynamically adjust the system parameters based on the evaluation results. However, we note that the existing literature only shows that the system is reliable, but few work has been done to analyze the reliability of the CFH system and the factors affecting the reliability. Therefore, by analyzing the effects of false alarm probability, missed detection probability, and communication link convergence delay, we obtain analytical expressions to evaluate reliability. Simulation results evaluate the proposed method.

Keywords: Cognitive frequency hopping (CFH), reliability capability, false alarm probability, missed detection probability, and communication link convergence delay

1 Introduction

Wireless communication, as a method of long-distance information transmission that can be carried without conductors or cables, has been widely used in many fields (such as cellular networks, industrial production, etc.). However, while the wireless communication technology is showing a booming trend, the risk that the information it brings is easily interfered or intercepted has also caused widespread concern [1–4]. Wireless communication is different from traditional wired communication, and it mainly relies on electromagnetic waves to transmit signals. Due to the open nature of wireless channels and the broadcast nature of radio transmission, information is extremely vulnerable to external interference during transmission. In addition to sources of malicious interference in wireless communication environments, signals from unauthorized users are also considered important sources of interference. Moreover, when multiple wireless devices
are working simultaneously, signals sent by other authorized users may also cause interference to the communication link, thereby reducing the security and reliability of wireless communication. With the increase of the amount of private information carried by communication networks, the interception technology has also been continuously improved [5–7], and the importance of information security has increased to an unprecedented level.

Focusing on these security threats and challenges, research on the anti-interference performance of wireless communications has received extensive attention in the past few years. In communication systems based on different application fields, there are many anti-interference schemes [8, 10–13], which can be divided into three categories. The first type comes from the spatial domain, such as beamforming technology for directional transmission [8, 7]. However, as the amount of high-rate multimedia wireless services carried by limited spectrum resources is rapidly increasing, the electromagnetic environment is changing in a more complex and changeable direction. These factors pose great challenges to the implementation of this method. The second type comes from the time domain, such as time-hopping spread spectrum communication [10, 11]. These methods achieve the purpose of anti-interference by expanding the signal frequency band, but at the cost of available resources in the time domain. The third type comes from the frequency domain, such as frequency hopping (FH) technology [12, 13]. Whereas its performance is limited by the fixed parameters (e.g., the frequency slot number, the frequency gap and bandwidth, etc.), it is difficult to resist the advanced dynamic interference signals.

Obviously, faced with the situation that the shortage of spectrum resources is becoming more and more serious, the aforementioned schemes cannot adapt to the highly complex and dynamic electromagnetic environment. In addition, a variety of new interference technologies have emerged, and the ability to interfere with devices has continued to increase. Therefore, in order to effectively resist the influence of interfering devices on the wireless communication system, a new method called cognitive frequency hopping (CHF) [16, 15, 14] has been recently proposed. This method can quickly determine the FH frequency slots affected by interference, and dynamically adjust parameters based on the evaluation results, thereby effectively improving the anti-interference performance of FH communications. However, because the existing literature only shows that the CFH system has anti-interference ability, the anti-interference ability of the system and the factors affecting the anti-interference performance are rarely analyzed. Therefore, we also propose a method for measuring the anti-interference ability of CFH communication systems.

In this paper, we propose a transmission scheme of the CFH communication system. Meanwhile, we derive the analytic expression of reliability tolerance by analyzing the effect of false alarm probability, missed detection probability, and communication link convergence delay. Compared with the literature that lacks CFH communication system performance analysis, our analysis provides a theoretical basis for studying complex and practical anti-interference communication systems. In addition, according to our theoretical derivation, the reliability ca-
pability of CFH systems can be optimized by adjusting relevant parameters to meet different requests.

The remainder of this paper is organized as follows. The infrastructure model is shown in Section II. Then Section III analyzes the anti-interference performance of CFH communication system. After that, the simulation results are provided in Section IV. Finally, Section V draws the conclusion of this paper.

2 Infrastructure Modeling

Based on the available FH communication system models and cognitive radio communication system models [15, 17, 16, 14], we construct a typical CFH communication system model in Fig. 1.

![System model](image)

Fig. 1. System model.

In our constructed communication scenario, multiple communication users $U_p (p = 1, 2, \cdots, N)$ and different types of interference devices $J_q (q = 1, 2, \cdots, M)$ coexist, which is consistent with the actual communication environment. Without desired and legitimate transmission between authorized users, the useless interference signal may be transmitted randomly by jammer devices, which will seriously affect the communication quality, and even cause information loss, communication link interruption and other negative influences. In order to effectively ensure the security of information transmission and enhance the anti-interference ability of the CFH communication system, a transmission scheme is shown in Fig. 2.

As shown in Fig. 2, the proposed CFH communication system has the communication bandwidth of $W$, which can be equally divided into $n$ frequency slots. Consequently, a set of frequency slots can be obtained $F = \{f_a | a = 1, 2, \cdots, n\}$, and the bandwidth of each frequency slot is $B = W/n$. The interference power, frequency hopping signal power, and noise power within each frequency slot are
Fig. 2. Transmission strategy.

\[ J_a, \ S_a, \text{ and } N_a, \text{ respectively. In the CFH system, before the communication user transmits the signal, the occupation of each frequency slot in the wireless channel can be analyzed through the spectrum sensing technology based on the high-order cumulant}\[18]. \text{ Therefore, the channel model can be expressed as}

\[
\begin{align*}
\{ H_0 : y(t) &= n(t), \\
H_1 : y(t) &= x(t) + n(t),
\}
\tag{1}
\]

where state \( H_0 \) and state \( H_1 \) represent the frequency slot is only occupied by the background noise and the presence of the signals transmitted by other users or devices, respectively. At time \( t \), the signals obtained by the sensing system, the transmitted signals of other users on the target frequency slot, and the background noise are indicated by \( y(t), \ x(t), \text{ and } n(t). \) In addition, \( n(t) \) is a complex Gaussian random variable which is characterized by the mean of zero and the variance of \( \alpha_n^2. \) It should be pointed out that no matter which sensing method is used, it is impossible to achieve completely accurate spectrum detection. Without loss of generality, there are false alarm (the state without interference is judged to be interference) probability and missed detection (the state with interference is judged as no interference) probability in CFH system’s perception of the interference signal in each frequency slot, which are represented by \( P_f \) and \( P_{md} \) respectively.

\[ \text{According to the high-order cumulant spectrum sensing method}[18], \text{ the corresponding decision metric value } T \text{ is compared with the set decision threshold } \gamma, \text{ and the detection probability } P_d \text{ and the false alarm probability } P_f \text{ can be expressed as}

\[
\begin{align*}
P_d &= \Pr(T > \gamma | H_1), \\
P_f &= \Pr(T > \gamma | H_0).
\end{align*}
\tag{2}
\]

Based on the results of spectrum sensing, it can be determined that \( k (0 \leq k \leq n) \) frequency slots in the spectrum set \( F \) are occupied by interference signals. The
CFH system removes the $k$ interfered frequency slots from the set $F$, and the set of available frequency slots $F_A$, which are not occupied by the interference signal, can be determined. To avoid the interference frequency slots and obtain a certain reliability tolerance, the transmitter makes the carrier of the modulated signal occupy the available frequency $F_A$ by up-conversion. The transmitter $U_1$ can select the frequency slots from $F_A$ to transmit the information by using FH method in each time slot $T = \{t_b \mid b = 1, 2, \ldots, m\}$.

### 3 Anti-interference performance analysis

According to the CFH communication system constructed in Section 2, its working cycle can be divided into four stages: spectrum sensing stage, time synchronization stage, cognitive decision-making stage, and communication stage. Once the FH communication user enters the system, this user begins to continuously perceive the spectrum environment to obtain the available frequency slots set $F_A$. Subsequently, all the FH communication users in this system adjust the local time of day (TOD) to complete the time synchronization. Combined with the existing method of generating a FH transmission strategy\[12–15\], we can clearly find that these strategy are controlled by the group of FH sequence families. Therefore, based on the idea of block cryptography, a group of basic FH sequences can be generated by controlling the network identification key of both communicating parties

$$S^0 = \{s_{a,b}^0, a = 1, 2, \ldots, n \text{ and } b = 1, 2, \ldots, m\}. \quad (3)$$

It is emphasized that the network identification key of the constructed system will be periodically updated from other systems in the wireless communication network, and CFH users can only join the cognitive system with the key of this system. Then, using the Round Function to $S^0$ for $y (y \leq n)$ consecutive iterations as

$$\begin{align*}
    s_{a,b}^y &= s_{a,b-1}^{y-1} \oplus \text{key}, \\
    s_{a,b-1}^y &= \text{Sbox}_j(s_{a,b}^y) \oplus s_{a,b}^{y-1}.
\end{align*} \quad (4)$$

Therefore, at time $b$, we can obtain the sequence families $\{S^0, S^1, \ldots, S^y\}$. Consequently, by using the remapping method, the sequence family that generates the control user transmission information can be presented in the form of an algorithm

$$\{X_{a,b}\} = g(TOD_b, key, f_a, \cdots). \quad (5)$$

When the communication user in the CFH system receives the feedback of the spectrum sensing result, they can identify the frequency slots corresponding to the FH sequence and transmit information independently from the available frequency slots $F_A$.

In addition, since the CFH communication system is obtained by adding cognitive module on the basis of conventional FH communication system, the reliability tolerance of CFH communication system can be expressed as

$$C = C_0 + C_1, \quad (6)$$
where $C_1$ is additional anti-interference gain by introducing a cognitive module.

### 3.1 Bit Error Rate (BER) of Conventional FH System

Encoding, modulation, and demodulation methods all affect the BER performance of FH communication systems. Since the transmitter in the conventional FH communication system does not consider whether there is an interference signal in the transmission environment, it directly selects the frequency slots from the set of frequency slots to send information.

Obviously, when the conventional FH communication system has $k$ frequency slots occupied by interference signals, the probability that the information transmits in the interfered frequency slots is

$$\Pr_{\text{int1}} = \frac{k}{n}. \quad (7)$$

Therefore, the BER of FH communication system with interference signals is

$$P_{e1} = \frac{n-k}{n} F\left(\frac{S}{N}\right) + \frac{k}{n} F\left(\frac{S}{N} + J\right), \ (0 \leq k \leq n) \quad (8)$$

where $F(\cdot)$ indicates a BER function of the FH communication system when transmitting information in one frequency slot.

### 3.2 BER of CFH System

**Without interference signal** If there is no interference signal in the CFH communication system, the BER of the system is the same as that of the conventional FH system, which can be given as

$$Q_{un} = F\left(\frac{S}{N}\right). \quad (9)$$

**Interference signal exists without missed detection** When the frequency slots occupied by the interference information is completely perceived, that is, the CFH communication system has no missed detection. The CFH communication system can select the frequency slots without interference for transmitting the information through the spectrum sensing results, effectively avoiding all interference frequency slots. Therefore, the probability of interference detection is

$$\Pr_{\text{no-miss}} = (1 - P_{md})^k. \quad (10)$$

Consequently, under the condition that the CFH communication system has interference and no missing detection, the BER of this system is

$$Q_1 = (1 - P_{md})^k F\left(\frac{S}{N}\right). \quad (11)$$
Interference signal exists with missed detection However, when the frequency slots occupied by the interference signals are not fully perceived, there is a miss detection in CFH communication system. Although the CFH communication system can select the frequency slots without interference based on the results of spectrum sensing, the frequency slots used to transmit information may still occupy the interfered frequency slots due to incorrect sensing results. Therefore, the influence of interference signals on the transmission process is inevitable. Accordingly, for CFH systems, the number of frequency slots occupied by interference signal is

\[ k^* = k - i + j, \tag{12} \]

where \(i (0 < i < n)\) indicates the number of miss detection of the frequency slots occupied by the interference signal, \(j (0 < j < n)\) represents the number of false alarms, and the number of \(i\) and \(j\) are independent of each other. Furthermore, the number of frequency slots without interference signal is

\[ n - k^* = n - k + i - j. \tag{13} \]

Therefore, when there is a missed detection and false alarm in the spectrum sensing results of the interference signals, the probability of the interference detection is

\[
\Pr_{ij} = C^i_k P_{md} (1 - P_{md})^{k-i} C^j_{n-k} P_f (1 - P_f)^{n-k-j}, \tag{14}
\]

where

\[
C^i_k = \frac{k(k-1) \cdots (k-i+2)(k-i+1)}{i(i-1)(i-2) \cdots 2 \cdot 1}. \tag{15}
\]

Under this condition, the probability of transmitting information in the interfered frequency slots is

\[
\Pr_{int2} = \sum_{i=1}^{k} \sum_{j=0}^{n-k} (P_{ij} \cdot \frac{i}{n-k^*}). \tag{16}
\]

In contrast, the probability of transmitting information in the frequency slots without interference is

\[
\Pr_{no-int2} = \sum_{i=1}^{k} \sum_{j=0}^{n-k} (P_{ij} \cdot \frac{n-k^*-i}{n-k^*}). \tag{17}
\]

Combined with (9), (16) and (17), the BER of the CFH communication system with the missed detection and interference signals is

\[
Q_2 = \sum_{i=1}^{k} \sum_{j=0}^{n-k} P_{ij} \left( \frac{i}{n-k^*} \cdot F\left( \frac{S}{N} \right) + \frac{n-k^*-i}{n-k^*} \cdot F\left( \frac{S}{N+J} \right) \right). \tag{18}
\]
Considering Section 3.2 and Section 3.2, the BER of CFH communication system with the interference signals can be concluded as

\[
Q = Q_1 + Q_2 \\
= \sum_{i=1}^{k} \sum_{j=0}^{n-k} P_{ij} \left( \frac{S}{N+J} \right)^{i-j} \cdot F(\frac{S}{N}) + \frac{i}{n-k+j} \cdot F(\frac{S}{N}) \\
+ F(\frac{S}{N}) \cdot (1 - P_{md})^k \\
= \sum_{i=1}^{k} \sum_{j=0}^{n-k} P_{ij} \left( \frac{(n-k-j)F(\frac{S}{N}+J) + iF(\frac{S}{N})}{n-k+i-j} \right) \\
+(1 - P_{md})^k F(\frac{S}{N}).
\]

Consequently, considering the two factors of false alarm and missed detection, the BER of the CFH communication system is

\[
P_2 = \begin{cases} 
F\left(\frac{S}{N}\right) & k = 0, \\
(1 - P_{md})^k F\left(\frac{S}{N}\right) + \sum_{i=1}^{k} \sum_{j=0}^{n-k} P_{ij} \left( iF\left(\frac{S}{N}\right) + (n-k-j)F\left(\frac{S}{N}+J\right) \right) & k \neq 0.
\end{cases}
\]

**Communication links convergence delay** The CFH communication system has no cognitive ability before the communication link converges. At this time, the BER of the CFH communication system is equal to the BER of the FH communication system. Once the communication links converge, the cognitive unit of the communication system begin to sense, analyze, and make decisions to generate frequency sets and communication scheme that can be used by the FH system. In addition, it is necessary to make this information pass through a highly reliable channel in time to achieve the synchronization of the FH frequency and the adjustment of the communication scheme at both ends of the transmitter and the receiver. Accordingly, the cognitive capability can be achieved in CFH communication systems. It is assumed that \(T_L\) is the period of interference change, and \(T_s\) \((T_L > T_s)\) is the link convergence time. It should be noted that \(T_s\) includes interference sensing time, the time required for transmitting sense information and adjusting the FH transmission scheme. Therefore, considering the effect of link convergence time, the BER of the CFH communication system is

\[
P_e = \frac{T_s}{T_L} P_1 + \frac{T_L - T_s}{T_L} P_2.
\]

Therefore, cognitive anti-interference gain can be expressed as

\[
C_1 = F^{-1}(P_e) - F^{-1}(P_1),
\]
where the system function $F^{-1}(\cdot)$ is the inverse of the BER function $F(\cdot)$. And the reliability tolerance of CFH communication system can be rewritten as

$$C = C_0 + F^{-1}(P_e) - F^{-1}(P_I).$$  \hspace{1cm} (23)

4 Numerical Results

This section presents the simulation results to evaluate the reliability tolerance of the CFH communication system. In our simulation, the frequency range used by the CFH communication system tested is from 2.65 GHz to 2.95 GHz. In addition, the occupied bandwidth $W$ is 300 MHz and the frequency hopping slots $n$ is 60. That is, the bandwidth occupied by each frequency slot is $B = 5$ MHz. Additionally, the signal-to-noise ratio ranges from 0 to 12 dB. Here, the signal-to-noise ratio (S/N) is converted to a normalized signal-to-noise ratio ($E_b/N_0$). Moreover, the false alarm probability $P_f$ is assumed to be 0.1. The number of frequency slots occupied by the interference signals is $k = 30$.

![Fig. 3. BER performance of three different types of FH communication systems with $P_{md} = 0.1$ and $I = -28$ dBw.](image)

When the probability of missed detection $P_{md}$ is 0.1 and the interference power $I$ is $-28$ dBw, Fig. 3 intuitively reflects the change of the BER performance of the CFH system, the conventional FH system without interference and the FH system with interference as the signal-to-noise ratio (SNR). It can be seen from the Fig. 3 that the simulation results of BER performance can be consistent with the theoretical results, which proves the rationality of the proposed reliability tolerance measurement method. Moreover, as the normalized...
signal-to-noise ratio increase, the BER performance decreases gradually. This shows that with a certain background noise power, as the transmission power increases, the possibility for an authorized user to receive complete information increases rapidly. We can also find that under the same signal-to-noise ratio, the performance curve of the CFH system is between the simulation curves of the other two systems, which proves that this CFH system can effectively resist the threat of interference signals.

![Fig. 4. BER performance of three different types of FH communication systems with $P_{md} = 0.01$ and $I = -32$ dBw.](image)

Different from the Fig. 3, Fig. 4 reflects the BER performance curves when the probability of missed detection $P_{md}$ is $0.01$ and the interference power $I$ is $-32$ dBw. By comparing the simulation results of the interference signal on the BER performance of the FH system in Fig. 3 and Fig. 4, it can be found that the reliability performance of the system decreases gradually as the increase of interference power. However, since the CFH communication system can acquire the channel conditions and update the frequency hopping sequence in time through the spectrum sensing method, the CFH system can effectively guarantee its reliability performance. At the same time, it also reflects that the reliability performance of the cognitive system is enhanced with the increase of the power of the interference signal.

On the basis of analyzing the influence of false alarm probability and missed detection probability on interference tolerance, Fig. 5 adds another consideration to the key factor of communication links convergence delay, which is a factor that cannot be ignored in the actual communication environment. It can be seen that the BER performance of the CFH system has decreased. This is due to the fact
that the cognitive system’s perception of interference signals, the transmission of sensing results, and the adjustment of FH sequences, which all cause time delays. Note that, it is still possible to ensure that the anti-interference gain of the CFH communication system is stable at 1.3 dB or more when the BER is $10^{-3}$.

5 Conclusion

Based on the lack of analysis for the reliability performance of the CFH communication system in the existing literature, this paper proposes an analysis method that can effectively measure the reliability tolerance of the system. We first have proposed a transmission scheme of the CFH, and then by analyzing the influences of false alarm probability, missed detection probability, and communication link convergence delay, an analytical expression of reliability tolerance has been derived. Simulation results reflect that this method can effectively measure the reliability of CFH communication systems, and provide a useful reference for studying complex and practical FH communication systems. In future work, based on the above theoretical derivation, the reliability of CFH systems can be optimized by adjusting relevant parameters to meet different requests.

Acknowledgment

Research and development leading to these results are funded by the National Natural Science Foundation for Distinguished Young Scholar 61825104, by the
Key project of National Natural Science Foundation of China 61631015, by the National Natural Science Foundation of China 61501354, 61501356, 61901328, by University Innovation Platform Project 201921815KYPT009JC011, and by China Postdoctoral Science Foundation 2019M653558.
References

Dynamic Machine Learning Algorithm for AODV Routing Attacks Detection

Md Raqibull Hasan\(^1\), Yanxiao Zhao\(^2\), Guodong Wang\(^3\), Yu Luo\(^4\), and Lina Pu\(^5\)

\(^1\) Contract Professionals Incorporation, Waterford Twp, MI 48329.
\(^2\) ECE Department, Virginia Commonwealth University, Richmond, VA 23284.
\(^3\) CS Department, Massachusetts College of Liberal Arts, North Adams, MA 01247.
\(^4\) ECE Department, Mississippi State University, Mississippi State, MS 39762.
\(^5\) School of CSCE, University of Southern Mississippi, Hattiesburg, MS 39406.
Email: mdraqibull.hasan.sdsmt@gmail.com, yzhao7@vcu.edu, guodong.wang@mcla.edu, yu.luo@ece.msstate.edu, lina.pu@ua.edu

Abstract. Ad hoc On-Demand Distance Vector (AODV) routing protocol is vulnerable to some routing attacks including blackhole attack and flooding attack. Typically, these two types of routing attacks are linked with two major malicious behaviors: fake Route Replies (RREPs) and fake Route Request (RREQ) flooding. In this paper, we develop a novel dynamic machine learning approach to detect blackhole and flooding attacks in AODV. The proposed solution primarily determines three distinct features by analyzing Hello, RREQ, and RREP packets in the AODV routing protocol. Then, these features are used to develop a mathematical model for the dynamic learning algorithm. Afterward, we generate the training set of data and assign a threshold for our machine learning model using these features. This training set of data is only valid for N time slots, which is regarded as one iteration. In the following iterations, it will update the latest valid outcomes from the dynamic learning model and determine an updated threshold for the model, which significantly increases the detection accuracy. Extensive simulations have been conducted to evaluate the accuracy and the time overhead of three classifiers, e.g., support vector machine, k-nearest neighbor, and decision tree. The simulation results show that the proposed algorithm can achieve very high accuracy with minimum time overhead to detect malicious behavior in the AODV routing protocol.

Keywords: AODV routing protocol, Routing attack, Smart meter network, Dynamic machine learning

1 Introduction

Smart meter networks are the major components of the smart grid, which are composed of smart meters and Data Aggregation Points (DAPs) \([1]\). A smart meter is an electronic device that is installed at houses or commercial sites to record consumption of electric energy and communicates that information...
back to the utility for monitoring and billing. Each smart meter equipped with network radio can transmit meter reading periodically or on request by utilities. DAPs are responsible for communicating information or data between smart meters and the utility company.

Since a smart meter network is essentially a multi-hop network, this means some smart meters serve as relay nodes to deliver information to the DAP through multiple hops. Similar to a typical multi-hop network, multiple routes exist in a smart meter network. In other words, there is more than one path between a smart meter and its associated DAP. Therefore, a routing protocol is required to find the appropriate route to DAPs for each smart meter.

In recent years, Ad hoc On-Demand Distance Vector (AODV) routing protocol is recommended for smart meter networks because it meets both the requirements of on-demand and periodic operations. However, the AODV routing protocol is vulnerable to different types of routing attacks. One typical routing attack is the denial of service including blackhole attack and flooding attack. In a blackhole attack, attackers broadcast numerous fake Route Replies (RREPs) with a tampered highest sequence number and minimum hop count. If these fake replies are accepted by other smart meters, blackholes will be created in a smart meter network. Consequently, legitimate data cannot be sent to DAPs. In a flooding attack, attackers broadcast Route Requests (RREQ) packages throughout the network. The entire network keeps busy to forward those fake RREQ packets. Under this attack, the data transmission can be delayed, corrupted, or even blocked. Therefore, how to effectively conduct routing attack detection and detect malicious meters in a smart meter network remains a critical challenge.

There are two problems in the existing work for AODV routing attack detections. The first problem is that the majority of work usually focuses on specific and known routing attacks, e.g., blackhole and flooding detection. It is known that different types of attacks exhibit distinctly. Most approaches detect malicious behaviors by targeting specific anomalous behaviors by adding new control packets to routing protocols. These approaches are not able to detect other types of attacks if present. The second problem is that a constant threshold is commonly used for detection. If an attacker is able to estimate the threshold, it can easily break the intrusion detection system and utilize the routing information to access data packets [2].

In this paper, we introduce a novel dynamic machine learning approach. First, we analyze the complete AODV routing packets including Hello, RREQ, and RREP control packets. Based on this analysis, we obtain three distinct features in the AODV routing protocol. Afterward, these features are utilized to develop a mathematical model of dynamic learning algorithm. Concurrently, these features are also considered in our machine learning model to correlate with the testing data as well as to generate an initial training set of data and a threshold. Every training set of data is only valid for N time slots, which is considered as one iteration. During this N time slots, our dynamic learning model will create new data points, which will be updated in the training set of data and be utilized to determine a new threshold for the next iteration. Therefore, during every N slot
intervals, the training data set will be updated based on the current state of the smart meter network and create a new threshold to detect malicious behaviors. The most promising part is that our proposed machine learning model updates parameters according to the changes of malicious behaviors, which achieves a high detection accuracy. The main contributions of this research are summarized below:

First, we develop formulas by analyzing AODV routing packets including Hello, RREQ, and RREP control packets. Leveraged by our proposed formulas, we obtain three distinct features including average one-hop neighbor distance, the dynamic range of sequence number, and the minimum hop count to analyze entire AODV routing packets.

Second, we derive a mathematical model for dynamic learning algorithm using the three identified features. Initially, these features are placed in a three-dimensional vector space. Therefore, it determines the N number of vectors for N time slots where all of them are considered in the first iteration. Then, we calculate the mean vector from N time slots. Finally, we find out the variance between input sample data and mean vector. If this variance is lower or equal to the current threshold, the input sample data will be considered as normal traffic. This data will be further forwarded to the training model in order to estimate a new threshold for the next iteration. In contrast, if the variance of sample data exceeds the current threshold limit, it will be identified as malicious traffic.

Third, we develop an adaptive machine learning model, where the training set of data is updated and a new threshold is calculated after N time slots interval. In the following iterations, the training model will only update the recent valid inputs of the previous iteration from the dynamic learning model. During the N time slots, the dynamic learning model will calculate the variance of each incoming data and pull the threshold from the memory block simultaneously. Afterward, these two data sets along with input features are evaluated using three classifiers including Support Vector Machine (SVM), k-Nearest Neighbors (k-NN), and Decision Trees (DTs) algorithm.

The rest of paper is organized as follows. Section 2 introduces related work. Section 3 introduces feature identification from AODV routing. Section 4 designs the mathematical model for the proposed dynamic machine learning method. Section 5 shows simulations results and Section 6 draws conclusions.

2 Related Work

In this section, we summarize the state-of-the-art for routing attack detection in AODV under smart meter network, wireless sensor network, or mobile ad hoc network. The followings are some recent work to tackle routing attacks in the AODV routing protocol using machine learning approaches or modified AODV routing mechanisms.

A very recent work [3] conducted a comprehensive survey on various attacks in AODV routing under MANETs. It also introduced some prospective techniques for detecting and predicting routing attacks, which are data mining,
SVM, genetic algorithms (GA), and some other machine-learning approaches. Machine learning and data mining methods for cyber analytic were investigated in [4] for intrusion detection. This work also analyzed the complexity of machine learning algorithms and discussed challenges for cyber-attack defense. In [5], a machine learning-based intrusion detection system was developed to protect critical infrastructures. Among various supervised machine learning classification techniques [6], the k-NN classification algorithm was utilized in wireless sensor networks to separate the anomalous node based on abnormal behaviors [7]. The authors also analyzed the relevant parameter selection and error rate of the intrusion detection system for AODV routing. In addition, an enhanced SVM for packet classification was proposed in [8] to provide unsupervised learning with low false alarm capability.

In [9], the authors proposed a new architecture for intrusion detection in mobile ad hoc networks using the machine learning approach to maximize detection accuracy. In this work, rough set and SVMs have been used for data reduction and classification respectively. The rough set reduces the size of features to simplify the complexity of SVM. In the following work, [10], a novel supervised learning framework was proposed by using a generative adversarial network for improving the performance of the classifier. This framework was utilized to continuously generate other complementary labeled samples for adversarial training and assisting the classification. In addition, several empirical training strategies were proposed to improve the stabilization of the supervised learning framework.

In the cluster wireless sensor network, a beta distribution based dynamic trust management has been proposed in [11]. The proposed method dynamically calculated the reputation and adopted a dynamic threshold to resist the on-off attack, bad-mouth attack, selective forwarding attack, and a mixed attack. In [12–14], some of the recent works have been emphasized on detection and prevention algorithms either for blackhole or flooding attacks under AODV routing. Specifically, in [14], a secure and lightweight routing protocol was proposed to prevent blackhole attacks in constraint-oriented networks. The proposed protocol is a hybrid of medium access control and AODV protocols. In this work, every node was registered with the nearest gateway/cluster head module through the MAC addressing scheme, and only registered nodes were allowed to communicate.

In [15], a logical scheme was proposed to tackle some common cyber-attacks in the smart grid. The hierarchy of the proposed system consisted of three remote terminal units (RTUs), a substation, and a control center, which communicated in two-way data flow in real time scenario. All the critical information of the smart grid, like the voltage, frequency, and voltage angle was encrypted through MD5 hash algorithm and later decrypted at the substation and control center using a key authentication method. In [16], the authors introduced the background of Advanced Metering Infrastructure (AMI) and identified major security requirements in AMI. Specifically, this work illustrated the energy-theft behaviors in AMI using an attack tree-based threat model.
3 Features Identification of AODV routing

As we mentioned previously, RREQ flooding and fake RREPs from blackhole attack are the major problems in AODV. To detect and prevent these attacks, we present a dynamic learning method in this section. In particular, three distinct features are obtained by analyzing Hello Packet, RREQ, and RREP in the AODV routing protocol. These three features are average one-hop neighbor distance, the dynamic range of sequence number, and minimum hop count. We will introduce how to determine these features as follows.

3.1 Calculate One-hop Neighbor Distance

In the AODV routing protocol, each smart meter sends Hello Packets to all of its one-hop neighbors before initiating control packets (RREQ and RREP) to find out the appropriate routes.

During Hello packets communications, smart meters calculate the distance of its one-hop neighbors based on the power label of received Hello Packets. The following equation is used to calculate the one-hop neighbor distance \[ P = \frac{4\pi D}{0.12476} \times 10^{-12.5}. \] (1)

After that, each smart meter estimates the average one-hop neighbor distance, using the following equation:

\[ D_n = \frac{\sum_{i=1}^{n} D_i}{n}. \] (2)

The average one-hop neighbor distance is regarded as the first distinct feature because Hello Packet communication happens before starting the actual routing packets. Using this method, a malicious attacker is being deprived of this information.

3.2 Ranges of Sequence Numbers

In the AODV routing protocol, every routing search deals with two sequence numbers during control packet communications. Initially, a source node initiates the RREQ control packet and includes a sequence number for the desired destination. This sequence number is also named as RREQ sequence number. After receiving RREQ by a destination, the destination acknowledges the source by sending the RREP control packet. This RREP control packet also includes a sequence number. Note that, at this time, the second sequence number is updated and is completely different from the previous one. Therefore, there is a difference between two sequence numbers for every routing setup. In other words, a range of sequence numbers exists for each valid routing, which is predefined by destination. The dynamic range of sequence numbers is regarded as the second distinct feature, which is denoted by \( S_d \) in our machine learning model.
3.3 Calculate the Minimum Possible Hop Count

For a smart meter network, every smart meter knows its own and its destination location. Based on the locations, it can predict the distance between source and destination. Under blackhole attacks, each source meter receives multiple RREPs, where fake replies always contain the minimum hop count. To avoid flows of fake replies, each smart meter calculates the minimum possible hop count, \( H_{\text{min}} \), from the predicted distance, \( P_{\text{dis}} \) and the average one-hop neighbor distance, \( D_n \). The equation (3) is used to determine the minimum hop count.

\[
H_{\text{min}} = \frac{P_{\text{dis}}}{D_n}
\]  

The minimum possible hop count is regarded as the third distinct feature in our dynamic machine learning model. Using this method, we can stop blackhole attackers from pretending one-hop neighbor in its fake reply.

4 Mathematical Model of Dynamic Machine Learning Method

In this section, we design a mathematical model for the proposed dynamic machine learning method. Each smart meter collects three above-mentioned distinct features, which are considered in a three-dimensional vector space, \( x_i = (x_{i1}, x_{i2}, x_{i3}) \) for \( i^{th} \) time slot. Here, each time slot contains a certain Active Routing Time (ART) and a Hello packet communication with one-hop neighbors. For \( N \) time slots, we calculate the mean vector of \( x \) using the equation (4).

\[
\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i.
\]  

Next, we calculate the distance from input data sample \( x \), to the mean vector from equation (5)

\[
d(x) = ||x - \bar{x}||^2.
\]

If the distance is larger than the threshold \( T_h \) (\( d(x) > T_h \)), that means it is out of range from normal traffic, so it will be regarded as an attack. Here, the projection distance with its maximum value is extracted from the learning data set:

\[
T_h = d(x_I), \text{where } I = \arg \max_i d(x_i), x_i \in D
\]
By implementing this mathematical model in AODV routing protocol, we calculate the threshold $T_h$, and generate training data for our machine learning model. This threshold is only valid for $N$ time slots. After that, it will update the threshold based on current network scenarios. Therefore, each smart meter can operate in a dynamic range of threshold. Now, for every following routing, Hello Packet, RREQ and RREP routing information are analyzed in a dynamic learning model to determine $d(x)$. The calculated $d(x)$ and its corresponding three distinct features are considered as testing data for our machine learning model, where three classifiers are employed including SVM, k-NN and DTs Algorithm. The complete flow chart of our proposed model is depicted in Fig. 1.

5 Simulation Results for Dynamic Learning Method

In our simulation, we consider three distinct features to identify routing attacks in AODV routing protocol, e.g., average one-hop neighbor distance during Hello packet communication, dynamic range of sequence number for each routing, and the minimum hop count. We generate all possible combinations of malicious data and normal data. Using training and cross-validation data, we test three default classifiers (SVM, k-NN and DTs Algorithm) under Python 3.6 Skit-learn module. Since we are dealing with supervised machine learning, accuracy and time overhead are two main performance metrics to evaluate the classifier.

The accuracy of each classifier represents the percentage of detection with our predefined and future randomized data. As shown in Fig. 2, SVM fluctuates its accuracy with the increase of independent variables $C$ as regularization parameter. In contrast, k-NN stabilizes its accuracy around 87%. The time overhead is depicted in Fig. 3, where we find that k-NN produces almost double delay (above
Fig. 2. Accuracy comparison for three classifiers.

Fig. 3. Time overhead comparison for three classifiers.
135ms) compared to SVM (around 70ms). To address the trade-off between time overhead and accuracy, we introduce a DTs algorithm. As shown in Fig. 2, the DTs algorithm obtains the maximum accuracy (100%) in a 12-minimum splits. In addition, the time overhead as shown in Fig. 3 achieves almost half (around 35ms) of the k-NN. Therefore, the DTs algorithm shows the maximum fitness to identify the routing attacks in AODV routing protocol.

6 Conclusions

In this paper, we proposed a novel dynamic machine learning approach to detect malicious behavior in the AODV routing protocol. To the best of our knowledge, the proposed solution is the first work that combines both the dynamic learning algorithm and the machine learning approach. We mainly focused on two major malicious behavior detection, e.g., RREQ flooding attack and blackhole attack. To detect those malicious behaviors, we developed a dynamic learning algorithm along with the machine learning model. The dynamic machine learning approach is implemented in Python 3.6 using the Scikit-Learn module. Three classifiers were used to evaluate the accuracy and time overhead of the proposed solution. Among those three classifiers, the DTs algorithm achieves 100% accuracy with minimum overhead (around 35ms) to detect the malicious behaviors in the AODV routing protocol.

References

Association rule mining of network security monitoring data based on time series

XUE Fang¹, LIU Rong²

{aaa21152@163.com¹, wz26677@163.com²}

¹Information Technology Center, Jimei University, XiaMen 361021, China;
²School of Information Engineering, Changsha Medical University, Changsha 410219, China

Abstract: The traditional network security monitoring number association rule mining technology has low mining accuracy, so a time series based network security monitoring data association rule mining technology is designed. The preprocessing of time series to construct the corresponding time series frequency set, using SWFI - tree structure data storage model is set up, get after filtering and reorder the transaction data set, data sets of will be clean and remove invalid data and the remaining data formatting, finally USES the particle swarm optimization (pso) algorithm with limited data flow, recursive calculation of particle movement, build sparse list, complete monitoring data mining of association rules. The designed mining technology was used in the experiment with the traditional technology, and the experimental results showed that the designed mining technology was 23.22% more accurate than the traditional technology.

Key words: Time series; Network security monitoring; Association rule mining; Data cleaning;
1 Introduction

With the rapid development of computer technology and the rapid popularization of the Internet, the role of computers in daily life and work is becoming more and more important. At the same time, the threats to network security are also increasing. Network viruses, Dos attacks, etc. are commonly used. Anti virus software, firewall, intrusion detection system and other network security defense technologies achieve passive security protection, which can not meet the requirements of network security. Therefore, real-time security monitoring of network security is needed [1-2], and association rules mining between security indicators Technology can implement a quantitative evaluation of the network status and monitor the network status in real time. It is a novel and superior network security technology. However, the traditional network security monitoring data association rule mining technology has low mining accuracy. Therefore, a time series-based network security monitoring data association rule mining technology is designed.

2 Association Rule Mining

Association rule mining is an important research direction of data mining. It excavates some interesting rules from a large amount of data. From these association rules, you can understand the association relationship between events and guide people Some actual meaningful behavior. This article is based on time series to mine the association rules of network security monitoring data. The relevant mining flowchart is shown below:

![Time series association rule mining flowchart](image)

The time series algorithm belongs to depth-first search. Recursive search for short
patterns instead of long frequent pattern mining requires only two database scans. The first scan of the database produces frequent itemsets, and the second scan of the database establishes a global time series. The analysis from the mining process is divided into two steps \cite{1-4}: First, compress the frequent sets in the database, retain the associated information between the data, and store them in the time series set; Second, because the time series contains all frequency sets, data mining only needs to work on the time series set, and separate it into the condition pattern database, respectively establish the condition time series set, and then carry out recursive mining.

2.1 Time series preprocessing

To apply time series to the mining of association rules between security indicators in network communication, we must first construct a corresponding time series set. However, since any monitoring data packet contains a large amount of network information, even if the time series set is fast and directly calculated, processing is also difficult, so it is necessary to preprocess the monitoring data \cite{5-6}. Information entropy is a better method of preprocessing. The detection data is packaged and the information is selected as a set of random events, as follows:

$$\ X = n_i \ (i = 1, 2, ..., N) \quad (1)$$

The above formula can represent the number of times the geology in the information packet occurs at different events, and the information entropy $H(X)$ can be obtained as:

$$H(X) = - \sum_{i=1}^{N} n_i \left( \frac{n_i}{S} \right) \log_2 \left( \frac{n_i}{S} \right) \quad (2)$$

In the above formula, $S$ represents the total number of times an address appears, which can be expressed as:

$$S = \sum_{i=1}^{N} n_i \quad (3)$$

$N$ is the total number of random events, $n_i$ represents a single event in a collection of events. It is a description of network uncertainty, which can intuitively and accurately reflect the distribution of system status. By preprocessing through information entropy, network security indicators can be expressed by the change in entropy of IP source and destination.
addresses and port numbers. In network communication, the information entropy is smaller where the data is more concentrated, and vice versa. The preprocessed data information is stored in the database. When index mining is performed, the database is scanned once to obtain the set of frequency sets and the corresponding minimum support. The frequency set is arranged in descending order to obtain set L. The entire process is shown in the following figure:

![Flow of frequency set generation](image)

Fig. 2. Flow of frequency set generation

Where P is the information entropy of the packet information and P is the relevant network information. The FP-tree frequency set insertion algorithm is called. If the time series N have the same item name, the count of N is increased by 1, otherwise a new time node N is created. And set the count to 1, while linking to the parent node T, the traversal node is linked to the node with the same time. If P is not the empty set, recursively call the generative insertion algorithm. This completes the preprocessing of the time series.

### 2.2 Building a data storage model

After the time series is preprocessed, the time series stream segment in the basic serial port is converted into a transaction set. In order to run the association rule mining algorithm from this transaction set, a data storage model needs to be established. Based on the characteristics of the association rules to be mined and the characteristics of the data set [7], this
paper uses the SWFI-tree model, which can maintain two trees at the same time, divided into global SWFI-tree and local SWFI-tree. The structure is shown below:

![SWFI-tree model structure](image)

*Fig.3* SWFI-tree model structure

With the established SWFI-tree model, it is not necessary to scan the data set multiple times. It is assumed that the original data set is shown in the following table:

<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Element items in a transaction</th>
<th>Minimum support</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>h,r,s,f,w</td>
<td>3</td>
</tr>
<tr>
<td>002</td>
<td>q,e,r,g,c,s,f,v,s</td>
<td>5</td>
</tr>
<tr>
<td>003</td>
<td>z</td>
<td>4</td>
</tr>
<tr>
<td>004</td>
<td>r,d,f,g,h</td>
<td>7</td>
</tr>
<tr>
<td>005</td>
<td>y,d,e,r,t,h</td>
<td>6</td>
</tr>
<tr>
<td>006</td>
<td>y,z,x,e,q,s,t,m</td>
<td>5</td>
</tr>
</tbody>
</table>

Using the SWFI-tree model only needs to scan the data set twice. The first scan counts the support of each item. The second scan sorts each transaction by the support count of the item, removing infrequent items, such as the table 2 shows:

Table 2 Scanned transaction data set
<table>
<thead>
<tr>
<th>Transaction ID</th>
<th>Element items in a transaction</th>
<th>Filtering and reordering of transactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>h,r,s,f,w</td>
<td>h,s,f,w</td>
</tr>
<tr>
<td>002</td>
<td>q,e,r,g,c,s,f,v,s</td>
<td>r,v,s,q</td>
</tr>
<tr>
<td>003</td>
<td>z</td>
<td>z</td>
</tr>
<tr>
<td>004</td>
<td>r,d,f,g,h</td>
<td>r,h,f</td>
</tr>
<tr>
<td>005</td>
<td>y,d,e,r,t,h</td>
<td>y,r,t</td>
</tr>
<tr>
<td>006</td>
<td>y,z,x,e,q,s,t,m</td>
<td>z,x,y,s,t</td>
</tr>
</tbody>
</table>

Using the filtered and reordered transaction data sets obtained after scanning, a frequent pattern tree is generated to mine frequent patterns. Complete the establishment of the data storage model.

**2.3 Cleaning monitoring data**

The data used in this paper are derived from the real data collected during the network security monitoring and maintenance process. The data comes from the Internet central control device, and all the collected fault information and network operation status information are transmitted to the ground in a timely manner. The data flow is shown in the following figure:
Under normal operating conditions, WTD sends real-time status data and screen display fault information collected by the network information collection and processing module every 60 seconds to the ground server. This real-time data does not require a ground response. When the main fault data is sent, the sent main fault record information is immediately sent to the ground, waiting for the response information from the ground, and the waiting time is generally set to seconds [8-9]. If it times out, the ground will be considered to have after receiving this major fault message, this major fault message will be resent. When a response is received, the data is considered to be sent successfully and the next data is sent.

Because network security generates a large amount of data during the monitoring process, it is difficult for rabbits to have some incomplete and dirty data. In addition, duplicate data, erroneous data, and garbled data are also important factors that reduce data quality. Therefore, in order to obtain more accurate mining results, this requires cleaning the original data, which can improve the accuracy of data mining. And effectiveness, on the other hand can also save the time required for mining. Therefore, two aspects have to be done in data cleaning: one is to remove invalid data; the other is data formatting. For the fault data set, mainly clean the fields of number type, null value and measurement type. During the cleaning
process, export the table data to a CSV format file in Oracle, upload the data file to HDFS using Hadoop commands, create a data table in Hive, and use HQL statements to clean the data. This completes the cleaning of the data.

2.4 Complete data association rule mining

After completing the steps of time series preprocessing, building a data storage model, and cleaning monitoring data, the time series data is symbolized and transformed into a transaction set. Because time series are fluid, particle swarms are used to limit the data flow. The particle swarm approach will have new data coming in and old data slipping out. Therefore, when mining association rules, it is necessary to be able to support incremental mining of increased data, and also to remove old data. The data set is dynamically changed, and the mining association rules are also dynamically changed. It is an association rule for multivariate time series data, and the amount of data is relatively large, so it is necessary to perform recursive calculations on the motion of ions:

\[
\overrightarrow{V}_{i}^{k+1} = \overrightarrow{V}_{i}^{k} + c_{i} \cdot r_{i} (\overrightarrow{P}_{i}^{k} - \overrightarrow{X}_{i}^{k}) \quad (4)
\]

In the above formula, the particle number is \( i = 1, 2, \ldots, m \), \( k \) is the number of iterations, \( \overrightarrow{V}_{i}^{k+1} \) is the particle speed, \( \overrightarrow{V}_{i}^{k} \) is the speed of the particle immediately before, \( c_{i} \) and \( r_{i} \) are learning factors, and \( \overrightarrow{P}_{i}^{k} \) and \( \overrightarrow{X}_{i}^{k} \) are motion vectors, among them.

\[
\overrightarrow{X}_{i}^{k+1} = \overrightarrow{X}_{i}^{k} + \overrightarrow{V}_{i}^{k+1} \quad (5)
\]

The above two formulas will work together on the particle's motion position in the next step. Taking two-dimensional space as an example, the process from the initial position to the new position is shown in the following figure:
Use particle traversal to obtain the support of the item set represented by the particle. Combining the characteristics of the data source and the basic algorithm, the linked list data structure is applied to the dynamic storage structure. Create a linked list header for each item in the full database, scan the set of items in order, and add the items in each item set to the end of the corresponding linked list \([10-13]\). For example, if the nth item set in the database is \((50, 180, 17)\), then add the nth item set to the end of the linked list 50, the end of the linked list 108, and so on, and finally create 200 linked lists. That is, a sparse linked list, the process is as follows

**Fig. 5 Particle movement process diagram**
Take the nth item set (50, 180, 17) and 200 linked list headers in the linked list as an example, and start the search from the linked list 50. Suppose that the 64th item set found contains the 50 item. Similarly, linked lists 108 and 17 are Corresponds to 88,24, that is, the items before the 88th item set do not contain the item corresponding to the particle, and the 88th item set is directly found. If it contains (50, 108, 17), the particle support is increased by 1, otherwise the search continues, and the linked list (50, 108, 17) finds the next data, assuming corresponding 121, 90, 65, respectively, and directly find the 121st item set. If it contains (50, 108, 17), the particle support is increased by 1, otherwise, the search continues. The linked list (50, 108, 17) finds the next data respectively, assuming that it corresponds to 121, 184, 121, and directly searches for the 184th Itemsets. If it contains (50, 108, 17), the particle support is increased by 1. This completes the mining of association rules for network security monitoring data based on time series.

3 Experiment

In order to verify the effectiveness of the association rule mining technology of network security monitoring data designed in this paper, it is necessary to design experiments to compare with the traditional association rule mining technology.

3.1 Experimental design

The experimental network structure is shown in the figure:
In the figure above, there are 5 nodes A, B, C, D, and E. Node A represents the intruder from the external network. The four nodes on the internal network inside the firewall are mainly responsible for providing network services, database services, server protection management, and storage. Important documents. The intranet access to the external network or communication between the intranet does not need to go through the firewall, but when the extranet nodes access the intranet nodes, they need to go through the firewall and can only access the network server. The data mining goal of this article is to analyze the network security monitoring data, and to attack the nodes of the internal network through some actual vulnerabilities in the internal network. The vulnerability settings of the internal network nodes are shown in Table 3:

<table>
<thead>
<tr>
<th>node</th>
<th>Hole number</th>
<th>Server or software</th>
<th>Elevated privileges</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network server B</td>
<td>Mc10-062</td>
<td>WWW</td>
<td>U-A</td>
</tr>
<tr>
<td>Network server B</td>
<td>Mc11-002</td>
<td>WWW</td>
<td>U-A</td>
</tr>
<tr>
<td>Database server C</td>
<td>Mc11-012</td>
<td>Oracle</td>
<td>O-A</td>
</tr>
</tbody>
</table>
Under the above conditions, all the data in the network security monitoring database are known conditions, and the mining slope $K$ value is estimated through experiments. The closer the $K$ value is to 1.85, the more accurate the association rule mining is.

### 3.2 Analysis of results

Under the above experimental conditions, the method of this paper and the traditional method are used for data mining experiments, and the $K$ value obtained by the two methods is compared with the set value, and the accuracy of data mining is obtained. The experimental results are shown in the following table:

<table>
<thead>
<tr>
<th>Number of training sets</th>
<th>Number of test sets</th>
<th>$K$ for traditional method</th>
<th>$K$ for Method of this paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>250</td>
<td>2.81</td>
<td>2.13</td>
</tr>
<tr>
<td>100</td>
<td>200</td>
<td>2.47</td>
<td>2.01</td>
</tr>
<tr>
<td>150</td>
<td>150</td>
<td>2.15</td>
<td>1.89</td>
</tr>
<tr>
<td>200</td>
<td>100</td>
<td>1.69</td>
<td>1.84</td>
</tr>
<tr>
<td>250</td>
<td>50</td>
<td>1.20</td>
<td>1.79</td>
</tr>
</tbody>
</table>

Through experiments on 300 data sets, in 5 experiments, the number of different training and test sets was set, the standard $K$ value was 1.85, and through calculation, the average accuracy of traditional technology was 67.84%. It is 91.06%, which shows that the technique in this paper has certain effectiveness in improving the accuracy of mining association rules.

### 4 Concluding remarks

The traditional network security monitoring data association rule mining technology has low mining accuracy. Therefore, a time series-based network security monitoring data association rule mining technology is designed. Preprocess the time series, use the SWFI-tree structure to build a data storage model, clean the obtained data set, and finally use particle
swarm algorithm to recursively calculate the particle motion to complete the mining of the association rules of the monitoring data. The designed mining technology is used to conduct experiments with traditional technology. The experimental results show that the designed technology mining accuracy is 23.22% higher than the traditional technology.

5 Fund projects


References


Design of encryption and decryption system of heterogeneous database based on Data Mining

BAI Yan-hui¹, MA Xi-jie²
{gaozhenyu101@2980.com¹, tiantai0200@tom.com²}

(1. School of Mathematics and Computer Science, Chifeng University, Chifeng 024000, China;
2. Nanjing Institute of Technology, Nanjing 211167 china)

Abstract: In view of the problem that the error rate is too high in the use of the original heterogeneous database encryption and decryption system, the original encryption and decryption system is optimized by data mining calculation, and the heterogeneous database encryption and decryption system based on data mining is designed. Through the data acquisition equipment, data storage equipment and system data transmission equipment to complete the hardware design of the system; according to the database user rights to design the database encryption dictionary, using data mining technology to complete the data preprocessing, using the form of homomorphic encryption to complete the encryption and decryption process of the database. At this point, the implementation of heterogeneous database encryption and decryption system design based on data mining. In the system test link comparing with the traditional database encryption and decryption system, by comparing the bit error rate, it is verified that the designed system can effectively reduce the bit error rate and improve the database security.

Key words: Data mining; Heterogeneous database; Data encryption; Data decryption;

1 Introduction

Database system is the most extensive and important carrier of data in the network era. Data security derives from database system security and becomes one of the basic supporting technologies in the future based on cloud computing. At present, all kinds of data stealing supported by groups or countries are more and more frequent. At the same time, data stealing behavior from internal personnel is also one of the main threats. Such unauthorized access to user access often leads to data information disclosure, user identity disclosure and authorized card number loss, which may cause a lot of economic losses, but may cause sensitive information or countries serious consequences of disclosure of confidential information [1-2].
The security of database system is related to whether the data involved in the database system is acquired by unauthorized users. Although the process of database intrusion is quite different, the consequences are the same, all of which are to obtain the information of database data in the unauthorized state, which leads to serious consequences. At the same time, it also results from the abuse of the rights granted to users, the rights granted by misoperation, and the malicious intrusion of not granting the right to steal, tamper and delete data [3]. In order to ensure the information security of database, heterogeneous database is designed on the basis of traditional database. Heterogeneous database is a collection of related databases, which can realize data sharing and transparent access. While improving the autonomy of database system and realizing data sharing, each database system still has its own application characteristics, integrity control and security control. Through the process of encrypting and decrypting the internal data of the database, the information security of the database is guaranteed. In the process of using the original heterogeneous database encryption and decryption system, the problem of high error rate of data information often occurs. Therefore, this paper uses data mining technology to optimize the original system.

2 Hardware design of heterogeneous database encryption and decryption system based on Data Mining

In view of the problems in the use of the original system, the hardware design of the system only needs to complete the data acquisition and storage process. In order to improve the data processing ability of database encryption and decryption system, the hardware part of this paper is introduced into the original system hardware. In order to ensure the normal operation of the system designed in the application paper, the new hardware architecture is set as follows.
According to the above hardware architecture, complete the hardware design. In view of the problems in the use of the original system, the information collector of 4-core processor is selected as the main part of the hardware design. The specific design process of the equipment is as follows.

2.1 Data acquisition equipment

In this design, high-precision data acquisition chip is used to complete the collection of data information in the database. The internal chip of data acquisition equipment is CAWDSZ6852 \textsuperscript{[4]}, the chip is introduced by TI company and has the characteristics of multi-functional multimedia application. The collector has 8 parallel processors, 5 data multiplexing interfaces and 3 multi form data interfaces, which is convenient for collection, compression and transportation. The data collector is automatically connected with the hardware network of the system. After data acquisition and synthesis, it is transmitted to the hardware data terminal of the system, stored in the terminal, and recorded in the hard disk. When data acquisition equipment collects data information, SCLK is used as clock to record all data signals. The internal VPO configuration mode of the collector is rawd mode, and the input data is the basis of data encryption. In this design, the data collector is connected with PCI and HPI, the bus interface is Ethernet interface, the data path is connected with B3 and B19 by PCI bus, and the data transceiver can receive and transmit 10-100m physical layer data. In order to realize the simultaneous processing and encryption of
various database data, AP685 chip is added in the system \cite{5}, the uplink rate is 3.5 Mbit / s and the downlink rate is 6.7 Mbit / s. In order to ensure the stable operation of the above chips in the system hardware, the optimization of the computer using the chips is carried out. The optimized host parameters are set as follows.

Table 1 Computer parameter setting after optimization

<table>
<thead>
<tr>
<th>Direction of use</th>
<th>equipment</th>
<th>parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware design</td>
<td>chip</td>
<td>CAWDSZ6852</td>
</tr>
<tr>
<td>Memory storage</td>
<td></td>
<td>16GB</td>
</tr>
<tr>
<td></td>
<td>storage</td>
<td>4GB RAM</td>
</tr>
<tr>
<td></td>
<td></td>
<td>20GB ROM</td>
</tr>
<tr>
<td></td>
<td>Expand</td>
<td>TF Card</td>
</tr>
<tr>
<td></td>
<td>storage</td>
<td></td>
</tr>
<tr>
<td>Software module</td>
<td>Decoding</td>
<td>EAN-8</td>
</tr>
<tr>
<td></td>
<td>mode</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>UPC-E</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAN12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UPC-A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>UCC/EAN128</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CODE BAR</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSI PLESSEY</td>
</tr>
</tbody>
</table>

The optimized host is connected to other hardware devices as the hardware foundation of the system.

2.2 Data storage device

In order to improve the storage efficiency of the memory, a flash memory with large storage range and low manufacturing cost is selected in this paper. An intelligent chip is added to the memory, which greatly increases the storage capacity and reduces the occupied area. The memory design results are shown in the figure below.
There are 6 buses outside the memory, each bus is connected to an FPGA interface, and different interfaces connect different signals. Bus 1 connects signals in I/O mode, with a link bit width of 52, and inputs and outputs in a two-way manner to realize two-way data transmission. The remark method is I/O. The bus 2 connects signals in the OUT mode with a link bit width of 84. It outputs data in a unidirectional manner and controls the data in and out of the database. The bus 3 connects data in BSC mode with a link bit width of 16 and outputs partial data information in a unidirectional manner. Bus 4 is connected to the database by clas, the link bit width is 6, and the data is input in one-way way to realize signal selection. The data input method of different types of database is different. Bus 5 is connected to the database in the BUSY mode, with a link bit width of 81, and a variety of data input methods are used for input. The bus 6 connects signals in the ADD mode, the link bit width is 27, and the address signal is output in a unidirectional manner [6-7].

2.3 System data transmission equipment

Serial port is the abbreviation of serial interface. In the process of data transmission, serial bit by bit transmission is adopted. The 9-pin COM port on the computer is the serial communication interface. According to the different communication modes, it can be divided into synchronous serial communication and asynchronous serial communication [8].

In asynchronous serial communication, the time interval between each bit in a single frame is fixed, while the time interval between adjacent frames is not. The
following four bits constitute a frame of asynchronous serial communication: start bit, data bit, check bit and stop bit. The maximum baud rate of asynchronous serial communication is 115200bps.

In the original serial port selection, although the data rate of the selected data serial port is relatively fast, its bit error rate is relatively high compared with other data transmission methods, and the transmission line is simple. Therefore, in this design, multi serial port is used to realize the two-way communication of the system. In this design, RS-232 9-pin serial port transmission line commonly used in embedded equipment is used to realize data transmission. The serial port is simple in structure, convenient in use, and effectively improves the security and integrity of data transmission. The specific data transmission lines and equipment settings are as follows.
In the fig. above, the 3011 L + high-precision positioning and orientation GPS compass of Magellan navigation and positioning company of the United States selected by the system in this paper is shown. It uses serial port for data transmission, has two RS232 cdb9 interfaces, and supports the transmission baud rate of 1200115200.

The above design hardware is used in the original system hardware framework to complete the hardware design process. In the subsequent software module design, the hardware framework will be the basis of the function realization of the software module.

3 Software design of heterogeneous database encryption and decryption system based on Data Mining

In view of the problem that the error rate of the original system is too high in the process of database encryption and decryption, the encryption dictionary module and data information processing module are added to the system function module to ensure the data integrity in the process of data encryption and decryption and avoid the generation of error rate. The optimized software modules are as follows.
According to the above framework, data mining technology is used to optimize the shortcomings of the original system, and the hardware designed in this paper is used to achieve the performance of the optimized module, to ensure the realization of the encryption and decryption function of the system.

3.1 Set database encryption dictionary

In this design, heterogeneous database is used to design encryption and decryption system. Heterogeneous database is a collection system of multiple structure databases. When encrypting this kind of database, it is necessary to set the corresponding database user authority table to control the users who can encrypt and decrypt the database. The user authority table is established and maintained by the central authority of the database, which specifies the safe operation authority of each user class to each data class in detail. Table 2 is the user permission table of four user classes to three data classes, in which the field UCID is the user class ID; D is the user class's access right to the i-th data class, 3 is full control, 2 is read-write, 1 is read-only, 0 is no access; E is the number of key exports of the i-th data, and the system exports this kind of data according to the user class key and the key of a certain kind...
of data Remark records some information about the user's permission. The data in Table 2 shows that user class 1 is a super user and has full control over all three types of data; user class 2 has permissions on all three types of data respectively: full control, read-write and read-only; user class 3 has only read permissions on all three types of data; user class 4 can only read the third type of data and has no permission to access other data. As shown in Table 2:

<table>
<thead>
<tr>
<th>UCID</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E1</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>D2</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E2</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>D3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E3</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>REMARK</td>
<td>infor</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Through the above-mentioned user encryption and decryption permission setting, the design process of heterogeneous database encryption dictionary is completed. The encryption dictionary is also a table, which stores the type, size, encryption and other information of each field in the data table. Through the dictionary, you can find the identity of the security data class to which the user access field belongs, and then you can find the number of key exports to export the access data class in the user permission table according to the identity. Use this encryption dictionary for effective control.

3.2 Heterogeneous database information processing based on Data Mining

Through the above processing, set the database encryption and decryption permission process. In this design, the encryption and decryption of heterogeneous database will be set. The data in heterogeneous database is too complex. In order to reduce the operation difficulty of database encryption and decryption process, data mining technology is used to preprocess the internal data of database. The specific pretreatment process is as follows.
In this design, we use the classified data mining method in data mining technology to complete the data preprocessing process in heterogeneous database. Firstly, the internal data of the database is formed into training set to select the classification features. According to the selected classification features, the training data set is trained to form a classifier of heterogeneous database. Then, the classifier is used to classify the sample data of the classification database to be carried out, and finally the classification mining is completed. The whole process is divided into two stages: building the classification data mining model and using the classification data mining model to predict the target data set.

Using the data after classification mining, using Gaussian filtering and other methods to complete the preprocessing process. The specific data preprocessing process can be summarized as follows: data collection, data classification mining, data cleaning, data filtering and data composition. Through the above process, the generation of error rate in data processing is reduced, and the integrity of encrypted database content is ensured [9].

3.3 Building data encryption and decryption model

The database content after preprocessing is transformed into plaintext, which is input into the encryption and decryption module [10]. In the encryption module, there are three algorithms included in the encryption model, which will work with other modules. The SQL statement rewriting module will extract the plaintext value.
involved in the database operation statement and pass it to the encryption and decryption module as input. The encryption and decryption module will determine the encryption model to be used according to the corresponding column, record location and operation in the current statement of the plaintext value in the database. If the working key needs to be generated in this process, the key management module will be called, if necessary, to get the existing key, then call the metadata management module to read the key from the database. In the process of data preprocessing, data mining technology is used to complete the classification of data. Therefore, homomorphic encryption algorithm is used in this encryption module to complete the encryption and decryption of database.

Set the whole homomorphic encryption process to $A$, the known homomorphic encryption process consists of key generation, data encryption, decryption and data evaluation, set generation key to $K$, encryption process set to $B$, decryption process set to $C$, the evaluation process is $D$. Then the formula can be expressed as:

$$ A = [K, B, C, D] \quad (1) $$

Suppose that the public key $mK$ and the private key $nK$ jointly generate the corresponding data security parameters, set to $\alpha$. Use $mK$ for encrypting plaintext, $nK$ used to decrypt ciphertext. Set plaintext $K \in n$, where $n$ is a positive number, then $\hat{K}n$ is the set of integers. Homomorphic encryption expressed as $A_{\hat{x}}(w)$, then the formula can be expressed as:

$$ A_{\hat{x}}(w_i + w_j) = A_{\hat{x}}(w_i) \oplus A_{\hat{x}}(w_j) \quad (2) $$

$$ A_{\hat{x}}(p w_i) = p \otimes A_{\hat{x}}(w_i) \quad (3) $$

At the same time, the data decryption process is set after the data is encrypted. Set data decryption process to $C$. Decrypt the ciphertext $\hat{U}$ through the private key, which can be expressed by the formula as follows:
Finally, the evaluation of the decryption results completes the calculation process of homomorphic encryption. Suppose $I$ is the evaluation function, and the ciphertext is set to $\hat{R}$. The evaluation algorithm evaluates the ciphertext $\hat{E}$ by evaluating the key $\hat{Y}$ evaluation function $i$.

$$E = C(Y, i, R) \quad (5)$$

Through the above formula to complete the encryption and decryption process of database, to ensure the security of heterogeneous database data.

Through the hardware part and software module part of the design, the design of heterogeneous database encryption and decryption system based on data mining is completed.

4 Experimental demonstration and analysis

Combined with the above hardware design results and software module design, the design of heterogeneous database encryption and decryption system based on data mining is completed. In order to ensure the effectiveness of the design, the performance differences between the original encryption and decryption system and the encryption and decryption system designed in this paper are compared by using the test link of the construction system.

4.1 System test platform

The hardware facilities of encryption and decryption system test of heterogeneous database based on data mining include server and other computer equipment. A system server, which is specifically configured as: CentOS version 6.5 64 bit operating system, high configuration processor, 16g memory, 4T hard disk; other computer equipment borrows computers from the existing network platform of a unit.

The system server configures the running environment and database environment required by the system, installs MySQL database, and opens the database security audit function. The development environment and language are java + mysql5.6, and the development tool is idea.
Using the above platform, the system test is completed, and the performance difference between the original system and the designed system is obtained. In order to ensure the validity of the system test results, set system test samples to provide data basis for system test. Specific system test samples are shown in the table below.

Table 3 Test sample data

<table>
<thead>
<tr>
<th>Test sample No</th>
<th>Sample data volume</th>
<th>Data form</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2000</td>
<td>TEXT</td>
</tr>
<tr>
<td>2</td>
<td>5000</td>
<td>TEXT</td>
</tr>
<tr>
<td>3</td>
<td>10000</td>
<td>TEXT</td>
</tr>
<tr>
<td>4</td>
<td>25000</td>
<td>TEXT</td>
</tr>
<tr>
<td>5</td>
<td>50000</td>
<td>TEXT</td>
</tr>
</tbody>
</table>

Using the above test sample data, using the original system and the design system to process the above data, to obtain the system test results.

4.2 Analysis of experimental results

Through the above design, the system test process is completed, and the specific system test results are as follows.

Table 4 Comparison of system test results

<table>
<thead>
<tr>
<th>Test sample No</th>
<th>Original system bit error rate %</th>
<th>Bit error rate of system</th>
</tr>
</thead>
</table>

Fig. 6 System test platform architecture
According to the above test results, the error rate of the designed system is significantly lower than the original system. Therefore, the data processing and encryption / decryption process of the system can effectively improve the data integrity in the data processing process and ensure the security and integrity of the database data content. The data processing process of the original system is relatively simple, and it is easy to create the situation of data missing, which does not meet the requirements of data security. In conclusion, the encryption and decryption performance of the design system in this paper is better than the original encryption and decryption system. Applying the design system in this paper to display life can effectively improve the internal security of the database.

5 Concluding remarks

Because the traditional heterogeneous database encryption and decryption system has a high error rate in the process of encryption and decryption, this paper designs a heterogeneous database encryption and decryption system based on data mining. Through the data acquisition equipment, data storage equipment and system data transmission equipment design system hardware; use data mining method to preprocess the data, in the form of homomorphic encryption, complete the database encryption and decryption system software design. The experimental results show that the error rate of this system is low, and the encryption and decryption performance is better than the original encryption and decryption system. We should popularize the design system in this paper, and improve the security of heterogeneous database in daily life. In this design, the description of the data processing part is relatively simple. In future applications, if there are data processing problems, they should be corrected in time to ensure the effectiveness of the data.
Reference

Blockchain-Based Vehicular Collaborative Computing

Qibin Zheng¹, Rongqing Zhang¹,²*, Shengjie Zhao¹,*
{zhengqb_2018@tongji.edu.cn, rongqingz@tongji.edu.cn, shengjiezhoao@tongji.edu.cn}

School of Software Engineering, Tongji University, Shanghai 201804, China¹
Shandong Provincial Key Laboratory of Wireless Communication Technologies, Shandong University, China²

Abstract. Collaborative computing among vehicles on the road has been regarded as a promising application for connected and intelligent vehicles as new computing and communication technologies emerge. Instead of using cloud computing services, a vehicle can complete computing-intensive tasks by making use of neighboring vehicles' idle computing resources and meanwhile, pay for this service to the assisting vehicles to increase their incentive to participate in such collaborative computing tasks. In collaborative computing among vehicles, how to balance the profits of both sides need to be investigated. What's more, the reliability of transactions is also a concerning problem. In this paper, we first introduce a collaborative computing scenario and propose a two-layer blockchain architecture to enhance the reliability of transactions. Then, we propose a coalition formation game-based collaborative computing algorithm to efficiently achieve effective computing coalition formation in a distributed manner. Simulation results show that the proposed algorithm can significantly improve the vehicles' utilities as well as the task performance.

Keywords: Internet of vehicles, Blockchain, collaborative computing.

1 Introduction

With the rapid development of computing and communication technologies, vehicles become more intelligent and are expected to support a rich and varied onboard applications, such as real-time navigation and vehicular entertainment. On the one hand, vehicles can get a precise and broad perception of their driving environment by the growing links with neighboring vehicles (vehicle-to-vehicle (V2V) links) or roadside units (vehicle-to-infrastructure (V2I) links). On the other hand, vehicles equipped with a more powerful on board computing unit (OBU) can perform independent computation better than before. In addition, a recent report shows that the global number of connected vehicles is increasing rapidly, and more than 286 million connected vehicles will be added globally during the 2019-2025 period [1]. These trends reflect the growing intelligence and complexity of vehicles.

Although the vehicles become more capable of computing, some emerging vehicular applications, such as augmented reality (AR) and virtual reality (VR), require completing a large amount of computations in a relatively short time, which cannot be supported by current computing ability of a single vehicle. One promising solution to this problem is to deliver the computation-intensive task from the vehicle (called task requesting vehicle) to other objects

* Corresponding authors: Rongqing Zhang and Shengjie Zhao.
with higher computing ability through the network to achieve higher performance. The target objects can be a cloud center [2], [3]. The cloud center has strong computing ability and can effectively reduce the computing overhead of requesting vehicles, but the long distance to a far-away center will also cause an extra delay. Another kind of target object is the roadside unit (RSU) or base station along with edge server which can complete the computation near to requesting vehicles [4], [5].

Different from the above two cases, a new solution tries to make use of idle resources of other vehicles to complete the task. That is, the vehicles around the requesting vehicle accept the task together and perform the computation in a collaborative computing way [6]-[10]. In [7], a learning-based task offloading framework was proposed using the multi-armed bandit (MAB) theory, which enables vehicles to learn the potential task offloading performance of its neighboring vehicles with excessive computing resources and minimizes the average offloading delay. In [8], a vehicular cloud network architecture was proposed to offload the computational burden of a centralized cloud by assigning more tasks to vehicular clouds formed near traffic lights. While [7], [8] both only considered the task performance (e.g., latency, cost) on behalf of the requesting vehicle, [9], [10] further considered the utilization of neighboring vehicles. An optimal computation resource allocation scheme was proposed in [9] to maximize the total long-term expected reward of the overall vehicles. In [10], a novel computation offloading marketplace was established in vehicular networks, where a Vickrey–Clarke–Groves based reverse auction mechanism was exploited, and a unilateral-matching-based algorithm was proposed to implement polynomial computational complexity.

However, these literature work only consider one side’s interest, lacking in making a tradeoff between task performance and utility of neighboring vehicles. What’s more, the increasing connectivity and decentralization of vehicular networks pose new challenges to the reliability of transactions when vehicles make deals with each other for tasks. It can be possible risky to make deal with strange vehicles. Blockchain is a promising technology to guarantee transaction reliability among multiple agents in a decentralized manner. To overcome the above challenges, we adopt the blockchain technology to create a reliable transaction mechanism for collaborative computing process.

The main contribution of this paper is to propose a two-layer vehicular blockchain architecture, which enables the neighboring vehicles to store their transaction records in both temporary and permanent manners to enhance the transaction reliability. What’s more, to balance both sides’ interests, we further employ the coalition formation game to model such a collaborative computing process and propose a distributed computing coalition formation algorithm of the vehicles that have the incentive to participate in collaborative computing. Simulations results verify the efficiency of our proposed scheme compared with other baselines in terms of utility and computing performance.

The rest of this paper is organized as follows: Section 2 presents the system model. In Section 3, we introduce the coalition formation game concepts and then provide the proposed algorithm. Section 4 presents the simulation results. Finally, the conclusion is drawn in Section 5.

2 System model

2.1 System architecture
As shown in Figure 1, the system architecture consists of the vehicular layer and the edge layer. The vehicular layer contains vehicles along the road. These vehicles equipped with OBU can perform specific computations and communicate with their neighboring vehicles through V2V communication links. The edge layer contains two types of nodes that are placed at the edge of the network compared to the cloud center: roadside units (RSUs) and base stations. The edge nodes have the stronger computing power and more storage resources than the vehicles. They communicate with the vehicles in the vehicular layer through V2I links, which can transfer and store the data generated during vehicle driving and interaction.

Since the vehicles also have computing capability, it is a feasible and efficient mechanism to complete a large computing task through collaborative computing among multiple vehicles. Therefore, the vehicles in the vehicular layer are further divided into two types in terms of different roles: requesting vehicles and assisting vehicles. The requesting vehicles issue computing tasks, whereas the assisting vehicles form coalitions to complete computing tasks in a cooperative manner. To ensure the reliability of transactions, a small scale blockchain network will be established by the formed vehicle coalition to record transaction data of the collaborative computing process. The data recorded on the blockchain include the payment of the requested...
task, the vehicle execution records of computing results, and the transactions of payment allocation.

After collaborative computing is completed, the vehicle group is dissolved, and the generated blockchain data is uploaded to the edge layer. Nodes at the edge layer also establish a blockchain network to permanently store the transactions.

Another class of data recorded in the permanent blockchain is the trust value of vehicles. Trust value reflects the credit conditions of vehicles during a certain period and changes based on vehicles’ activities over time. To enhance the reliability, requesting vehicles can set a trust value threshold for their tasks, which makes assisting vehicles ignore the ones with low trust value during computing coalition formation.

2.2 Computing model

In this paper, we consider a collaborative computing scenario in vehicular networks, where there is a requesting vehicle (denoted as v₀) on the road, and N nearby vehicles around v₀ in its communication range, denoted as the set \( \mathcal{N} = \{v_1, v_2, \ldots, v_N\} \). All these vehicles can communicate through V2V links. When v₀ needs to complete a computing task \( Task_{v_0} \), v₀ sends the request message to the vehicles in the set \( \mathcal{N} \). The content of the request message is represented as \( \{A_{v_0}, T_{v_0}, R_{v_0}\} \), where \( A_{v_0} \) is the computing amount of \( Task_{v_0} \), \( T_{v_0} \) is the completion time limit of \( Task_{v_0} \), and \( R_{v_0} \) is the reward payment for the completion of the task.

It is assumed that due to the limitations of \( A_{v_0} \) and \( T_{v_0} \), a single vehicle in \( \mathcal{N} \) usually cannot complete the whole task alone. To complete the task in time, the minimum computing power (denoted as \( x_0 \)) is:

\[
 x_0 = \frac{A_{v_0}}{T_{v_0}}. \tag{1}
\]

After receiving the message, each vehicle in \( \mathcal{N} \) starts to find nearby vehicles to form the coalition for completing the task \( Task_{v_0} \) in a collaborative manner. Let \( S \subset \mathcal{N} \) be a coalition consisting of \( |S| \) vehicles. Assume the vehicle \( v_i \in \mathcal{N} \) has the computing power \( x_i \), and the total computing power of vehicles in \( S \) is given by \( x_S \):

\[
 x_S = \sum_{v_i \in S} x_i. \tag{2}
\]

In order to complete the task in a shorter time, the task payment \( R_{v_0} \) is designed as a function of \( x_S \):

\[
 R_{v_0} = \begin{cases} 
 b \cdot R_0 & x_S > b \cdot x_0 \\
 \frac{x_S}{x_0} \cdot R_0 & x_S \leq b \cdot x_0 \end{cases} \tag{3}
\]

where \( R_0 \) is the given payment when \( S \) spends \( T_{v_0} \) to complete the task (that is, \( x_S = x_0 \)). When \( x_S > x_0 \), the payment keeps increasing until reaching the maximum payment \( b \cdot R_0 \). To earn a considerable profit, the vehicles in \( \mathcal{N} \) will select suitable partners to form the coalition. The profit measurement and the method of selecting will be discussed further in section III.

Denoted \( P = \{S_1, S_2, \ldots, S_k\} \) is the set consisting of several disjoint coalitions formed by the vehicles in \( \mathcal{N} \). v₀ will select the appropriate coalition (or coalitions) from \( P \) and delegate
the task to it. The selected coalition starts the collaborative computing and records related data in the temporary blockchain. When the computing task is completed, the selected coalition will allocate the payment $R_{v_0}$ to the member vehicles through the blockchain network. The selected coalition uploads the data on the temporary blockchain to the permanent blockchain network at the edge layer for preservation periodically or before dissolution.

3 Computing coalition formation algorithm

To realize efficient coalition formation in a distributed manner, we employ the coalition formation game[11] to obtain an effective solution to the above problems, which can not only meet the computing requirements of $v_0$, but also encourage the vehicles to participate in collaborative computing and gain a considerable profit.

3.1 Coalition game concept

**Definition 1:** Transferable utility $v(S)$. $u(S)$: $u(S)$ is the revenue of the coalition $S$ in the game. It represents the reward the whole coalition will get by forming this coalition.
$c(S)$: $c(S)$ is the cost of the coalition $S$ in the game. Coalition formation game theory considers the presence of cost in the game. In such an assumption, the coalition will cause some kind of cost during the forming operation, e.g., the cost of exchanging information between members.
$v(S)$: $v(S)$ is the difference between the revenue and the cost of a coalition which is: $v(S) = u(S) - c(S)$. $v(S)$ is a real value that can be divided in any manner between the coalition members. We will define $v(S)$ for the collaborative computing scenario in section III-B.

**Definition 2:** Coalition formation game for the vehicular collaborative computing problem. A coalition formation game for the vehicular collaborative computing problem is defined by the pair $(\mathcal{N}, v)$ which involves a set of assisting vehicles, denoted by $\mathcal{N} = \{1, \ldots, \mathcal{N}\}$ who seek to form cooperative computing coalitions to compute the tasks. And $v$ or $v(S)$ is the profit of the coalition $S$.

**Definition 3:** A coalition formation game $(\mathcal{N}, v)$ with transferable utility is superadditive if for any two disjoint coalitions $S_1, S_2 \subseteq \mathcal{N}$, $v(S_1 \cup S_2) \geq v(S_1) + v(S_2)$.

**Theorem 1:** The coalition formation game for the collaborative computing scenario is not superadditive.
**Proof:** According to the definition of $v(S)$ in section III-B, if the game is superadditive, then the coalition $\mathcal{N}$ containing all the vehicles will obtain the maximum transferable utility. However, due to the second cost item $c(S)$ defined in section III-B are growing at second-order of coalition size which is faster than the revenue $u(S)$, so the transferable utility will decrease when coalition size surpasses a certain value (the value won’t be big and are reachable in our simulation).

**Definition 4:** Coalition partition. A partition is defined as the set $P := \{S_1, \ldots, S_k\}$, where $S_k \subseteq \mathcal{N}$ are mutually disjoint coalitions and $\bigcup_{i=1}^{k} S_i = \mathcal{N}$.

**Definition 5:** Preference order. For any vehicle $v_i \in \mathcal{N}$, we define preference order $\triangleright_1$ as a kind of comparison relation, where $B \triangleright_1 A$ means that $v_i$ prefers to join coalition $B$ instead of coalition $A$. We use the preference order as follow:
\[ B \triangleright i A \iff x_{B'} > x_A \text{ and } v(B') \geq v(B). \]  

(4)

where \( B' = B \cup \{v_i\} \). This definition means vehicle \( v_i \) prefers to join coalition \( B \) by leaving coalition \( A \) when the new computing power of \( B \) will be larger than the origin computing power of \( A \), meanwhile the utility of \( B \) is not decreasing.

**Definition 6: Switch rule.** We introduce the simple switch rule as the basic operation of our algorithm as follow [12]:

**Switch:** \[ A, B \rightarrow A \setminus \{v_i\}, B \cup \{v_i\}, \]

where \( A \subseteq \mathcal{N}, B \subseteq \mathcal{N} \cup \emptyset, B \neq A, v_i \in A, B \triangleright i A \cdot \]

(5)

**Definition 7: Payoff scheme.** The coalition \( S \) needs to divide the utility \( v(S) \) for its members, and every vehicle \( v_i \) get its payoff \( \phi_i \). We use the following division scheme:

\[ \phi_i = \frac{x_i}{x_S} \cdot v(S). \]

(6)

**Definition 8: Partition convergence.** A partition \( P \) is considered convergence if for any \( A \in P, B \in P \cup \emptyset \) and \( B \neq A \), no \( v_i \in A \) satisfying \( B \triangleright_i A \) exists.

### 3.2 Transferable utility of the computing coalition

In this section, we design the transferable utility mentioned in section III-A for coalition formation game of collaborative computing.

The revenue for forming coalition \( S \) is defined as \( R_{v_0} \), that is:

\[ u(S) = R_{v_0}. \]

(7)

In addition to getting paid, \( S \) also generates some overhead, including computing cost and communication cost. The computing cost is defined as:

\[ c_1(S) = \sum_{v_i \in S} x_i \cdot T_S \cdot p_i \]

(8)

where \( T_S \) is the real computing time of \( S \) which is defined as:

\[ T_S = \begin{cases} \frac{A_{v_0}}{x_S} & \text{if } x_S > x_0 \\ T_{v_0} & \text{otherwise} \end{cases} \]

(9)

and \( p_i \) is the unit computing cost of vehicle \( v_i \) (such as the power consumption per unit time of CPU).

In the process of forming a coalition, vehicles need to transfer messages directly or by using blockchain, which brings communication overhead. The amount of the communication overhead is related to the size of the coalition and the distance among the vehicles. It is represented by \( c_2(S) \) and \( c_3(S) \), respectively:
\[ c_2(S) = \alpha \cdot |S|^2 \]  
where \( \alpha \) is a weight factor and \(|S|\) is the size of the coalition \( S \).

\[ c_3(S) = \beta \cdot \max_{v_i, v_j \in S} d(v_i, v_j) \]  
where \( \beta \) is a weight factor, and \( d(v_i, v_j) \) is the distance between \( v_i \) and \( v_j \) in coalition \( S \).

From the above formulas (7) - (11), the utility of the coalition is obtained:

\[ v(S) = u(S) - c_1(S) - c_2(S) - c_3(S) . \]

Due to \( c_2(S) \) and \( c_3(S) \) increasing when the coalition expands, forming a grand coalition may not be the solution with maximum \( v(S) \), and the vehicles in \( N \) will form several disjoint coalitions instead.

### 3.3 Coalition formation algorithm

Instead of using an optimal centric algorithm, we adopt a distributed algorithm that consists of two phases as shown in Table 1. During the first phase, the vehicles perform switch operations in a random way and form coalitions until no more switch operation can happen, and then the partition \( P \) reaches a convergence. In the second phase, \( v_0 \) selects the appropriate coalition (or coalitions) from \( P \). In most cases, the coalition with maximum computing power will be chosen to execute the task to minimize the latency. However, sometimes no coalition in \( P \) can provide computing power more than \( x_0 \), then \( v_0 \) will choose more coalitions to guarantee the sum of computing power to be greater than the required threshold \( x_0 \). In such a case, \( v_0 \) needs to divide the whole task into several subtasks and deliver them to different coalitions. Compared with the no-task-division case, \( v_0 \) has to coordinate and mix together the work of more coalitions which may bring extra cost and increase the risk for damaging the performance of the task. In the task-division case, the reward for each coalition is calculated with a punishing factor:

\[ R'_{v_0,S_i} = \frac{x_{S_i}}{\sum_{S_k \in F} x_{S_k}} \cdot R_{v_0,S_i} . \]

where \( R_{v_0,S_i} \) is the origin reward of coalition \( S_i \) according to (3). The punishing factor considers the contribution of \( S_i \) to the set \( F \). After completing the task, the coalition divides its utility to the vehicles within as described in Definition 7.

### 4 Simulation results

In our conducted simulations, vehicle \( v_0 \) is placed at the center of a 100 m × 10 m road, and the other vehicles are placed randomly on the road. The number of assisting vehicles \( N \) varies from 5 to 15 in the experiment. For each value of \( N \), we run the simulation for 150 times and take the average of the results (utility or computing power described below) as the final value. The other simulation parameters are listed in Table 2.
Table 1. Coalition formation algorithm.

**Initialize** the vehicles are non-cooperative. The partition of the network is $P = \{v_1, v_2, \ldots, v_N\}$.

**Phase I - coalition formation:**
1 repeat
2 randomly select coalition $A, B$ and $v_i$, where $A \in P, B \in P \cup \emptyset$,
3 $B \neq A, v_i \in A$.
4 if $B \triangleright v_i A$ then
5 $P = P \setminus \{A, B\} \cup \{A \setminus \{i\}, B \cup \{i\}\}$.
6 until $P$ reach convergence.
7 suppose now $P = \{A_1, A_2, \ldots, A_K\}$.

**Phase II - coalition selection:**
8 $\emptyset \rightarrow F$.
9 sort $P$ and get $P = \{B_1, B_2, \ldots, B_K\}$, where $x_{B_i} > x_{B_j}$ if $i < j$.
10 if $x_{B_i} > x_0$ then
11 $\{B_1\} \rightarrow F$.
12 else
13 find $k$ where $\sum_{i=1}^{k} x_{B_i} >= x_0$ and $\sum_{i=1}^{k-1} x_{B_i} < x_0$.
14 $\{B_1, \ldots, B_k\} \rightarrow F$.
15 $v_0$ delegate $Task_{v_0}$ to $F$.
16 $F$ completes the task and divide the utility.

---

**Table 2. Simulation parameters.**

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Computing Amount $A_{v_0}$</td>
<td>100 unit</td>
</tr>
<tr>
<td>Basic Reward $R_{v_0}$</td>
<td>200 unit</td>
</tr>
<tr>
<td>Maximum Reward Ratio $b$</td>
<td>1.65</td>
</tr>
<tr>
<td>Minimum Task Computing Power $x_0$</td>
<td>192 unit</td>
</tr>
<tr>
<td>Cost Weight for Coalition Size $\alpha$</td>
<td>3</td>
</tr>
<tr>
<td>Cost Weight for Vehicles’ Distance $\beta$</td>
<td>6</td>
</tr>
<tr>
<td>Computing Power of Assisting Vehicle $x_i$</td>
<td>generated randomly between 48-80 unit</td>
</tr>
<tr>
<td>Unit Computing Cost $p_i$</td>
<td>generated randomly between 0.1-0.15 unit</td>
</tr>
</tbody>
</table>

We will compare our coalition formation game-based algorithm (denoted as CF) with two baseline schemes as follows:

1) **Maximum utility scheme (MU).** This scheme searches the coalition with maximum utility by using the exhaustive method. However when the number of vehicles is relatively small ($N = 5$ or $6$), the coalition with maximum utility can’t provide computing power more than $v_0$ which is similar to our proposed scheme. At this time, $v_0$ also selects more coalitions beside the one before to fulfill the requirement.
2) Nearest selection scheme (NS). This scheme sorts the vehicles by the distance to $v_0$ and selects the nearest ones whose total computing power is more than $x_0$. To improve the utility, it will keep adding new vehicles in the selection until the utility of the coalition begins to decrease.

**Figure 2.** A snapshot example of coalition formation.

Figure 2 shows a snapshot of a network with 10 assisting vehicles. By using the coalition formation game-based algorithm, the vehicles form three coalitions in this example. The utility and computing power are shown in Table 3. $v_0$ will select coalition 3 because it has the maximum computing power and satisfies the requirement of $x_0$.

**Table 3.** Coalition parameter result.

<table>
<thead>
<tr>
<th></th>
<th>Coalition 1</th>
<th>Coalition 2</th>
<th>Coalition 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utility</td>
<td>103.80</td>
<td><strong>109.68</strong></td>
<td>101.37</td>
</tr>
<tr>
<td>Computing Power</td>
<td>216.14</td>
<td>173.39</td>
<td><strong>218.25</strong></td>
</tr>
</tbody>
</table>

**Figure 3** compares the utilities of the selected coalition(s) generated by three different schemes. From **Figure 3**, we can find that the utility obtained by CF is very close to the one by MU, which indicates that CF can guarantee the assisting vehicles getting well paid from the task. Moreover, we can also see that NS performs much worse than CF and MU. This is because NS forms the coalition in a greedy manner to get enough computing power which includes “join operation” only and does not include “leave operation”.

In **Figure 4**, the computing power of the three schemes are compared. We can find that the computing power of CF and MU are close when $N$ varies from 5 to 10, and then CF begins to perform better than MU. This is because when $N$ is relatively small, CF often finds the optimal coalition with the maximum utility which is the same as MU, and sometimes MU even gets better results by exhaust search (both the utility and computing power are higher) which reduces the difference between CF and MU. But as $N$ keeps growing, CF can perform better by seeking greater computing power at an acceptable cost of utility. It indicates that CF can reach better or at least the same task performance compared with MU. And we also have that NS performs better than CF and MU when $N$ is 5 or 6. This is because NS tries to gather enough computing...
power without considering utility at first, and may form the larger coalition which is hard for CF and MU when $N$ is relatively small.

**Fig.3.** The utility of selected coalition(s).

**Fig.4.** The computing power of selected coalition(s).
5 Conclusions

In this paper, we constructed a two-layer blockchain architecture suitable for reliable collaborative computing among vehicles. We proposed a coalition formation game-based algorithm to seek the appropriate vehicles for collaborative computing, so as to make a tradeoff between task performance and system utility of vehicles. Simulation results show that our algorithm reaches a good performance close to the exhaustive search method.

Acknowledgments. This work is supported in part by the National Key Research and Development Project under Grant 2017YFE0119300, 2019YFB2102300 and 2019YFB2102301, in part by the National Natural Science Foundation of China under Grant 61936014 and 61901302, in part by the Scientific Research Project of Shanghai Science and Technology Committee under Grant 19511103302, in part by the open research fund from Shandong Provincial Key Laboratory of Wireless Communication Technologies (No. SDKLWCT-2019-02), and in part by the Fundamental Research Funds for the Central Universities (China).

References

Blockchain-Based Collaborative Decision-Making in Vehicular Networks

Dezhen Wang¹, Rongqing Zhang¹,²*, Shengjie Zhao¹*,
{wdztju@163.com, rongqingz@tongji.edu.cn, shengjiezhao@tongji.edu.cn}

School of Software Engineering, Tongji University, Shanghai 201804, China¹
Shandong Provincial Key Laboratory of Wireless Communication Technologies, Shandong University, China²

Abstract. Collaborative decision-making (CDM) in vehicular networks can greatly improve the driving efficiency of vehicles. However, vehicle clusters often have serious security threats. To solve this issue, we propose a blockchain-based collaborative decision-making (BCDM) model, which is divided into two parts: the architecture level and the algorithm level. At the architectural level, we employ blockchain into vehicular networks and propose a layered blockchain network architecture (LBNA) that not only eases the data calculation and storage pressure of vehicular networks, but also further guarantees the security of the system. At the algorithm level, a BCDM algorithm combining direct trust and indirect trust is provided to determine the occurrence of traffic events and identify false messages. Simulation results reveal that the proposed system is effective and feasible in processing and storing trust information in vehicular networks.

Keywords: Blockchain, collaborative decision-making, vehicular networks, data credibility.

1 Introduction

In the field of vehicular networks, the development of communication technology among vehicles has achieved great results. In an increasingly complex road environment, vehicular networks need to face a variety of traffic events [1], [2]. The Traffic Incident Management Handbook [3] defines an event as “any non-recurring event that causes a reduction of roadway capacity or an abnormal increase in demand.” The 2000 Highway Capacity [4] defines an event as being “any occurrence on a roadway that impedes normal traffic flow.” In order to take safe and efficient driving actions according to different traffic events, cooperative decision-making (CDM) is regarded as a promising solution for connected and intelligent vehicles in vehicular networks. In the process of CDM, there are many security threats such as in-vehicle sensor recognition errors and malicious vehicles’ false information pouring. In the case of limited hardware, software, and energy resources of vehicular networks, enhancing vehicle safety and proposing effective solutions to potential safety hazards are major challenges [5].

At the level of the event discrimination algorithm, Ahmad et al. [6] summarized several trust models (TMs) applicable to vehicular networks, which can be divided into three categories: Entity-oriented Trust Models (ETM), Data-oriented Trust Models (DTM), and Hybrid Trust Models (HTM). Kerrache et al. [7] proposed an opportunistic alert dissemination mechanism

* Corresponding authors: Rongqing Zhang and Shengjie Zhao.
based on trust relationships between vehicles. Chen et al. [8] proposed a security scheme for evidence combination in CDM. This scheme can combine the direct trust value from local data with the indirect trust value from neighboring vehicles to obtain traffic incident decision results. Li et al. [9] proposed an Attack-Resistant Trust (ART) management scheme. This scheme can detect malicious attacks, evaluate the credibility of data and mobile nodes in vehicular networks.

However, with the development of information technology, algorithm-level security solutions are no longer sufficient to ensure the reliability and robustness of vehicular networks. Therefore, we propose to exploit blockchain to adjust and optimize the vehicular network architecture [10]. The blockchain has inherent characteristics such as fault tolerance, transparency, tamper resistance, and traceability. Based on this, Kang et al. [11], [12] proposed that blockchain technology can ensure secure data sharing in vehicular edge computing and networks (VECONs). What’s more, blockchain-enabled Internet of Vehicles (BlIoV) can enhance soft security performance through miner selection and block verification. Ali et al. [13] proposed a multi-tier blockchain network architecture, which has good horizontal extensibility. Yang et al. [14] proposed a Blockchain-based Traffic Event Validation (BTEV) framework, which collects traffic data by roadside units (RSUs), and vehicles can use these data to verify the occurrence of traffic incidents. Yang et al. [15] proposed a decentralized management system for vehicular networks based on blockchain technology. In this system, vehicles can use Bayesian inference models to verify messages received from neighboring vehicles. Besides, each vehicle that sends a message will get a corresponding trust value, and these values will be stored in the blockchain network formed by RSUs.

Therefore, in this paper, we propose a Blockchain-based Collaborative Decision-Making (BCDM) model applied to vehicular networks. At the network level, the BCDM model includes an innovative layered blockchain network architecture (LBNA), which can protect the system’s security while avoiding on-board memory and computing resources that are occupied by blockchain data. At the algorithm level, we further propose a BCDM algorithm for vehicle clusters based on hierarchical events. The algorithm has the functions of event determination, malicious node identification, and reputation rating, which can be perfectly integrated with the blockchain network architecture.

The remainder of this paper is organized as follows. Section 2 describes the system model of BCDM. In Section 3, we propose a BCDM algorithm combining direct trust, indirect trust, and reputation rating. The numerical results and conclusions are drawn in Section 4 and Section 5, respectively.

2 System model

2.1 Network architecture model

As illustrated in Figure 1, the LBNA is divided into two layers, namely the center-layer and the edge-layer.

The edge-layer includes multiple groups of temporary blockchain networks (TBN), and each group is composed of vehicles in the same traffic environment. Because the vehicle itself has inherent characteristics such as high mobility, relatively small on-board memory, and relatively low on-board computing power, the TBN will not exist for a long time. Before the TBN disintegrates, its last miner node will upload the necessary data for the entire chain to the center-layer.
The center-layer is a permanent blockchain network (PBN), which is composed of RSUs, base stations, and databases in a large geographical area. These nodes have stronger computing and data storage capabilities, which can process and save massive data from edge-layer in order to query historical records or trace information sources when necessary. In addition, the PBN will perform statistics and analysis on the historical quality of messages broadcast by each vehicle, and thereby identify malicious vehicles. Identified malicious vehicles will be notified online for criticism and punished by prohibiting them from broadcasting messages.

Fig. 1. The layered blockchain network architecture (LBNA) in vehicular networks.

2.2 Event model

While driving, vehicles often face a variety of complex traffic events, such as road construction, road congestion, or different types of traffic accidents. These traffic events make up the collection: $\mathcal{E} = \{E_1, E_2, \ldots, E_n\}$. In this paper, it is assumed that traffic events have been pre-classified, and each type of traffic event is divided into different levels. Each level is independent of each other and all levels together constitute a complete set representing the corresponding event. For example, road congestion can be divided into five levels: (I) roads are clear, (II) lightly congested, (III) moderately congested, (IV) severely congested, and (V) impassable.
In addition, the BCDM model will set a corresponding prior probability for each level of a specific event based on empirical statistics on traffic big data: 

\[ \text{Prior}_i = \{p_{r_1}, p_{r_2}, \ldots, p_{r_\omega}\} \]

Where, \( p_{r_i} \) represents the prior probability of level \( (I) \) in \( E_i \), \( \omega \) represents the total number of levels included in \( E_i \).

During the driving process, the surrounding environment will be detected by on-board sensors. Once a specific traffic event is observed, the event will be evaluated according to the preset event levels, and a cooperative awareness message (CAM) will be generated based on this assessment result. Each CAM will contain the identity of the sender, the generation time, and a level recommendation for a certain traffic event. It will be broadcast to other vehicles in the same TBN.

2.3 Test model

All the CAMs received by a vehicle \( V_m \) constitute a set \( \text{CAM} = \{\text{cam}_1, \ldots, \text{cam}_j, \ldots\} \). This set may contain false contents from malicious vehicles or untrusted vehicles. Malicious vehicles are vehicles that intentionally broadcast false CAMs, whereas untrusted vehicles are vehicles that accidentally broadcast an error CAM because they are far away from where the event occurred. In the hypothesis of our paper, only a few of benign vehicles become untrusted due to occasional errors. Among them, \( \text{cam}_j = \{\text{cam}_j^1, \ldots, \text{cam}_j^k, \ldots\} \) refers to the set of all CAMs related to \( E_j \) received by \( V_m \), \( \text{cam}_j^k \) in the set denotes the CAM that \( V_k \) broadcasts to \( V_m \) about \( E_j \).

Each vehicle in the TBN will infer and judge the true situation of events based on two principles of direct confidence \( T_{\text{dir}}() \) and indirect confidence \( T_{\text{ind}}() \). The calculation of \( T_{\text{dir}}() \) is based on the trustworthiness of CAMs received by the vehicle. The calculation of \( T_{\text{ind}}() \) is based on the historical reputation records of the vehicle that issued the CAM. The historical reputation record is the quality judgment of all CAMs that the vehicle has ever issued. The vehicle receiving the message, \( V_m \), will calculate the comprehensive trust \( T_{\text{cam}} \) for the broadcast vehicle \( V_k \) based on \( T_{\text{dir}}() \) and \( T_{\text{ind}}() \).

In the end, \( V_m \) uses Dempster-Shafer theory (DST) to fuse all comprehensive trusts about \( E_j \), and obtains the event’s final discrimination result, \( T_{\text{final}} \). Meanwhile, \( V_m \) will give a reputation score to the broadcast vehicle \( V_k \) based on \( T_{\text{final}} \). If \( T_{\text{final}} \) proves the CAM is false, the corresponding vehicle \( V_k \) is given a negative reputation score (−1). This reputation score will affect the vehicle’s \( T_{\text{ind}}() \) and further affect the determination of whether the vehicle is malicious.

3 The proposed algorithms

3.1 The BCDM algorithm

**Step 1 Direct trust.** The vehicle’s direct trust \( T_{\text{dir}}() \) is an assessment of the reliability of vehicles’ broadcast CAMs.

The closer the vehicle is to the event, the more reliable the information it detects. Therefore, the distance trust of a certain CAM is defined as follows:

\[ d_j^k = e^{-a_j^k} + \beta \]  

(1)
where $d_j^k$ is the distance trust of $cam_j^k$ sent by vehicle $V_k$, $l_j^k$ is the distance between $V_k$ and the event location. $\alpha$ and $\beta$ are two present parameters, which respectively control the change rate and lower bound of the distance trust.

Suppose that the total number of CAMs received by $V_m$ for a certain event $E_j$ is $N$, that is, vehicle $V_m$ has a distance trust set for $E_j$:

$$Tot = \{d_1^j, d_2^j, ..., d_N^j\}.$$  

(2)

In the set $Tot$, there are $M$ CAMs of the same level as $cam_j^k$ about $E_j$, which will form a subset:

$$Sub = \{d_1^{(i)}, d_2^{(i)}, ..., d^{(M)}_j\}.$$  

(3)

Each element in the set $Sub$ also belongs to the set $Tot$.

Therefore, the direct trust of the CAM about $E_j$, which received by $V_m$ from $V_k$ is defined as follows:

$$T_{dir} = \frac{\sum_{i=1}^{M} d_j^{(i)}}{M}.$$  

(4)

**Step 2 Indirect trust.** The vehicle’s indirect trust $T_{ind}$ is calculated based on its own historical reputation.

The historical reputation $T_k^{b_i}$ is the accuracy of all the CAMs $V_k$ has broadcast when it joined the TBN for the $i-th$ time.

$$T_k^{b_i} = \frac{\text{Reliable}_{k}^{b_i}}{\text{Reliable}_{k}^{b_i} + \text{Unreliable}_{k}^{b_i}}.$$  

(5)

where $\text{Reliable}_{k}^{b_i}$ is the number of reliable CAMs that $V_k$ have been broadcasted in its $i-th$ TBN, while $\text{Unreliable}_{k}^{b_i}$ is the number of unreliable CAMs.

In order to ensure the reasonableness of the historical reputation evaluation of vehicles, we will focus on examining the performance of CAMs when vehicles join in the TBN this or the penultimate time. Therefore, the weighted aggregation method is used to calculate the indirect trust of $V_k$:

$$T_{ind} = \begin{cases} 
\sigma \cdot \frac{\sum_{i=1}^{H-2} T_k^{b_i} + T_k^{b_{H-1}}}{\sigma + 1}, & \text{num} \leq \delta \\
\sigma \cdot \frac{\sum_{i=1}^{H-1} T_k^{b_i} + T_k^{b_{H}}}{\sigma + 1}, & \text{num} > \delta 
\end{cases}.$$  

(6)

where $H$ is the total number of times $V_k$ has added to the TBN, $T_k^{b_i}$ is the historical reputation of $V_k$ joining in TBN this time, $\text{num}$ is the number of times $V_k$ broadcasted in the current TBN,
δ is a positive integer parameter which denotes the threshold for switching formulas, σ is the factor between 0 and 1 which denotes the weight given to the previous TBN’s historical reputation.

**Step 3 Comprehensive trust.** Based on the step 1~2, the comprehensive trust of the CAM from $V_k$ to $V_m$ is:

$$T_{com} = γ \cdot T_{dir} + η \cdot T_{ind}, \quad γ + η = 1$$  \hspace{1cm} (7)

where γ and η are the weights of $T_{dir}$ and $T_{ind}$, respectively.

**Step 4 Final trust.** In this work, Dempster-Shafer theory (DST) is able to fuse $T_{com}$ of multiple vehicles on the event $E_j$, and even if there are non-true discrimination results among them, accurate final judgment can be obtained. In DST, the recognition frame $Ω$ of $E_j$ is composed of all its levels:

$$Ω = \{ r_1, r_2, ..., r_x, ..., r_ω \}$$  \hspace{1cm} (8)

where $r_1$ indicates that the level of $E_j$ is (I), and $ω$ indicates the total number of levels in $E_j$.

When $V_m$ receives the CAMs sent by any other vehicle about $E_j$, it can calculate the corresponding comprehensive trust and obtain the set $\{ T_{com}^1, ..., T_{com}^N \}$. Based on this set, the basic probability assignment (BPA) for each level in $Ω$ can be completely obtained.

Here are the examples: suppose that $V_k$ judges $E_j$ as level $r_x$, and $V_m$ gets the corresponding $T_{com}^k(r_x) = m_k(x)$. According to the prior probability set $Priori = \{ pr_1, pr_2, ..., pr_ω \}$, other levels of BPA in $Ω$ can be calculated:

$$m_k(y) = \frac{pr_y \cdot [1 - m_k(x)]}{1 - pr_x} \quad r_y \in Priori, y \neq x.$$  \hspace{1cm} (9)

In DST, the probability is replaced by an uncertainty interval bounded by belief (bel) and plausibility (pl). bel is the lower bound of this interval and represents the supporting evidence. pl is the upper bound of this interval and represents non-denied evidence. Trust interval $[bel(r_x), pl(r_x)]$ represents the value range of $T_{final}$, while $pl(r_x) - bel(r_x)$ represents the uncertainty of the judgement about $r_x$. When $pl(r_x) - bel(r_x) = 0$, it means that the degree of trust in the judgement about $r_x$ is completely determined.

The belief function and plausibility function with regard to the level $r_x$ of $E_j$ obtained by $V_m$ are calculated as follows:

$$bel(r_x) = \sum_{r_z \in r_x} mass(r_z)$$  \hspace{1cm} (10)

and

$$pl(r_x) = \sum_{r_z \notin r_x \neq \phi} mass(r_z) = 1 - bel(\bar{r}_x).$$  \hspace{1cm} (11)
Here $r_x$ are all the basic elements that compose the level $r_x$. Since levels of $E_j$ in our hypothesis are single-element propositions and mutually exclusive, we have the following formulas:

$$\text{bel}(r_x) = \text{pl}(r_x) = \text{mass}(r_x) \quad \forall r_x \subseteq \Omega.$$  

$$\text{mass}(r_x) = \bigoplus_{n=1}^{N} m_n(r_x).$$  

Here \(\text{mass}(r_x)\) denotes the fusion of discrimination results about \(r_x\) for a total of \(N\) broadcast vehicles. We can combine these discrimination results by applying the Dempster’s rule, which is defined as follows:

$$m_1(r_x) \oplus m_2(r_x) = \frac{\sum_{a,b:R_a \cap R_b = r_x} m_1(R_a) \cdot m_2(R_b)}{1 - \sum_{a,b:R_a \cap R_b \neq a,b} m_1(R_a) \cdot m_2(R_b)}.$$

The final result of a particular level \(r_x\) about \(E_j\) is:

$$T_{\text{final}}(r_x) = \text{bel}(r_x) = \text{pl}(r_x).$$

### 3.2 Distributed consensus algorithm

Vehicle clusters store data in the BCDM process through a TBN. As a decentralized system, the TBN needs to choose an appropriate consensus mechanism to ensure that all vehicles can follow the established protocol rules. We choose to adopt the proof-of-stake (PoS) miner election method. Compared with proof-of-work (PoW), PoS is more suitable for applications in the field of vehicular networks, it can greatly shorten the time to reach consensus in each block, and does not require energy consumption for mining [16], [17]. In PoS protocols, instead of computational power resources, miners are selected based on their stakes:

$$P_i = \frac{s_i}{\sum_{j=1}^{N} s_j}.$$  

In our research, the weight of a vehicle \(s_i\) is the number of false CAMs it receives. In this way, most false CAMs and related data can be stored in TBN promptly, thereby ensuring the rapid identification of malicious nodes.

Considering the scalability of the entire architecture model, the PBN will use the PoW algorithm to be publicly deployed in a permissionless manner so that TBN in the edge layer can upload data freely.

### 3.3 Identification and punishment mechanism of malicious vehicles

The BCDM model will identify and punish malicious vehicles based on the data stored in the blockchain. The specific execution rule is to calculate the proportion of false CAMs to the total CAMs broadcasted by each vehicle in the current TBN. If the proportion of false CAMs exceeds the threshold \(Thr_1\), the corresponding vehicle will be warned by broadcast. If the proportion of false CAMs exceeds the threshold \(Thr_2\) (\(Thr_2 > Thr_1\)), the corresponding vehicle will be punished by banning broadcast CAMs.
4 Simulation results

In order to evaluate the efficiency of the proposed BCDM model, we conduct the simulations based on MATLAB. The configurations of key parameters are listed in Table 1. We classify all vehicles into three categories: trusted, untrusted, and malicious vehicles. Malicious vehicles are vehicles that intentionally broadcast false CAMs. Untrusted vehicles are vehicles that accidentally broadcast an error CAM because they are far away from where the event occurred. In order to simulate a real traffic scene, a certain percentage (broadcast ratio) of random vehicles will detect traffic events and broadcast them during each iteration. The number of iterations is 150 in the simulation.

Experimental strategies in the simulations: In the simulations, we evaluate the performance of the BCDM model by event discrimination accuracy, Precision \( P \) and Recall \( R \).

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \omega )</td>
<td>4</td>
</tr>
<tr>
<td>( P )</td>
<td>( pr_1 = pr_2 = pr_3 = pr_4 = 0.25 )</td>
</tr>
<tr>
<td>( N )</td>
<td>50</td>
</tr>
<tr>
<td>trusted to untrusted vehicle ratio</td>
<td>8:1</td>
</tr>
<tr>
<td>broadcast ratio</td>
<td>0.6</td>
</tr>
<tr>
<td>( \alpha )</td>
<td>0.015</td>
</tr>
<tr>
<td>( \beta )</td>
<td>0.0723</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>0.7</td>
</tr>
</tbody>
</table>

4.1 Event discrimination accuracy

We compare our accuracy performance with Yang et al. [15], in which they calculated the credibility of the vehicle broadcast message through Bayesian Inference (BI) to determine whether the event occurred. Figure 2 shows the impact of malicious vehicles on accuracy.

From Figure 2, we can see that as the proportion of malicious vehicles gradually increases, the accuracy of our BCDM algorithm is always higher than the scheme in [15], and when the proportion of malicious vehicles is less than 55\%, the BCDM algorithm can keep the discrimination accuracy as 100\%.

4.2 Precision and Recall

We use Precision and Recall as evaluation parameters, which are widely used in the CDM scenario.

In this paper, \( P \) and its baseline are defined as follows:

\[
P = \frac{\text{malicious vehicles detected}}{\text{wrong broadcast vehicles detected}}
\]

(17)

\[
\text{baseline} = \frac{\text{real malicious vehicles}}{\text{real wrong broadcast vehicles}}
\]

(18)
where *wrong broadcast vehicles* means malicious vehicles and untrusted vehicles. **Figure 3** shows the impact of malicious vehicles on Precision. When the proportion of malicious vehicles is 65% or less, the Precision of BCDM exactly matches the baseline.

**Fig. 2.** The impact of malicious vehicles on Accuracy.

**Fig. 3.** The impact of malicious vehicles on Precision.
In this paper, $R$ is defined as follows:

$$R = \frac{\text{malicious vehicles detected}}{\text{real malicious vehicles}}. \quad (19)$$

**Figure 4** shows the impact of malicious vehicles on Recall with different proportions of $T_{\text{dir}}$. In $T_{\text{com}}$, the more malicious vehicles in TBN leads to the smaller proportion of $T_{\text{dir}}$ and the better Recall performance. However, $T_{\text{dir}}$ can measure the trustworthiness of CAM itself and play an indispensable role in the BCDM model when there are fewer malicious vehicles. Therefore, the proportion of $T_{\text{dir}}$ in $T_{\text{com}}$ cannot be too small in practical applications.

![Figure 4. The impact of malicious vehicles on Recall.](image-url)

## 5 Conclusions

In this paper, we proposed a BCDM model. With this model, vehicles can collect CAMs in the same TBN, and process the information according to both direct and indirect trust. The BCDM model will discriminate events based on the processing results and score each vehicle’s reputation. These results and scores will be stored in the TBN at the edge layer and uploaded to the PBN before it disintegrates. Architecture analysis and simulation results show that our proposed model has better decision-making performance, and also makes practical contributions to the innovation of cooperative vehicular network architecture.
Acknowledgments. This work is supported in part by the National Key Research and Development Project under Grant 2017YFE0119300, 2019YFB2102300 and 2019YFB2102301, in part by the National Natural Science Foundation of China under Grant 61936014 and 61901302, in part by the Scientific Research Project of Shanghai Science and Technology Committee under Grant 19511103302, in part by the open research fund from Shandong Provincial Key Laboratory of Wireless Communication Technologies (No. SDKLWCT-2019-02), and in part by the Fundamental Research Funds for the Central Universities (China).

References


Data transmission deployment and reliability analysis in enterprise informatization construction

LI Jun¹, LIU Xiao-Ling²
{LIjun3812@163.com¹, LiuXiaoLing3812@163.com²}

(1. Institute of Education and Physical Education, Jiangxi Institute of Applied Science and Technology, Nanchang 330100, China; 2. Institute of Education and Physical Education, Jiangxi Institute of Applied Science and Technology, Nanchang 330100, China)

Abstract: In the construction of enterprise informatization, the traditional reliability analysis method of data transmission deployment has the problem of high data packet loss rate, which affects the reliability analysis results. Therefore, data transmission deployment and reliability analysis in enterprise informatization construction are proposed. Firstly, the wireless transmission module is designed to ensure the safe and barrier free transmission of data, and the can interface is designed to connect all links in the information construction with the wireless transmission module. On this basis, data is divided and deployed by data deployment algorithm; path loss of data transmission channel is taken as the basis of reliability analysis, and reliability analysis is completed by calculating various losses. The experimental results show that: compared with the traditional method, the data transmission deployment and reliability analysis method designed in the enterprise information construction has no packet loss, and the data is safe and reliable. This method is suitable for the application in the enterprise information construction.

Key words: Information technology; Data transmission; Reliability analysis; Data transmission deployment;

1 Introduction

Enterprise is the representative of a country's economic strength, which plays an important role in the process of national economic construction and development. The effective development of enterprises is an important guarantee for the country to realize modernization and strengthen the country[1]. With the development of information technology, enterprise information construction is a common way to improve enterprise efficiency in the 21st century. Therefore, how to do well in large-scale enterprise information construction has become one of the focuses of scholars in recent years[2].

Enterprise informatization construction refers to the establishment of enterprise informatization planning blueprint based on the planning objectives of the enterprise, combined with the practice of enterprise informatization and the mastery of the development
trend of information technology, so as to guide the enterprise informatization process comprehensively and systematically. Coordinate the application of information technology, make full and effective use of information resources of enterprises, and meet the needs of business development of enterprises[3]. It is a gradual derivation and definition process from the goal and business process of enterprise information construction to the framework of information system and the implementation strategy of information system. It takes the realization of enterprise information construction goal as the ultimate goal of information and system implementation[4].

In the past, there was data packet loss in the data transmission deployment of enterprise informatization construction, which led to the lack of strong basis for reliability analysis. Therefore, the data transmission deployment and reliability analysis in the enterprise information construction are proposed. Because the path loss in data transmission is an important index to measure the reliability of enterprise information construction, the evaluation of path loss in data transmission has been widely concerned by researchers. Most of the research on path loss evaluation in data transmission is based on the premise that data is transmitted through disjoint paths. However, in real life, data will inevitably be transmitted through paths containing the same link. Therefore, the reliability of path loss in data transmission is analyzed to judge whether enterprise information construction is reliable.

2 Data transmission deployment in enterprise information construction

2.1 Data transmission deployment in enterprise information construction

In the enterprise information construction, due to the bad working environment and strong interference, the data transmission module with strong anti-interference ability and error detection and correction ability is selected to ensure the accuracy of data transmission. After experimental comparison and test, str-35 industrial low power wireless data transmission module meets the requirements. The channel center frequency of str-35 wireless data transmission module is 900MHz, which has high anti-interference ability and low bit error rate. The modulation mode based on GFSK and efficient forward error correction channel coding technology are adopted to improve the ability of data anti burst interference and random interference[5]. The module can automatically filter out the false data generated in the air (received is sent). The interface baud rate is 19200bps. Flexible programming, no need to prepare redundant programs, as long as the data from the interface can be sent and received.

The technical specifications of STR-35 transmission module are shown in Table 1 below:
<table>
<thead>
<tr>
<th>Number</th>
<th>Technical Index</th>
<th>Parameter</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Modulation</td>
<td>MSK</td>
<td>User set</td>
</tr>
<tr>
<td>2</td>
<td>Working frequency</td>
<td>300-365MHz, 400-480MHz, 826-968MHz</td>
<td>User set</td>
</tr>
<tr>
<td>3</td>
<td>Transmit power</td>
<td>0dbm, 5dbm, 10dbm</td>
<td>User set</td>
</tr>
<tr>
<td>4</td>
<td>Receiving sensitivity</td>
<td>-118dBm</td>
<td>User set</td>
</tr>
<tr>
<td>5</td>
<td>Emission current</td>
<td>≤27mA</td>
<td>User set</td>
</tr>
<tr>
<td>6</td>
<td>Receive current</td>
<td>≤18mA</td>
<td>User set</td>
</tr>
<tr>
<td>7</td>
<td>Wireless wake-up current</td>
<td>≤8mA</td>
<td>User set</td>
</tr>
<tr>
<td>8</td>
<td>Stand-by current</td>
<td>≤20μA</td>
<td>User set</td>
</tr>
<tr>
<td>9</td>
<td>Interface speed</td>
<td>1200-115200bps</td>
<td>User set</td>
</tr>
<tr>
<td>10</td>
<td>Interface Type</td>
<td>UART</td>
<td>User set</td>
</tr>
<tr>
<td>11</td>
<td>Operating Voltage</td>
<td>+4.5~5.5V</td>
<td>User set</td>
</tr>
<tr>
<td>12</td>
<td>Operating temperature</td>
<td>-40℃~85℃</td>
<td>User set</td>
</tr>
<tr>
<td>13</td>
<td>Storage temperature</td>
<td>-65℃~150℃</td>
<td>User set</td>
</tr>
<tr>
<td>14</td>
<td>Working humidity</td>
<td>10%~90% Relative humidity</td>
<td>User set</td>
</tr>
<tr>
<td>15</td>
<td>Number of switches</td>
<td>4 channels, expandable to 16</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Maximum controllable switching voltage</td>
<td>AC 220V, DC 30V</td>
<td>User set</td>
</tr>
<tr>
<td>17</td>
<td>Maximum controllable switching current</td>
<td>AC 0.5A, DC 2A</td>
<td>User set</td>
</tr>
</tbody>
</table>

Based on the designed wireless transmission module, the CAN bus interface is designed to connect all links in the enterprise information construction through the CAN bus interface, so as to realize the data transmission between each link.

2.2 CAN interface design

The internal link of enterprise information construction is connected by CAN bus, and P82C250 chip is used as CAN bus transceiver in the design [6]. P82C250 is the interface between the CAN protocol controller and the physical bus. The device provides differential transmission capability to the bus and differential reception capability to the CAN controller.

P82C250 has 8 pins, dip package, and the function definition of each pin is shown in Table 2. The eighth pin rs of p82c250 can be grounded or connected to high level through
different connection modes, so that p82c250 can work in one of three modes of high speed, standby and slope control [7].

<table>
<thead>
<tr>
<th>Pin</th>
<th>name</th>
<th>Pin</th>
<th>name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TXD</td>
<td>5</td>
<td>VREF</td>
</tr>
<tr>
<td>2</td>
<td>GND</td>
<td>6</td>
<td>CANL</td>
</tr>
<tr>
<td>3</td>
<td>VCC</td>
<td>7</td>
<td>CANH</td>
</tr>
<tr>
<td>4</td>
<td>RXD</td>
<td>8</td>
<td>RS</td>
</tr>
</tbody>
</table>

Table 2 Pin function definition of P82C250

P87C591 is equipped with CAN controller and CAN interface circuit is designed. The interface between P82C250 and CAN bus adopts certain safety and anti-interference measures. The canh and canl pins of 82C250 are respectively connected with CAN bus through a 5Ω resistance, which CAN play a role of current limiting and protect 82C250 from over-current impact. Two 30pF small capacitors are connected in parallel between CANH and CANL and the ground, which can filter the high-frequency interference on the bus and have certain anti electromagnetic radiation ability [8]. The Rs pin of 82C250 is connected with a slope resistance, which is adjusted appropriately according to the bus communication speed.

2.3 Data deployment placement algorithm

In order to reduce the data transmission overhead across the data center after the data deployment, the data is divided according to the strength of the association dependency between the data, so that the data with strong association relationship can be divided into the same data subset and deployed to the same data center [9]. Due to the large scale of data in enterprise information construction and the limited storage space of data center, it is necessary to consider whether the data subset scale exceeds the maximum storage capacity of a single data center when dividing data [10]. In addition, for the data with fixed deployment location restrictions, since the data center to be deployed has been determined, the deployment location of the data subset composed of the data closely associated with it can also be determined, and the size of the subset can also be determined by the storage space of the data center.

Based on the above analysis, the data center storage space is sorted in descending order. Assuming that the sorted data center sequence is expressed as \(\{u_1, u_2, \ldots, u_n\}\), the storage space in the data center is sorted as \(Q(u_1) \geq Q(u_2) \geq \ldots \geq Q(u_n)\).

On the basis of the sorting results, the flow of data partition algorithm with limited scale is as follows. For data \(r_\epsilon \in R_\epsilon\) with fixed deployment restrictions, the size of data subset deployed to data center \(cen(r_\epsilon)\) is determined according to its storage capacity of deployed data center \(cen(r_\epsilon)\); According to the relationship between data and \(r_\epsilon\) and the size of data, select the members of the data subset from the data set that needs to be deployed to the same
location and the data set that does not have the limitation of deployment location. Repeat the above process until all data in the $R_\varepsilon$ set is divided. Secondly, for the data $r_\tau \in R_\tau$ without the limitation of deployment location, the size of data subset deployed to a single data center is determined according to the maximum storage capacity of the current data center; According to the relationship between data and $r_\tau$ and the size of data, the members of the data subset are selected from the data set $R_\tau$. Repeat the above process until all data in the $R_\tau$ set is divided. At this point, all data is divided into corresponding data subsets. Repeat the above process until all data members are determined, then the data members that data subset $q_1$ should contain are obtained. Then continue to divide and exchange the data of the remaining data members in the original data, which can complete the division of other limited size data subset members, and also ensure the small amount of data transmission between data subsets.

After the data partition, deploy the data subset to the data center. In data deployment, in order to ensure that the data access delay does not exceed the limit, the candidate data center sets available for deployment for each data set are different. Let $F=\{r_1, r_2, ..., r_n\}$ be the set of all data. Each data in $F$ is represented by a tuple, where $r_\tau=\langle size, cen \rangle$ represents the size of the data, $size(r_\tau)$ contains multiple candidate data $r_\tau$ centers, and $cen(r_\tau)=u_1, u_2, ..., u_n$ represents the set of candidate data centers that the data can be deployed. Because of the synchronization or data association between different data items in data sets, they need to transmit data frequently to form a relatively stable data traffic feature. The data transmission is deployed in the wireless transmission module designed above. Definition $F \rightarrow G$ is the mapping from data set $F$ to $G$. for any data $r_\varepsilon, r_\tau \in F$, the mapping meets the following requirements:

$$size(r_\varepsilon) < Q(u_n) \text{ and } u_n \in cen(r_\tau)$$

$size(r_\tau) < Q(u_n)$ and $u_n \in cen(r_\tau)$ are guaranteed to be met, and these two limits guarantee that data set $r_\varepsilon$ does not exceed the storage capacity limit of data center $u_n$, and that it is placed within the range of its alternative data center. The specific data transmission deployment scheme is shown in Figure 1.
The data placement strategy in Figure 1 is \{a→u_2, b→u_3, c→u_6, d→u_1, e→u_6,\}. As long as the data center capacity limit is not exceeded, data \(c\) and data \(e\) can be deployed to the same data center \(u_6\).

After the completion of data transmission deployment in enterprise information construction, the reliability analysis is carried out.

### 3 Reliability analysis of enterprise information construction

The reliability of enterprise information construction is analyzed from the path loss of data transmission channel. Path loss is defined as the difference between the effective transmit power and the received power, which represents the attenuation of the signal, and the unit is a positive value of dB. Path loss consists of three parts: transmission loss, shadow fading, reflection loss and diffraction loss.

In practical work, such as deploying a data transmission channel, in order to ensure the reliability of the enterprise information process, the channel is generally calculated. One of the important contents is to calculate the transmission loss of the channel to measure the degree of path loss in the transmission process of data. With the increasing distance between the
transmitting end and the receiving end, the intensity of electromagnetic wave will continue to
decline. The path loss in free space refers to the attenuation degree of the transmitting
electromagnetic wave when there is a completely unobstructed sight distance path between the
receiving end and the transmitting end. The so-called free space refers to an infinite space full
of even and lossless medium, which is isotropic, with conductivity of 0 and average relative
permittivity and relative permeability of 1. Assuming that the distance between the sender and
the receiver in the free space is $\eta$, the input power of the data transmission process is $W_1$, the
output power is $W_2$, and $\sigma$ represents the wavelength, the transmission loss is:

$$ Y_s = \frac{w_1}{w_2} = \left( \frac{4\pi\eta}{\sigma} \right)^2 \rho $$

In the process of data transmission, the free space is a vacuum, and the transmission loss
is the loss caused by the natural diffusion of energy with the increase of transmission distance.

The shadow effect is the loss caused by the influence of natural or human factors in the
enterprise information construction. Shadow effect is easy to cause the level fading of sender
and receiver, which is called shadow fading. The probability of shadow fading is as follows:

$$ L(\psi \geq x) = \frac{1}{c\sqrt{2\pi}} \int_0^x 2e^{-u^2} du \rho $$

In the formula, $\psi$ represents the decibel of the fading amount, and $c$ represents the
standard deviation of the fading amount.

In the construction of enterprise information, reflection, diffraction and scattering will
affect the transmission. In the process of data transmission, reflection occurs on the wall or
obstacle surface. In the wireless channel, the influence of the ground reflection should also be
considered. The ground reflection dual line model is shown in Figure 2 below.

![Figure 2 Two line model of ground reflection](Image)

From the two line model of ground reflection:
In the formula, $h_1$ represents the distance between the transmitting end and the ground, and $h_2$ represents the distance between the receiving end and the ground. $Y_f$ is the path loss during reflection.

Diffraction occurs when the moving path between the transmitter and the receiver is blocked by sharp edges. The secondary wave generated by the barrier surface spreads in the space, even around the back of the barrier. If the secondary wave bypasses the barrier and reaches the shadow area, this area is called "Fresnel area".

\[ Y_f = w_1 \frac{h_1^2 h_2^2}{\eta} \quad (4) \]

$w_1$ in the formula represents the path loss caused by diffraction.

The transmission loss, shadow fading probability and path loss in the process of reflection and diffraction are taken as the basis of reliability analysis. The standard range of path loss is determined according to the actual project size of enterprise informatization, so as to judge the reliability of enterprise informatization construction and complete reliability analysis.

4 Design and analysis of simulation experiment

4.1 Experimental environment

The hardware platform includes AT91 RM9200 development board, Siemens MC55 GPRS wireless communication module, and Linux host which is easy to develop and compile. While the method of building the software platform is embedded Linux as the experimental environment, the statistics of the experimental results are based on Windows system and the
third party software.

The experiment mainly includes the packet loss in the process of data transmission. Before data transmission, basic settings must be completed. The client / server structure mode is adopted in the communication. The ARM board is used as the client PC as the Server, and the TCP protocol is used for data transmission. First, download the compiled program from the host computer to the development board, and set the corresponding execution authority. Set the IP in the Ping function to 211.129.11.102, the IP of the analog front-end computer to 211.86.39.54, and the port number to 80801.

4.2 Experiment setup

The network topology between the experimental data centers is shown in Figure 4. After the unified simulation of the data deployment location, the unit data transmission cost of the link between data centers.

![Experimental network topology](image)

**Figure 4** experimental network topology

As shown in Figure 4, it contains 11 data center nodes and 17 bidirectional links. For each bidirectional link, the free link capacity of the uplink and downlink is the same, the free bandwidth resource of each link is 10GB, and the transmission cost of each link is shown in Table 3. The free storage space of data center nodes is 100TB.

<table>
<thead>
<tr>
<th>Link</th>
<th>Overhead</th>
<th>Link</th>
<th>Overhead</th>
</tr>
</thead>
<tbody>
<tr>
<td>{a₁, a₂}</td>
<td>30</td>
<td>{a₅, a₁₀}</td>
<td>25</td>
</tr>
<tr>
<td>{a₁, a₄}</td>
<td>3</td>
<td>{a₃, a₆}</td>
<td>1</td>
</tr>
<tr>
<td>{a₁, a₇}</td>
<td>25</td>
<td>{a₆, a₁₁}</td>
<td>50</td>
</tr>
<tr>
<td>{a₂, a₄}</td>
<td>6</td>
<td>{a₇, a₅}</td>
<td>10</td>
</tr>
<tr>
<td>{a₂, a₃}</td>
<td>15</td>
<td>{a₇, a₆}</td>
<td>8</td>
</tr>
<tr>
<td>{a₃, a₆}</td>
<td>35</td>
<td>{a₈, a₉}</td>
<td>35</td>
</tr>
<tr>
<td>{a₄, a₇}</td>
<td>15</td>
<td>{a₉, a₁₀}</td>
<td>25</td>
</tr>
</tbody>
</table>
The scale of the data set used in the experiment increased from 10 to 100, with a total of 10 groups. The size of each data set is randomly generated between 0.1tb and 1TB, obeying the law of uniform distribution, and the traffic between data is randomly generated between 0.1gb and 1GB, obeying the law of uniform distribution.

4.3 Experimental results and analysis

In the above experimental environment and experimental settings, the experimental results of different methods are shown in Figure 5.

<table>
<thead>
<tr>
<th>General method 1</th>
<th>General method 2</th>
<th>Design method</th>
</tr>
</thead>
<tbody>
<tr>
<td>98.75%</td>
<td>43.26%</td>
<td>64.83%</td>
</tr>
</tbody>
</table>

Figure 5 experimental comparison results of different methods

Observing the results in Figure 5, it is shown that data transmission deployment and reliability analysis are carried out by using conventional methods 1 and 2, and packet loss rate is low; while data transmission deployment and reliability are carried out by using the designed method, no packet loss occurs, and the packet loss rate is 0%. Compared with different methods, the design of data transmission deployment and reliability analysis method is better than the conventional method.

5 Conclusion

In the enterprise information construction, data transmission deployment is an important part. Its good or bad deployment directly affects the completion of enterprise information and the subsequent development of the enterprise. It is of great significance for the development of the enterprise to use certain means to analyze the reliability of enterprise information.
construction and strictly control the quality of enterprise information construction. In view of the problems existing in the traditional methods, after the completion of the method design in this paper, a comparative experiment is designed to prove that the design method solves the problems existing in the traditional methods and provides some help for the enterprise information construction.

In the future, the wireless network will deploy more than 10 times more wireless nodes than the existing stations. Within the coverage area of the macro station, the distance between stations will be kept within 10 m, and support 25000 users in every 1km2. At the same time, the ratio of the number of active users and the number of stations may reach 1:1, that is, users and service nodes correspond one by one. The densely deployed network shortens the distance between terminals and nodes, greatly improves the power and spectrum efficiency of the network, expands the network coverage, expands the system capacity, and enhances the flexibility of services in different access technologies and coverage levels. Although the super dense heterogeneous network architecture has a great development prospect in 5g, the reduction of the distance between nodes, the more dense network deployment will make the network topology more complex, which can be further studied in the future.

Reference

Signal-level Honeypot: A Covert Communication and Interference Collection System

Cheng Chang¹, Xingsheng Zhu², Yue Gu³, Zhijun Deng⁴, and Sheng Liu⁵
{loebx@163.com¹, zhuxsh@hotmail.com², guyue01@163.com³}
China Academy of Launch Vehicle Technology, 1 South Dahongmen Road, Beijing, China¹,²,³

Abstract. Honeypot is originally a network trap that can attract hostile attackers and collect their attack behaviors to protect the real cyber systems and resources. Signal-level honeypot is a covert communication system with signal-level trap, which utilizes the basic idea of traditional honeypot and transform domain communication technology to attract and confront hostile interferences for reliable communication and interferences recording. In the transmitter, actual modulated signal is hidden underneath a well-camouflaged “target” signal. The actual modulated signal is designed to be noise-like, low power spectrum density, and orthogonal with the “target” signal to engage covert communication. In the receiver, a band-pass transform domain filter is used to separate signals to demodulate the actual modulated signal and collect the interferences. The proposed system can supply a high reliable communication approach with an “active” passive defense mode.

Keywords: Signal-level honeypot, Transform domain communication system, Covert communication, Reliable communication, Interferences collection

1 Introduction

With the fast development of information technology, wireless communications bring infinite convenience to the world. However, corresponding countermeasures are also continuously evolving, wireless communication systems have been plagued by various interferences and invasions [1]. Signal-level reliability that involves physical layer and data link layer is the foundation of communication systems. Most of the existing signal-level reliability measures are sorted into resisting, hiding, offsetting, and eluding 4 categories [2]. Resisting measures utilize power benefit to extract data signals from received signals, such as direct sequence spread spectrum technology [3]; hiding measures conceal data signals or their characteristics, such as embedding and noise-like technologies [4]; offsetting measures can restrain interferences or reinforce data signals, such as interference cancellation technology [5]; eluding measures utilize orthogonal or other characteristics to ensure the data signal and interferences being separated [6]. Nowadays, wireless communications can be regarded as a gaming between users and attackers. Therefore, some more active measures should be considered to achieve reliable communications.

Honeypot technology is an active cyber resources protection technology, which can build a fraudulent cyber environment to attract hostile cyber attackers to detect, attack, and capture it, so as to record their hostile behaviors [7]. Honeypot itself can be regarded as a strictly monitored computing resources. Every accessing behavior is suspectable, and the value of the
honeypot is measured by the recorded information. Honeypot technology is widely used in computer network and its related field to collect cyber-attacks [8]. However, it has not received intensive consideration in the literature for signal-level reliability.

In this paper, transform domain communication technology [9] is combined with the basic idea of traditional honeypot to propose a signal-level honeypot to attract hostile interferences for reliable communication and interferences collection. In the transmitter, the actual modulated signal is hidden underneath a well-camouflaged “target” signal. The actual modulated signal is designed to be noise-like, low power spectrum density (PSD), and orthogonal with the “target” signal to engage covert communication. In the receiver, a band-pass transform domain filter is used to separate two signals to demodulate the actual signal and collect the interferences. The differences between traditional honeypot and the proposed system is showed in Table 1. The proposed system can supply a high reliable communication approach with an “active” passive defense mode. The next section of this paper will briefly review transform domain communication technology and the basic idea of honeypot. The signal-level honeypot is proposed detailedly in Section 3. Simulations and analysis are presented in Section 4. The paper is then concluded in Section 5.

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Traditional honeypot</th>
<th>Signal-level honeypot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose</td>
<td>Recording cyber-attacks, protecting cyber systems and resources</td>
<td>Covert communication, recording signal-level interferences and invasions</td>
</tr>
<tr>
<td>Action range</td>
<td>From network layer to application layer</td>
<td>Physical layer and data link layer</td>
</tr>
<tr>
<td>Entity</td>
<td>Trap</td>
<td>Communication system with trap</td>
</tr>
</tbody>
</table>

2 Preliminary

2.1 Transform domain communication technology

Transform domain communication system (TDCS) provides reliable communications with spectrum spreading in the unoccupied frequency bins of the real-time environment [10]. The transmitter senses spectrum to get spectrum mask $A(k)$, which is a 1-D matrix composed by 0 and 1 if the $k$ th frequency bin is occupied or unoccupied. The spectrum mask and a pseudo-random $\theta_i$ are applied element by element to get frequency domain basis waveform $B(k)$. After inverse fast Fourier transform and normalization, $B(k)$ turns to time domain basis waveform $b(n)$, which is used to generate modulating symbols with cyclic code shift keying (CCSK). Then data is modulated in Gray code. The $i$ th transmitting symbol is deduced as

$$s_{TDCS, i}(n) = \frac{1}{\sqrt{NN}} \sum_{k=0}^{K-1} A(k) e^{j\theta_i} e^{-j2\pi nk/M} e^{j2\pi n/N}$$

(1)
In Equation (1), \( N \) and \( N_1 \) are the numbers of the total and the unoccupied frequency bins, respectively. \( m_i \in [1, M] \) is the \( i \) th transmitting data. In the receiver, local modulating symbols are generated as the transmitter, data is demodulated with maximum peak detection of the correlations between the received signal and local modulating symbols.

### 2.2 Honeypot technology

The protection process of the traditional honeypot is generally divided into 3 stages. The first stage is trap construction. Fraudulent data and files are built to improve the “sweetness” of the trap to attract cyber-attackers and facilitate interactions. The degree of the interactions depends on the fidelity between the trap and the actual system. The second stage is intrusion behavior detection and recording. Specific objects such as flows, ports, permissions, bugs, and documents are monitored and recorded to prevent damages. The last stage is post processing. The records of the attack behaviors are processed with data visualization, flow analysis, attack identification, alert generation, and traceability to supply further improvement for the actual systems.

### 3 Signal-level honeypot model

#### 3.1 System structure

The diagram of the signal-level honeypot is showed in Fig. 1. The transmitter generates a false “target” signal and conceals the modulated signal beneath it. “Target” mask \( A_i(k) \) is the same as spectrum mask in TDCS, with \( k = 1 \) representing the spectrum bins that the “target” signal occupied. To improve the “sweetness” of the trap, “target” signals \( b_1(n) \) are composed by common QPSK modulated signals. Transmitting mask \( A_i(k) \) is complementary with the “target” mask to generate the \( i \) th modulated signal \( b_{2,i}(n) \). As shown in Equation (2) and (3), \( I(k) \) is an all 1 matrix, \( N_2 \) is the number of 1 in \( A_i(k) \). Then the “target” and the modulated signals are superimposed with power adjustment to control the covert communication ability. The transmitting signal is showed as Equation (4), \( \gamma \) is the factor of the power adjustment.

\[
\begin{align*}
A_i(k) + A_i(k) &= I(k) \quad (2) \\
b_{2,i}(n) &= \frac{1}{\sqrt{NN_2}} \sum_{k=0}^{N-1} A_i(k)e^{j2\pi kn}/M e^{j2\pi kn}/N \quad (3) \\
x_i(n) &= \gamma b_{2,i}(n) + b_1(n) \quad (4)
\end{align*}
\]

In the receiver, received signal is firstly passed through a transform domain filter, whose frequency range coincides exactly with predetermined “target” mask \( A_i(k) \), to separate the received “target” part \( c_i(n) \) and received data part \( c_{2,i}(n) \). Then the “target” part is recorded.
and analyzed, while the data part is demodulated with maximum peak detection of the correlations between the received signal and local modulating symbols. In Equation (5) and (6), $C_i(k)$ and $B_r(k)$ are the frequency forms of the received data part and local generated modulating signal $b_n(n)$. $\text{conj}()$ and $\text{real}()$ are the conjugate and the real part of the complex signal.

\[
R_i(\tau) = \text{IFFT}[C_i(k) \cdot \text{conj} (B_r(k))] = \frac{1}{\sqrt{N_2}} \sum_{i=0}^{N-1} e^{-j2\pi \tau (r-c) / N}
\]

\[
data i = \arg \max_{r \in \{0,1,...,N-1\}} \text{real} (R_i(\tau))
\]  

**Fig. 1.** The diagram of signal-level honeypot

### 3.2 Performance analysis

According to the Shannon theory [11], transmission rate is in direct proportion to available bandwidth and signal noise ratio (SNR). For the power of $b_n(n)$ is equally distributed in unoccupied frequency bins, the wider bandwidth of the modulated signal, the lower PSD of the system can achieve.

\[
PSD(b_n(n)) \propto \frac{1}{W} \propto \frac{1}{N_2}, R_i \propto W \propto N_2
\]
Time domain low detection performance depends on the performance of the pseudo-random sequences. The longer sequence, the better performance system achieved [12].

Bit error rate (BER) with certain transmission rate and additive white Gaussian noise (AWGN) is the key indicator of the proposed system. The lower BER, the better transmitting performance achieved.

Bandwidth of the “target” signal is used to evaluate the collection range. The wider bandwidth of the “target” signal, the larger of the collection range, but the narrower bandwidth of the modulated signal, the lower transmission rate of the covert communication. Interference identification is not considered in this paper.

4 Simulations

To verify the proposed method, a semi-physical system is built. The “target” signal is generated by a signal source with QPSK modulation of 8MHz bandwidth. The modulated signal is generated by a modified TDCS transmitter with CCSK modulation of 62MHz bandwidth to fully use total 70MHz bandwidth. The original output powers of the two signals above are the same, and the power adjustment factor γ is set as -35dB. In other words, the power of the “target” signal is 35dB larger than that of the modulated signal. A modified TDCS receiver samples with 8 times oversampling, 560MHz, to obtain the received signal in Fig. 2. In consideration of the path loss and the sensitivity of the receiver, actual received signal power is greatly lowered beneath the noise.

![Fig. 2. The frequency domain form of the transmitting signal](image)

The BER of the proposed method with different power adjustment factors is showed in Fig. 3. The modulation is set as 64-ary CCSK. A BER without “target” signal is run as the
reference. Power adjustment factor $\gamma$ is set as -20dB, -30dB, and -40dB to get 3 typical BERs with AWGN. When $\gamma$ is -20dB, “target” signal and AWGN collectively influence BER, system needs extra 2.5dB SNR to compensate the BER deterioration caused by the “target” signal at BER=10^{-4}. When $\gamma$ is -30dB and -40dB with $E_b/N_0$ less than 5dB, “target” signal and AWGN collectively influence BER. If $E_b/N_0$ is more than 6dB, system cannot fully compensate the BER deterioration, platforms appear at BER=10^{-3}. Therefore, when $\gamma$ is more than -20dB, the proposed system can well achieve covert communication and interferences collection.

![Fig. 3. The BER of the proposed method with different power adjustment factors](image)

5 Conclusions

This paper introduces a signal-level honeypot, which utilizes the basic idea of traditional honeypot and transform domain communication technology to attract and confront hostile interferences for reliable communication and interferences collection. The proposed system can achieve low detection probability with time domain noise-like and frequency domain low PSD. Besides, by sacrificing affordable $E_b/N_0$, the system can process wideband signal trap to collect hostile interferences. The proposed system supplies a high reliable communication approach with an “active” passive defense mode.

References

Research on data transmission encryption algorithm of wireless sensor network in cloud storage

ZHOU Xian-chun1, WU Qi-xiang2
{wangyang19111@163.com, ainiyiwantian255@tom.com}

(1. School of Information and Intelligence Engineering, Sanya University, Hainan Analysis and Application Engineering Research Center of Software Gene Virus, Sanya 572022, China;
2. College of Information Engineering, Fuyang Normal University, Fuyang 236041, China)

Abstract: In cloud storage, the traditional wireless sensor network data transmission encryption algorithm has the problem of low pseudo-randomicity of key flow, resulting in weak encryption security. Therefore, a data transmission encryption algorithm for wireless sensor network in cloud storage is proposed. Formulate the encryption strategy and the initial key strategy, use the basic mode of developing the debugging board to execute the encryption algorithm, and control the input of the encryption algorithm. Control the iteration times of the algorithm, encrypt the data transmission through the encryption wheel transformation and key expansion, and use the pseudo-code to execute the encryption algorithm to complete the data transmission encryption. The test results show that compared with the traditional encryption algorithm, the key flow of wireless sensor network data transmission encryption algorithm designed for cloud storage has higher pseudo-randomicity and stronger security.

Keywords: cloud storage; data transmission; encryption algorithm;

1 Introduction

Cloud storage technology uses the Internet to get the services needed by users with the advantages of on-demand, dynamic and easy expansion. As an online network service, it uses server cluster application technology, grid technology and distributed file system technology to gather data centers for providing storage services through the Internet[1]. So that these data centers can jointly provide users with data storage and business access services[2]. Cloud storage technology is a new technology developed from parallel storage, distributed storage and grid storage. It has great development potential and wide application prospect.

On this basis, the wireless sensor network can realize the base wood functions such as data acquisition, data processing and sending and gathering[3]. Due to the large data processing density and forwarding volume of wireless sensor networks, the limited energy of a single sensor, and the outsourcing storage mode of cloud storage, it is likely to expose some or all of the data to cloud storage service providers or privileged users. Privileged users have the ability of unauthorized access to users' private data, which is likely to be used by cloud storage service providers, or data leakage due to poor internal management of service providers, the
security of user sensitive data has been completely out of the control of the data owner, which is likely to lead to internal attacks such as the disclosure of user data and privacy information.

For the above problems, literature [4] studies the node encryption of wireless sensor network, first introduces the current situation of wireless sensor network encryption. Then, based on the study of S-box and chaos theory, a multi chaos S-box encryption algorithm is proposed, the mapping relationship between them is established, and the constant chaos iteration is realized. This paper has a good encryption effect. However, this encryption algorithm is easily affected by the amount of wireless sensor network data in cloud storage, which makes the generated key stream less pseudo-random and less secure. Therefore, this paper proposes the research of data erasure hybrid encryption algorithm for wireless sensor network in cloud storage to solve the problems existing in the traditional algorithm.

2 Design of data transmission encryption algorithm for wireless sensor network in cloud storage

2.1 Develop encryption strategy

In the data transmission of wireless sensor network in cloud storage, the gateway server encrypts the transmission data packet and uses the basic mode of ssx31-b development and debugging board to implement the encryption algorithm[5]. The settings of each bit segment of PE control word of ssx31-b security chip are shown in Table 1.

<table>
<thead>
<tr>
<th>BIT</th>
<th>Function</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>0~6</td>
<td>Keep</td>
<td></td>
</tr>
<tr>
<td>7~8</td>
<td>Packet processing mode</td>
<td>00</td>
</tr>
<tr>
<td>9~24</td>
<td>Keep</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Send data packet encryption processing, position is 1;</td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>Receive packet decryption processing, position is 2</td>
<td>0 or 1</td>
</tr>
<tr>
<td>26</td>
<td>Operating mode</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>Output message</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>Keep</td>
<td></td>
</tr>
</tbody>
</table>
In the kernel encryption module of the driver, using the kernel interface of ssx31b, the basic encryption and decryption based on the encryption algorithm is carried out for the data segments of IP packets. The SA function of ssx31b security chip is to specify specific encryption algorithm and corresponding key\(^6\). The encryption card completes the corresponding encryption processing according to the specified SA. Table 2 shows the settings of each bit segment of SA command word of ssx31b security chip.

<table>
<thead>
<tr>
<th>BIT</th>
<th>Function</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>0~9</td>
<td>Keep</td>
<td></td>
</tr>
<tr>
<td></td>
<td>00: ECB</td>
<td></td>
</tr>
<tr>
<td></td>
<td>01: CBC</td>
<td></td>
</tr>
<tr>
<td>10~11</td>
<td>10: AES_CTR</td>
<td>01</td>
</tr>
<tr>
<td></td>
<td>11: Keep</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0000: DES</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0001: 3DES</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0010: AES128</td>
<td></td>
</tr>
<tr>
<td>12~15</td>
<td>0011: AES192</td>
<td>0001</td>
</tr>
<tr>
<td></td>
<td>0100: AES256</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1111: NULL</td>
<td></td>
</tr>
<tr>
<td></td>
<td>000: HMAC-MD5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>001: HMAC-SHA-1</td>
<td></td>
</tr>
<tr>
<td>16~19</td>
<td>010: MD5</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>011: SHA-1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>100: NULL</td>
<td></td>
</tr>
<tr>
<td></td>
<td>111: NULL</td>
<td></td>
</tr>
<tr>
<td>19~31</td>
<td>Keep</td>
<td></td>
</tr>
</tbody>
</table>

The SA structure contains eight doubleword key fields. For DES key, only key is used; for 3DES key, key ~ Key3 is used; for AES key, all 8 doublewords of key domain may be
used.

In the encrypted IP packets, an encryption mark should be added to identify the encrypted IP packets, but the mark should not change the data segments in the IP packets, so that the gateway server receiving the IP packets can identify and decrypt them.

The implementation of data transmission encryption in wireless sensor network is inseparable from the key. Under the above encryption strategy, the initial key strategy is formulated\(^7\). A random initial key obtained by both sides of data transmission is the security foundation of the whole encryption process. Random selection of a large prime number \(Q\), Smaller number \(q\), random number \(i\), Public key obtained \(k = q^i \mod Q\), Large prime number \(Q\) and Smaller number \(q\) through negotiation between both parties of data transmission. The server randomly generates a smaller number \(q\) calculation \(k = q^i \mod Q\), Client randomly generated \(i\), and calculate \(k' = q^i \mod Q\), exchange between two parties \(k\) and \(k'\). Successfully obtained the same random key. Because there are only large prime numbers \(Q\), Smaller number \(q\), \(k\) and \(k'\) may be intercepted by eavesdroppers, So unless privileged users can compute discrete logarithm to recover random number \(i\) and \(i\). Otherwise, the computer will not be able to exit \(k\) and \(k'\). Therefore \(k\) and \(k'\) can be used in encryption algorithm as the same secret key which is distributed to both sides of communication.

2.2 Control the input of encryption algorithm

Set an iterative block cipher. The length of plaintext block and key is one of the following three values: 128bit, 192bit, 256bit. The encryption algorithm is run on a 4x4 matrix called "state". Every step of the algorithm acts on this matrix. The plaintext packet and the key are expressed as 4x \(H_a\) respectively matrix of 4x \(H_b\), \(H_a\) and \(H_b\) one thirty second of plaintext packet and key length respectively\(^8\). For plaintext packets or key matrices, each element corresponds to a value on GF (28). The number of iteration rounds of encryption
algorithm $H_c$ is related to $H_a$ and $H_b$.

In the encryption process, the input of the control encryption algorithm is carried out on the 4X4 matrix, and the plaintext grouping and key length are controlled at the same time, so as to control the number of iterations of the algorithm and avoid too many times making the algorithm too heavy.

### 2.3 Encryption round transformation and key extension

The data transmission encryption process of cloud storage wireless sensor network is determined by the round transformation function, which consists of four modules: byte transformation, row displacement, column obfuscation and round key addition\(^9\). Each module performs different operations on the state matrix. The radiative transformation formula of state matrix is as follows:

\[
\begin{bmatrix}
  u_0 \\
  u_1 \\
  u_2 \\
  u_3 \\
  u_4 \\
  u_5 \\
  u_6 \\
  u_7 \\
\end{bmatrix} = \begin{bmatrix}
  1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
  1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \\
  1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\
  1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\
  1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\
  0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\
  0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\
  0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\
\end{bmatrix}
\begin{bmatrix}
  v_0 \\
  v_1 \\
  v_2 \\
  v_3 \\
  v_4 \\
  v_5 \\
  v_6 \\
  v_7 \\
\end{bmatrix} + \begin{bmatrix}
  1 \\
  1 \\
  0 \\
  0 \\
  0 \\
  1 \\
  1 \\
  0 \\
\end{bmatrix} \tag{1}
\]

The above affine transformation is composed of two transformations. The two transformations are implemented with the S-box of 1-byte input / 1-byte output, and the principle is shown in the figure below.

![S-BOX](image)

**Figure 1** Schematic diagram of byte realization nonlinear transformation

When decrypting, the module corresponding to SubBytes is called InvSubBytes, and
they are the inverse process of each other. The specific process is as follows: For each byte in the state matrix, a new byte is obtained by using the inverse affine change, and then it is multiplied by the inverse, so that the inverse S-box can be obtained. Finally, each byte of the state matrix is transformed by using the inverse S-box\[10\].

Row shift is mainly a cyclic shift for each row of the state matrix, and each row shift is different. So what I'm going to do is, line 0 doesn't change, line 1 goes around by $a_1$ to the left. Row 2 loops $a_2$ to the left. Line 3 loops $a_3$ to the left. Among them, the selection of displacement $a_1, a_2$ and $a_3$ is related to $H_a$.

When decrypting, the module corresponding to ShiftRow is called InvShiftRows, and they are the inverse process of each other, so the process is exactly the same.

The third step is the mixed transformation of the columns. Each column on the state matrix is treated as a polynomial, where each coefficient of the polynomial is on GF(28) and the degree of the polynomial is less than 4. Then, the polynomial and the expression are:

$$a(v) = '03'v^3 + '01'v^2 + '01'v + '02'$$

In the formula, the coefficient is represented by base 8, $(x)='03'x3+'01'x2+'01'x+ '02'$ reversible polynomials multiply over $v^4 +1$. We get the final result. The specific transformation process is shown below

$$\begin{bmatrix}
 u_0 \\
 u_1 \\
 u_2 \\
 u_3 \\
\end{bmatrix} =
\begin{bmatrix}
 02 & 03 & 01 & 01 \\
 01 & 02 & 03 & 01 \\
 01 & 01 & 02 & 03 \\
 03 & 01 & 01 & 02 \\
\end{bmatrix} \begin{bmatrix}
 v_0 \\
 v_1 \\
 v_2 \\
 v_3 \\
\end{bmatrix}$$

The sum in formula (3) represents the terms on the state matrix of $u$ and $v$ in the column mixture. Its decryption and row displacement decryption transformation mode is the same.

The round key encryption transformation is just to add the round key and the state matrix. The addition refers to the operation of different fields on GF (28). The round key is obtained through key extension, and the specific extension process is shown in the figure below.
Iteration through the seed key produces a key with a total length of \(4H_a (H_a + 1)\) bytes required for the entire encryption. So you can think of the key as the number of groups from \(*N_1,+1\). Each of these elements is 4 bytes in size. Through the above process encryption wheel transformation and key expansion, for the subsequent implementation of encryption preparation.

2.4 Encryption of data transmission

The key distribution strategy is adopted to randomly initialize the key, and the public key of the key exchange is generated through the initialization function of the server in the wireless sensor network, and the public key data stored in the pointer variable is assigned to the public key array to facilitate communication and transmission. The public key data is transmitted to the client, and at the same time the server receives the array of the client public key from the client. The public key of the client is used to generate the Shared key, and the wireless sensor network data transmission is encrypted through the encryption wheel transformation process mentioned above.

3 Simulation test and analysis

3.1 test preparation

The pseudo-randomicity of the key flow is tested by test simulation. In the test, MATLAB tool and in the Windows 7 environment were used to obtain data groups of different data sizes to test and analyze the encryption algorithm.

In the above environment, multiple sets of test data prepared are used to more accurately analyze the pseudo-random test results of key flows of different encryption algorithms.
3.2 The test data

Using the open source text data set SemEval, the relevant parameters of the ten sets of data required for the test are shown in the table below.

<table>
<thead>
<tr>
<th>name</th>
<th>The amount of data (MB)</th>
<th>Number of text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data-1</td>
<td>26</td>
<td>5</td>
</tr>
<tr>
<td>Data-2</td>
<td>54</td>
<td>9</td>
</tr>
<tr>
<td>Data-3</td>
<td>147</td>
<td>17</td>
</tr>
<tr>
<td>Data-4</td>
<td>89</td>
<td>8</td>
</tr>
<tr>
<td>Data-5</td>
<td>441</td>
<td>23</td>
</tr>
<tr>
<td>Data-6</td>
<td>562</td>
<td>16</td>
</tr>
<tr>
<td>Data-7</td>
<td>631</td>
<td>29</td>
</tr>
<tr>
<td>Data-8</td>
<td>847</td>
<td>41</td>
</tr>
<tr>
<td>Data-9</td>
<td>906</td>
<td>53</td>
</tr>
<tr>
<td>Data-10</td>
<td>547</td>
<td>21</td>
</tr>
</tbody>
</table>

Using the above ten sets of data to test the pseudo-random strength of the generated key flow under different encryption algorithms.

3.3 Test results and analysis of pseudo-randomness of key flow

MATLAB software is used to simulate the run test. The purpose is to test the pseudo-randomness of the key flow by calculating the number of uninterrupted sub-sequences composed of the same bits in the key flow sequence.

The test group using the designed encryption algorithm was set as the experimental group, and the test group using the traditional encryption algorithm was set as the control group. The test results are as follows:

<table>
<thead>
<tr>
<th>Data set</th>
<th>The amount of data (MB)</th>
<th>Statistic ( \delta )</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26</td>
<td>0.5006</td>
<td>0.3056</td>
</tr>
<tr>
<td>2</td>
<td>54</td>
<td>0.4997</td>
<td>0.4125</td>
</tr>
<tr>
<td>3</td>
<td>147</td>
<td>0.4996</td>
<td>0.1741</td>
</tr>
<tr>
<td>4</td>
<td>89</td>
<td>0.4991</td>
<td>0.2635</td>
</tr>
<tr>
<td>5</td>
<td>441</td>
<td>0.5001</td>
<td>0.2584</td>
</tr>
<tr>
<td>6</td>
<td>562</td>
<td>0.5003</td>
<td>0.2633</td>
</tr>
<tr>
<td>7</td>
<td>631</td>
<td>0.5007</td>
<td>0.3147</td>
</tr>
</tbody>
</table>
Table 5: Test results of control group

<table>
<thead>
<tr>
<th>Data set</th>
<th>The amount of data (MB)</th>
<th>Statistics δ</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26</td>
<td>0.5011</td>
<td>0.1746</td>
</tr>
<tr>
<td>2</td>
<td>54</td>
<td>0.4983</td>
<td>0.1596</td>
</tr>
<tr>
<td>3</td>
<td>147</td>
<td>0.4991</td>
<td>0.1006</td>
</tr>
<tr>
<td>4</td>
<td>89</td>
<td>0.4995</td>
<td>0.1079</td>
</tr>
<tr>
<td>5</td>
<td>441</td>
<td>0.5021</td>
<td>0.0214</td>
</tr>
<tr>
<td>6</td>
<td>562</td>
<td>0.3019</td>
<td>0.0365</td>
</tr>
<tr>
<td>7</td>
<td>631</td>
<td>0.2027</td>
<td>0.0694</td>
</tr>
<tr>
<td>8</td>
<td>847</td>
<td>0.2015</td>
<td>0.0613</td>
</tr>
<tr>
<td>9</td>
<td>906</td>
<td>0.1087</td>
<td>0.0749</td>
</tr>
<tr>
<td>10</td>
<td>547</td>
<td>0.3092</td>
<td>0.0893</td>
</tr>
</tbody>
</table>

The statistical value δ in the table represents the proportion of statistical 0 and 1 in the whole sequence, and the standard is \( |δ - 0.5| < 0.1 \). In this interval, the smaller the value is, the stronger the pseudo-randomicity of the key flow generated by the encryption algorithm is and the higher the security is. P-value represents the distribution of 0 and 1 in the sequence, and the standard value is P-value < 0.1.

According to the results in the above standard observation table, table 4 shows that for different data volumes, the statistic δ is always in the standard range. P-values are all greater than 0.1. As shown in table 5, if the data volume is less than 500MB, the statistical values δ and p-value obtained in the test are within the normal range. However, as the data volume increases and exceeds 500MB, the statistical values δ and p-value are abnormal, which are not within the standard range. Combined with the above results, the proposed encryption algorithm for wireless sensor network data transmission in cloud storage has higher pseudo-randomicity of key flow and stronger security, which is better than the traditional data transmission encryption algorithm.

4 Conclusion
With the development of computer and Internet technology, cloud storage is applied in more and more fields. All kinds of important data are saved and transmitted in the form of cloud data. Due to the divergence and randomness of the Internet, data are copied, tampered and forged in the process of transmission, and data security is seriously threatened. Therefore, an encryption algorithm for data transmission in wireless sensor network in cloud storage is proposed. Through comparison and simulation test, it is verified that the designed encryption algorithm can effectively solve the problems existing in the traditional encryption algorithm and ensure the security of data information in data transmission. As an important research hotspot, data transmission security has great economic benefits and social significance.

5 Fund projects
Sanya City Science and Technology Cooperation Project 2019YD26

Reference
Analysis of Transmission Delay of Wireless Sensor Network Based on GNSS Signal

LIU Rong¹, XUE Fang²
{wz26677@163.com¹, aaa21152@163.com²}
¹School of Information Engineering, Changsha Medical University, Changsha 410219, China;
²Information Technology Center, Jimei University, XiaMen 361021, China

Abstracts: Aiming at the problem that the traditional network transmission delay analysis method occupies too much memory of the operating platform during operation, a GNSS (Global Navigation Satellite System) signal-based transmission delay analysis method for wireless sensor networks is proposed. Establish a cluster structure and use the method based on network fusion and clustering algorithms to perceive the transmission delay of wireless sensor networks. Delay parameterized processing of GNSS signals, and the establishment of a wireless sensor network data transmission model, complete the study of wireless sensor network transmission delay analysis methods. The comparison experiment between the design and the traditional delay analysis method proves that the GNSS signal-based delay analysis method occupies less memory of the operating platform and performs better under the same task.

Keywords: GNSS signal; Wireless sensor network; Transmission delay; Analytical method

1 introduction

GNSS is the unified name for single satellite navigation and positioning of Beidou system, GPS (Global Positioning System), GLONASS (Global Navigation Satellite System), and Galileo system. It can also refer to the enhanced system of single satellite navigation and positioning, and also refers to the addition of all satellite navigation and positioning and its enhanced system. GNSS is a large system composed of multiple satellite navigation and positioning systems and their enhanced systems[1]. GNSS uses satellites as star-level radio navigation systems for navigation stations, and provides all-weather, high-precision position, speed, and time information for all types of military and civilian carriers on land, sea, air, and sky. With the progress and development of GNSS technology of global satellite
navigation system, the application fields of the system are becoming more and more extensive, especially in navigation and positioning and atmospheric measurement.

A wireless sensor network is a distributed sensor network. Its tip is a sensor that can sense and check the outside world. The sensor in wireless sensor network is wireless communication, and the network setting is flexible. It can change the location of the device at any time, or connect to the Internet by wire or wireless, and form a multi hop self-organized network through wireless communication [2]. There are two types of nodes in wireless sensor networks, one is the convergence node and the other is the sensor node. The convergence node mainly refers to the gateway's ability to remove false report data from the sensor nodes and combine it with related reports to fuse the data and judge the events that occur. The connection between the sink node and the user node can be directly communicated through a wide area network or satellite, and the collected data is processed [3]. Due to the placement of nodes in a wireless sensor network, there is a certain transmission delay between the networks. Network transmission delay is a phenomenon in which certain steps take longer than normal transmission time when data is transmitted in the network. Literature [4] method takes up too many resources of the operating platform when analyzing the network with too many nodes, which has great limitations, resulting in too much memory of the operating platform. To solve this problem, this paper proposes a transmission delay analysis method based on GNSS signal in wireless sensor network.

2 Analysis method of transmission delay of wireless sensor network based on GNSS signal

2.1 Wireless sensor network transmission delay perception

In order to analyze the transmission delay of wireless sensor networks and reduce the energy consumption of wireless sensor networks, this paper uses a method based on network fusion and clustering algorithm to detect the transmission delay of wireless sensor networks.

First, establish the cluster structure shown in the following figure. Nodes A and B are cluster members and node C is the cluster head.
A and B nodes transmit data packet $\alpha$ to cluster head C. Assume that the unit time required to transmit data $\alpha$ from one node to another node is $t_{TX}$, and the corresponding transmitter and receiver energy consumption is $e_{TX}$ and $e_{RX}$, respectively. To simplify the calculation process, energy consumption is ignored during data fusion. At the same time, the fusion center (FC) can be regarded as a device node with sufficient energy. If nodes A, B and C can be fused into a single data packet size $\alpha'$ ($\alpha' = \alpha$), the total delay $M$ and $N$ is the smallest.

When the output data $\alpha' > \alpha$, the delay and energy consumption of the wireless sensor network will increase. The calculation formula is as follows:

$$
\begin{align*}
    t_{TOT} &= 2t_{TX} + \frac{\alpha'}{\alpha} t_{TX} \\
    e_{TOT} &= 2(e_{TX} + e_{RX}) + \frac{\alpha'}{\alpha} e_{TX}
\end{align*}
$$

According to the above analysis process, assuming that there are $N$ wireless sensor nodes in the wireless sensor network, a network with $k$ clusters will be constructed. According to the basic properties of the tree structure shown in Figure 1, there are:

Fig. 1 Schematic diagram of cluster structure
In formula (2), \( m_i \) is the largest number of nodes in the \( i \)-th cluster, and can be expressed as:

\[
m_j = 2^j - 1 \quad (3)
\]

\( m_j = 2^j - 1 \) can be obtained using the recursive formula. Then formula (2) can be converted into the following form:

\[
2^{k-1} - 1 < N \leq 2^k - 1 \quad (4)
\]

Analysis of the cluster structure shown in Figure 1 shows that smaller clusters can shorten the data aggregation process to reduce the delay time [5]. Therefore, before the \( k \)-th cluster is filled with cluster members, the \( k-1 \)-th cluster has been filled with cluster members. As can be seen from formula (4), taking the number of cluster structure nodes \( N = 7 \) as an example, \( k = 3 \) clusters are constructed. This structure is a multilayer cluster structure constructed by multiple clusters, that is, a multi-hop structure exists in one cluster. Multi-hop may cause additional data fusion delay and network energy consumption.

Therefore, the aggregation structure shown in the figure below is improved, in which the virtual circle represents the cluster head node, the solid circle represents the cluster member node, FC represents the center node of cluster structure fusion, and the number represents the interleaving communication sequence of cluster head node and FC.
As can be seen from Figure 2, the improved structure has multiple clusters consisting of multiple single-hop cluster structures of different sizes, that is, there is no multi-hop structure in a cluster [6]. Based on the above analysis process, it can be determined that before the larger cluster is ready, the smaller cluster can complete communication with the FC. Therefore, the delay time of the entire wireless sensor network is determined by its largest cluster. Therefore, a clustering algorithm is used to cluster nodes in a wireless sensor network. When the largest cluster after clustering is determined, the perception of the transmission delay of the wireless sensor network is completed. When the transmission delay is perceived in the wireless sensor network, the delay parameter haul processing is performed on the GNSS signal.

### 2.2 GNSS signal delay parameter processing

When electromagnetic waves pass through the air, there will be a certain degree of refraction. Not only will the speed change, but the propagation path will also change due to bending. The propagation velocity $v$ of the electromagnetic wave in the medium can be expressed as:

$$v = \frac{c}{n} \tag{5}$$

In formula (5), $c$ is the speed of light in a vacuum, and the unit is m/s; $n$ is the refractive index of the medium. When the GNSS signal propagates in the air, the ranging code propagates at the group velocity $v_g$, and the carrier phase propagates at the phase velocity $v_p$.

When the signal frequency is $f$ and the wavelength is $\lambda$, the phase velocity $v_p$ and group velocity $v_g$ can be defined according to the following formula.

$$\begin{align*}
  v_p &= f \times \lambda \\
  v_g &= -\frac{df}{d\lambda} \times \lambda \tag{6}
\end{align*}$$
According to the well-known Appleton-Hartree formula, when electromagnetic waves propagate in wireless sensor networks, assuming that ionic collision and thermal motion are not considered, the phase refractive index can be expressed according to the following formula [7].

\[
\eta^2 = 1 - \frac{X}{1 - \frac{Y_f^2}{2(1-X)^2} + \sqrt{\frac{Y_f^2}{4(1-X)^2}}} \tag{7}
\]

In formula (7), it is the displacement transmitted in the horizontal direction in the wireless sensor network and the displacement transmitted in the vertical direction in the wireless sensor network. The calculation formula of the above three parameters is as follows:

\[
\begin{align*}
X &= \frac{e^2 N_e}{4\pi^2 \varepsilon_0 m \nu} \\
Y_f &= \frac{\mu_0 \varepsilon_0 (H_0 \sin \theta)}{2\pi m f} \\
Y_v &= \frac{\mu_0 \varepsilon_0 (H_0 \cos \theta)}{2\pi m f}
\end{align*}
\tag{8}
\]

In the above formula, \( e \) is the amount of charge carried by the electron, and the value of \( e = 1.6022 \times 10^{-19} \text{C} \); \( N_e \) is the electron density of the electromagnetic wave in the wireless sensor network, and its unit is \( \text{m}^{-3} \); \( \varepsilon_0 \) is the dielectric constant of the electromagnetic wave in the air; \( m \) is the mass of the electron, and the value of \( m = 9.1096 \times 10^{-31} \text{kg} \); \( \nu \) is the frequency of the electromagnetic wave signal; \( \mu_0 \) is the magnetic permeability of the electromagnetic wave in the vacuum; \( \mu_0 = 4\pi \times 10^{-7} \text{H/m} \) of the electromagnetic wave in the vacuum; \( H_0 \) is the magnetic field strength of the geomagnetic field; \( \theta \) is the angle between the direction of the geomagnetic field and the direction of propagation of the electromagnetic wave signal [8]. The above
analysis process is mathematically processed, and a physical quantity $T_c$ is introduced, where $T_c$ represents the total electron content of the GNSS signal transmitted in the wireless sensor network. The calculation formula is as follows:

$$T_c = \int_\alpha N_e ds$$ (9)

The unit of the physical quantity $T_c$ is unit / m2. The meaning of $T_c$ is the unit area as the base area. After the above processing, the parameter delay processing of GNSS signal transmission in wireless sensor network is completed. In order to analyze the transmission delay of wireless sensor networks, based on the above research, a data transmission model of wireless sensor networks based on Markov’s principle is established.

2.3 Building a data transmission model for wireless sensor networks

According to Markov’s principle, let $P$ be a matrix composed of transition probabilities $P_{ij}$ and state space $I = \{1, 2, \ldots\}$, then use the following formula to represent the transition probability matrix of the system state.

$$P = \begin{bmatrix}
P_{11} & P_{12} & \cdots & P_{1\rho} \\
P_{21} & P_{22} & \cdots & P_{2\rho} \\
\vdots & \vdots & \ddots & \vdots \\
P_{\rho1} & P_{\rho2} & \cdots & P_{\rho\rho}
\end{bmatrix} \text{ (10)}$$

$$P_{ij} \geq 0, i, j = 1, 2, \ldots, N$$

$$\sum_{j=1}^{N} P_{ij} = 1, i = 1, 2, \ldots, N$$

In the above formula, $P_{ij}$ is the conditional probability that the communication transmission state of the wireless sensor network changes from the $i$-state to the $j$-state. In the wireless sensor network, channel transfer occurs during transmission. According to the Markov model, the wireless sensor network is divided into channels
according to the signal-to-noise ratio of the transmitted signal. The signal-to-noise ratio $\gamma$ is divided into $K+1$ parts, and the segmentation point of the signal-to-noise ratio division is $\gamma_{k}^{k+1}$. When $\gamma \subseteq \{\gamma_{k}^{k+1}\}$, this channel belongs to the state $k$, so the channel has a total of $K$ states. To avoid deep fading, it is assumed that no transmission occurs during state $k$, and the rate at this time is $R_{0} = 0$ bit/symbol. According to the principle of Markov model, when the wireless sensor network channel is divided, the signal-to-noise ratio $\gamma$ is used as the threshold value and divided into several channel states. Different channel states correspond to their corresponding transmission modes, and the transition between channel states is achieved by transition probability.

Therefore, the threshold $\{\gamma_{k}^{k+1}\}_{k=0}^{k}$ of the SNR division interval can be obtained according to the following calculation process:

Let $k = K$, $\gamma_{k+1} = +\infty$, for each state $\gamma_{k}$ of the wireless sensor network, find a unique $\gamma_{k} \subseteq [0, \gamma_{k+1}]$, and $\gamma_{k}$ satisfies $P_{ER_{k}} = P_{0}$. Among them, $P_{0}$ is the probability of data packet transmission error, and $P_{ER_{k}}$ is the average error probability of data packet transmission when the wireless sensor network is in a channel state. $P_{ER_{k}}$ is related to the average signal-to-noise ratio, channel state parameters and channel state received by the receiving nodes in the wireless sensor network.

At that time $k > 1$, ordered $k = k - 1$, continue to look for the only $\gamma_{k} \subseteq [0, \gamma_{k+1}]$; If $k \leq 1$, let $\gamma_{0} = 0$, when the signal-to-noise ratio of the GNSS signal transmitted by the wireless sensor network is within the $[\gamma_{k}, \gamma_{k+1}]$ interval, the corresponding channel state is, and the signal-to-noise ratio at this time is the
Establish a Markov transmission model of various channel states as shown in the figure below.

![Data transmission model of wireless sensor network](image)

**Figure 3** Data transmission model of wireless sensor network

In the figure above, $T_{k,k'}$ represents the transition probability between channels.

Under slow fading channel conditions, channel state transitions only occur between adjacent channels, and the transition probability between non-adjacent channels is $T_{k,k'} = 0, |k - k'| \geq 2$. The transition probability of adjacent states can be calculated according to the following formula.

$$
\begin{align*}
T_{k,k+1} &= \frac{N_{k+1}T^\ell}{P(k)}, \quad k = 0, 1, 2, \ldots, K - 1 \\
T_{k,k-1} &= \frac{N_kT^\ell}{P(k)}, \quad k = 1, 2, \ldots, K
\end{align*}
$$

(11)

In formula (11), $T^\ell_j$ represents the transmission time of a data packet, and $P(k)$ is the probability of an error in data packet transmission. The probability that
the channel state does not change in adjacent time slots can be calculated according to the following formula:

\[
T_{i,j} = \begin{cases} 
1 - T_{i,j-1} - T_{i,j+1}, & 0 < k < K \\
1 - T_{i,j}, & k = 0 \\
1 - T_{i,j-1}, & k = K 
\end{cases}
\]  

(12)

According to the above calculation process, the transition probability of adjacent states can be calculated. This can determine the transition probability matrix of the wireless network sensor network, complete the construction of the wireless sensor network transmission model, and analyze the wireless sensor network transmission delay according to the model.

2.4 Analyze Wireless Sensor Network Transmission Delay

After establishing a wireless sensor network transmission model, use GNSS signals to analyze the wireless sensor network transmission delay. In the wireless sensor network transmission model, if the GNSS signal flows from the \( k \)-th channel \( j \) in the network to the \( k - 1 \)-th channel \( j \), then:

\[
\begin{align*}
L_T(i,j) &= L_T(i,j) + \frac{1}{L} + W_T(i,j) \\
S_V(i,j) &= S_V(i,j) + W_T(i,j)
\end{align*}
\]

(13)

In formula (13), \( L_T(i,j) \) is the delay time required to obtain data of the \( k \)-th channel \( j \) from the \( k - 1 \)-th channel \( i \); \( S_V(i,j) \) is the service required to obtain data of the \( k - 1 \)-th channel \( j \) from the \( k \)-th channel \( i \); \( \frac{1}{L} \) is the time required to transmit a data envelope between two adjacent stages; \( W_T(i,j) \) is the average waiting time to obtain the \( k \)-th channel \( j \) from the \( k - 1 \)-th channel \( i \).

Since the relevant parameters of the GNSS signal are known, according to the above process, the delay time when the wireless sensor network transmits data can be determined. At present, the research of transmission delay analysis method based on GNSS signal has been completed.
3 Test experiment

In this paper, a method for analyzing transmission delay of wireless sensor networks based on GNSS signals is studied. In order to verify the effectiveness of the method, an experimental procedure is designed. By analyzing the experimental data, the corresponding experimental conclusions are drawn.

3.1 Experimental content

This experiment is in the form of a comparative experiment. The experimental comparison group is the traditional network transmission delay analysis method. The experimental group is the GNSS signal-based wireless sensor network transmission delay analysis method studied in this paper. The comparison index of the experiment is the ratio of the memory occupied by the loading platform to the memory occupied by the two analysis methods when analyzing the same experimental network transmission delay.

The better performing transfer latency analysis model uses less running memory. In order to ensure the scientific and effective experimental data, the experimental process controls only the experimental variables, and the experiment is completed in accordance with the experimental process. Analyze experimental data and draw experimental conclusions.

3.2 Experiment preparation and process

The delay analysis methods of the experimental group and the comparison group were respectively loaded on the identically configured runners. Before the experiment, the two runners were tested to run well. After the two sets of analysis methods were loaded, the relevant data monitoring software could be run normally.

Set up a wireless sensor network as shown in the figure in the laboratory.
In order to test the two analysis methods, 10 experiments were performed. During the experiment, the relevant parameters of the experimental network were set according to the following table.

**Table 1** Related parameters of the experimental network

<table>
<thead>
<tr>
<th>The serial number</th>
<th>Number of nodes</th>
<th>Transmitted data volume/GB</th>
<th>Transmission distance /m</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>45</td>
<td>50</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>45</td>
<td>60</td>
<td>100</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>120</td>
<td>100</td>
</tr>
<tr>
<td>6</td>
<td>80</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>7</td>
<td>100</td>
<td>185</td>
<td>150</td>
</tr>
<tr>
<td>8</td>
<td>75</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>9</td>
<td>90</td>
<td>220</td>
<td>200</td>
</tr>
</tbody>
</table>

**Figure 4** Experimental environment
After setting the parameters of the experimental network, run the experimental network. Related software and equipment are used to control the network transmission delay, and the experimental group and comparative group analysis methods are used to analyze the transmission delay in the experimental network. The monitoring software of the experimental running platform occupies the platform's memory during the running of the two methods and outputs monitoring data. Analyze the experimental data and draw relevant conclusions.

### 3.3 Experimental results

The experimental results are shown in the following table. The data in the table is analyzed to obtain corresponding conclusions.

<table>
<thead>
<tr>
<th>The serial number</th>
<th>Experimental method/%</th>
<th>Comparison group method/%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.8</td>
<td>49.4</td>
</tr>
<tr>
<td>2</td>
<td>21.7</td>
<td>454</td>
</tr>
<tr>
<td>3</td>
<td>24.5</td>
<td>52.3</td>
</tr>
<tr>
<td>4</td>
<td>25.2</td>
<td>50.3</td>
</tr>
<tr>
<td>5</td>
<td>25.4</td>
<td>56.3</td>
</tr>
<tr>
<td>6</td>
<td>28.5</td>
<td>56.4</td>
</tr>
<tr>
<td>7</td>
<td>27.5</td>
<td>50.4</td>
</tr>
<tr>
<td>8</td>
<td>28.4</td>
<td>53.5</td>
</tr>
<tr>
<td>9</td>
<td>24.4</td>
<td>41.2</td>
</tr>
<tr>
<td>10</td>
<td>29.9</td>
<td>46.9</td>
</tr>
</tbody>
</table>

It can be seen from the analysis of the above table that the proportion of memory occupied by the analysis method of the experimental group by the experimental group is much lower than the proportion of memory occupied by the comparison group analysis method by the operation platform. Calculate the average proportion of
memory occupied by the two analysis methods during the experiment. The experimental group is 26.13%, the comparison group is 50.81%, and the comparison group is about 1.94 times the experimental group. It shows that the experimental group method can take up less memory to complete the same task. In summary, the performance of the GNSS signal-based wireless sensor network transmission delay analysis method studied in this paper is better.

4 Concluding remarks

The traditional network transmission delay analysis method occupies too much memory of the operating platform during operation, resulting in poor analysis performance. Therefore, this paper proposes a method for analyzing the transmission delay of wireless sensor networks based on GNSS signals. The comparison experiment with the traditional transmission delay analysis method proves that the analysis method studied in this paper takes up less memory and has better performance than the traditional method.

5 Fund projects


References

A resilient multi-path routing for the SDN based space-based network

Feihu Dong¹, Xi Chen², Yasheng Zhang³, Zhongyuan Jiang⁴, Ci He⁵
{6891hfd@163.com¹, chenqian1997@stu.xidian.edu.cn², zys163@163.com³,
zyjiang@xidian.edu.cn⁴, heci1027@sina.com⁵}

Beijing University of Posts and Telecommunications, Beijing 100876, China¹; Xidian University, Xi’an, Shaanxi 710071, China²,³; Science and Technology on Communication Networks Laboratory, Shijiazhuang, Hebei 050081, China¹,²,³,⁴; The 54th Research Institute of CETC, Shijiazhuang, Hebei 050081, China⁵

Abstract. To adapt the development of global communication and the rapid upgradation of communication technology (e.g. 4G to 5G), the space-based network (SBN) is regarded as a very important networking architecture to improve the efficiency of data communication. However, with the increasing need of bandwidth, the SBN faces many open challenges: 1) high cost of update and deployment and complex implementations; 2) low robustness or resilience against the dynamic network structure changes. In this work, we first incorporate the SDN (Software Defined Network) architecture into the SBN to ensure the feasible implementation by separating the control panel and the data panel. Secondly, we present a new multi-path routing (MPR) to adapt the fast network state change and enhance the network resilience. Finally, we compose extensive experiments on both random networks and scale-free networks to verify the effectiveness of our MPR mechanism which has potential applications in the future deployment of SBN.

Keywords: Multi-path routing, SDN, Space-based network, Resilience.

1 Introduction

With the rapid generation upgradation of communication technology (e.g. 4G to 5G), the bandwidth needs are enlarging, and the user applications are enriching. It requires the larger bandwidth and more timely communication. Moreover, data communication is in urgent need of lots of wide areas (e.g. the remote oceans and depopulated zones) which are not covered by the data network yet. In other words, it needs to achieve global communication in the next or future generation of communication technologies. However, the traditional communication network cannot deploy base stations to some harsh environment such as desert and ocean. Therefore a lot of research work concentrates on the space-based network (SBN) which can be
deployed into the space to achieve full signal coverage and provide data communication service.

However, there are three characters in the SBN: (1) the SBN topology is dynamic because of the high speed of satellite nodes; (2) the space-based network is cyclical; (3) the space-based system is vulnerable to interference and damage. Therefore a lot of research concentrate on the network routing algorithm to increase the robustness, resilience of data transmission in the space-based network [1-2].

In view of the problems such as the difficulty of the satellite topology changes frequently and the difficulties in network management caused by the characters of space-based network nodes and network deployment, software defined network (SDN) is utilized to solve the problem. In our research, we adopt the architecture of SDN to improve communications efficiency of the space-based network.

Concerned about the resilience and robustness of communication, there are a lot of routing algorithm is proposed. In the past decade, many researchers concentrate on the satellite networks routing. Lu et al. [3] propose a novel inter-satellite routing based on link recognizing to improve the network routing problem and reduce the packet loss and delay. Li et al.[4] propose a novel routing strategy based on fuzzy theory for satellite network to evaluate the satellite congestion and then decrease the network congestion and packet loss. Liu et al.[5] analyze the efficiency of routing algorithms for satellite network under a single satellite node fault and random node fault. Wang et al.[6] propose two-layer compressed search algorithm which is a routing search algorithm in satellite IP switches. Feng et al.[7] designs a inter-satellite routing based on Fibonacci Heap to improve the routing efficiency of inter-satellite. Xu et al.[7-8] impose the integrated routing algorithm for multi-satellites, multi-ground-stations and multi-processing-centers which improves routing performance of the satellite and ground station. Considering about the topology of SBN is dynamic. Some packet could be lost in the process of data transmission because of the topology changes. So it is necessary to design the efficient routing to achieve the robustness in the process of data transmission [9-13]. However, some innovative routing algorithms lead to networking performance degradation such as increasing the average path length [14-17] and the transmission time and even leading the network congestion. Take into account these research, we propose a novel routing algorithm.

In this work, we introduce an innovative multi-path routing algorithm to improve the space-based network performance. In the paper, the multi-path routing mechanism is proposed, the resilience of a space-based network is studied when a certain proportion of links are destroyed and a large number of delay verification experiments is done.

2 SDN based Space-based Network (SBN)

In general, the framework of SDN separates the underlying infrastructure from application programs and network services, makes network control directly
programmable, and only forwards data according to the control information of the control layer. Therefore the SDN is cost-efficient, dynamic and controllable. But due to the deployment cost of space-based network is high. Therefore the combination of space-based network and SDN increases network efficiency and scalability and reduces the overhead on the satellite. And space-based network based SDN is easy to manage and the utilization rate of network resources is high [18].

As the above shows, the framework of SDN is adapted to the satellite network. Firstly, the operation of the satellite network is periodic and regular. Therefore the controller can make full use of these information to compute and manage the network. Secondly, satellites move fast and the topology is dynamic. The SDN controller performs real-time according to the topology information, so as to make the network run efficiently and increase the network stability. Thirdly, the deployment cost of the satellite is expensive, so updating and reconfiguring the system of satellite is difficult. SDN architecture could reduce the expenditure.

As the Fig.1 shows, the SBN consists of low earth orbit (LEO) satellite and geosynchronous Earth orbit (GEO) satellite. SDN divides the network into two layers: control layer and data layer. Part of ground and GEO satellite is belonging to the control layer. The data layer consists of relay nodes on the ground and LEO satellites which transmit the information of the flow table. The control layer is responsible for the calculation and management of the network. The control layer has the information and
traffic distribution of all communication nodes in the network, so it can manage the whole network more flexibly. The data layer just needs to transmit the packet.

Therefore we deploy the time-varying topology on the control layer and when switch over time slice, the control layer update route of the data layer. In the paper, we study an innovative routing algorithm and deploy it on the control layer. Then the data packet of data layer follow the rule to transmit and the control layer collect the information of all the satellite in the data layer.

3 The Multi-path Routing (MPR)

In the future of SBN deployment, the routing robustness is still a key problem to achieve the communication immediacy and security in the global communication network[13]. In the paper, we propose a Multi-path Routing (MPR) to compute multiple paths for every node pair and enhance the routing resilience against the changing network topology. The process is 1) we compute multiple paths for every node pair; 2) the first path is used for the data communication; 3) when the first path is unusable, we randomly select one available path from the alternative paths. In case of network failure, one of the multiple paths will be used as standby path and the communication will be restored.

Here we propose a novel multi-path routing with the following calculation flow:

(1) Given a network topology, path set \( P \) and the link between node \( i \) and \( j \) in the topology has a weight \( w_{ij} \) which is equal to 1 or the reciprocal of bandwidth or distance.

(2) Computing a path from source node to destination node with the lowest weight by Dijkstra algorithm [16].

(3) To judge whether the path already exists in the path set \( P \), and eliminate duplicate paths.

(4) Adding an increment \( \Delta w (\Delta w = 1) \) to each link weight \( W \) on the calculated path and updating the \( W \).

(5) Calculating single path repeatedly until \( k (k \text{ is an integer}) \) paths are obtained.

For example we compute the multi-path from node 1 to node 6 in the Figure 2:(1) the weight of links which connect two reliable nodes is 1, the weight of links which connect one reliable node is 2, the other is 3. And the path set between node 1 and node 6 is empty (2) we compute the path which weight is smallest: \{1,2,4,6\}, \{1,2,5,6\} and \{1,3,5,6\}(the smallest weight is 3;\{1,2,4,6\}), (3) judge the path \{1,2,4,6\} is not in the path set. (4) add a \( \Delta w = 1 \) to the weight of the link \( (1,2) \), (2,4) and (4,6), update the weight \( w_{(1,2)} = w_{(2,4)} = w_{(4,6)} = 2 \). (5) calculate the second path \( w_{1,2,5,6} = 2 + 2 + 2 = 6 \), \( w_{1,3,5,6} = 2 + 3 + 2 = 7 \), the \( w_{1,2,5,6} \) is smaller and joins in a path set, then update the weight. And then continues the step (2)-(4). Finally we compute three path weight \( w_{1,2,4,6} = 3 \), \( w_{1,2,5,6} = 6 \), \( w_{1,3,5,6} = 8 \). When we
choose a path from source to destination, we will choose the path which the weight is smallest to reduce the rate of loss packets and increase the stability of the network.

As the Fig.2 shows that different algorithm get available paths between source node and destination node compared with MPR, shortest path routing(SPR), node disjoint multi-path SPR(NDMP-SPR), efficient routing (ER) [10] and node disjoint multi-path ER(NDMP-ER). And the number of paths of MPR is more than the other three algorithms.

![Fig.2. The application and comparison of multi-path routing](image)

In the network, there are always some unreliable nodes which are vulnerable to attack or link failure as shown in Fig.2. The multi-path routing we designed is a routing algorithm with increasing weight, which can find out $k$ different paths for data transmission to avoid unreliable nodes and links and improve communication security and resilience.

### 4 The Resilience of SBN

According to the periodic changes of LEO satellite network topology, the SBN will calculate the network topology information of each time slice. When the time slice is switched, the LEO satellite will call the routing information of the corresponding time slice and transmit the packet according to the routing table, and adopt multi-path backup to forward the packet in case of failure. In the process of time slice switching, the link failure occurs when the link corresponding to the previous time slice does not exist on the current time slice. Check whether the multi-path information of the satellite node can reach. The multi-path routing ensures the stability of data transmission and reduces packet loss during time slice switching. In the case of multi-node and multi-link failure, the network ensures the immediacy of space-based network communication and the robustness of network performance.

There are three cases of space-based network failure: (1) single node failure: if the time slice is not switched, but the node fails, we will find available path which weight is smallest from the backup path to transmit the packet; (2) time slice switching: if the time slice is switched to $T$. We will choose a route which can reach the destination node from the backup paths; (3) multi-node failure: if there is a multi-node failure in time slice $T$, the node will find
an available path from its multiple backup paths, then based on the acyclic and forced forwarding, if the node has no path available and the data packet will timeout retransmission.

Here we test the resilience of SBN in different routing algorithms on the successful forwarding rate and the average path length of maximal connected subgraphs. In this experiment, we use the network of 200 satellite nodes to do the elastic experiment of space-based network. When the failure rate is upto 20%, the multi-path routing technology can recover the data forwarding of the route, and in the failure caused by the single node, multi-node and time slice switching, achieve the immediacy of the communication within 50ms.

In the novel of MPR, we set the node failure ratio as \( f \), and the path budget in the multi-path routing algorithm as \( k \) which is the maximum value of a path set.

5 Experiments and results

As we known, the traffic efficiency is related to the successful forwarding rate and average path length. Therefore, we test these parameters in the paper. Here we focus on the resilience of the network in one time slice because of the destruction of nodes and links in the BA network and ER network. The network scale is 200, and the average degree is 4. The result is that the average value of 10 networks with the same network size.

5.1 Network model

Here we adopt two kinds of network models in the experiment: 1) the ER network model, 2) the BA network model.

ER network, one kind of random network, starts with \( N \) nodes and connects each pair of nodes with probability \( p \), which builds a graph with approximately \( pN(N-1)/2 \) randomly placed links. Node degree of ER network follows a Poisson distribution, which indicates that most nodes have approximately the same number of links. In the recent research, most simulation adopt the ER network to simulate the network architecture [19].

BA network is one kind of scale-free networks, which is characterized by a power-law degree distribution, the probability that a node has \( k \) links follows \( P(k) \sim k^{-\gamma} \), where \( \gamma \) is the degree exponent. A new node tends to connect to an old node of high degrees in the BA network which conforms to the law of real network. Same as the ER network, the architecture of the BA network is used to be simulated in some network research [20].

5.2 Evaluation Metrics

In the paper, the successful forwarding rate represents the network resilience when network nodes fail, and the average path length shows the network performance with different routing methods. Therefore, we evaluate the successful forwarding rate and the average path length of different routing methods with two kinds of network models.
The successful forwarding rate is the ratio of packets successfully arriving at the destination node to all packets sent between the source to destination, which evaluates the performance of routing algorithm. Here we compute the successful forwarding rate as

$$\text{SFR} = \frac{n_s}{n_a}.$$  \hfill (1)

which \(n_s\) represents the number of packets which successfully arriving at the destination node and \(n_a\) is the number of all packets sent.

The average path length is the average hops between origin and destination as

$$\text{APL} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} h(i, j)}{N(N-1)}.$$  \hfill (2)

where \(h\) depicts as the hops for a source-destination pair in the routing matrix of the network and \(N\) is the number of nodes in the network.

5.3 Successful forwarding rate in ER network

Fig. 3. The evolution of successful forwarding rate as a function of \(k\) on the ER networks.

As we know, the successful forwarding rate of the network can evaluate the communication recovery efficiency of the network. Therefore, we test it with different
numbers of backup paths $k$ and several failure rates $f$ as Fig.3 shows. In the Fig.3 (a), we investigate the comparisons of two routing algorithms (MPR and SPR) with different $k \in (1,10)$ and the same $f=0.05$ under ER network. In the ER networks, one can see that the MPR have greater successful forwarding rate than the SPR. And when $k$ is greater than 5, the success forwarding rate of MPR is equal 100% which mean all the packets will arrive their destinations. However, there is nearly no change in SPR.

In the Fig.3 (b), we increase the failure rate to 0.1, and the rest of the parameters remains unchanged. We find that when the successful forwarding rate $f$ of MPR reaches 100%, the number of backup path $k$ become 7 which is larger than the $k$ when $f=0.05$. The successful forwarding rate of SPR is reduced from 88% to 74%.

In the Fig.3 (c), we increase the $f$ again, the failure rate reaches 20% and the successful forwarding rate can reach 98% most. Meanwhile, SPR is reduced from 74% to 56%. In general, the successful forwarding rate of MPR is higher than that of SPR. Moreover, when the node failure rate reaches 20%, the success rate of MPR keeps a high value.

In the Fig.3(d), we can see that the MPR have good performance in different node failure rate. The higher node failure rate is, the more backup paths achieved. Our MPR routing appears to achieve high robustness against the node failure.

5.4 Successful forwarding in BA network

![Fig. 4. The evolution of successful forwarding rate as a function of $k$ on the BA networks.](image-url)
In the Fig.4 we set $k \in (1,10)$ and $f=0.05, 0.1$ in BA networks and 0.2 and compare the successful forwarding rate of MPR and SPR. In the Fig.4 (a), we investigate the comparisons of the two routing algorithms with different $k \in (1,10)$ and the same $f=0.05$ under the BA network. In the BA networks, we can see that the MPR has better successful forwarding rate than the SPR. And when $k$ is greater than 6 and node failure rate is 5%, all the packets will arrive successfully their destinations. However the successful forwarding rate is nearly no change in SPR which is equal to 89%.

In the Fig.4 (b), we increase the failure rate to 0.1, and the rest of the parameters remains unchanged. We find that when the successful forwarding rate $f$ of MPR reaches 100%, the number of backup path $k$ become 7 which is larger than the $k$ when $f=0.05$. The successful forwarding rate of SPR is reduced from 89% to 82%.

In the Fig.4 (c), we increase the $f$ again, the failure rate reaches 20% and the successful forwarding rate can only reach 98% most. However SPR is reduced from 82% to 66%. In general, the successful forwarding rate of MPR is higher than that of SPR. And even when the node failure rate reaches 20%, the success rate of MPR can reach 98%.

In the Fig.4 (d), we can see that the MPR has good performance in different node failure rate. The higher node failure rate, the more backup paths.

5.5 Comparisons of successful forwarding rate for BA network and ER network

Fig. 5. The evolution of successful forwarding rate as a function of $f$ on the BA and ER network.

In the Fig.5(a)-(b), we can see the effect of $k$ value for the successful forwarding rate. The $k$ value is larger and the successful forwarding rate is larger. Especially with the $f$ is becoming larger, more nodes lost and the number of backup paths is more important. As the figure shows, when $k=5$, even the node failure rate reaches 20%, we can restore more than 90% of communications in the BA and ER networks.

5.6 The average path length with different $k$ in ER and BA networks
As we know, it is easy to increase the average path length while optimizing the routing algorithm. Therefore we test it with different number of backup paths $k$ in Fig.6. In the Fig.6 (a), we investigate the comparisons of three different values of $f$ with different $k \in (1,10)$ under ER network. In the ER network, with the value of $k$ is larger, and the average path length is smaller. Because larger $k$ means more feasible options for the alternative path switch. And when $k$ is equal 4, the average path length just is nearly unchanged.

In the Fig.6 (b), the BA network is complicated, however the average path length tends to be stable and the value is lower than that in the ER network. As the Fig.6 (b) shows, although there are some fluctuations, the average path length is becoming smaller as a whole with the $k$ is becoming larger. With higher failure rate, the available alternative path number reduces, and larger average path length is needed.

![Fig. 6](image6.png) The evolution of the average path length as a function of $k$ in ER and BA networks.

### 5.7 The average path length with different $f$ in ER and BA networks

![Fig. 7](image7.png) The evolution of the average path length as a function of $f$ in ER and BA networks.
In the Fig.7 (a), we investigate the comparisons of three different values of $f$ with different $k \in \{1, 2, 10\}$ in ER network. In the ER network, with the value of $f$ is larger, and the average path length is larger which varies from 3.9 to 5.3 because the $f$ is larger and there are more links disconnect. Therefore, packets arriving at the destination node will pass through more nodes. When the number of the backup path is larger, the average path length is shorter.

In the Fig.7 (b), the average path length of the BA network varies from 3.3 to 4 when the $f$ changes from 0.05 to 0.2. However when $k=2$ and $k=10$, the average path length of the BA network has not been an obvious change with the same $f$.

**Conclusion**

To sum up, this work studied the resilient multi-path routing for the SDN based space-based network. The multi-path routing increases the successful forwarding rate and reduces the average path length when the nodes failure occur. When the rate of nodes failure of SBN is high, the resilience of the network is still high when the number of backup paths is big enough. Therefore the multi-path routing is meaning for the resilience of a space-based network based on SDN, and has potential applications in the future generation of communication technology such as the NFV (Network Functions Virtualization) architecture [21-22].

**Acknowledgements**

This work was supported by the Natural Science Basis Research Plan in Shaanxi Province of China (2020JM-203), the research program of Science and Technology on Communication Networks Laboratory (6142104190306), and the National Natural Science Foundation of China (61502375).

**References**


Passive Radar based on the Specific Sequence of Wireless Network

Qian Zhao, Hui Zhao, etc
{qzhao¹, hzhao²}@bupt.edu.cn
Wireless Signal Processing and Networks (WSPN) Lab,
Key Laboratory of Universal Wireless Communications, Ministry of Education,
Beijing University of Posts and Telecommunications (BUPT), Beijing, 100876, China

Abstract. Passive radar locates the UAV target by processing the reflected signal from non-collaborative illuminators of opportunity (IOs). It will inevitably be affected by the direct-path interference and multipath interference. The conventional passive radar uses the reference channel to receive the direct-path signal as the reference signal for interference cancellation, which is complicated in hardware and susceptible to noise. However, most of the wireless signals carry the prior information used for synchronization or channel estimation, such as the preamble of WiFi signal or the CRS of LTE signal. This paper presents a passive radar system based on the specific sequence in the wireless network, which is the prior information of IOs. Firstly, we analyze the ambiguity function of several typical specific sequences and prove their feasibility as the reference signal. Then we adopt a coherent approach for interference cancellation. The system availability is proved by simulation. A 9.64 dB reduction of interference is observed.

Keywords: Unmanned Aerial Vehicles (UAVs) localization, interference cancellation, passive radar, wireless network.

1 Introduction

Unmanned aerial vehicle (UAV) applications are rapidly expanding[1], but they also bring numerous challenges and security risks. In some sensitive areas, it is particularly important to quickly detect and locate the UAVs [2]. Common active radars detect targets by emitting radar signals themselves, but it costs a lot and not suitable for large scale applications. Passive radar [3] detects and locates the target by processing reflected signals from non-collaborative illuminators of opportunity (IOs). It spends less by utilizing the existing communication equipments, and works in covert operation. In a word, it is more suitable for positioning the non-cooperative UAV.

In a conventional passive radar system[4], the receiver uses two receiving channels, in which one for surveillance and the other for reference. The antenna of the reference channel points to the IO and receives the direct wave signal from the IO; the antenna of the surveillance channel points to the sky and receives the target echo signal. Both channels are susceptible to noise while receiving valid signals. At the same time, the target echo signal in the surveillance channel will also be subject to interference such as the strong direct-path signal and the multipath signals. J. Tong [5] completed a detailed study on joint estimation of target position and velocity in the passive radar system. The CRLB of the target position depends not only on the IO waveform parameters, but also on the values of SNR and INR (interference-to-noise).
Many broadcast and communication signals have been considered and analyzed as the possible passive radar sources, including FM radio, DVB-T, GSM, LTE and WiFi [6][7][8]. These wireless signals usually carry signals used for channel estimation or synchronization, which can be considered as prior knowledge at the receiver. For example, the Long-Term Evolution (LTE) needs cell specific reference signals (CRS) for channel estimation and correlation demodulation. And the 802.11n WiFi standard specifies the preamble part for frame synchronization. Different from the existing passive radar solutions, this paper will use the sequence carrying prior knowledge in the IO signal as the reference signal. So the receiver can only use a single channel to receive the reflected signal from IOs. It can also avoid the delay and noise impact caused by the reference channel.

On the other hand, affected by the sidelobes of the surveillance antenna, direct path interference (DPI) and multipath interference (MPI) are often too higher than the target echo to accurately detect UAV target in passive radar systems[9]. If the interference is not eliminated, it is impossible to detect the target only by the conventional processing methods such as specifying the antenna direction and pulse compression[10]. Therefore, interference cancellation becomes a key issue in passive radar systems. At present, adaptive algorithms for interference cancellation mainly include: Least Mean Square (LMS), Recursive Least Squares (RLS), Extended Cancellation Algorithm (ECA)[11][12], Sequential Cancellation Algorithm (SCA), and so on[13]. The ECA has been shown to be effective against typical scenarios with a limited number of iterations.

In this paper, we propose a passive radar system based on a wireless network. The specific sequence in non-service signal is used as the reference signal, which is known to both transmitter and receiver. Firstly, we analyze the ambiguity function[14], and it is proved that the specific sequence carrying prior information is sufficient to replace the reference signal received by the reference channel. Then, the interference subspace matrix generated by the specific sequence filters the surveillance signal for interference cancellation. After cancellation, the time delay and frequency offset estimates of the UAV target can be obtained by the Range-Doppler processing [15] of the target echo signal. Therefore, the target position can be estimated through the classic time difference-of-arrival algorithm.

The paper is organized as follows: Section 2 analyzes the ambiguity functions of CRS in LTE and preamble in WiFi. Section 3 introduces the passive radar system model based on wireless networks. Section 4 derives an algorithm for extended interference cancellation. Simulation and analysis are discussed in section 5, in the scenarios which parameters were generated by Winprop propagation model software. Finally, conclusions are drawn in section 6.

## 2 Ambiguity function analysis

The sequence carrying prior knowledge in wireless network signals has a very specific structure designed for channel estimation or synchronization. It is called the specific signal in this paper. A brief analysis of the relevant factors influencing a passive radar system is provided here.

Range resolution and Doppler resolution are important parameters used in radar systems to indicate the ability to distinguish two or more targets. Range resolution $\Delta R$ can be written as:

$$\Delta R = \frac{c}{2B \cos(\beta / 2)}$$  \hfill (1)
where \( c, B \) and \( \beta \) are the speed of light, the bandwidth of the signal and the bistatic angle separately. The length of the pulse \( T \) determines the Doppler resolution for that single pulse, which can be written as:

\[
\Delta f_d = \frac{1}{T} \tag{2}
\]

In this section, we analyze the ambiguity function of the 801.11n WiFi signal and the LTE signal separately. Considering \( \beta = 45^\circ \) and \( T = 0.05s \), **Tables.1** shows parameters of these two IO signals in theory.

**Table 1.** Parameters of transmitted signals.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>802.11n</th>
<th>LTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bandwidth ( B ) (Hz)</td>
<td>20M</td>
<td>10M</td>
</tr>
<tr>
<td>Subcarrier spacing ( \Delta f ) (Hz)</td>
<td>312.5k</td>
<td>15k</td>
</tr>
<tr>
<td>Sampling frequency ( f_s ) (Hz)</td>
<td>40M</td>
<td>15.36M</td>
</tr>
<tr>
<td>Range resolution ( \Delta R ) (m)</td>
<td>7.5</td>
<td>9.76</td>
</tr>
<tr>
<td>Doppler resolution ( \Delta f_d ) (Hz)</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

Characteristics determined by the nature of the transmitted waveform in **Tables.1** can be evaluated by ambiguity function in simulation:

\[
|\xi (\tau, f_d)| = \left| \int_{-\infty}^{\infty} s'(t+\tau)s(t)e^{j2\pi f_d t} dt \right| \tag{3}
\]

where \(|\xi (\tau, f_d)|\) is the value of ambiguity function of the transmitted signal \( s(t) \), \( \tau \) denotes the delay, \( f_d \) is the Doppler frequency, and \( s'(t) \) is the complex conjugate of \( s(t) \). In practice, the ambiguity function is often digitally implemented on a computer and the equation becomes:

\[
|\xi (\tau, f_d)| = \left| \sum_{n=0}^{N-1} s[n]s^*[n+\tau]e^{j2\pi f_d nT_s} \right| \tag{4}
\]

where \( s[n] = s(nT_s) \), \( T_s \) is the sample period and \( N \) is the number of samples in the integration time.

A simulated ambiguity function of the preamble in the 802.11n signal is shown in **Figure 1(a)**. As can be seen, the ambiguity function has a relatively high peak of 25.05dB at the origin. **Figure 1(b)** shows the range profile (\( f_d = 0 \)) of the ambiguity function, where the sidelobe levels are 21.41dB and the range ambiguity is 0.8 \( \mu s \). The corresponding distance is 1.2km. Other range ambiguities are multiple of 0.8 \( \mu s \), which don't make much sense because the corresponding distances are beyond radar coverage. **Figure 1(c)** shows the Doppler profile (\( \tau = 0 \)) of the ambiguity function, where the sidelobe levels are 8.66dB.
Similarly, the ambiguity function of the CRS in LTE signal is shown in Figure 2(a). The ambiguity function has a relatively high peak of 26.11 dB at the origin. Figure 2(b) shows the range profile ($f_d = 0$) of the ambiguity function, where the sidelobe levels are 20.2 dB and the range ambiguity is 0.39 $\mu$s. The corresponding distance is 0.58 km. Figure 1(c) shows the Doppler profile ($\tau = 0$) of the ambiguity function, where the sidelobe levels are 19.63 dB.

Through the above description and discussion, the ambiguity caused by the sidelobe in the ambiguity function will not affect the positioning in a certain small range (< 500 m). This paper proves the feasibility of using WiFi preamble and CRS signal as passive radar non-cooperative IOs, and estimates the detection capabilities of passive radar systems based on the wireless network.
Fig. 2. (a) Ambiguity function of the CRS in LTE signal (b) Range profile ($f_r = 0$) (c) Doppler profile ($\tau = 0$).

3 System model

We consider the signal model in a typical passive radar geometry. The signal collected from the echo channel contains not only the target echo signal, but also DPI, MPI and noise signal. While the reference channel of conventional radar also contains the direct-path signal and noise signal. In this paper, the specific sequence with prior information was seen as the reference signal.
As seen in Figure 3, assuming that multipath echoes can be expressed as the collection of the signal reflected by multiple fixed positions, the complex envelope of the signal at the receiver is given by:

\[
\begin{align*}
s_g(t) &= A_d d(t) + \sum_{i=1}^{N_c} c_i(t) d(t - \tau_i) + \sum_{j=1}^{M} \alpha_j d(t - \tau_j) e^{i 2\pi f_d t} + n_g(t) \tag{5}
\end{align*}
\]

where \(A_d d(t)\), \(\sum_{i=1}^{N_c} c_i(t) d(t - \tau_i)\) and \(\sum_{j=1}^{M} \alpha_j d(t - \tau_j) e^{i 2\pi f_d t}\) denote the direct-path signal, the stationary scatterers and the target echoes respectively. \(d(t)\) and \(A_d\) are the complex envelope and complex amplitude of the direct-path signal; \(c_i(t)\) and \(\tau_i\) are the complex amplitude and the delay of the \(i\)-th stationary scatterer \((i = 1, 2, \ldots, N_c)\), \(N_c\) is the number of scatterers; \(\alpha_j\), \(\tau_j\) and \(f_d\) are the complex amplitude, the delay (with respect to the direct-path signal) and the Doppler frequency of the \(j\)-th target \((j = 1, 2, \ldots, M)\), \(M\) is the number of targets and the value is one in this paper; \(n_g(t)\) is the thermal noise contribution at the receiver.

The complex amplitude \(c_i(t)\) are considered slowly varying functions of time, so that they can be represented by only a few frequency components around zero Doppler:

\[
\begin{align*}
c_i(t) &= c_i e^{i 2\pi f_{\alpha} t} \tag{6}
\end{align*}
\]

where \(f_{\alpha}\) and \(c_i\) are the Doppler shift and complex amplitude of the \(i\)-th multipath echo.

Assuming operation with a sampling frequency \(f_s\) which satisfies the Nyquist theorem, the samples \(s_g\) collected at the surveillance channel at time instants \(t_i = i / f_s = i T_s, i = 0, \ldots, N - 1\), are arranged in an \(N \times 1\) vector:

\[
\begin{align*}
s_g &= [s_g[0], s_g[1], s_g[2], \ldots, s_g[N-1]]^T \tag{7}
\end{align*}
\]

Similarly, we obtain \(N\) samples of the reference signal in the following vector:

\[
\begin{align*}
s_{\text{ref}} &= [s_{\text{ref}}[-R+1], \ldots, s_{\text{ref}}[0], \ldots, s_{\text{ref}}[N-1]]^T \tag{8}
\end{align*}
\]

where \(R\) is the extra simple points to ensure direct path in \(s_g\) and \(s_{\text{ref}}\) line up exactly.

The detection process in passive radar is based on the evaluation of the delay-Doppler cross-correlation function (CCF) between the surveillance and the reference signal:
\[ \xi(l, p) = \sum_{i=0}^{N-1} s_k[i] \cdot s_{\text{ref}}^*[i-l] e^{-\frac{i2\pi p}{N}} \]  

where \( l = 0, \ldots, R-1 \) is the time bin representing the time delay \( \tau[l] = lT_s \), \( p \) is the Doppler bin representing the Doppler frequency \( f_d[p] = p/(NT_s) \).

While in the Delay-Doppler cross-correlation function of the received signal \( s_k \) and the reference signal, peak of target echo is masked by the sidelobes of the direct-path signal, so it is necessary to pay attention to the cancellation of interference in practice.

### 4 Interference cancellation algorithm

An effective cancellation filter for passive radar can be obtained by an Least Square approach. The algorithm searches for a minimum residual signal power after cancellation of the interference (direct-path signal and multipath echoes), thus:

\[
\min_{\alpha} \left\{ \| s_k - X\alpha \|^2 \right\} 
\]

where \( X \) is the interference subspace matrix and \( \alpha \) is the adaptive coefficient.

\[
X = B[\Lambda_p S_{\text{ref}} \cdots \Lambda_p S_{\text{ref}}^N] \Lambda_{\text{ref}} [\Lambda_p S_{\text{ref}} \cdots \Lambda_p S_{\text{ref}}^N] 
\]

where \( B \) is an incidence matrix that selects only the last \( N \) rows of the following matrix:

\[
B = \{ b_i \mid i \in [1,N], j \in [1,N+R-1] \} 
\]

\[
b_i = \begin{bmatrix} 1 & 0 & \cdots & 0 \\
0 & e^{i2\pi p} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & e^{i2\pi p(N+R-1)} 
\end{bmatrix} 
\]

\[
\Lambda_p = \begin{bmatrix} 1 & 0 & \cdots & 0 \\
0 & e^{i2\pi p} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & e^{i2\pi p(N+R-1)} 
\end{bmatrix} 
\]

\[
S_{\text{ref}} = \begin{bmatrix} s_{\text{ref}}^1 & D s_{\text{ref}}^2 & D^2 s_{\text{ref}} & \cdots & D^{K-1}s_{\text{ref}} \end{bmatrix} 
\]

\[
D = \{ d_{ij} \mid i,j \in [1,N+R-1] \} 
\]

\[
d_{ij} = \begin{bmatrix} 1 & i = j+1 \\
0 & \text{otherwise} 
\end{bmatrix} 
\]

The columns of matrix \( X \) in (10) define a basis for an \( M \)-dimensional interference subspace, where \( M = (2P + 1)K \). Solving (10) yields:

\[
\alpha = \left( X^H X \right)^{-1} X^H s_r 
\]

Therefore, the surveillance signal after cancellation becomes:

\[
\hat{s}_r = s_r - X\alpha = \left[ I - X \left( X^H X \right)^{-1} X^H \right] s_r = Ps_r 
\]
where the projection matrix $P$ projects the received vector $s_n$ in the subspace orthogonal to the interference subspace.

5 Simulation and analysis

In order to evaluate the performance of the passive radar system mentioned above, we consider a study case in the open filed of suburb without the dense buildings. A 802.11n WiFi transmitter is used as IO with a bandwidth of 20 MHz. The transmit power of the transmitter is 0.1W. The noise spectral density $N_0$ is set as $-101$ dBm. Based on the transmission distance of the 802.11n WiFi signal in a suburb area, the distance between the transmitter and the receiver is set to 300m. Assuming that the speed of the UAV target is 50Km/h, and the flight direction is consistent with the direction of the direct wave signal from the transmitter, the value of the Doppler frequency is calculated to be 111 Hz. The propagation modeling software WinProp provides a more realistic simulation: depending on the scenario, predictions are based on topographical, clutter, and building databases; different transmission modes can be defined and the coverage maps are computed individually for each transmission mode.

WinProp outputs the time delay, Doppler frequency and power information of each path in the scene. Considering the UAV target is in 70m height, the parameters are shown in Table. 2.

Table 2. Parameters of received signals (70m height).

<table>
<thead>
<tr>
<th>Signal</th>
<th>Direct-path</th>
<th>Multipath echo1</th>
<th>Multipath echo2</th>
<th>Target echo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Delay (ns)</td>
<td>1003.97</td>
<td>1056.44</td>
<td>1055.89</td>
<td>1101.17</td>
</tr>
<tr>
<td>Doppler frequency (Hz)</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>111</td>
</tr>
<tr>
<td>Power(dBm)</td>
<td>-84.48</td>
<td>-88.69</td>
<td>-88.36</td>
<td>-98.18</td>
</tr>
</tbody>
</table>

Firstly, considering the sampling frequency at receiver is 40MHz, we assume that cancel range $K = 4$ and Doppler area $p = 100$. Then we apply the interference cancellation algorithm mentioned in Section 4. The delay-Doppler cross-correlation function between the surveillance and the reference signal was shown in Figure. 4.

![Figure 4](image)

Fig. 4. (a) CCF before cancellation (b) CCF after cancellation.

Figure.4(a) shows the CCF output of the surveillance signal and the reference signal before cancellation. And Figure.4(b) shows the CCF output after removing the direct-path interference
and all multipath interferences laying in the first K range bins (K = 4 and P = 100). Here, the weak UAV target now appears as a strong peak. In Figure 4(b), the location of the weak target is shown obviously. The most prominent peak represents the target echo whose amplitude is 17.42 dB. At origin, the value of the function is about 14.26 dB, which means that the amplitude of the direct-path signal was eliminated more than 9.64 dB. After cancellation, the amplitude of the direct-path signal is 3.16 dB less than the target echo. Then the UAV target can simply be detected by using a simple constant false alarm rate (CFAR) detector such as the CA-CFAR detector.

Here, the peak to average power ratio (PAPR) tests are introduced for evaluating the performance of the algorithm. For calculating the PAPR, a 50m height UAV target echo and the multipath echo are introduced with characteristics tabulated in Tables 3.

<table>
<thead>
<tr>
<th>Signal</th>
<th>Direct-path</th>
<th>Multipath echo</th>
<th>Target echo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Delay (ns)</td>
<td>1003.97</td>
<td>1022.78</td>
<td>1053.41</td>
</tr>
<tr>
<td>Doppler frequency (Hz)</td>
<td>0</td>
<td>0</td>
<td>111</td>
</tr>
<tr>
<td>Power(dBm)</td>
<td>-84.48</td>
<td>-87.56</td>
<td>-96.00</td>
</tr>
</tbody>
</table>

We analyze the performance of the algorithm, assuming that cancel range K = 4 and Doppler area p = 100.

![Fig. 5. Diagram of PAPR versus the number of IO signal samples.](image)

In Figure 5, a simulated PAPR curve of the interference cancellation algorithm versus the number of IO signal samples is shown in the comparison of different height. As can be seen, the amplitude of the target after cancellation will be increased by increasing the number of samples, and reduced by increasing the height. This is consistent with the Cramér–Rao lower bound analysis. One can observe that there exists a significant performance degradation when the IO signal samples is under 256.
6 Conclusion

For UAV localization, this paper proposed a passive radar system based on the specific sequence of the wireless network. Ambiguity function valued the detection performance of the specific sequence with prior information. The cancellation algorithm was based on the projections of the received signals in a subspace orthogonal to the interference. Using the specific sequence as the reference signal makes the system more economical and effective. By using the propagation model generated by Winprop, the simulation results showed that the approach counteracted interference more effectively. It is of practical significance to detect UAV targets by using this radar system in the future.
Acknowledgments. This work was supported in part by the Beijing Municipal Science and Technology Commission Research under Project(Z181100003218015).

References

Data transmission pattern recognition method for communication network with terminal band constraint

CHEN Xin Yuan¹, LI Jun²
{chenxinyuan7559@163.com¹, lijun73812@163.com²}  
¹. Institute of artificial intelligence, Jiangxi University of Applied Science and technology, Nanchang 330100, Jiangxi;  
². Institute of artificial intelligence, Jiangxi University of Applied Science and technology, Nanchang 330100, Jiangxi

Abstract: In view of the problem of large recognition error and low accuracy of traditional pattern recognition methods, a new method of data transmission pattern recognition based on terminal band constraint is proposed. The transmission mode is analyzed, the coupling mechanism mutual inductance model is used to extract the characteristic parameters, and the Spark-KNN fast pattern recognition algorithm is used to realize the communication network data transmission pattern recognition with terminal number band constraints. Design a control experiment, compared with the traditional method analysis, the experimental results confirmed that the proposed method identification is more accurate.

Keywords: communication network; data transmission; pattern recognition; support tree

1 Introduction

A large number of studies have shown that the support tree is the most basic and optimal topological structure of the communication network, the edge on the support tree represents the connection line between the communication devices. Generally, this connection can be represented by the amount of information in communication, but it can also be defined by the communication cost or time. Thus, the optimal communication line connection among the various alternative connections is equivalent to finding the minimum support tree topology of the communication network. Prim algorithm, Kruskal algorithm is an effective algorithm to solve these problems [1].

However, in the actual communication network design, there are often some additional design requirements, because each node represents different communication devices, their role in the network is different. Obviously, it represents the hub, the middle node of the multiplexer, because of the connection with multiple nodes, both the sufficient hardware and software are more complex and much higher equipment cost than the terminal equipment of a single node. Therefore, before the design, the number of terminals is often clearly fixed, the design of the communication network topology can be attributed to the terminal node band
constraints of the minimum tree problem, is a difficult combination optimization problem. At present, there is no very effective algorithm to solve this problem, although when the problem is proposed, a heuristic algorithm is designed, obviously, it is very meaningful to further study the effective algorithm to talk about the problem. As a new modern heuristic algorithm, genetic algorithm has been widely and successfully applied in many fields of science and technology [2]. This paper will use the new technology of genetic algorithm to realize the pattern recognition of communication network data transmission.

2 Data Transfer Pattern Recognition for Terminal Band Constraint Communication Network

The frequency variation of the communication network data signal cannot be obtained by Fourier transforming the signal directly. This is mainly due to the time domain information of the signal is eliminated after the fourier change maps the signal to the frequency domain. the signal spectrum reaction obtained by the transformation is the statistics of the energy of the signal at different frequencies without the time combination information of the signal. For a disturbance signal, the information of the signal mainly exists in the combination of a certain energy of the signal and its time of occurrence. Therefore, a signal processing method is needed to show the signal frequency and time information in order to facilitate pattern recognition.

2.1 Transport Mode Analysis

Radio energy transmission technologies are classified according to the energy transmission mechanism and can be divided into the following three categories: First, electromagnetic radiation type of wireless power transmission technology. At present, this kind of technology usually uses laser energy transmission technology and microwave energy transmission technology to realize the long-distance transmission of small power. However, because of the strong transmission directionality of the technology, complex tracking and positioning systems are needed during operation, and there can be no obstacles in the transmission path [3]. Obviously, the limitation of the technique is relatively large and the application is small. Second, electromagnetic induction wireless power transmission technology. This technology belongs to the electromagnetic field near-field coupled radio energy transmission technology, based on the law of electromagnetic induction, through the original coil and secondary coil coupling electromagnetic field to achieve radio energy transmission. The electromagnetic induction wireless energy transmission technology can increase the transmission power and improve the transmission efficiency by adding high magnetic conductivity material to the air magnetic circuit. However, the effective transmission distance of this technology is limited, which is usually used in cases where the distance is less
than 10cm and the original and secondary coils are relatively fixed, and is generally suitable for charging and supplying power to small portable electronic devices or household appliances, etc. Third, magnetic resonance radio power transmission technology [4]. Magnetic resonance wireless energy transmission technology also belongs to electromagnetic field near field coupled radio energy transmission technology, it adds two resonant coils with high quality factor Q to the electromagnetic induction wireless power transmission technology, through the resonant coils of these two high q values, a higher intensity magnetic field can be generated in space, enabling the efficient transmission of electric energy at a further distance. The technology is characterized by no obvious directionality, the transmission distance is far more than the electromagnetic induction, the theory can achieve thousands of watts of power transmission at medium distance, and can be transmitted through non-metallic materials, the impact on the human body and the surrounding environment is relatively small, more safe and reliable.

The theory of mutual inductance coupling is the basic method for analyzing the electromagnetic field near-field coupled radio energy transmission technology. Its basic principle is magnetic field coupling:

$$\begin{align*}
\text{where } L_p \text{ is the primary coil, } L_s \text{ is the secondary coil, } I_p \text{ is the current value in the primary coil, } I_s \text{ is the current value in the secondary coil, and } M \text{ is the mutual inductance between the primary coil and the secondary coil. When the high frequency AC voltage with the angular frequency } \omega \text{ is connected in the primary coil, the high frequency electromagnetic field will be generated in the space around the coil. Because of the coupling between the primary and secondary coils, assuming that the mutual inductance value is } M, \\
\text{when the secondary coil senses the high-frequency electromagnetic field generated by the original stage coil, the high-frequency inductive voltage will be generated } V_s. \text{ The inductive voltage on the secondary coil is determined by the original stage coil current } I_p. \end{align*}$$

![Fig. 1 Coupled mechanism mutual inductance model](image-url)
\( j\omega MI_p \). At the same time, the primary coil will also produce a high-frequency induction voltage, determined by the secondary coil current \( I_s \), equal to \(-j\omega MI_s\). Obviously, the transmission power and efficiency of the coupling mechanism can be improved by increasing the angular frequency of the AC input to the primary coil. However, the impedance of the coil will also become very large with the increase of the angular frequency. In order to reduce the reactive power loss and make the transmission power maximum, the capacitance will generally be added to compensate the coupling mechanism to work in the resonant state\(^6\). There are two kinds of resonant compensation methods commonly used, namely series compensation resonance and parallel compensation resonance. When the primary coil and the secondary coil adopt series compensation resonance or parallel compensation resonance respectively, they can constitute four basic coupling topologies: SS, SP, PS and PP, as shown below:

![Fig. 2 Basic Topology](image)

The equivalent impedance of the secondary coil acting on the primary coil, as we call it, is usually expressed in \( Z_r \):
Fig. 3 Schematic diagram of the original stage circuit with reflected impedance

$Z_r$ is determined by the angular frequency $\omega$ of the AC voltage in the primary coil and the mutual inductance $M$ between the primary coil and the secondary coil:

$$Z_r = \frac{\omega^2 M^2}{Z_s}$$  \hspace{1cm} (1)

$Z_s$ is the equivalent impedance of the secondary coil, which can be calculated according to formula (2):

$$Z_s = \begin{cases} 
  j\omega L_s + \frac{1}{j\omega C_s} + R \\
  j\omega L_s + \frac{1}{j\omega C_s} + 1/R 
\end{cases}$$  \hspace{1cm} (2)

The real part of the reflected impedance reflects the active power transferred from the original stage coil to the secondary coil, and the imaginary part reflects the reactive power transferred from the original stage coil to the secondary coil.

### 2.2 Feature parameter extraction

Feature extraction needs to select the appropriate network as the feature extraction network of the algorithm according to the actual needs. A feature extraction network common in Faster R-CNN algorithm is ZF-NET, whose network structure is shown in the following...
Artificial neural network is one of the main research contents of pattern recognition processing. The artificial neural network based pattern recognition method has obvious advantages over the traditional pattern recognition method, which is embodied in the following aspects: (1) It has strong fault tolerance and anti-interference ability and is not easily affected by noise input; (2) Adaptive learning ability, can achieve automatic adjustment to approximate the objective function; (3) The process of pre-processing and identification can be carried out simultaneously; (4) Adopting a parallel mode of operation; (5) Using distributed memory for information, information is not easy to lose and robust [7]. Typical neuronal models are shown in the figure:

$x_j (j = 1, 2, ..., N)$ is the input signal of the neuron. $w_{ij}$ is the connection weight.
The activation function is used to perform the transformation between network input and output in a variety of forms, the most common of which are threshold function, piecewise linear function and sigmoid function. The threshold function is also called a step function, as shown in the following figure:

![Fig.6 Step function](image)

The symbolic functions shown below are also often used as neuronal excitation functions:

![Fig.7 Neural excitation function](image)

Feature parameter extraction is the process of extracting data that can reflect pattern features from input signals. Different patterns contain their own inherent characteristics, which are intrinsic factors and prerequisites for distinguishing different patterns [8]. The square difference of each time period is connected to the square difference curve of the whole disturbed time period, and the square difference of each time period represents the degree of data fluctuation in the time period. The variance curves of backward Rayleigh scattering caused by different disturbance modes are different, so the characteristic curves obtained by
squared difference processing are as follows:

![Characteristics curves obtained by squared difference processing](image)

As shown above, external disturbance signals affect the phase of the sensing signal, zero crossing rate is a basic method to extract the signal frequency-time, reflecting the frequency of the signal vibration near the zero level per unit time. The main advantage is that the operation is relatively simple and can suppress the subtle frequency change of the signal, so the short-term average frequency of the signal can be extracted to achieve the purpose of analyzing the external disturbance signal [9]. Over-level rate refers to the number of times a signal passes through a certain level when it changes near the level. It is a basic method of extracting the frequency-time of the signal, reflecting the frequency of the disturbance of the signal at a certain level per unit time.

\[
L_{CR} = \frac{\sum_{n=0}^{N-1} \delta(I(n) \geq \beta)}{N}
\]  

(3)

Where \( I(n) \) is the amplitude of the input signal point, \( \beta \) is the set level threshold, \( \delta \) is the indicator function, and the value is 1 when the condition in parentheses holds, otherwise 0. The short-time over-level rate will change with the change of the short-time average frequency and the short-time phase fluctuation of the sensor output signal. Therefore, the characteristic information of the disturbance signal can be understood by analyzing the short-time over-level rate of the signal.

2.3 Spark-KNN Fast Pattern Recognition Algorithm

The basic idea of the KNN algorithm is that if most of the K most similar samples of a sample in the feature space belong to a certain category, then that sample also belongs to that
category. Since the KNN method mainly relies on the surrounding limited adjacent samples, rather than the discriminant domain method, therefore, the knn method is more suitable than other methods for the pending sample sets with more crossover or overlap of the class domains. The establishment process of the pattern is as follows: establish and maintain a priority queue of size K by distance from large to small for storing nearest neighbor training samples \(^{10}\). The missing samples were randomly selected from the training samples as the initial nearest neighbor samples, the distance from the test samples to this k samples was calculated separately, and the training sample labels and distances were stored in the priority queue. Traversing the training sample set, calculating the distance between the current training sample and the test sample, comparing the resulting distance \(L\) with the maximum distance \(L_{\text{max}}\) in the priority queue. If \(L \geq L_{\text{max}}\) discards the sample and traverses the next sample. If \(L < L_{\text{max}}\), remove the sample with the maximum distance in the priority queue and store the current training sample in the priority queue until the traversal is complete. After the priority queue is updated and determined, the majority class of K samples in the priority queue is calculated and taken as the category of the test sample to complete the pattern recognition process.

3 Simulation experiment

In order to verify the validity of the communication network data transmission pattern recognition method, use traditional method 1 and traditional method 2 to compare with this method, a control experiment is proposed to obtain the test results.

3.1 Preparation process

Because the experiment needs to process a lot of data information, the requirements of the experimental platform are high, so the T330 model data processor is selected to complete the experiment. The specific parameters are as follows:

<table>
<thead>
<tr>
<th>Name</th>
<th>Specific parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>T330</td>
</tr>
<tr>
<td>structure</td>
<td>Tower type</td>
</tr>
<tr>
<td>category</td>
<td>Scalable single tower server</td>
</tr>
<tr>
<td>Number of internal hard disks</td>
<td>8</td>
</tr>
<tr>
<td>Disk array card</td>
<td>S130、H330、H730、</td>
</tr>
</tbody>
</table>
In order to ensure the validity of the experiment process, the advantages and disadvantages of traditional method 1 and traditional method 2 and the design method are analyzed in the form of comparative experiments. A test environment that needs to be designed to match the test method. A good test environment is beneficial to the implementation of the test method, and makes the test results more accurate and more efficient. The experiment uses socket to simulate the communication network data transmission process, which can be regarded as one of the endpoints in the communication connection when two network applications communicate. When communicating, one of the network applications writes a piece of information to be transmitted to the socket of its host, which sends this piece of information to the socket of another host via the transmission medium of the network interface card, enabling it to be transmitted to other programs. In network application design, because the core content of TCP/IP is encapsulated in the operating system, if the application is to use TCP/IP, it can be implemented through the programming interface of TCP/IP provided by the system. In the Windows environment, the network application programming interface is called Windows socket. To support users in developing application-oriented communication programs, most systems provide a set of application programming interfaces.
based on TCP or UDP, often in the form of a set of functions, also called sockets. The details are as follows:

Fig. 9 Operation process

After the above preparation is completed, the experiment is carried out to draw the conclusion.

3.2 Interpretation of result

Specific control results are shown below:

Fig. 10 Experimental results

According to Fig. 10, when the time passes 0.2s, the error classification rate using traditional method 1 is 19%, the error classification rate using traditional method 2 is 16%, and the error classification rate using this method is 14.8%. It can be seen that the error classification rate using this method is the lowest, and its data transmission pattern recognition error is low.

The following table is available:

Table 2 Experimental results

<table>
<thead>
<tr>
<th>fau</th>
<th>traini</th>
<th>Test</th>
<th>Accuracy rate%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sampling</td>
<td>Traditional Method 1</td>
<td>Traditional Method 2</td>
</tr>
<tr>
<td>-----</td>
<td>----------</td>
<td>----------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Z</td>
<td>25</td>
<td>84.00</td>
<td>84.00</td>
</tr>
<tr>
<td>N</td>
<td>25</td>
<td>81.48</td>
<td>85.19</td>
</tr>
<tr>
<td>X</td>
<td>25</td>
<td>83.67</td>
<td>87.76</td>
</tr>
<tr>
<td>Y</td>
<td>25</td>
<td>82.46</td>
<td>84.21</td>
</tr>
</tbody>
</table>

From the above results, the average accuracy rate using traditional method 1 is 82.9%, the average accuracy rate using traditional method 2 is 85.3%, and the average accuracy rate using this method is 88.7%. From this, the accuracy of the proposed pattern recognition method is obviously higher than that of the other two traditional methods.

4 Conclusion

In order to solve the problems of large error and low precision in data transmission pattern recognition of communication network, this paper proposes a method of data transmission pattern recognition based on the number of terminals with constraints. Using the mutual inductance model of the coupling mechanism to extract the characteristic parameters, and using Spark-KNN fast pattern recognition calculation, the data transmission pattern recognition of the communication network with constraints on the number of terminals is realized. The recognition method has high precision and small error, and can provide new ideas for research in related fields.

Reference


Fast Topology Inference of Wireless Networks Based on Hawkes Process

Yehui Song¹, Jiachen Sun², Guoru Ding³
{Songyh564@sina.com¹, sun_jiachen@outlook.com², dr.guoru.ding@ieee.org³}

College of Communications Engineering, Army Engineering University, Nanjing 210007, China¹,²,³

Abstract. Based on the reasonable equivalent assumption of communication events, using the Hawkes process to model the information interaction in wireless communication networks is an emerging direction in the field of non-cooperative topology inference. At present, topology inference algorithms based on the Hawkes process mainly use a fixed sample size for inference, considering only its reliability, but not regarding its effectiveness. In this paper, we consider introducing a sample size as a new performance indicator. For small sample size scenarios in wireless networks, a kind of fast topology inference algorithm is proposed, which uniformly represents parameters belonging to different dimensions, and thoroughly mines topological information from different batches to increase the speed and effectiveness of inference. Experimental simulations show that compared with the existing algorithm, our algorithm has better performance in small sample size scenarios.

Keywords: Fast topology inference, Wireless networks, Hawkes process, Effectiveness

1 Introduction

With the widespread application of wireless networks, the importance of intelligent analysis of network behaviors is becoming increasingly prominent. In the analysis of networks behaviors, learning and reasoning about the connectivity of unknown networks is a fundamental problem [1]. Through topology inference, we can mine the connection relationships between nodes to achieve the visual network's topology, and further improve our knowledge of the observed networks. For wireless networks management, topology inference can provide information support for managers to troubleshoot and enhance network security and robustness. For monitoring uncooperative wireless networks, especially in the field of electronic warfare, topology inference can help us identify and combat key nodes and critical links in enemy networks. Further, we can mine intelligence from topological information, infer the enemy's combat intentions, and realize the information superiority on the battlefield. But whether it is the topology inference for our network or the enemy networks, this requires timeliness, and we need to get the topology of the observed networks as quickly as possible. These observations motivate us to research fast topology inference in the framework of wireless networks.

The existing research on topology inference can be divided into two categories: cooperative topology inference and non-cooperative topology inference. Cooperative topology inference refers to the fact that sensor nodes are part of the perceived network and can reason out the topology of the network through message exchange within the network. Non-cooperative
topology inference refers to the fact that sensor nodes do not belong to the perceived network, and can reason out the topology of the network just only through passive detection of the network.

With the development of wireless networks, in recent years, researches on non-cooperative topology inference are receiving more and more attention [2]. Based on a reasonable equivalent assumption for communication events: the receiving nodes will act accordingly when they receive the signal from the sending nodes, using the Hawkes process to model the information interaction in wireless communication networks is an emerging direction in the field of non-cooperative topology inference. The Hawkes process is widely used in financial transactions, social network analysis, bioinformatics and many other fields [3]. In the field of communication, [4] models the information transmission process as the Hawkes process based on reasonable assumptions of communication equivalence. Based on [4], [5] further proposes a method named Low Cost Paths for Acyclic Graphs (LCPAG) to discover event chains. Then, [6] considers the wireless channel, develops a physical model for external topology awareness, and proposes a wireless channel-oriented topology sensing method based on the Hawkes process. However, the above works still have limitations: firstly, the fixed sample size is used for inference, only considering the reliability of the algorithm, without its effectiveness; secondly, there is a lack of research on small sample size scenarios in wireless networks, so it is necessary to improve the use efficiency of sample size. Thirdly, the choice of the global threshold will greatly affect the reasoning effect, and there are problems in selection in practical application.

Therefore, in view of the above problems, this paper reconsiders topology inference algorithm based on Hawkes process by considering both reliability and effectiveness. The main contributions of this paper are summarized as follows:

- Considering the effectiveness of the topological inference algorithm, the sample size is introduced, and a new performance index for the topological inference algorithm is proposed.
- Based on the idea of sequential detection, we propose a fast topology inference algorithm that uniformly represents parameters belonging to different dimensions and fully mines the topological information obtained from inference between different batches to improve the speed and effectiveness of inference.
- Through experimental simulations, we prove that the proposed algorithm has better performance in small sample size scenarios compared with existing algorithms.

The rest of this paper is organized as follows. Section 2 introduces the system model. In Section 3, the fast topology inference algorithm based on Hawkes process for wireless channels is proposed. Then, in Section 4, we present simulation results that demonstrate the advantages of the algorithm in small sample size scenarios. In addition, Section 5 presents conclusions and future prospects.

2 System Model

2.1 Single-node Sensing Model

As shown in Figure 1, this paper considers a single-node sensing model. The target wireless network $V$ is a wireless communication network of Unmanned Aerial Vehicles (UAVs), where the sensor $S$ has the ability of signal sensing and topology inference.
The working frame of the sensor S is shown in Figure 2. It means that there are \( N \) sample slots in a sensing interval and the data in a sensing interval is used for once topology inference. Once a sensing is over, the sensor will make inferences based on all existing data. If it meets the requirements, stop perception, otherwise continue until the inference effect reaches the requirements or the maximum number of perceptions is reached.

To simplify the problem, we make assumptions [6]: the transmission channel in the target wireless network \( V \) is the ideal channel, and the sensing channel outside the network is the AWGN channel, and the sensor S as an external observer has only two kinds of information available: the node transmitting signal (called the event sequence) and the time at which the signal was transmitted (called the time sequence).

For this model, our goal is to use the sensed event sequence and time sequence to reason about the topology of the target wireless network.
2.2 Mathematical Model of Information Interaction

The Hawkes process is a point process that has autoregressive dependence on past events, where the intensity of an event at any time is a function of the most recent event in the process. From this, we can discover the causality or correlation between events through the Hawkes process. In the field of wireless communications, it is natural that information exchange is performed by each node by transmitting signals. Now we will emit the signal as an event. Then, we can assume that when there is information interaction between two nodes, events on one node are likely to cause a response on the other node. Through such equivalent assumptions, we can model communication events between nodes as Hawkes processes, and then infer the topology of the communication network by solving key parameters.

For an event, given the time when the event occurred in the past \( t_k \), then the intensity of the occurrence of the event at the moment \( t \) is

\[
\lambda(t) = \mu(t) + \sum_{k=1}^{K} \gamma(t-t_k),
\]

where the parameter \( \mu(t) \geq 0 \) represents the basic intensity of the occurrence of the event; the parameter \( A \geq 0 \) represents the degree of self-motivation of the event, that is, how much the event occurs at the moment \( t_k \) has an impact on the event occurs at the moment \( t \); the kernel function \( \gamma(t) \) represents a time relationship between the corresponding events, which is a known, causal, non-negative, and integrable function. Under normal circumstances, we believe that the basic basis of an event does not change with time, that is, the probability of a new event occurring at any time is certain, so the formula (1) can become

\[
\lambda(t) = \mu + \sum_{k=1}^{K} \gamma(t-t_k).
\]

In a wireless communication network, for a node, the event of transmitting a signal is not only related to its own behavior, but also related to the behavior of other corresponding nodes. The one-dimensional Hawkes model can be easily extended to a process that contains multiple subprocesses. At this time, the intensity of the occurrence of a subprocess is not only affected by its own behavior, but also by the behavior of other subprocesses. Then, we can use the multidimensional Hawkes process to model the communication process in a wireless network, that is, a node transmitting a signal in the network can be see as a subprocess. Further, for a process with \( N \) subprocesses, according to the formula (2), the occurrence intensity of the \( n^{th} \) subprocess is

\[
\lambda_n(t) = \mu_n + \sum_{j=1}^{N} A_{nj} \sum_{k \in K_j} \gamma(t-t_k),
\]

where \( K_j \) represents the set of events in the \( j^{th} \) subprocess; \( \mu_n \) represents the basic intensity of the \( n^{th} \) subprocess, we also think that it does not change with time; \( A_{nj} \) represents the degree of response of the \( n^{th} \) subprocess to the \( j^{th} \) subprocess, and \( A_{nj}=0 \) represents the occurrence of the \( j^{th} \)
subprocess has no effect on the \( i \)th subprocess; and the larger \( A_{ij} > 0 \) becomes, the more likely the \( i \)th subprocess would occur because of the \( j \)th subprocess. But it should be noted that \( A_{ij} \) is not a probability value. In addition, we can call \( A \) adjacency matrix.

3 Fast topology inference in wireless networks

3.1 Determination of Hawkes Process Parameters

For the topology inference, the most critical thing is to get the parameters \( A_{ij} \) and \( \mu_i \), which we can use to infer whether there is a communication relationship between the two nodes. In many cases, the parameters are unknown, so we need to reason about the parameters based on the observations within a certain period, that is, time sequences and event sequences, to estimate the parameters. Naturally, we choose the maximum likelihood estimation method to determine the parameters. Then, in \( t \in [0, T] \), the negative likelihood function of the \( i \)th subprocess is

\[
L_i(\mu_i, A_{ij}) = \int_0^T \hat{\lambda}_i(t) - \sum_{k=k_i} \log \hat{\lambda}_i(t_k) \quad j = 1, 2, \ldots, N. \tag{4}
\]

The maximum likelihood estimation method requires us to minimize this convex function to estimate the parameters \( A_{ij} \) and \( \mu_i \). From [5], we choose to use quasi-Newton method for iterative solution. The data required for the solution of \( L_i(\mu_i, A_{ij}) \) is only derived from the independent data set related to the \( i \)th subprocess, so the parameter estimate of each subprocess is independent. Therefore, in \( t \in [0, T] \), the negative likelihood function of a process containing \( N \) subprocesses is

\[
L(\mu, A) = \sum_{i=1}^{N} L_i(\mu_i, A_{ij}) \quad j = 1, 2, \ldots, N. \tag{5}
\]

3.2 Fast topology inference

The solution of the maximum likelihood estimate for each node depends only on an independent subset of the data associated with it. That is, the data set used for the solution of \( A_{ij}(j = 1, 2, \ldots, N) \) is the same and is independent of the data set used for the solution of \( A_{ij}(k \neq i, j = 1, 2, \ldots, N) \). Therefore, \( A_{ij}(j = 1, 2, \ldots, N) \) of different nodes are not numerically comparable, so it is more reasonable to perform optimization individually rather than global optimization. For a node, the closer \( A_{ij} \) is to 0, the weaker the connection with another node is, and vice versa. Based on this premise, the traditional binary hypothesis is:

\[
\begin{align*}
H_0 & : A_{ij} \leq \alpha \quad (j=1,\ldots,N) \\
H_1 & : A_{ij} > \alpha \quad (j=1,\ldots,N)
\end{align*}
\tag{6}
\]
where $H_0$ represents there is no connection between the node $i$ and the node $j$, and $H_1$ represents connection exists between the two nodes. In this case, it is more challenging to select the global threshold. This paper is based on the idea of expanding the tendency by accumulation, for a fixed node, and put forward a unified statistic for the same batch:

$$Z_{ij}^k = \frac{A_{ij}^k - \pi_{ij}^k}{\pi_{ij}^k}, \quad (7)$$

where $A_{ij}^k$ indicates the communication intensity value obtained from the sample size of the node $i$ in the $k^{th}$ batch, and $\pi_{ij}^k$ indicates the average communication intensity value between the node $i$ and other nodes which may have the link with the node $i$ in the $k^{th}$ batch. When threshold of the binary hypothesis is set to 0, $Z_{ij}^k$ indicates the tendency of that $A_{ij}^k$ obey $H_0$ or $H_1$. Then, by accumulation and calculation:

$$Z_{ij}^k = \sum_{m=1}^{k} Z_{ij}^m, \quad (8)$$

this tendency will continue to become apparent while avoiding the limitations of setting global thresholds by artificial experience and reducing the need for sample size.

Since it is impossible to know the actual communication relationship inside the wireless network in advance, it isn’t very elementary to precisely quantify the performance of the inferred topology. For this problem, in the actual operation, we propose to analyze the adjacency matrix $A_{k-1}$ obtained from the previous batch as the actual adjacency matrix and then compare the adjacency matrix obtained from the current batch $A_k$ with $A_{k-1}$ to calculate the false alarm and missed alarm of the current batch. When the false alarm and missed alarm have little change compared with the previous batch, the sampling can be stopped, and the topology obtained by reasoning can be given.

The fast topology inference algorithm is summarized as follows: Algorithm 1. The main components of the algorithm are: Hawkes process parameter determination (lines 4-6), the unification of parameter values and the accumulation of the expansion tendency (lines 7-8), the decision of adjacency matrix (lines 10-16) and determining whether to stop sampling (line 3). As for the traditional topology inference algorithm, please refer to the paper [5].

Specifically introduce the key functions in Algorithm 1. The objective function in the line 5-6 is to use the pre-processed data set to estimate the required parameters by the maximum likelihood estimation method. It should be emphasized again that the parameter set of each node is estimated separately. $\pi_{ij}^k$ in the line 7 refers to the mean of all non-zero values in the parameter set related to node $i$ in the $k^{th}$ batch. Due to the unified and cumulative processing of the parameters, we can fixedly choose 0 as the decision threshold, as shown in lines 10-16.
Algorithm 1 Fast topology inference algorithm based on Hawkes process

**Input:** The number of nodes in the network, \( N_{\text{node}} \); The set of events (the source node ID), \( E_{\text{node}} \); The set of slots in which there is a signal sent, \( E_{\text{slot}} \); Number of samples, \( N_{\text{sample}} \); Threshold of false alarm probability ratio, \( \alpha \); Threshold of miss alarm probability, ratio, \( \beta \);

**Output:** Recovered influence matrix, \( A \);

1: Initialize influence matrix, \( \mathcal{A} \).
2: Sort \( E_{\text{node}} \) and \( E_{\text{slot}} \);

3: for each \( k \in [1, N_{\text{sample}}] \) and \( \frac{P_f^k}{P_f^{k-1}} < \alpha \) and \( \frac{P_m^k}{P_m^{k-1}} < \beta \)
4: for each \( i \in [1, N_{\text{node}}] \) do
5: \( (\mu_i^k, A_{ij}^k) = \arg \min_{\mu, A} L_i(\mu, A) \)
6: where \( L_i(\mu, A) = \int_0^T \hat{\lambda}_i(t)dt - \sum_{k \in K_i} \log \hat{\lambda}_i(n_k) \)
7: calculate \( \pi_i^k \), \( Z_{ij}^k = Z_{ij}^{k-1} + \frac{A_{ij}^k - \pi_i^k}{\pi_i^k} \)
8: \( A_{ij}^k = Z_{ij}^k \)
9: end for
10: for each \( i, j \in [1, N_{\text{node}}] \) do
11: if \( A_{ij}^k > 0 \) then
12: \( A_{ij}^k = 1 \)
13: Else if \( A_{ij}^k \leq 0 \) then
14: \( A_{ij}^k = 0 \)
15: end if
16: end for
17: compare \( A^k \) and \( A^{k-1} \) then calculate \( P_f^k \) and \( P_m^k \)
18: end for

### 4 Performance Evaluation

#### 4.1 Simulation Preparation

This paper considers that the target wireless network has a complex distributed topology, as shown in Figure 3. The network has 25 nodes. A fact can be concluded from the topology
diagram that there are only 8% node pairs in communication relationship. Because there is no actual data, we chose to use MATLAB to simulate the process of signal transmission between nodes and sensing of sensor to generate the event sequence and time sequence required for inference. It should be noted that, because AWGN is considered in sensor sensing channel, the sample data obtained has noise samples, that is, there are random event values and time values. This paper mainly considers the small sample size scenario. Set the number of times that 24 pairs of nodes communicate with each other in 96 in 1 batch, and there are 10 batches in all. Other specific simulation parameters are shown in Table 1.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distribution range of the network</td>
<td>1km x 1km</td>
</tr>
<tr>
<td>Noise spectral density</td>
<td>-174dBm/Hz</td>
</tr>
<tr>
<td>Noise bandwidth</td>
<td>10M</td>
</tr>
<tr>
<td>Signal power</td>
<td>-5 dBm</td>
</tr>
<tr>
<td>Signal decision threshold</td>
<td>$4 \times 10^{-14}$ W</td>
</tr>
</tbody>
</table>

Table 1. Simulation Parameters

Fig. 3. Target wireless network topology in simulation

4.2 Simulation Results

In order to facilitate observation and analysis, a total of five batches of samples were taken for reasoning. The traditional topology inference algorithm and fast topology inference algorithm were used for simulation, and the results are shown in Figure 4. It can be found that, under the same sample size, the inference effect of the fast topology inference algorithm is significantly better than the traditional topology inference algorithm. In other words, the proposed algorithm significantly improves the effectiveness of samples and the speed of inference.
Fig. 4. Comparison of simulation results based on different algorithms, in which the yellow represents that the two nodes are connected and the blue represents that the two nodes are disconnected. The subfigure (a) is the true topology; the subfigure (b) is the recovered topology by fast topology inference algorithm; the subfigure (c) is the recovered topology by traditional topology inference algorithm.

This paper firstly uses the probability of missed alarm and false alarm to measure the effect of topology inference, namely:

$$FNR = \frac{FN}{FN + TP} = \frac{FN}{P}$$

and

$$FPR = \frac{FP}{FP + TN} = \frac{FP}{N(N-1)-P}.$$  \hspace{1cm} (10)

Among them, $FNR$ represents the probability of missed alarm and $FPR$ represents the probability of false alarm. When there are $N$ nodes in the target wireless network, assuming that the total number of links in which there is a communication relationship is $P$, the total number of links in which there is no communication relationship is naturally $N(N-1)-P$.

Given the actual topology, the MonteCarlo simulation method is used to calculate the false alarm and missed alarm, and the number of repetitions is 1000 times. The results are shown in Figure 5. It can be found by comparison that the proposed fast topology inference algorithm is significantly superior to traditional topology inference algorithm in both false alarm and missed alarm indicators.

Based on Bayesian criteria, a comprehensive cost function $C$ can be introduced as follow:

$$C = C_{10} \times FPR \times P_0 + C_{01} \times FNR \times P_1,$$  \hspace{1cm} (11)

where $C_{10}$ represents the cost of the false alarm, and $C_{01}$ represents the cost of missed alarm. These two parameters generally take 1. And $P_0$ and $P_1$ are the prior probabilities of link and non-link between nodes in the network. It can be understood that the smaller the cost become, the better the inference performance.

From Figure 6, it can be more intuitively found that in a small sample size scenario, the cost of the fast topology inference algorithm is far less than the traditional topology inference algorithm, that is, the comprehensive performance of the fast topology inference algorithm is greatly improved.
Fig. 5. Comparison of specific performance of different algorithms, in which the blue represents the traditional topology inference algorithm and the yellow represents the fast topology inference algorithm. The subfigure (a) shows the false alarm varies with sample size; the subfigure (b) shows the missed alarm varies with sample size.

Fig. 6. Comparison of the comprehensive cost of different algorithms, in which the blue represents the traditional topology inference algorithm and the yellow represents the fast topology inference algorithm.

4.3 Performance analysis

For fast topology inference algorithms, it is necessary to make a balance between inference performance and sample size. Therefore, the sample size is introduced to form a new comprehensive cost function. Assuming that the sampling is homogeneous, each batch of
samples will consume a same cost, then the total cost function of topology inference after the $k$th batch of samples becomes:

$$C = C_{10} * P_f^k * P_0 + C_{01} * P_m^k * P_1 + cE\{k\},$$

(12)

where $P_f^k$ is the false alarm after the $k$th inference, $P_m^k$ is the missed alarm after the $k$th inference, and $E\{\bullet\}$ represents the average value under the average distribution. Therefore, the optimal choice for the sample size is:

$$\inf_{k \in \mathbb{K}}[C_{10} * P_f^k * P_0 + C_{01} * P_m^k * P_1 + cE\{k\}].$$

(13)

Constraint (13) provides a way to find the fastest topology inference.

5 Summary and Outlook

In this paper, we study the fast topology inference for small sample size scenarios in wireless networks. Under the single-node sensing model, by considering both reliability and effectiveness, we propose a fast topology inference algorithm based on the Hawkes process, and give a sample selection formula for the fastest topology inference. Simulation results show that the proposed algorithm outperforms traditional recognition algorithms in small sample size scenarios. In the next step, we can conduct research in the framework of distributed sensing and further explore the mathematical logic of the fastest topology inference.

References


Deep \(N\)-ary Error Correcting Output Codes

Hao Zhang\(^{1,}\)*, Joey Tianyi Zhou\(^{1}\), Tianying Wang\(^{1}\), Ivor W. Tsang\(^{2}\), Rick Siow Mong Goh\(^{1}\)
{zhang_hao, joey_zhou, wang_tianying, gohsm}@ihpc.a-star.edu.sg, ivor.tsang@uts.edu.au

Institute of High Performance Computing, A*STAR\(^{1}\); AII, University of Technology Sydney\(^{2}\)

**Abstract.** Ensemble learning consistently improves the performance of multi-class classification through aggregating a series of base classifiers. To this end, data-independent ensemble methods like Error Correcting Output Codes (ECOC) attract increasing attention due to its easiness of implementation and parallelization. Specifically, traditional ECOCs and its general extension \(N\)-ary ECOC decompose the original multi-class classification problem into a series of independent simpler classification subproblems. Unfortunately, integrating ECOCs, especially \(N\)-ary ECOC with deep neural networks, termed as Deep \(N\)-ary ECOC, is not straightforward and yet fully exploited in the literature, due to the high expense of training base learners. To facilitate the training of \(N\)-ary ECOC with deep learning base learners, we further propose three different variants of parameter sharing architectures for deep \(N\)-ary ECOC.

To verify the generalization ability of deep \(N\)-ary ECOC, we conduct experiments by varying the backbone with different deep neural network architectures for both image and text classification tasks. Furthermore, extensive ablation studies on deep \(N\)-ary ECOC show its superior performance over other deep data-independent ensemble methods.

**Keywords:** deep \(N\)-ary ecoc; ensemble learning; multi-class classification.

1 **Introduction**

Multi-class classification is one of the fundamental problems in machine learning and data mining communities, where one trains a model with labeled data of different classes for classification purposes. The multi-class classification exists diverse real-world applications from computer vision tasks such as object recognition [1], [2], [3], face verification [4], to natural language processing tasks like sentiment classification [5], [6].

To handle multi-class classification problems, existing approaches could be mainly divided into two groups. One group focuses on solving the multi-class problems directly by extending its corresponding binary classification algorithm. These approaches include decision tree-based methods [7], multi-class linear discriminant analysis [8], multi-layer perceptron [9], multi-class support vector machines (SVM) [10] and etc. Another research direction focuses on the decomposition of a multi-class problem into multiple binary subproblems so that one can reuse the well-studied binary classification algorithms for their simplicity and efficiency. Most of these methods can be reinterpreted in the framework of error correcting output codes (ECOC) [11], [12]. For example, Allwein et al. [13] show one-versus-one (OVO), one-versus-all (OVA) could be incorporated into the framework of ECOC.

* Corresponding Author
where all the classes are reassigned with either binary codes \{-1,1\} or ternary codes \{-1,0,1\} for each base learners (1/−1 represents positive/negative class, 0 represents non-considered class). Zhou et al. [14] further extend traditional ECOCs into N-ary ECOC by introducing N meta-classes rather than binary classification for each base learner. The final results are determined by the ensemble of a series of base learners. The biggest advantage of ECOCs methods is their easiness of implementation and parallelization.

Most traditional ECOCs methods are based on the pre-defined hand-craft features and focus on how to ensemble the results of base learners on these features. Recently, deep learning methods significantly advance the multi-class classification performance through learning features in an end-to-end fashion. For example, a single AlexNet [15] outperforms the second place at the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) by more than 10%. To further improve performance, Goodfellow et al. [16] demonstrate that a simple ensemble of seven AlexNet models with different random initiations could significantly reduce an error rate from 18.2% to 15.3%. In most high profile competitions, e.g. ImageNet\(^1\) or Kaggle\(^2\), ensembles techniques often appear in the winner solution. Traditional ensemble methods usually assume that the base learners for binary classification are inexpensive to train, such as SVMs and decision trees. Unfortunately, this assumption appears to be invalid with deep learning algorithms. For example, AlexNet consisting of more than 60 millions of parameters [15] takes between five and six days to train on two GTX 580 3GB GPUs. Therefore, the expensive learning procedure hinders the use of the ensemble of deep neural networks on a large scale.

In this paper, we focus on addressing the ensemble of deep neural networks in the framework of ECOC. The biggest reason to choose ECOC rather than other ensemble techniques such as Boosting [18] is that ECOC is easy to parallel due to the independence of base learners. In contrast, boosting trains a number of models sequentially and continuously compensates the mistakes made by the earlier models, which results in that each base models in boosting are highly dependent on each other. At this point, ECOC exhibits a large advantage in large-scale real-world applications since all the base learners could be trained independently and simultaneously.

Specifically, we choose N-ary ECOC, an extension of ECOC, which shows significant improvement over OVA, OVO, and traditional ECOCs [14]. As known, most existing works did not investigate the influence of deep learning on ECOCs or N-ary ECOC. In this paper, we make a marriage between N-ary ECOC to investigate such an influence. In the sequence, we term this problem as Deep N-ary ECOC. The main contributions of this paper are as follows:

- We investigate a new problem named Deep N-ary ECOC where we mainly discuss how to effectively and efficiently leverage advantages of deep learning models in the framework of ECOC.
- To facilitate the training procedure, we further propose three different parameter sharing strategies for Deep N-ary ECOC framework, i.e., full parameters share, partial parameters share, and no parameter share. Specifically, the full share model shares all the feature learning parameters except for top classifier; the partial share model shares part of feature learning parameters; the no parameter share means all the base learners are learned from scratch.

---

\(^1\) ImageNet: [http://www.image-net.org/](http://www.image-net.org/)

\(^2\) Kaggle: [https://www.kaggle.com/](https://www.kaggle.com/)
We explore the influence of two crucial hyper-parameters of $N$-ary ECOC, i.e., $N_L$ and $N$, with deep neural networks for improving the accuracy. We also give specific suggestions for choosing those two hyper-parameters.

We conduct extensive experiments and compare with several ensemble strategies, i.e., an ensemble of random initialization (ERI), ECOC and $N$-ary ECOC, on both image and text classification tasks to analyze the advantages and disadvantages of each ensemble strategy.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 presents Deep $N$-ary ECOC. Finally, Section 4 discusses our empirical studies and Section 5 concludes this work.

2 Related Work

Our proposed deep $N$-ary ECOC is highly related to the following topics, including ECOCs, ensemble learning, and deep neural networks.

2.1 ECOCs

Many ECOC approaches [13], [19], [20], [21] have been proposed to design a good coding matrix in recent years. Most of them are fallen into the following two categories. The first one is data-independent coding, such as OVO, OVA, and ECOCs [20]. Their coding matrix design is not optimized for the training dataset nor the instance labels such that all the base learners could be independently learned. For example, the sparse ECOC coding approach aims to construct the ECOC matrix $M \in (-1,0,1)^{N_C \times N_L}$, where $N_C$ is the number of classes, $N_L$ is the code length, and its elements are randomly chosen as either $-1$, $1$, or $0$ [20]. In ECOCs, the classes corresponding to $1$, $-1$ are considered as positive and negative classes, respectively, and $0$ are not considered in the learning process. More recently, Zhou et al. [14] extend the existing ECOCs into $N$-ary ECOC to enable the construction of $N$ metaclasses. Both theoretical and empirical findings validate the superiority of $N$-ary ECOC over traditional ECOCs.

Another direction is data-dependent ECOCs where the data are considered in the learning coding matrix, such as discriminant ECOC (D-ECOC) [19], ECOC-ONE [22], subspace ECOC [21], Adaptive ECOC [23], etc. In this way, different base learners interact with each other during training phrases, which is also similar to Boosting [18] methods, such as AdaBoost [24]. In Boosting methods, a series of models are sequentially trained with latter models correcting mistakes committed in previous models. Compared to data-independent ECOCs, these methods require sophisticated algorithm design and are difficult to be parallelled.

To our best knowledge, there is little research to investigate the combination of ECOCs and deep learning. In this paper, we take a step further to analyze the performance of combining our previous work $N$-ary ECOC with deep learning.

2.2 Deep Ensemble Learning

A lot of studies show that deep neural network models are nonlinear and have a high variance, which can be frustrating when preparing a final model for making predictions [16].
Deep ensemble learning appears to one of the solutions that combine the predictions from multiple neural network models to reduce the variance of predictions and reduce generalization error. Recently, there are some studies to integrate base learners of deep neural networks with ensemble learning in three major ways. The first one is ensemble training data including re-sampling [25], bootstrap aggregation [26], where the choice of data is varied for training different base models in the ensemble. The second one is to ensemble models where different base models are used in the ensemble, including different random initialization, a random selection of mini-batches, differences in hyper-parameters, etc [16]. The third way is varying combinations where one vary the choice of combining outcomes from ensemble members. The most famous method is a model averaging ensemble and weighted average ensemble. Different from the aforementioned deep ensemble learning methods, deep N-ary ECOC serves a complementary piece for existing methods.

2.3 Deep Neural Networks

In recent years, a lot of different deep neural networks are proposed for different applications. For computer vision tasks, the most dominating model comes from Convolutional Neural Networks (CNNs), and its follow-up works such as AlexNet [15], VGGs [27], ResNet [28] and DenseNet [29]. For natural language processing tasks, most popular networks belong to Recurrent Neural Network (RNNs), or its many variants such as Long Short-Term Memory (LSTM) [30], Gated Recurrent Unit (GRU) [31] and etc. In the experiment, we validate deep N-ary ECOC in both CNNs and LSTMs architecture for vision and text datasets respectively.

3 Deep N-ary ECOC

In this section, we first introduce the concept of N-ary ECOCs. To facilitate the training procedure, we further propose three different parameter sharing architectures, namely full, partial and no sharing.

3.1 N-ary Ensemble for Multi-class Classification

Error correcting output codes (ECOC) constructs an ensemble of binary base classifiers by randomly and independently assigning positive/negative pseudo labels (i.e., 1/−1 in the coding matrix) for each base task. The results of all the base learners are combined to make a prediction. ECOC consists of two main steps: 1) encoding 2) decoding. In encoding, we create an encoding matrix to encode each class into a unique code that is as different as possible from the codes of the remaining classes. One example of the encoding matrix is illustrated in Table 1(a). A row of the coding matrix represents the code of each class, while a column of the coding matrix represents the binary classes to be considered when learning a base classifier. In decoding, ECOC first computes the prediction vector that is the concatenation of results of all the base tasks. The final label is determined by assigning the label with the “closest” label vector in encoding matrix $M \in \{1,2, \ldots, N\}^{N_c \times N_l}$, where $N_c$ denotes the number of classes and $N_l$ denotes the number of base learners. As proved in many research works [12], [14], the capability of error correction relies on the minimum distance, $\Delta_{\text{min}}(M)$, between any distinct
pair of rows in the coding matrix $M$. In this way, the trained base classifiers could be sufficiently differentiated from each other.

To achieve this goal, ECOC is extended to a new framework of $N$-ary ECOC [14] where the original classes are decomposed into $N$ meta-class ($3 \leq N \leq N_c$). Table 1(b) shows an example of $N$-ary ECOC encoding matrix. Zhou et al. [14] both empirically and theoretically showed that $N$-ary ECOC is able to achieve larger row separation and lower column correlation. It is interesting to note that $N$-ary ECOC is a more general framework for ECOC since traditional coding schemes could be treated as special cases of $N$-ary ECOC. For example, when $N = 2$, $N$-ary ECOC corresponds to the binary coding scheme; when $N = 3$, $N$-ary ECOC corresponds to the ternary coding scheme. Furthermore, recent works [16] showed that an Ensemble of models with different Random Initialization (ERI) is able to improve multi-class classification performance. This deep ensemble learning strategy could be also viewed as a special case in the framework of $N$-ary ECOC if we keep the original label assignment, namely $N = N_c$.

**Table 1. Example of ECOC and $N$-ary coding matrix.**

(a) ECOC

<table>
<thead>
<tr>
<th>$C_1$</th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M_3$</th>
<th>$M_4$</th>
<th>$M_5$</th>
<th>$M_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_2$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>$C_3$</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$C_4$</td>
<td>-1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
</tr>
<tr>
<td>$C_5$</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>$C_6$</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>$C_7$</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
</tr>
<tr>
<td>$C_8$</td>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>$C_9$</td>
<td>-1</td>
<td>1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
<td>-1</td>
</tr>
</tbody>
</table>

(b) $N$-ary ECOC with $N = 4$

<table>
<thead>
<tr>
<th>$C_1$</th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M_3$</th>
<th>$M_4$</th>
<th>$M_5$</th>
<th>$M_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C_2$</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$C_3$</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$C_4$</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>$C_5$</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>$C_6$</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>$C_7$</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>$C_8$</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>$C_9$</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>$C_{10}$</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

On the other side, most existing work on ECOCs including our previous work on $N$-ary ECOC is constrained to classifier training with pre-defined features. With a significant advance of deep learning, the performance of various machine learning tasks has been improved. There is few works to discuss how to extend ECOC in the scenario of deep learning. In this work, we specifically study this open problem in the framework of $N$-ary ECOC, termed Deep $N$-ary ECOC, and propose several approaches to address it. In this paper, we mainly investigate the following three questions:

1. Do we necessarily independently train all the deep base learners from scratch for all the situation?
2. Whether the $N$-ary ECOC framework still retains the advantages over other data-independent ensemble approaches with deep neural network?
3. Any new suggestion on the choice of the meta-class number $N$ and base learners number $N$?
For the first question, we are going to propose three different parameter sharing architectures, which is described in more details next section. For the remaining two questions, we delay the investigation in the experiment section.

### 3.2 Efficient Implementation for Deep N-ary ECOCs

Different from traditional ECOC problems with pre-defined features, Deep ECOCs should consider deep feature learning as well as the classifier construction during training. This increases the difficulty of deploying ECOCs in real-world scenarios, since even training a single deep neural network is also expensive. Fortunately, thanks to the nature of ECOCs, all the base deep neural networks could be trained simultaneously. Furthermore, in this paper, we investigate a more efficient realization and propose three different parameter sharing strategies, namely, no share, partial share, and full share, which is depicted in Figure 1.

Typically, we take the model of the CIFAR dataset, termed CIFAR-CNNs (explain in detail later), as an example to illustrate the three strategies. For the no parameter sharing strategy, as shown in Figure 1(a), we trained $N_b$ base learners independently, which means that the feature encode layers of each base learner are trained by the inputs directly and do not interact with other learners. The partial parameter sharing strategy contains shared and task-specific layers, as in Figure 1(b), the first three feature encoder layers are shared by all the base learners while the top encoder layer is task-specific, which is only optimized by the corresponding meta-class objectives. The full parameter sharing strategy is simply set all the feature encode layers to be shared by all the base learners except the top classifiers (see Figure 1(c)). The top layer classifiers of all the sharing strategies are trained independently with its meta-class objectives. Note that, all the base learners of no parameters sharing strategy are trained from scratch while the shared layers of partial and full parameters sharing strategies are initialized by a pre-trained single model and fine-tuned through training to accelerate the model convergence rate. Obviously, the no parameter sharing strategy contains most parameters ($N_n$), then the partial sharing strategy ($N_p$) and the full sharing strategy ($N_f$) is least, say, $N_n > N_p > N_f$.

![Figure 1](image)

**Figure 1.** An example of three different parameters sharing strategies.

### 4 Experiments

#### 4.1 Datasets

We conduct the experiments on 4 image datasets and 2 text datasets. The image datasets contain MNIST [32], CIFAR-10 [33], CIFAR-100 [33], and FLOWER-102 [34], which are
widely used image classification datasets in the computer vision community. While the text datasets are Text REtrieval Conference (TREC) [35] dataset and Stanford Sentiment Treebank (SST) [36] dataset. The TREC is the question and answering dataset which involves classifying question sentences into 6 question types, say, whether the question is about person, location, numeric information and etc. The SST is the sentiment analysis sentence data with 5 classes that range from 0 (most negative) to 5 (most positive). The statistics of these datasets are described in Table 2. Note that we do not utilize the K-fold cross-validation method, but simply use the split of train/validation/test sets. If the datasets do not contain development part, we randomly split 10% training samples as the development dataset.

<table>
<thead>
<tr>
<th>Image Dataset</th>
<th>Dataset</th>
<th>Image Size</th>
<th># Train Sample</th>
<th># Dev Sample</th>
<th># Test Sample</th>
<th># Classes ($N_c$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td></td>
<td>28×28</td>
<td>60,000</td>
<td>N/A</td>
<td>10,000</td>
<td>10</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td></td>
<td>32×32</td>
<td>50,000</td>
<td>N/A</td>
<td>10,000</td>
<td>10</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td></td>
<td>32×32</td>
<td>50,000</td>
<td>N/A</td>
<td>10,000</td>
<td>100</td>
</tr>
<tr>
<td>FLOWER-102</td>
<td></td>
<td>256×256</td>
<td>6,552</td>
<td>N/A</td>
<td>818</td>
<td>102</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Text Dataset</th>
<th>Dataset</th>
<th>Avg. Sent. Len.</th>
<th># Train Sample</th>
<th># Dev Sample</th>
<th># Test Sample</th>
<th># Classes ($N_c$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TREC</td>
<td></td>
<td>10</td>
<td>5,500</td>
<td>N/A</td>
<td>500</td>
<td>6</td>
</tr>
<tr>
<td>SST</td>
<td></td>
<td>18</td>
<td>11,855</td>
<td>N/A</td>
<td>2,210</td>
<td>5</td>
</tr>
</tbody>
</table>

### 4.2 Experimental Setup

**Deep Neural Networks.** We employ different neural network-based models for different datasets. Specifically, we use LeNet [32] for the MNIST dataset and the FLOWER-102 dataset is trained by AlexNet [15]. Note that, due to the difficulty for the AlexNet model to learn consequential and representative features from the small training dataset of FLOWER-102 directly, the AlexNet is not trained from scratch but obtained by fine-tuning the pre-trained AlexNet model\(^3\), which is trained on ILSVRC dataset. For CIFAR-10/100 datasets, we build a model with eight convolutional layers and two full-connected layers, named as **CIFAR-CNNs**, as shown in Figure 2(a), where the eight convolutional layers are divided into four groups, they share the same structure with the different numbers of filters and kernel widths. The architecture of each group is structured as follows: one convolutional layer following the batch normalization [37] and dropout [38] layer, another convolutional layer with batch normalization and max-pooling is applied. And the ELU [39] activation function is used for each convolutional layer.

To train the TREC and SST text datasets, we construct a three-layer bidirectional LSTM model with character-level CNN [40] and self-attention [41] mechanism, termed **Bi-LSTMs**, as shown in Figure 2(b), where the character-level CNN learned the character features to represent a word from the character sequences of such word, which can help to enrich the meaning of word features, especially for rare and out-of-vocabulary words, and boost the performance by capturing morphological and semantic information, and the self-attention mechanism encodes the learned contextual affluent word-level feature sequence of

\(^3\) Pre-trained AlexNet: [http://www.cs.toronto.edu/~guerzhoy/tf_alexnet/](http://www.cs.toronto.edu/~guerzhoy/tf_alexnet/)
bidirectional LSTM into a single vector by considering the importance of each word-level feature.

**Parameters Setup.** For LeNet of MNIST dataset, we follow the same settings as LeCun et al. [32] and RMSProp [42] is chosen as the parameters optimization method with a learning rate of 0.001 and decay rate of 0.9, we also introduce Dropout [38] strategy with a drop rate of 0.5 at each convolutional layer and the first full-connected layer to prevent over-fitting. While the AlexNet for FLOWER-102 dataset, we utilize exactly the same structure, parameter setting, and optimization method as Alex et al. [15]. For CIFAR-CNNs model of CIFAR-10/100 datasets, we set the number of filters for each convolutional block as 32, 64, 128, and 256, respectively, the kernel sizes of (3, 3) and pool size of (2, 2) for all the blocks. The hidden size of the first fully-connected layer is 512, while the second depends on the class size. To avoid over-fitting, we apply $\ell_2$ regularization with weight decay rate of 0.0005 for all the weight parameters and Dropout strategy, the drop rate is 0.3, 0.4, 0.4, 0.4 for each convolutional block respectively, and 0.5 for the first fully-connected layer. Parameters optimization is performed by Adam optimizer [43] with gradient clipping of 5.0 and learning rate decay strategy. We set the initial learning rate of $\beta_0 = 0.002$ and fixed it for the first 5000 training iterations, then the learning rate $\beta_i$ is updated by $\beta_i = \beta_0 / (1 + \rho \times \frac{t}{5000})$, where $t$ is the decay step of 500 and $\rho$ is the decay rate of 0.05. Meanwhile, in order to improve the performance, the data augmentation is also utilized.

![CIFAR-CNNs model](image1.png) ![Bi-LSTMs model](image2.png)

**Figure 2.** The general architecture of CIFAR-CNNs and Bi-LSTMs models.

For the Bi-LSTMs model of TREC and SST text datasets, we use the 300-dimensional publicly available pre-trained word embeddings as the word-level feature representation, which is trained by fastText package on Common Crawl and Wikipedia [44], [45], and the 50-dimensional randomly initialized task-specific character embeddings. The word embeddings are fixed and character embeddings are learned during training. We use three different

4 fastText: [https://github.com/facebookresearch/fastText](https://github.com/facebookresearch/fastText)
convolutional layers with widths 2, 3, 4, respectively, for character-level CNN encoder and set the filter number of each layer as 20, the learned character features of each layer are concatenated and then optimized by a two-layer highway network [46] before concatenating with the corresponding word embeddings. The dimension of hidden states of LSTM layers are set as 200. Parameters optimization is performed by Adam optimizer [43] with gradient clipping of 5.0 and learning rate decay strategy. We set the initial learning rate of $\beta_0 = 0.001$, at each epoch $t$, learning rate $\beta_t$ is updated by $\beta_t = \beta_0/(1 + \rho \times t)$, where $\rho$ is the decay rate with 0.05. To reduce overfitting, we also apply Dropout [38] at the embedding layer and the output of each LSTM layer with the drop rate of 0.2 and 0.3, respectively.

**N-ary ECOC Coding Matrix Setup.** For $N$-ary ECOCs, including ECOCs (which is a special issue of $N$-ary ECOC when $N = 2$), we train the $N_L$ base learners based on the coding matrix and use the predicted code sequence of each class and generated coding matrix to make a prediction based on distance measurement. Zhou et al. [14] introduced several coding matrix construction methods and distance measurements designed for general or task-specific applications. For simplicity, we utilize the random dense encoding method to randomly split the original classes $N_c$ into $N$ subsets and make sure that the number of classes in each subset should be approximately balanced, simultaneously. For the decoding method, we adopt the minimum Hamming distance due to its simplicity and effectiveness. In our experiments, we experiment on the different numbers of meta-class $N$ and number of base learners $N_L$ for different datasets, as described in Table 3. Note that we do not experiment on all the possible meta-classes for each dataset, because of the limitations of computing resources and we only trained 60 base learners for MNIST, FLOWER-102, TREC, and SST datasets and 100 base learners for CIFAR-10/100 datasets, respectively. Specifically, in order to evaluate the effects of number of base learners on the ensemble learning performance, we trained another 300 classifiers for FLOWER-102 and TREC datasets, respectively.

<table>
<thead>
<tr>
<th>Dataset</th>
<th># Classes ($N_c$)</th>
<th>Tested # Meta-Class ($N$)</th>
<th>Tested # Base Learners* ($N_L$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>10</td>
<td>2, 4, 5, 8, 10</td>
<td>60</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>10</td>
<td>2, 4, 5, 8, 10</td>
<td>100</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>100</td>
<td>2, 5, 10, 30, 50, 75, 95, 100</td>
<td>100</td>
</tr>
<tr>
<td>FLOWER-102</td>
<td>102</td>
<td>2, 3, 5, 10, 20, 40, 60, 80, 90, 95, 102</td>
<td>60</td>
</tr>
<tr>
<td>TREC</td>
<td>6</td>
<td>2, 3, 4, 5, 6</td>
<td>60</td>
</tr>
<tr>
<td>SST</td>
<td>5</td>
<td>2, 3, 4, 5</td>
<td>60</td>
</tr>
</tbody>
</table>

* It indicates the maximal number of classifiers is used for training.

### 4.3 Experimental Results

#### 4.3.1 Comparison with Different Ensemble Methods

In this section, we compare the performance of different ensemble methods on the aforementioned image and text datasets. In the experiment, we trained a single model and ensemble models of three coding schemes for each dataset, i.e., Ensemble with Random Initializations (ERI), ECOC, $N$-ary ECOC, then report their (ensemble) accuracy with standard deviations. Note that we only report the highest score under a specific meta-class $N$ for $N$-ary ECOC. For the MNIST, FLOWER-102, TREC, and SST datasets, we use 60 base
learners for each scheme, while 100 base learners for CIFAR-10/100 datasets. The results are summarized in Table 4. Generally, we observe that most ensemble models show relatively significant improvements, compared with the single model, on the given datasets with different deep neural networks.

We observe two interesting results in Table 4. First, comparing the single model with $N$-ary ECOC, we find that the improvement ratio of $N$-ary ECOC is inverse relation with single model performance, i.e., the improvement of $N$-ary ECOC scheme is more prominent if the performance of the single model is lower. For example, it is obvious that the baseline model performance, i.e., the improvement of $N$-ary ECOC, we find that the improvement ratio of $N$-ary ECOC is higher on MNIST, CIFAR-10, TREC and FLOWER-102 datasets, respectively, while the improvement ratios of CIFAR-100 and SST (almost $<60\%$) than on CIFAR-100 and SST (almost $>80\%$), then the improvement ratios are $0.59\%$, $5.54\%$, $5.64\%$ and $5.80\%$ from the single model to $N$-ary ECOC on MNIST, CIFAR-10, TREC, and FLOWER-102 datasets, respectively, while the improvement ratios of CIFAR-100 dataset are $13.72\%$ and $15.15\%$ on SST dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>Single Model</th>
<th>ERI</th>
<th>Ensemble Model*</th>
<th>N-ary ECOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>LeNet [32]</td>
<td>98.98 ± 0.07%</td>
<td>99.11 ± 0.11%</td>
<td>99.23 ± 0.08%</td>
<td>99.57 ± 0.09%</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>CIFAR-CNNs</td>
<td>87.12 ± 0.43%</td>
<td>90.54 ± 0.31%</td>
<td>89.37 ± 0.54%</td>
<td>91.95 ± 0.24%</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>CIFAR-CNNs</td>
<td>61.50 ± 0.57%</td>
<td>69.57 ± 0.29%</td>
<td>34.26 ± 2.42%</td>
<td>69.94 ± 0.32%</td>
</tr>
<tr>
<td>FLOWER-102</td>
<td>AlexNet [15]</td>
<td>83.12 ± 0.29%</td>
<td>86.32 ± 0.60%</td>
<td>77.05 ± 0.73%</td>
<td>87.94 ± 0.28%</td>
</tr>
<tr>
<td>TREC</td>
<td>Bi-LSTMs</td>
<td>90.50 ± 0.12%</td>
<td>94.80 ± 0.09%</td>
<td>95.80 ± 0.08%</td>
<td>95.60 ± 0.10%</td>
</tr>
<tr>
<td>SST</td>
<td>Bi-LSTMs</td>
<td>44.17 ± 0.92%</td>
<td>48.69 ± 0.18%</td>
<td>48.91 ± 0.26%</td>
<td>50.86 ± 0.13%</td>
</tr>
</tbody>
</table>

* Here $N_c$ are 60, 100, 100, 60, 60 and 60, respectively, for the ensemble models from top to bottom row. While $N$ are 3, 4, 95, 95, 3, 4, respectively, for the $N$-ary ECOC.

Second, the $N$-ary ECOC scheme outperforms ECOC and ERI ensemble methods on most image and text datasets, except for the TREC text dataset. Specifically, $N$-ary ECOC always performs better than ERI. This is due to that $N$-ary ECOC varies the predicted classes for each base learner and makes them more diverse than ERI, where the diverse forecast errors made by base learners of $N$-ary ECOC are more beneficial to the ensemble learning in comparison to the similar base learner errors of ERI. Meanwhile, compared with ECOC, $N$-ary ECOC also shows its superiority in most cases, especially when the number of classes is large (i.e., $N_c \geq 100$ in our experiments). It is primarily due to, as mentioned by Zhou et al. [14], the better quality of the coding matrix and the higher discriminative ability (in terms of how many meta-classes a base learner tries to discriminative) of $N$-ary ECOC than ECOC.

In fact, we find the contribution of class merge degree to the ensemble accuracy of $N$-ary ECOC replies on the dataset, say, the datasets with a different number of classes require different class merge degree strategy, as discuss in Section 4.3.2. Note the class merge degree, which is measured by $\frac{N_c-N}{N_c}$, is the ratio of class numbers reduced when the classes are merged into meta-classes.

### 4.3.2 Evaluation on the Effect of Meta-class Number $N$

In this section, we investigate the influence of meta-class number $N$, which is one of the crucial hyper-parameters of $N$-ary ECOC. For the datasets with a small value of $N_c$, we experiment on all the possible meta-class numbers, i.e., from $2$ to $N_c$ ($N = 2$ denotes ECOC...
and \( N = N_c \) denotes ERI), while for the datasets with a large value of \( N_c \), we select several representative meta-class numbers for the experiment. The ensemble accuracies with respect to \( N \) are depicted in Figure 3.

From Figure 3(a), we observe that the performances of ensemble models with different \( N \) are relatively stable, the highest ensemble accuracies of MNIST, CIFAR-10, and SST achieve when \( N = 3 \), \( N = 4 \), and \( N = 4 \) respectively, and the best performance of TREC is obtained at \( N = 3 \) if we do not consider the ECOC. After that, the performance of each dataset is gradually decreased with small fluctuations with an increasing number of meta-class \( N \). It is interesting to see that \( N \)-ary ECOC for datasets with a small value of \( N_c \) always tends to arrive the best performance with small value of \( N \), i.e., large class merge degree. Specifically, the class merge degree for MNIST, CIFAR-10, TREC and SST are 0.7, 0.6, 0.5 and 0.2 respectively.

![Figure 3](image)

**Figure 3.** Ensemble accuracies with respect to \( N_c \), where the first point of each line represents ECOC (\( N = 2 \)), the last represents ERI (\( N = N_c \)) and the rest is \( N \)-ary ECOC with various \( N \).

However, as shown in Figure 3(b), the performance of ensemble models with different \( N \) fluctuates significantly on the datasets with a large value of \( N_c \). For the ECOC scheme, it only achieves 34.26\% ensemble accuracy on the CIFAR-100 dataset and 77.05\% on the FLOWER-102 dataset. For the \( N \)-ary ECOC with \( N = 3 \), it obtains 83.52\% and 59.75\% on FLOWER-102 and CIFAR-100 datasets, respectively. Then the ensemble accuracy improves gradually with the increase of meta-class \( N \) and reaches the summit with accuracies of 87.94\% and 69.94\%, when \( N = 95 \), for FLOWER-102 and CIFAR-100 datasets, and mildly decreases after the optimal performances. The ensemble accuracies of the ERI scheme (\( N = N_c \)) on these two datasets are slightly lower than that of \( N \)-ary ECOC. Obviously, the ECOC fails to address the datasets with a large value of \( N_c \), while the higher ensemble performance of \( N \)-ary ECOC needs a large value of \( N \), namely, a small class merge degree. For the FLOWER-102 and CIFAR-100 datasets, the \( N \)-ary obtains good results after \( N \geq 75 \) and achieve best at \( N = 95 \). In particular, the class merge degree for FLOWER-102 is 0.069 and CIFAR-100 is 0.05.

In general, we conclude from the experiment that the ensemble performance is relatively stable for the datasets with a small value of \( N_c \), which slightly improves until the peak and then decreases a bit, or just slightly decreases with the increase of \( N \) after the peak. While the performance, for the datasets with a large value of \( N_c \), boosts significantly at the very beginning, then it saturates as \( N \) continues increasing and reaching the optimum when \( N \) is
close to \(N_c\). This could be explained by that the base learners with large \(N\) has stronger discriminability [14].

Thus, our suggestions for the choice of \(N\) are: 1) For the dataset with small \(N_c\), the large class merge degree strategy, i.e. small \(N\), is better for achieving good performance, such as \(N = 3\) or \(4\) for the dataset with \(N_c \leq 10\). 2) Reversely, for the dataset with large \(N_c\), the small class merge degree strategy should be applied, e.g. \(75 \leq N \leq 95\) for \(N_c\) is around 100.

### 4.3.3 Evaluation on the Effect of Base Learner Number \(N_L\)

In this experiment, we further explore another crucial hyper-parameter of \(N\)-ary ECOC, namely the number of base learner \(N_L\) (also equivalent to the code length), and study its influence on the ensemble accuracy. We first report the ensemble accuracies of different \(N_L\) for each dataset with the optimal meta-class number \(N\), as described in Table 5. Then, we study the ensemble accuracies of different meta-class \(N\) with respect to \(N_L\) (see Figure 4 and 5).

From Table 5, we observe that one requires a smaller number of base learners \(N_L\) for datasets with small \(N_c\) than that for datasets with large \(N_c\) to reach the optimal ensemble accuracies generally. For example, MNIST and TREC only need 50 base learners to get the optimum, while SST obtains best accuracies with 60 base learners and CIFAR-10 requires 80. In comparison, it reaches the optimal ensemble accuracies with the help of 100 base learners on CIFAR-100 (note that it first reaches optimum when \(N_L = 90\)). There is a special issue that FLOWER-102 holds a large \(N_c\) (102 classes), but only requires 60 base learners to derive the optimal ensemble accuracies. It is because the pre-trained model on the large-scale dataset (ILSVRC dataset in our experiment) is utilized and the pre-trained model already encodes a variety of abstractly and typically well-learned features. Moreover, we also find that the requirement of \(N_L\) is related to the single model performance to some degree, say, the single model achieves better performance, then its ensemble model requires fewer base learners to achieve the optimal result.

![Table 5](image)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>(N)</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>45</th>
<th>50</th>
<th>60</th>
<th>80</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>3</td>
<td>99.14%</td>
<td>99.20%</td>
<td>99.35%</td>
<td>99.48%</td>
<td>99.57%</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>4</td>
<td>87.45%</td>
<td>89.76%</td>
<td>91.78%</td>
<td>91.83%</td>
<td>91.82%</td>
<td>91.92%</td>
<td>91.95%</td>
<td>91.93%</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>95</td>
<td>67.94%</td>
<td>69.12%</td>
<td>69.11%</td>
<td>69.33%</td>
<td>69.34%</td>
<td>69.46%</td>
<td>69.67%</td>
<td>69.94%</td>
</tr>
<tr>
<td>FLOWER-102</td>
<td>95</td>
<td>86.06%</td>
<td>86.45%</td>
<td>86.45%</td>
<td>87.06%</td>
<td>87.16%</td>
<td>87.94%</td>
<td>87.46%</td>
<td>87.59%</td>
</tr>
<tr>
<td>TREC</td>
<td>3</td>
<td>93.80%</td>
<td>94.00%</td>
<td>95.20%</td>
<td>95.20%</td>
<td>95.60%</td>
<td>95.60%</td>
<td>95.50%</td>
<td>95.60%</td>
</tr>
<tr>
<td>SST</td>
<td>4</td>
<td>46.74%</td>
<td>48.19%</td>
<td>49.41%</td>
<td>50.18%</td>
<td>50.45%</td>
<td>50.86%</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

In addition to the observations from Table 5, we also study the impact of \(N_L\) on ensemble performance under different meta-class \(N\). Obviously, the optimal number of base learners \(N_L\) for achieving the best accuracy is related to the meta-class \(N\), as shown in Figure 4 and 5. Normally, an ensemble model with small meta-class \(N\) needs more base learners \(N_L\) to achieve the same result compared to an ensemble model with large meta-class. It is because the discriminative ability of the codes for small \(N\) is worse than that for large \(N\).
Considering the definition of ECOC, $N$-ary ECOC and ERI schemes, the discriminative ability of ECOC is worst due to its small meta-class ($N = 2$), which means ECOC needs relatively most base learners to reach the optimal performance compared with $N$-ary ECOC and ERI, where ERI holds the best discriminative ability. Thus, we conclude that $N_l$ for ECOC is greater than or equal to $N_l$ for $N$-ary ECOC, while $N_l$ for $N$-ary ECOC is greater than or equal to $N_l$ for ERI. Note that we use “greater than or equal to” since there is no guarantee that the optimal $N_l$ for a small $N$ must be larger than that for a large $N$, especially for some extreme situations such as $N = 2$ (ECOC) versus $N = 3$ ($N$-ary ECOC) or $N = 99$ ($N$-ary ECOC) versus $N = 100$ (ERI) for the dataset with 100 classes ($N_c$).

In Figure 4, the experiment results on all the datasets show similar trends that ensemble accuracies of larger $N$ converge faster than that of smaller $N$ as the increasing of $N_l$, which means larger $N$ requires less $N_l$ and vice versa. For example, as shown in Figure 4(a), ECOC reaches optimal ensemble accuracies at $N_l = 100$, while $N$-ary ECOC with $N = 4$ and 5 optima at $N_l = 80$, then $N_l = 60$ for $N = 8$ and ERI peaks at $N_l = 55$. The patterns on TREC and SST datasets are consistent as CIFAR-10. Typically, such patterns are more distinct for datasets with large $N_c$ (ref. Figure 4(d) and 4(e)). For instance, in Figure 4(d), the ensemble accuracies of $N = 10$ are highest at $N_l = 100$, $N_l = 95$ for $N$-ary ECOC with $N = 30, 50, 75$, while 90 base learners are required for $N = 95$ and ERI needs $N_l = 85$. Here we do not take ECOC into consideration, since it fails to improve the ensemble accuracy with only 34.26%. For the Figure 4(e), we see that ensemble accuracies of $N = 2, 3, 5, 10$
converge at $N_L = 60$, $N = 20, 40$, converges at around $N_L = 50$, $N = 60, 80, 90$ at approximately $N_L = 45$ and $40$ base learners are needed for $N$-ary ECOC with $N = 95$ and ERI to reach the optimal ensemble accuracy.

Apart from the optimal $N_L$ for each meta-class $N$ to reach optimal ensemble accuracy, we also observe that using 15~25 base learners for ERI is good enough for datasets with small $N_c$ while 20~40 base learners for large $N_c$. For ECOC, it fails with the large $N_c$, and on the dataset with small $N_c$. Although ECOC performs comparably to $N$-ary ECOC and ERI, it still needs more base learners to converge, which is different from the conclusion in [13] that ECOC requires $N_L = 10 \log_2(N_c)$ on traditional classifiers. For $N$-ary ECOC, the optimal performance is highly related to the choice of $N$. If the choice of $N$ follows suggestions in Section 4.3.2, 40~60 base learners for small $N_c$ are enough to achieve good performance, while large $N_c$ needs around 60~100 base learners.

![Figure 5](image.png)

**Figure 5.** Ensemble accuracies with respect to large $N_L (= 300)$ for three coding schemes on FLOWER-102 and TREC datasets.

We further extend number of base learners to 300 and experiment on the FLOWER-102 and TREC datasets to investigate ensemble performances with the increasing $N_L$, as in Figure 5. From Figure 5(a), we find that the performance of ECOC improves significantly when $N_L$ increases, then keep relatively stable with a slight increase after $N_L = 100$ and reach the optimal accuracy of 82.17% at around $N_L = 270$. However, the best performance of ECOC derived by using a large number of base learners is still lower than $N$-ary ECOC with $N = 95$ and ERI with only 5 base learners used, which indicates that ECOC is not suitable for the large $N_c$ case. For the $N$-ary ECOC and ERI, they obtain good scores with only small numbers of base learners and slightly improve to the optimal accuracy at around $N_L = 40$. After that, the performance remains stable with the increase of $N_L$ and it drops when $N_L$ continues to increase, which indicates that increasing $N_L$ monotonously has no impact on performance. Similar observations could be found in Figure 5(b).

Generally, there is no concrete conclusion for the choice of the number of base learners $N_L$, but some helpful guidelines can be summarized for experiments: 1) The choice of meta-class $N$ is more important than the number of base learners $N_L$ for the performance of $N$-ary ECOC, especially for the dataset with large $N_c$. Since the increase of $N_L$ cannot compensate for the negative effects caused by a badly selected $N$ (e.g., $N = 10$ for CIFAR-100). 2) Albeit the optimal number of base learners $N_L$ varies along $N_c$, the suggested $N_L$ is in the range of $[10 \log_2(N_c)], [10 \log_{1.5}(N_c)]$. For example, the optimal $N_L$ ranges in [30, 58] for $N_c = 10$ and [59, 110] for $N_c = 100$, which aligns with the observations in our experiments.
4.3.4 Comparison with Three Parameter Sharing Strategies

In this Section, we study the effect of three different parameter sharing strategies in the framework of ECOC, $N$-ary ECOC, and ERI. Note that, for the $N$-ary ECOC framework, we only select the optimal meta-class $N$ of each dataset for display except for the CIFAR-100 dataset which four different $N$ are chosen for display. We first study the performance of three different parameter sharing strategies on each tested dataset.

![Figure 6. Parameter sharing strategies in ECOC, $N$-ary ECOC and ERI for TREC dataset.](image)

From the experimental results on the TREC dataset (see Figure 6), we observe that no parameter sharing strategy performs better than partial and full parameter sharing strategy for ECOC, $N$-ary ECOC, and ERI. When the number of base learner $N_L$ is small, the performance of no share is not satisfactory. Then it improves significantly with the increase of $N_L$, while the performances of partial and full share are relatively stable with respect to $N_L$. Moreover, when the number of meta-class $N$ is small, partial share outperforms the full share and the performance of no share is much better than partial and full share. However, when $N$ is large, full share is better than partial share and the performance of no share is just slightly higher than partial and full share.

![Figure 7. Parameter sharing strategies in ECOC, $N$-ary ECOC and ERI for SST dataset.](image)

From Figure 7, we have the following observations. First, when the number of meta-class $N$ is small, both partial and no share models improve significantly with the increase of $N_L$. The partial share generally outperforms the no and full share except when $N_L$ is less. Second, when the number of meta-class $N$ is large, as shown in Figure 7(b) and 7(c), the performance of the
three strategies are stable, and the improvement of no share is most significant with the increase of $N_L$. No share strategy governs the best performance with $N = 4$ while partial share strategy always performs best for ERI situation.

**Figure 8.** Parameter sharing strategies in ECOC, $N$-ary ECOC and ERI for CIFAR-10 dataset.

**Figure 9.** Parameter sharing strategies in ECOC, $N$-ary ECOC and ERI for CIFAR-100 dataset.

In Figure 8, the performances of no, partial, and full share strategies are more stable. When the number of base learners $N_L$ is small, we see that the performance of no share is worst with ECOC and $N$-ary ECOC, and partial share performs better with $N$-ary ECOC and ERI situations. With the increase of $N_L$, for ECOC, all the strategies improve significantly, partial share outperforms another two strategies at the beginning, and then no share comes closer to partial share and reaches slightly higher performance than partial share. For $N$-ary ECOC, partial and full share strategies do not show significant improvement, while no share
improves obviously and outperforms the partially and full share despite its lower ensemble accuracy at the very beginning. For the ERI, all these three strategies perform stable while no share always performs best and the performance of full share stays the bottom.

In the last experiment, we study the parameter sharing strategies in ECOC, N-ary ECOC, and ERI for the dataset with a large number of classes, as shown in Figure 9. For N-ary ECOC situation, we experiment on four different meta-class with \( N = 10, 30, 50, 95 \).

First, we observe that ECOC model with no share strategy fails to achieve satisfactory performance, while partial and full share strategies with the ECOC improve significantly with the increase of \( N_L \). Moreover, partial share always outperforms full share.

Secondly, for the N-ary ECOC with small number of meta-class, we observe that partial share strategy outperforms no and full share always. No share improves most significantly and its performance is comparable to that of partial share with the increase of \( N_L \). The performance of full share always maintains the worst. With an increasing number of meta-class \( N \), partial share strategy outperforms no share strategy at the beginning, but its performance is gradually surpassed by no share when number of base learners \( N_L \) increases. For \( N = 50 \) and \( 95 \), the performance of no share is comparable to that of partial share when the number of base learners \( N_L \) is small. No share outperforms partial share with the increases of \( N_L \). Moreover, for the N-ary ECOC, full share strategy consistently performs worst.

Thirdly, for the ERI model, the observations are similar to the N-ary ECOC with large meta-class \( N \) and the no share strategy is comparable to partial share when \( N_L \) is small. It always performs best when \( N_L \) increases, meanwhile, the performance of full share is worst.

Finally, we conclude that: 1) In general, for the dataset with the small number of classes, the performance of no share model is better than or equal to that of the partial share model, thus no share strategy is suggested to be chosen. 2) For the dataset with the small number of classes, when the number of meta-class \( N \) is large, these three strategies perform stable. 3) For the dataset with a large amount of classes, when the number of meta-class is small, the performance of partial share model is the best. 4) For the dataset with large amount of classes, when the number of meta-class is large, no share strategy model outperforms partial and full share models in most cases. Thus no share strategy should be preferred in such a case. 5) If the number of meta-class is large, the performance difference between three sharing strategies is marginal. Then full share could be suggested due to its parameter efficiency.

5 Conclusion

In this paper, we mainly investigate how to effectively integrate deep learning with the N-ary ECOC framework, also termed Deep N-ary ECOC. To achieve this goal, we give three different realizations. We further carry out extensive experiments to show the superiority of deep N-ary ECOC over existing data-independent deep ensemble strategies.

Acknowledgement. The paper is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project No. A18A1b0045). Ivor W. Tsang was supported by ARC DP180100106 and DP200101328.
References

A Machine Learning Method for Optimizing Partition of Prediction Block in Coding Unit in H.265/HEVC

Wenchan Jiang¹, Ming Yang², Ying Xie¹, and Zhigang Li²
wenchan.jiang@americold.com, ming.yang@kennesaw.edu, ying.xie@kennesaw.edu, zhigang.li@kennesaw.edu

¹Americold Logistics, Atlanta, GA, USA
²College of Computing and Software Engineering
Kennesaw State University, Marietta, GA 30060, USA

Abstract: In the latest generation of video coding standard – H.265/HEVC, the partition of Coding Tree Unit (CTU) into CU (Coding Unit) is a very critical yet time-consuming component. Traditional methods find the optimum partition mode for each CTU through iterative and exhaustive search, which is a very time-consuming process and hinders its application to real-time video streaming scenarios. In this research, we explored and implemented a machine learning based method to avoid the exhaustive search and to improve the performance of encode/decode by optimizing the partition of prediction block in the coding unit. Our results in coding unit split pattern prediction show a significant performance improvement in terms of processing time.

Keywords: Machine Learning, CNN, CTU, Partitioning, H.265/HEVC, Coding Unit.

1 Introduction

Following the previous video coding standard H.264/AVC, High Efficiency Video Coding (H.265/HEVC) has been deemed as the newest video coding standard of the ITU-T Video Coding Experts Group and the ISO/IEC Moving Picture Experts Group [1]. HEVC has the potential to deliver better performance than earlier standards such as H.264/AVC. H.265/HEVC has introduced a refined content-adaptive approach for Coding Block Partitioning (corresponding to fix-sized macroblock in previous standards), which has significantly improved the compression efficiency. However, the implementation in H.265/HEVC reference software HM explore all the possibilities in a traversal and exhaustive manner to find the best partition and merge pattern for a specific prediction unit. It is a time-consuming process and will be difficult, if not impossible, to stream ultra HD video contents in real time using the new HEVC standard.

Over the past decade, machine learning has become one of the top trending information technologies deeply integrated into our life. With the ever-increasing amounts of data becoming available, it’s reasonable to believe that smart data analysis will become even more pervasive as a necessary ingredient for technological progress. [2]. In this paper, we utilized the machine learning technique in Keras framework to speed up the process of encoding by improving the partition speed of prediction block in coding block. This approach avoids the
exhaustive searching process and effectively speeds up the coding process to enable its application in real-time video streaming scenarios.

2 Background

2.1 High Efficiency Video Coding (H.265)

The new standard of HEVC, or H.265, promises a 50% storage reduction as its algorithm uses efficient coding by encoding video at the lowest possible bit rate while maintaining a high image quality level. H.265 still uses the widely accepted hybrid coding framework since H.264, including intra-frame prediction, inter-frame prediction based on motion compensation, transformation, entropy coding, and quantization. Table 1 summarizes the improvements of H.265/HEVC has made over H.264/AVC [3].

<table>
<thead>
<tr>
<th>Category</th>
<th>H.264</th>
<th>H.265</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partition Size</td>
<td>Macroblock 16 × 16</td>
<td>Coding Unit 8 × 8 to 64 × 64</td>
</tr>
<tr>
<td>Partitioning</td>
<td>Sub-block down to 4 × 4</td>
<td>Prediction Unit Quadtree down to 4 × 4</td>
</tr>
<tr>
<td>Intra Prediction</td>
<td>Up to 9 predictions</td>
<td>35 predictions</td>
</tr>
<tr>
<td>Transform</td>
<td>Integer DCT 8×, 4×4</td>
<td>Transform Unit square IDCT from 32×32 to</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4×4 + DST Luma intra 4×4</td>
</tr>
<tr>
<td>Filters</td>
<td>Deblocking filter</td>
<td>Deblocking filter, Sample Adaptive Offset</td>
</tr>
<tr>
<td>Motion Prediction</td>
<td>Spatial Median (3 blocks)</td>
<td>Advanced Motion Neighbor Vector Prediction (AMVP) for both spatial and</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Entropy Coding</td>
<td>CABAC, CAVLC</td>
<td>CABAC</td>
</tr>
</tbody>
</table>

2.2 CU Partitioning in HM

Coding Tree Units (CTU) and Coding Unit (CU) are the most important concepts in our research. The CTU is the basic processing unit similar to MacroBlock (MB) in previous generation of coding standards such as H.264/AVC. The size N of the CTUs is chosen by the encoder, ranging from 16 × 16, 32 × 32, to 64 × 64, with 64x64 usually being the default. CTU may be too big to decide whether we should perform inter-picture prediction or intra-picture prediction. Thus, each CTU can be differently split into multiple CUs and each CU becomes the decision-making point of prediction type. The size of the CU can range from the same size as the CTU to a minimum size of (8 × 8). CTU corresponds to MB in previous standards, which has a fixed size of 16x16. In the past decade, we have much higher frame sizes to deal with since 4K production became practical. Therefore, larger macroblocks are
needed to efficiently implement motion estimation and motion compensation for these frame sizes. On the other hand, blocks at the granularity of 4×4 are also essential to process prediction and transformation of small details.

As mentioned above, the CTU is further partitioned into multiple CUs to adapt to various local characteristics. Rate-Distortion analysis is utilized to determine the optimum partition. Such recursive procedure will be repeated for all the CUs in a CTU, as well as for all CTUs in one frame, and finally the optimal split mode for each CU within one video sequence will be obtained. This processing order of CUs can be interpreted as a depth-first traversing in a Zig-Zag order in the coding tree structure as shown in Figure 1 below. Figure 2 also illustrates a broader view of how CTU partitioning works within a frame.

Fig. 1. Example of CTU partitioning and processing order when size of CTU is equal to 64×64 and minimum CU size is equal to 8×8. (a) CTU partitioning. (b) Corresponding coding tree structure [4]
The split pattern of a specific CTU in the reference software HM is exhaustive and recursive. It means HM will calculate the rate distortion cost of every possible split pattern (i.e., 17 combinations in total), and further compare these cost values recursively to obtain the best split pattern based on the minimum rate distortion cost. Understandably, it is a very time-consuming process, and it will be unrealistic to compress and transfer videos of ultra-high definition integrating H.265 standard in a real-time streaming manner. There will be a lot of overhead upfront that is unbearable and unfortunate to apply HEVC new standard in a broader spectrum.

3 Literature Review

Different approaches have been proposed to speed up the coding process of H.265/HEVC. The common goal is to avoid the exhaustive search process in coding decision making and speed up the encoding process. Momcilovic et al proposed a novel fast Coding Tree Unit partitioning for H.265 encoder [6]. The proposed approach decouples the requirement of any pre-training and yields a high adaptivity to the dynamic changes in video contents. The proposed methodology has reduced the encoding time for up to 65% with negligible rate-distortion penalties. In another study the authors proposed a machine learning based method using features that describe CU statistics and sub-CU homogeneity [7]. Comparatively, the experiment results can achieve 36.8% complexity reduction on average with only 3.0% bitrate increase. In Alam’s work [8], a fast Convolutional-Neural-Network (CNN) based quantization strategy for HEVC was proposed. They utilized the contrast gain control model to develop a structural facilitation model to capture effects of recognizable structures on distortion visibility. Liu et al. proposed a fast algorithm based on convolution neural network to decrease no less than two CU partition modes in each CTU for full rate-distortion optimization (RDO) processing, therefore reducing the encoder’s hardware complexity [9] [10]. The proposed algorithm can save 63% Intra encoding time at the cost of average 2.66% BDBR increase. A fast coding unit (CU) depth decision algorithm for intra coding of HEVC using an artificial neural network (ANN) and a support vector machine (SVM) was proposed by Chen et al [11]. In their methodology, machine learning provided a systematic approach for developing a fast algorithm for early CU splitting or termination to reduce intra coding computational complexity.

4 Proposed Approach and Experimental Settings

The procedures for splitting CTUs into CUs in the reference software HM for HEVC is very ineffective and insufficient, thus making it unsuited for real-world streaming service. In this study, we implemented CNN in place of the split procedure of CTUs in the reference software.

4.1 Training Data Set Generation
The training and testing data consist of 13 video clips. For each video, 80% of frames are used as training dataset and the rest 20% as testing dataset. Each frame is divided into $32 \times 32$ sized CTUs. For a typical frame in our experiment, it is divided into 99 CTUs.

The luma samples for each CTU will be used as the input of the CNN, and the output of CNN will be the split pattern/intra prediction mode. For a $32 \times 32$ CTU with minimum $4 \times 4$ CU, there are altogether 17 possible split patterns. The CNN network will be trained to generate a compiled model that will be used to predict the labels for new inputs, and exhaustive search for the optimum split patterns will be avoided. As long as comparable levels of accuracy can be achieved, the CNN will be deemed as fully trained.

One of the most powerful and easy-to-use Python libraries for developing and evaluating deep learning models is Keras. It wraps the efficient numerical computation libraries Theano and TensorFlow. The advantage of this is mainly that you can get started with neural networks in an easy and fun way. In our experiment, we use Python as the language and Theano as the backend.

4.2 Convolutional Neural Network Design

In deep learning, selection of optimal number of layers and neurons is also one of the hyperparameters that can be fine-tuned. A method is to add layers until it starts to overfit the training set. Then it is time to add dropout or another regularization method. The idea is that once your network overfits you're sure that it is powerful enough for your task. The dropout helps to prevent feature co-adaptation and therefore avoid over-fitting. The number of neurons in each layer is not really sensible. Usually a bit more or as much neurons should be put on the first layer than inputs, and the number should decrease slowly as we approach the output layer.

In our research, we tried different combinations of parameters and keep the one with the lowest loss value or better accuracy on the validation set. We tried two possibilities of 2 and 3 fully connected layers. For 2 layers, it has 512 and 17 hidden units for the fully connected layers. Whereas for 3 layers, it has 512, 128, and 17 hidden units for each layer.

For the dataset that we trained, the accuracy is slightly higher for all video clips except one when we use 2 fully connected layers versus 3 layers. Another benefit of using less fully connected layers is less amount of execution time. In our case, average execution time for 3-layer is 7 seconds per epoch, versus 5 seconds per epoch for 2-layer network, which results in 28% decrease of time when we choose 2-layer. Therefore, we will use 2 fully connected layers across our experiments.

After fine tuning the parameters, we have decided on the CNN’s structure shown in Figure 4 below. There are 2 main layers in the network with 1 convolution layer and 1 max pooling layer in each main layer. Each convolution layer consists of 32 filters in the size of $3 \times 3$ to extract the feature map. The activation function is ReLU across the boarder. The border mode is set as “valid”, which means there is no padding around input or feature map. For max pooling layer, the pool size is $2 \times 2$ with strides of $2 \times 2$. After two main layers, it comes with the fully connected layers where the hidden units will be “flattened” and directed to the output of 17 labels.
5 Results and Discussion

5.1 Max Pooling

The purpose of max pooling is to down-sample an input representation to reduce its dimensionality and allow for assumptions to be made about features contained in the sub-regions binned. Max pooling is employed in our experiments for split pattern prediction after each layer by using a $2 \times 2$ filter. Although the network can achieve similar accuracy either with or without employing max pooling layer, with max pooling layer in place, it only needs about half of execution time. Similar trends are observed in other clips, and we take one of the clips as an example and summarize the observations in Table 2.

Table 2. Comparison of execution time for max pooling

<table>
<thead>
<tr>
<th></th>
<th>with max pooling</th>
<th>w/o max pooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>time of each epoch (s)</td>
<td>10</td>
<td>23</td>
</tr>
<tr>
<td>accuracy of 30 epochs (%)</td>
<td>70.8</td>
<td>71.4</td>
</tr>
</tbody>
</table>

5.2 Dropout

In our experiment, dropout is applied to reduce the complexity of the model and prevent overfitting. Also, training will be faster with dropout set. We tuned dropout ratios in our experiments. Three different dropout ratios were tested: 0.5, 0.25, and 0.1. Dropout ratio of 0.5 generates the lowest accuracy, while dropout ratio 0.1 generates the highest accuracy. Small dropout ratio obviously has higher level of computational overhead, compared to the cases of greater dropout ratio. However, in our experiments, such kind of cost is marginal compared to the extra accuracy achieved. Thus, we set dropout ratio as 0.1 by taking all the factors into account.
It should also be noted that dropout is only applied during training, and we need to rescale the remaining neuron activations. Specifically, if 50% of the activations in a given layer is set to zero, we need to scale up the remaining ones by a factor of 2.

5.3 Training and Testing Results

For a typical sample video clip with 300 frames and each frame with size of 352 × 288 pixel, there will be $11 \times 9 \times 300 = 29700$ CTUs generated by reference software HM and available to use. Specifically, the input will be the matrix of pixel value of 32 × 32 CTU as demonstrated in figure below. During the process of building the model, such big data set will be split into 80/20 portion of training/test data set. Therefore, the test data set can be used to validate the training results.

During the training process, the input data is fed into the Keras model with the parameters configured as mentioned in the previous section. For some sample video clips, 50 iterations, which takes roughly 25 minutes with the test machine can yield ~90% training accuracy. In the experiment, the video clips of bus, mobile, tempete, and flower can quickly achieve high training accuracy as summarized below (Figure 6). The reason is because these video clips have relatively smooth scenes, without many variations in terms of pixel changes and ranges. So, the model can quickly learn the relationship between the input pattern and split pattern with relatively high accuracy.

![Figure 2. Training accuracy of ~90% in 50 cycles](image)

For the other video clips, it generally requires about 100, sometimes even 200 iterations to achieve ~90% training accuracy (Figure 7). These video clips generally have more variations and more complex scenes. Thus, it takes longer time for the CNN to adapt to these changes and train itself to achieve a high level of prediction accuracy.
After the network is trained, we start to use the CNN to predict split patterns for the testing data set. Most of the video sample clips achieved reasonably high accuracy of around 90% (Figure 8). Another observation is that more training data set in each video sequence will in general lead to higher accuracy. The reason is because more frames will help the network to adapt to the variation in the video sequence.

![Testing Accuracies](image)

**Fig. 3.** Training accuracy of above 90% in 200 cycles

**Fig. 4.** Comparison of testing accuracies for different sample video
In addition to the good training accuracy for split pattern prediction, the trained network also significantly improved the speed of splitting the CTUs. The results from our experiment indicated that, the trained model is able to predict the split pattern for a specific CTU with over 90% accuracy within 400 microseconds, which is a significant improvement from 4000 microseconds per CTU in HM.

6 Conclusion

As discussed earlier in this paper, the CU split pattern prediction method in the HEVC reference software is ineffective and not well-suited for real-world use cases. In this study, we took advantage of the convolutional neural network to improve the prediction speed of the CU split pattern. Our results demonstrated a prediction accuracy over 90% and a significant 90% improvement in prediction speed. In the future, we are planning to apply CNN based methodology to replace other computationally intensive modules in H.265/HEVC encoder and speed up the coding process of Ultra-HD video.

7 References


Study of Synthetic Airspeed Algorithm Based on Machine Learning for Lift Coefficient Curve Fitting

Dianzhong Chen¹, Yue Xu², Lei Wang³
{dc2e12@163.com¹, xuyue2@comac.cc², wanglei13@comac.cc³}

COMAC Shanghai Aircraft Design & Research Institute, No. 5188, Jinke Road, Shanghai, China¹,²,³

Abstract. Traditional air data system of airplane utilizes pitot probe for airspeed measurement. However, problems such as icing and bird strike will lead to failure of pitot probe. Airspeed display loss is rated as disastrous loss status. Airspeed calculation algorithm based on inertial data and movable surface positions (status of flaps and slats) has been studied by the Boeing Company and the Airbus Company and applied in airplane models of Boeing 787 and Airbus A350. Commercial Airplane of China has been dedicated in studying algorithm of airspeed calculation. Study indicates the importance of accurate lift coefficient identification for different flight configurations under certain attack angles. Theoretical analysis indicates the relationship of piecewise linearity between lift coefficient and attack angle. Based on the above relationship, machine learning algorithm of support vector regression (SVR) is applied to process air data. Furthermore, synthetic airspeed algorithm is proposed and verified.

Keywords: synthetic airspeed algorithm, support vector regression (SVR).

1 Introduction

Airspeed is an important air data provided in air data system (ADS) [1]. ADS, a multiple input multiple output (MIMO) system, measures air data of total pressure, static pressure, attack angle, sideslip angle, total temperature through sensors of pitot probe, static port, attack angle sensor, total temperature sensor, and provides data for subsystems, such as flight control system, flight management system, fuel system [2,3]. Accurate airspeed data is the basis for operation of flight control system [4,5]. However, accidents such as icing and bird strike will cause failure of pitot probe.

To avoid disastrous loss status of airspeed display loss, airspeed calculation algorithm without total pressure data is studied, which can provide a non similar data source, as a supervision for pitot probe state and a substitute in case of pitot probe system (two or three pitot probes with voting algorithm) failure. Chen proposes an airspeed calculation method based on inertial navigation data and wind field data [6]. While, wind field changes rapidly and accuracy of data from weather forecast is not high enough. A method for Mach number and true airspeed calculation using data of altitude from global positioning system (GPS) and...
inertial reference system (IRS) is put forward in patent [7]. However, it is known that neither of these two systems can give accurate altitude data with enough precision, for problems of positioning algorithm and accumulative error, respectively. Airspeed calculation through data from inertial navigation system (INS) and flight control system (FCS) exhibits acceptable results, reported in [8] and [9]. Up till now, mature airspeed calculation is only applied at advanced airplane models of Boeing 787 and Airbus A350.

In the paper, a synthetic airspeed algorithm based on INS and FCS data is to be proposed. To deal with nonlinear characteristic of attack angle-lift coefficient curve, machine learning algorithm of SVR is to be applied. Air data from two flight tests will be used to verify the accuracy of the proposed synthetic airspeed algorithm.

2 Synthetic Airspeed Modeling

Synthetic airspeed calculation relies on Expression (1) between impact pressure $q_c$ and calibrated airspeed $V_{CAS}$.

$$V_{CAS} = C_{S0} \left[ 5 \left( \frac{q_c}{P_{S0}} + 1 \right)^{7/2} - 1 \right]^{1/2},$$  \hspace{1cm} (1)

in which $C_{S0}$ is the conversion ratio from Mach to Knot, $P_{S0}$ is the normal atmosphere pressure, $q_c$ is the impact pressure, denoting actual pressure sensed by a moving creature. According to Expression (1), synthetic calibrated airspeed $V_{CAS}$ is the function of parameter impact pressure $q_c$. When airspeed is smaller than 0.3 M, impact pressure $q_c$ is approximately equal to dynamic pressure $q_{bar}$, which is defined under the condition of impessible fluid, calculated by:

$$q_c \approx q_{bar} = q_t - q_s,$$  \hspace{1cm} (2)

where $q_t$ and $q_s$ are total pressure and static pressure, respectively. Total pressure $q_t$ and static pressure $q_s$ can be sensed by pitot probes and static port sensors. While, when airspeed is larger than 0.3 M, air is compressible, and impact pressure is unequal to dynamic pressure. Mach number will reflect the compressibility of air. The difference between $q_c$ and $q_{bar}$ increase with the rise in Mach number. However, a more accurate $q_c$ could be obtained through correction on $q_{bar}$ by Mach.

Except for operation between $q_c$ and $q_{bar}$, dynamic pressure $q_{bar}$ can be derived through the following equation:
\[
q_{\text{bar}} = \frac{mN_z}{SC_L},
\]

in which \(m, N_z, S, C_L\) respect gross weight, load factor (Z-axis acceleration), reference wing area, lift coefficient, respectively. The above four parameters on the right of Equation (3) are real-time values. Gross weight change of an airplane mainly comes from fuel consumption and is monitored by flight management computer. Z-axis acceleration is sensed by inertial navigation module. Movable surface positions (status of flaps and slats) influence reference wing area (parameter \(S\)) and lift coefficient \(C_L\) simultaneously. In the following study, given the multiplication relationship between \(S, C_L\), and relatively small area change of \(S\) caused by different movable surface positions, \(S\) is approximately regarded as a constant value. Influence by movable surface area is reflected in the value of \(C_L\). Patent [10] reveals that \(C_L\) is modeled as:

\[
C_L = C_{L0} + \Delta C_L + C_{La} \cdot \alpha,
\]

in which \(C_{L0}\) is the lift coefficient with movable surfaces at 0 position when angle of attack \(\alpha\) equals to 0, \(\Delta C_L\) is the lift coefficient increase caused by movable surfaces, \(C_{La}\) is the derivative of lift coefficient \(C_L\) to angle of attack \(\alpha\).

For the designed airplane (Model XXX), finite element analysis is conducted to study the relationship between lift coefficient \(C_L\) and attack angle \(\alpha\). Simulation result (Figure 1) reveals characteristics of nonlinearity and piecewise linearity. Based on the characteristics, machine learning method of Support Vector Machine (SVM) is to be studied to fit the curve.

![Fig. 1. Lift coefficient \(C_L\) when attack angle \(\alpha\) changes from -2\(^{\circ}\) to 5\(^{\circ}\).](image-url)
3 Piecewise Fitting Based on Support Vector Regression

Common nonlinear fitting methods include cubic polynomial fit, least square fit, nonlinear approximation based on neural network. SVM, a machine learning method developed from statistical learning theory, is widely applied in model fitting of sensor outputs [11,12]. SVM improves generalization capacity through structural risk minimization principle, solving problems such as curse of dimensionality, small-sample learning [13,14]. Advantage of SVM compared with other fitting methods is that SVM outputs globally optimal solution with good generalization capacity [15].

For the lift coefficient fitting problem as Figure 1, piecewise linearity function fitting is a simple solution, with fast computation speed. However, high fitting precision cannot be achieved. For methods such as polynomial fit and neural network fit, piecewise linearity may lead to the problem of overfitting. Therefore, machine learning method of multiple support vector regression machine (MSVRM) is selected as a promising method.

Fitting principle is shown as Figure 2, in which $X_i$ is the input vector of the $i$th support vector regression machine (SVRM), and $Y_i$ is the output of the $i$th SVRM [16]. According to the lift coefficient curve of Figure 1, attack angle input space is segmented into subspaces of $X_1, X_2, \cdots, X_n$. Input space cover the whole input range of attack angles and subspaces do not intersect between each other. A SVRM is constructed for each subspace according to the relationship between attack angle and lift coefficient. All SVRMs have the same structure but with different parameters. For each SVRM, training samples are collected at corresponding subspace, and SVRM is trained independently to approach actual output characteristic of the subspace. After constructing SVRMs, the final OUTPUT $Y (C_L)$ at the entire input range ($\alpha$) is obtained.

![Fig. 2. Principle of fitting based on MSVRM.](image-url)
According to the theory of statistical learning, SVM is to look for the optimum separating hyperplane satisfying classification requirements, which results in the maximum distance between samples nearest to the separating hyperplane. In most cases, sample set is linearly indivisible, and a kernel function is selected to map the original function for nonlinear transformation. Nonlinear transformation \( \phi : x \rightarrow \phi(x) \) transforms the given samples to certain high-dimensional feature space. Thus, separating hyperplane is constructed in high-dimensional feature space. A separating hyperplane can be expressed in original space as:

\[
\alpha^T \phi(x) + b = 0, \tag{5}
\]

in which \( \alpha, \phi(x) \) are n-dimension vectors, \( b \) is the threshold, \( (\cdot) \) respects inner product.

The problem of looking for optimum separating hyperplane can be transferred to a problem of convex quadratic optimization:

\[
\min \left( \frac{1}{2} \| \alpha \|^2 + c \sum_{i=1}^{n} \xi_i \right), \tag{6}
\]

which satisfies constraint condition:

\[
y_i(\alpha \cdot \phi(x_i) + b) \geq 1 - \xi_i \quad i = 1, 2, \cdots, n, \tag{7}
\]

in which \( c \) is the penalty coefficient and \( \xi \) is the relaxation factor.

Penalty coefficient \( c \) reflects the degree of compromise between classification precision and model complexity. The larger \( c \) is, the higher fitting degree is, however, leading to higher complexity of SVM and the problem of overfitting. On the contrary, if \( c \) is too small, punishment to empirical error is not enough, leading to SVM of low complexity but the problem of underfitting. For the problem of lift coefficient curve fitting, penalty coefficient \( c \) is selected from an index series \([2^{-12}, 2^{11}]\). Experiments reveal that, if \( c \) is larger than \( 2^9 \), increase of \( c \) has little effect on the performance of SVM, while complexity of SVM model puts too much load on computer. Cross validation tests are conducted for \( c \) value selection. According to test results (Table 1), when penalty coefficient \( c \) is confirmed as \( 2^5 \), accuracy of cross validation (83%) is the highest.
Table 1. Penalty coefficient $c$ and corresponding accuracy of cross validation.

<table>
<thead>
<tr>
<th>Penalty coefficient $c$</th>
<th>Accuracy of cross validation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2^{-12}$</td>
<td>62.3%</td>
</tr>
<tr>
<td>$2^{-11}$</td>
<td>68.5%</td>
</tr>
<tr>
<td>$2^{-10}$</td>
<td>71.2%</td>
</tr>
<tr>
<td>$2^{-9}$</td>
<td>73.6%</td>
</tr>
<tr>
<td>$2^{-8}$</td>
<td>78.1%</td>
</tr>
<tr>
<td>$2^{-7}$</td>
<td>81.1%</td>
</tr>
<tr>
<td>$2^{-6}$</td>
<td>83.0%</td>
</tr>
<tr>
<td>$2^{-5.5}$</td>
<td>81.2%</td>
</tr>
<tr>
<td>$2^{-6}$</td>
<td>79.3%</td>
</tr>
<tr>
<td>$2^{-5}$</td>
<td>79.5%</td>
</tr>
<tr>
<td>$2^{-4}$</td>
<td>79.6%</td>
</tr>
<tr>
<td>$2^{-3}$</td>
<td>79.6%</td>
</tr>
</tbody>
</table>

To improve SVM calculating speed and the algorithm stability, an improved SVM based on least square method is proposed. Based on structural risk minimization principle, solving of $\omega$ and $b$ comes down to minimizing:

$$\min \left( \frac{1}{2} \| \omega \|^2 + c \cdot E_c \right),$$  \hspace{1cm} (8)

in which $E_c$ represents loss function.

Loss function is defined as:
\[ e_e = \frac{1}{2k} \sum_{i=1}^{k} \lambda_i e_i^2, \]  

(9)

in which \( k, e_i, \lambda_i \) are sample size, error, independent weighing coefficient, respectively.

Independent weighing coefficient \( \lambda_i \) is applied for different sample point, which means a major weighing coefficient is adopted for the sample point with minor nonlinear error. Then, the optimization problem is transferred to:

\[ \min J(\omega, b, e) = \frac{1}{2} \| \omega \|^2 + \frac{c}{2k} \sum_{i=1}^{k} \lambda_i e_i^2, \]  

(10)

satisfying the condition:

\[ y_i [\omega \varphi(x) + b] = 1 - e_i. \]  

(11)

Lagrange function is applied to resolve the above constrained extreme value problem:

\[ L(w, b, e, a) = J(w, b, e) - \sum_{i=1}^{k} \alpha_i \{ y_i [\omega \varphi(x_i) + b] - 1 + e_i \}, \]  

(12)

in which \( \alpha_i (i=1,2, \cdots k) \) represents Lagrange multiplier, \( e \) is the vector expressed as \( (e_1, e_2, \cdots, e_k) \).

The fitting model finally obtained to conduct curve fitting is:

\[ f(x) = \sum_{i=1}^{l} (\alpha_i^* - a_i) k(x_i, x) + b, \]  

(13)

in which kernel function \( k(x_i, x) \) is to replace mapping of inner product.
Proper choice of kernel function is the precondition of correct recognition, which is conducted as: a number of common kernel functions such as polynomial kernel, Gaussian kernel, RBF kernel, Fourier kernel, are selected and SVM of different types are constructed; fitting effects in terms of calculating speed, fitting precision are compared; finally, Fourier kernel is selected to be the kernel function of MSVRM for the fitting of the attack angle-lift coefficient curve.

Based on the study of MSVRM above, piecewise fitting is conducted as the following procedures:

(1) Match values of attack angle with corresponding lift coefficient as sample couples;
(2) Divide the attack angle-lift coefficient curve into linear regions;
(3) According to the region division result, divide sample space into corresponding subspaces and train 100 sample couples in each subspace respectively;
(4) For each subspace, based-on different kernel functions, model identification is conducted, and parameters of $\alpha_i, b$ are confirmed;
(5) To compare SVRM fitting performance with different kernel functions, another 100 sample couples are tested to compare fitting results. Results reveal that Fourier kernel outputs the best fitting result.

4 Synthetic Airspeed Calculation Based on Impact Pressure

Airspeed synthetization based on air data (without using air pressure data) is conducted as the following procedures:

(1) Transform air data of real time fuel quality, Z axis acceleration, from binary number to decimal number according to encoding rules of corresponding storages;
(2) On the basis of the study in Section 3, fit attack angle-lift coefficient curve based on SVRM. SVRM produces expressions of lift coefficient with attack angles as input parameter;
(3) The attack angle-lift coefficient curve is based-on the configuration with flaps and slats at 0 position (clean configuration). For different configurations during two flight tests (flight test A and flight test B), gains of lift coefficient compared with clean configuration are identified through data processing. The gain is also related to the difference between impact pressure and dynamic pressure as analyzed in Equation (2).

Obtaining of gains of lift coefficient for different configurations makes it possible for airspeed estimation based on Equations (1) ~ (3), as a secondary airspeed source. Air data during flight test A with a high altitude of around 24000 feet (configuration 1), a low altitude of around 4000 feet (configuration 2) and flight test B with a mid-altitude of 150000 feet (configuration 3) are extracted and airspeed calculation is conducted based on the data.
Concrete flap and slat positions for different configurations are not given here for sake of confidentiality.

Figure 3(a) and Figure 3(d) illustrate measured dynamic pressure and calculated dynamic pressure with gains of k1 (corresponding to configuration 1) and k2 (corresponding to configuration 2) in flight test A, respectively. Figure 3(b), 3(c), 3(e), 3(f) indicate that calculated airspeeds approximate measured airspeeds respectively at steady flight stages with errors within 10 knots. For it is a primary study of airspeed calculation, emphasis is put on steady flight stage (with small acceleration). Figure 3(g) reveals ascent and descending processes are sharp with large acceleration, which is to be researched further.

Figure 4(a) ~ Figure 4(g) illustrate results of airspeed calculation for flight test B. In the test, airplane is steady at a mid altitude and with a relatively slow ascent process. Airspeed calculation is conducted in both processes. Gains for slow ascent process and steady process at the mid altitude are k3 and k4, respectively. An obvious error exists between time from 4000 s to 5200 s. From flight management computer, it is known that attack angle adjustment and configuration variation are conducted during the period, which result in a dynamic lift coefficient $C_L$. Except for this period, airspeed calculation error is within 10 knots (impulse noises are not considered).

In conclusion, airspeed calculation to flight test A and flight test B as procedures 1~3 outputs synthetic airspeeds, based on inputs from flight management computer (real-time fuel quality, configuration information), inertial navigation system (acceleration), with errors within 10 knots for steady flight stage and slow ascent stage.

(a) Measured and calculated dynamic pressure with gain k1.
(b) Measured and calculated airspeed with gain k1.

(c) Airspeed calculation error with gain k1.
(d) Measured and calculated dynamic pressure with gain k2.

(e) Measured and calculated airspeed with gain k2.
(f) Airspeed calculation error with gain $k_2$.

(g) Pressure altitude from traditional air data system.
Fig. 3. Airspeed calculation during flight test A.

(a) Measured and calculated dynamic pressure with gain k3.

(b) Measured and calculated airspeed with gain k3.
(c) Airspeed calculation error with gain $k_3$.

(d) Measured and calculated dynamic pressure with gain $k_4$. 
(e) Measured and calculated airspeed with gain k4.

(f) Airspeed calculation error with gain k4.
5 Conclusion

An airspeed calculation algorithm is proposed in the paper. To deal with the problem of attack angle-lift coefficient curve nonlinearity, machine learning algorithm of SVR is studied and applied. To verify the accuracy of the proposed algorithm, air data from two flight tests is utilized for airspeed calculation. Results reveal that for steady flight stage and slow ascent stage, errors are within 10 knots. Further study will focus on airspeed calculation of steady flight stages in other configurations and motion stages with large accelerations. Moreover, other machine learning algorithms will be studied to deal with the problem of attack angle-lift coefficient curve nonlinearity.
References


[7] Nathan V T, Anandappan T: Determination method for calculating at least one of first true airspeed of first Mach number of aircraft in flight determines true airspeed of first Mach number from group consisting of e.g. standard atmospheric data, thrust data. (2015).


Does More Bandwidth Really Not Matter (Much)?

Seraj Al Mahmud Mostafa\textsuperscript{1}, Mike P. Wittie\textsuperscript{1}, and Utkarsh Goel\textsuperscript{2}

\textsuperscript{1} Montana State University, Bozeman, MT, USA
\textsuperscript{2} Akamai

Abstract. The prevailing wisdom is that more network bandwidth does not matter much and that website performance is primarily limited by network latency. However, as mobile websites become more complex and mobile network performance improves, does this adage continue to hold? To understand the effects of small changes in network bandwidth and latency on website performance, we propose a novel webpage characterization metrics – Critical Path of Improvement (CPI). We compute CPI for 45 websites and analyze it against the network performance of four mobile ISPs in 57 US cities. Our results show that 18\% of websites are primarily limited by bandwidth with others limited by bandwidth to some extent. These results show that contrary to accepted wisdom, insufficient bandwidth is a limiting factor in some website/network combinations. We also offer a discussion of approaches website developers and mobile network administrators can follow to understand and mitigate bandwidth limitations to website performance.

Keywords: Mobile web, measurement, web performance metrics

1 Introduction

The evolution of mobile web services goes hand-in-hand with the improvements in mobile network performance. As mobile networks offer lower latency and higher bandwidth, mobile websites may transfer more content to offer a richer experience, while maintaining good response times. But what are there instances where website design and network performance do not match up well?

The prevailing wisdom is that network latency is the limiting factor to website performance. Two key articles driving this idea came from Mike Belshe \cite{1} and Ilya Grigorik \cite{2}. They observe that beyond some point, adding more bandwidth results in diminishing marginal returns of website load time. Incremental reduction of latency, on the other hand, results in corresponding improvement of website performance. It would seem then that beyond a certain minimal bandwidth website performance is limited by network latency.

While that may have been true a decade ago for landline internet and desktop websites, much of today’s web traffic involves mobile networks and smartphones. The mobile networks themselves have undergone significant improvement in the last several years. OpenSignal reports that from February, 2016 to January, 2020
4G latency (averaged across providers) improved from 75.5 ms to 52.9 ms (by 29.9%) and bandwidth from 9.7 Mbps to 25.6 Mbps (by 163%) [3, 4]. At the same time, there remain pronounced regional differences in mobile network performance. Specifically, in January, 2020 the variation of mean latency between US cities was as large as 37.8 ms and the variation of mean bandwidth more than 28.9 Mbps [3]. Further, as websites become more complex they may act differently in networks of varying performance. Considering these differences, does the adage of network latency being the limiting factor website performance continue to hold?

To answer this question, we take a fresh look at characterizing mobile website network performance requirements. We propose a new website characterization method we dub the Critical Path of Improvement (CPI), illustrated in Figure 1. CPI allows us to understand how small changes to network bandwidth and latency affect website performance as measured by the Perceptual Speed Index (PSI) [5], or any other web performance metric [6]. We base CPI on the observation that at any given point of network performance on the bandwidth/latency plane, PSI will improve on average with small increases to bandwidth, or small reductions to latency. Figure 1 illustrates PSI improvement by lighter points along the CPI. We evaluate the impact on PSI of these small changes in network performance and record CPI as the path of the fastest PSI improvement (details in Section 3). Effectively, CPI allows us to understand when a website performance benefits more from additional bandwidth or from lower latency.

![Fig. 1. Illustration of CPI and network conditions.](image-url)
To understand whether latency or bandwidth of mobile network is the limiting factor for a given website, we analyze the relationship between a network’s performance and the CPI. In Figure 1, the rectangular regions represent network performance envelopes in terms of minimum and maximum latency and bandwidth measurements. For a network whose performance envelope lies in region D, the path to better PSI points downwards to the CPI and so via lower latency. On the other hand, if the network performance envelope corresponds to region C, the path to the CPI and better PSI lies towards higher bandwidth. We want to understand how common are scenarios where a website could benefit from higher bandwidth, or lower latency, than what is available in a given network deployment. In other words, how often is website performance limited by insufficient bandwidth versus by high latency.

This paper offers the following contributions:

– We propose CPI – a new website performance characterization method. We show how developers may use CPI to understand the impact of small changes in network performance on the performance of their websites.
– We analyse the CPI of 45 websites from the Alexa-100 list in four mobile networks across 57 US cities to identify cases where websites are bandwidth- and latency-limited.
– We analyse regional variation in network performance to show areas of the US where network bandwidth tends to be the limiting factor to website performance.
– We point to factors of webpage design that contribute to websites being limited by bandwidth.
– We offer a discussion of how developers and network operators may use these results to improve the match between website design and network performance.

The rest of this paper is organized as follows. In Section 2 we discuss related work on website performance bottlenecks. Section 3 details CPI measurements and comparisons with network performance. In Section 4 we describe our measurement methodology and discuss measurement results. Finally, we offer a discussion of the applicability of our findings to website developers and network administrators in Section 5 and conclude in Section 6.

2 Related Work

A number of papers quantified performance bottlenecks in mobile websites. Nejati and Balasubramanian found that device processing speed as well as delays in transferring objects on the critical path of rendering are the main causes of lower website performance on mobile devices [7]. While some researchers have focused on reducing the computational load of rendering [8,9], others have focused on eliminating loading bottlenecks.

One approach to eliminating loading bottlenecks is to reorder objects on the critical path [6,10]. Even with object reordering, Furtak et al. have shown that
increased bandwidth improves page Speed Index by permitting large images to load more quickly [6]. However, others have observed that the impact of extra bandwidth is less pronounced on Page Load Time (PLT) [11]. Han et al. have shown that extra bandwidth improves the performance of large pages [12]. They loaded pages using MultiPath TCP (MPTCP) over cellular and WiFi connections in parallel and showed that the loading process can use the bandwidth of the additional link effectively, when the performance of the links is comparable in terms of latency.

These results, however, focus on the impact of network performance in general, but do not consider the correspondence of specific web pages loaded in specific mobile networks. While researchers agree that additional network bandwidth may be helpful in some cases, its not clear to what degree such cases are prevalent. In this paper we aim to answer that question.

3 Critical Path of Improvement

CPI measures how small changes to network performance affect website performance. To obtain the CPI, we load a given website under different combinations of latency and download bandwidth controlled using NetEm [13]. For each set of network conditions we measure website PSI using PWMetrics [14].

Using these tools, we measure CPI as follows:

– As a starting point we chose a set of atrocious network conditions, specifically bandwidth of 256 Kbps and latency of 180 ms. We load a website over this constrained link and measure its PSI. For example, in Figure 2a these low bandwidth, high latency conditions might result in the PSI of 1000.
– We then proceed to find the next CPI point. We measure the PSI of two candidate points. For the first candidate point we improve network latency by 10 ms, but keep the bandwidth of the starting point. For the second candidate point, we improve network bandwidth, but keep the latency of
the starting point. Since improvements in network bandwidth have a greater impact on website performance in the lower ranges [1, 2], we double the network bandwidth up to 8 Mbps and then increase it by a constant interval of 8 Mbps. We then measure the PSI of the candidate points. In Figure 2a the candidate points have the PSI of 900 and 800 respectively.

Finally, we select the next point on the CPI as the candidate point with the lower PSI of the two, as shown in Figure 2b. It’s important to note that website PSI varies somewhat across loads. When computing the CPI of the starting and candidate points we average the PSI over seven trials.

Figure 1 shows an example of a complete CPI measurement. As we move along the CPI from the starting point, we can see improvements to PSI represented as lighter points. The CPI shows how small changes to network performance improve website performance. The CPI also shows a website’s preference for higher bandwidth, or lower latency at any point. A CPI with more points further away from the x-axis show a preference for more bandwidth, while those with more points close to the x-axis show preference for lower latency.

We also want to consider the relationship between the CPI and the performance of a network to understand whether a website is bandwidth- or latency-limited in a given network. In Figure 1 we plot four areas: A, B, C, and D. These areas represent some possible envelopes of a network’s performance represented as minimum and maximum latency and bandwidth. While a network’s performance varies during a given time period, it will in general be bounded some maximum and minimum values, as reported by OpenSignal measurements [3, 15, 16]. As such, a network’s performance constrains page PSI to a particular region of the bandwidth/latency region.

Referring to the regions in Figure 1 identify four cases of these constraints.

– Case A: the CPI exits network performance bounds through the minimum latency boundary of area A. This means that page PSI would improve more if A extended to better (lower) latency, rather than towards higher bandwidth.

– Case B: the CPI exits network performance bounds through the maximum bandwidth boundary of area B. This means that page PSI would improve more if B extended to better (higher) bandwidth, rather than towards lower latency.

– Case C: the CPI lies to the right of the network performance bounds of area C. This means that page PSI would improve more if C extended to better (higher) bandwidth, rather than towards lower latency. The rationale for this claim is that a CPI measurement started in area C tends towards the shown CPI and so by improving network bandwidth to the right.

– Case D: the CPI lies below the network performance bounds of area D. This means that page PSI would improve more if D extended to better (lower) latency, rather than towards higher bandwidth. The rationale for this claim is that a CPI measurement started in area D tends towards the shown CPI and so by improving network latency downward.

Our goal is to understand how common are the B and C cases. Their presence indicates that web pages could benefit from additional network bandwidth, which
stands in opposition to the accepted wisdom that web performance is constrained by insufficient latency.

4 Evaluation

Our goal is to understand the relationship between page CPI and network performance envelopes. To frame our results, we first describe our measurement methodology.

4.1 Measurement Methodology

We chose 43 pages from the list of Alexa top 100 pages, excluding adult content [17]. To satisfy our curiosity, we also added major news sites from the authors home countries: prothom-alo.com and wyborcza.pl. For each site, we measure the CPI using the process described in the previous section. We load each page from within the Montana State University’s wired network, while constraining the performance of the internet link with NetEm [13]. Using speedtest.net we measured the performance of our campus network to the regional ISP hosting the Content Distribution Network (CDN) servers that serve web content. Our network is limited to 18.14 ms of latency and 307.64 Mbps of bandwidth, and so NetEm constrains our network in real terms up to these numbers.

We then find the network performance regions of network providers using data from OpenSignal. Specifically we extract 4G mobile network performance for four major network providers (AT&T, Sprint, T-Mobile, and Verizon) across the US for January 2020, 2019, and July 2018 [3, 15, 16] using WebPlotDigitizer [18]. These data contain the mean latency and bandwidth measurements in a 57 cities along with measurement confidence intervals [19]. These confidence intervals describe the likely upper and lower bounds on network performance and we treat them as the upper and lower bounds of the network performance envelopes.

4.2 Results

Figure 3 shows a summary of our dataset. The x and y-axis respectively show the link bandwidth and latency constraints of our CPI measurements. The figure includes CPIs for all the websites in our dataset. Each CPI is represented by a semi-transparent gray line. Segments where these lines are darker represent overlaps between CPIs of different websites. The points along CPI lines represent normalized PSI values – we illustrate improvements in PSI more clearly further down in this section. Finally, the figure represents network performance envelopes for the four provider networks.

Overall we observe that as the network performance improves, so does page PSI, as CPI points become lighter in the low right quadrant of Figure 3. We also see that CPIs tend to concentrate in the area 80 ms - 180 ms and 256 Kbps - 4 Mbps, as overlapping CPI lines become darker, but spread out at latencies
lower than 60 ms. This tells us that while the performance of many websites initially benefits primarily from lower latency, once latency becomes sufficiently low, the benefits from improving bandwidth and latency become comparable.

We also observe that the performance of network providers tends to concentrate in a region of 40 ms-80 ms and 8 Mbps-32 Mbps. The CPIs of different websites cross this region, and so we expect to see that these websites are limited by latency and by bandwidth with respect to the different networks. In the following subsections we delve deeper into this phenomenon by characterizing the relationships between individual CPIs and individual network performance regions into A, B, C, and D cases.

**How often are individual websites limited by bandwidth?** Figure 4 shows the frequency of the A, B, C, and D cases on the y-axis against website URL on the x-axis. The figure shows cases based on the 2019 network performance data across all network providers and regions. We observe that most of the cases are in the A and D categories, which implies that website performance would improve with lower network latency. However, there is also a significant percentage of B and C cases. Specifically, B and C cases represent 5.69% and 9.02% of cases respectively. For 8 out of 45 (18%) of the websites B and C cases represent the majority of cases, which means that these websites are limited by insufficient bandwidth in most network/region combinations. This result shows
Fig. 4. A, B, C, D cases across websites.
a contradiction to the accepted wisdom that website performance is limited by insufficient latency in general. While that is true for most of the measured websites, it is not true for all. In fact 15 out of 45 (33%) of these websites are limited by network bandwidth in at least some network/region performance envelopes.

Our recommendation is that developers evaluate the CPI for their websites and compare it against network performance to determine whether the websites are bottlenecked on latency or bandwidth in a given network. In Section 5 we discuss techniques they may apply, depending on its limiting factor in a given website, to take advantage of available network performance.

Are certain providers more bandwidth limited? We also wanted to understand whether the frequency of B and C cases varies by network provider, or over time. Figure 5 shows the number of A, B, C, and D cases on the y-axis and network providers on the x-axis. The three panels show the number of occurrences for the different years of network performance data.

In general, we observe that for all providers most of the cases are in the D and A categories. At the same time all providers show a number of B and C cases. The frequency of B and C cases does not vary substantially across the years and in fact has decreased over the years as network providers upgrade their infrastructure. Still in the 2020 network data, B and C cases represent 12.12% of cases. The number of B and C cases does not differ substantially across providers, with the difference between 371 on T-Mobile with the most B and C cases in 2020 and 314 on Verizon with the fewest being less than 16%.

Do bandwidth limitations vary geographically? We also wanted to understand whether website performance is limited by bandwidth more often in certain areas of the country. In Figure 6 each point shows the ratio of the occurrence of B and C cases to A and D cases for different US cities. Each point shows the fraction of \( \frac{B+C}{A+D} \) cases with darker/larger points showing higher values, in other words the prevalence of bandwidth-limited websites. The rows of maps shows data for different years of network performance, while the columns show data for different network providers.

In general we observe that instances of websites being bandwidth limited (darker points showing higher \( \frac{B+C}{A+D} \) ratios) as local phenomena that vary by year and provider. In 2018, websites were bandwidth limited primarily in Oklahoma in the Sprint network and in different regions in the Verizon network. In 2019, websites were bandwidth limited to a larger degree across all providers and regions. In 2020, as network performance improved, bandwidth bottlenecks were more prevalent in North East and the South West regions of the country for all providers.

While providers in strive to improve network performance both in terms of latency and bandwidth, they may not be aware how different websites perform in their networks regionally. By understanding the prevalence of the different
Fig. 5. A, B, C, and D cases across network providers.
types of cases in different regional network deployments, providers may adjust
network resource allocation to better meet website needs.

**Do small changes to bandwidth and latency improve PSI?** We also
wanted to understand how small changes to network latency and bandwidth
alone improve website performance. In Figure 7 we show the effect of network
latency on the x-axis on PSI on the y-axis. In Figure 8 we show the effect of
network bandwidth on the x-axis on PSI on the y-axis. The lines in the graphs
show page CPIs.

In general we observe that as network performance improves so does the
PSI. Somewhat contrary to results observed by Belshe and Grigoric [1, 2] we
observe that both improvements in latency and improvements in bandwidth have
a steady effect on page PSI. In other words we do not observe the diminishing
improvements to page performance of additional bandwidth for all pages. This
result is interesting in that as modern pages become more complex, there is no
clear path to improving page performance by simply decreasing network latency.
Instead a careful analysis of page CPI and network performance gives more
insight.

**What makes websites limited by bandwidth?** We also wanted to un-
derstand what aspects of a page make it more likely to fall into the B and
C cases. To do so, we use our measurement data to predict the ratio of B
and C cases to A and D cases \( \frac{B+C}{A+D} \) using page characteristics. To construct
Fig. 7. PSI versus latency.

Fig. 8. PSI versus bandwidth.
the data set for our model we compute the $\frac{B+C}{A+D}$ ratio for each page, provider network, and year. Next, we compute page characteristics as the number of script (.js), Cascading Style Sheets (CSS) (.css), image (.png, .jpg, .gif, .svg), and text (.html, .json) files as well as their total size in KB, from HTTP Archive (HAR) files for each page. Based on these data we train a linear regression model (sklearn.linear_model.LinearRegression) using an 80/20 training/testing split in a 100 trials.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Coefficient Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSS count</td>
<td>0.261606</td>
</tr>
<tr>
<td>image count</td>
<td>0.041035</td>
</tr>
<tr>
<td>script count</td>
<td>0.037109</td>
</tr>
<tr>
<td>CSS KB</td>
<td>0.025595</td>
</tr>
<tr>
<td>text count</td>
<td>0.019713</td>
</tr>
<tr>
<td>text KB</td>
<td>0.000432</td>
</tr>
<tr>
<td>script KB</td>
<td>0.000123</td>
</tr>
<tr>
<td>image KB</td>
<td>0.0000123</td>
</tr>
</tbody>
</table>

Table 1. Page factors in linear regression of the $\frac{B+C}{A+D}$ ratio.

On average we are able to predict the $\frac{B+C}{A+D}$ ratio within about a factor of 2.4 on average. While the behavior of pages is complex and depends on their structure, this prediction nevertheless gives us some insight into the influence of the page characteristics that results in particular $\frac{B+C}{A+D}$ ratios. Table 1 shows the model factors (page characteristics) and their sorted coefficient values. We observe that the number of CSS, image, and script objects have the greatest effect on the $\frac{B+C}{A+D}$ ratio. When comparing page $\frac{B+C}{A+D}$ ratio to the number of CSS, image, and script objects on a page we observe that page/network combinations with high $\frac{B+C}{A+D}$ ratios tend to have more of these objects. However, there are also pages which fall into D cases in all networks with similar number of objects. Thus, we conclude that a more in-depth analysis of page structure, which results in higher $\frac{B+C}{A+D}$ ratios, should be a part of future work. In the meantime, we reiterate, that developers should measure the performance of their pages in different network conditions directly to understand if there are network scenarios where their pages are limited by bandwidth.

5 Discussion

Our results show a mismatch between website needs in terms of network performance, expressed through the CPI, and the network performance offered by network operators. Within each network performance envelope a website is limited by either insufficient bandwidth or latency. If a website is bandwidth-limited within a network performance envelope, it could improve user experience by re-
shaping its CPI through page optimization. Alternatively, if a website is latency-limited it could be redesigned to change its CPI and take advantage of the extra bandwidth available in a given network. Below we offer a discussion of techniques website developers could apply to their sites to change their CPI to better fit available network performance. While the optimization approaches discussed below are already used by website developers and CDNs, we think that grouping them by impact is of value to design strategies for reshaping page CPIs.

If a website is limited by high latency, website developers could leverage different CDNs to serve content from replica servers closer to users in a given network [20]. If possible, they could move origin logic onto the CDN for even faster delivery. Developers could also use Transport Layer Security 1.3 for faster connection setup times [21]. Finally, they could deliver their content using HTTP/3 over the QUIC protocol to tackle lossy networks and retransmission [22].

If a website is limited by low bandwidth, developers could use HTTP link headers to preload CSS and script assets critical for rendering embedded objects without having to wait for the browser to identify these resources and prioritize their HTTP requests [23]. This approach helps to reduce network idle time as the browser parses page HTML. Developers could also defer loading of non-critical assets after the onload event, which will improve page interactivity, or defer below-the-fold assets, since they do not contribute to lower the PSI. Developers may also adapt image sizes to the quality of the network. For example, Akamai Image Manage and Adaptive Image Compression dynamically adjusts image sizes in response to network quality [24, 25].

6 Conclusions and Future Work

We investigated the degree to which latency continues to be the limiting factor to mobile website performance. Through a novel webpage performance characterization method, Critical Path of Improvement (CPI), we show the sensitivity of websites to small changes in network bandwidth and latency performance. Comparing against measurements of mobile network performance we show instances where websites do suffer from a lack of bandwidth in mobile networks. Finally, we suggest techniques that web developers and network administrators can use to create better matches between website requirements and network performance by reshaping page CPI.

To be sure, CPI should be used to investigate factors other than bandwidth/latency on website performance. Two interesting directions are to investigate the impact of website design features on CPI and of web server configurations, including different network protocols used to serve web content. While we consider such investigations a part of future work and hope that CPI will be a useful tool there, we believe they are out of scope for this paper, which focuses on taking a fresh look at the predominance of latency as the limiting factor in website performance.

Another interesting direction for future investigations is to incorporate the effects of jitter and packet loss in CPI characterization. Jitter and loss could arise
Due to packet queuing, mobile network scheduler decisions, and route switching. To keep our discussion focused, we do not explicitly control for these factors, choosing instead to observe them in wild through measurements of page PSI in live networks. However, CPI could be measured on axes other than latency/bandwidth and opens a path to including jitter and loss in future work.

Disclosure

The positions, strategies, or opinions reflected in this article are those of the authors and do not necessarily represent the positions, strategies, or opinions of Akamai.

References

Research on Non-stationary Blind Separation Method for Unfixed Signal Based on Extended Joint Diagonalization

LI Jun¹, Chen Xin-Yuan²

{lijun73812@163.com¹, chenxinyuan7559@163.com²}

(1. College of Artificial Intelligence, Jiangxi Institute of Applied Science and Technology, Nanchang 330100, China; 2. College of Artificial Intelligence, Jiangxi Institute of Applied Science and Technology, Nanchang 330100, China)

Abstract: In order to solve the problem of signal instability caused by the number of iterations of the traditional nonstationary signal underfixed blind separation method, a nonstationary signal underfixed blind separation method based on extended combined diagonalization method is studied. Firstly, the second-order correlation matrix is extended to the fourth-order accumulation by the extended joint diagonalization method, the linear time-frequency variation of the nonstationary signal is analyzed, the uncertain part of the nonstationary signal is calculated, the function expression of the source signal is obtained, and the time-domain and frequency-domain changes in the function expression of the iterative source signal are obtained. The experimental results show that compared with the traditional separation method, the nonstationary signal underfixed blind separation method based on extended joint diagonalization method has the least number of iterations, and the isolated signal is the most stable.

Keywords: Extended joint diagonalization; nonstationary signal; underfixed blind separation; number of iterations

1 Introduction

With the development of information technology, radar, communication and other electronic equipment are increasingly widely used, and the frequency band of use is also expanding. In addition, the complex and changeable topography, the movement of signal source caused by the transfer of personnel and equipment, and the unintentional interference caused by artificial intentional interference and various natural radiation, these factors form a complex electromagnetic environment with
high density in time domain, severe aliasing of spectrum, interlaced space and
dynamic variation of time \[^1\]. In the complex electromagnetic environment, there are
other unknown interference signals in the mixed signals received by radar,
communication and other cooperative systems, besides their own signals, which
seriously affect the normal operation of the system \[^2\]. For non-cooperative systems,
because the receiver receives wide beam and wide bandwidth coverage, the
intercepted multiple signals are often overlapped in time domain, frequency domain
and even airspace, so it is difficult to complete the reconnaissance processing tasks
such as parameter estimation and information extraction directly at this time.
Moreover, because the number of potential source signals is unknown and the number
of array elements is limited, the number of sources in the mixed signal is often larger
than the number of array elements \[^3\]. Therefore, either the cooperative system or the
non-cooperative system usually need to be separated by docking the received hybrid
signal first.

In the complex electromagnetic environment, because the mixed signals are
overlapped in time domain and frequency domain, the separation method based on
traditional time domain or frequency domain filter cannot adapt to this complex
situation \[^4\]. With the development of array signal processing technology, spatial
filtering based on signal wave direction estimation and beamforming has become a
major means of multi-signal separation \[^5\]. However, the method of spatial filtering
has specific arrangement requirements for antenna arrays and must have known
accurate array parameters. However, it is difficult to obtain the exact parameters about
the array in a non-cooperative environment, which limits the practical application of
the algorithm. When the number of sources in the hybrid signal is larger than the
number of array elements, even if the array parameters are known, the method of
spatial filtering cannot complete the separation of the signals.

At present, blind signal separation is the main method to solve this problem. It
can recover the waveform of the source signal from the observed hybrid signal under
the condition of unknown mixing process and source signal. Separation technology,
which originated from the study of cocktail parties. Cocktail party sound separation
technology refers to in the noisy environment, the voice source signals issued by
different speakers, mixed together through space transmission, the human ear can
accurately capture the voice of concern and interest. The separation technology only
needs the source signal to satisfy the few assumption conditions (e.g. sparsity,
independence, non-Gaussianity, etc.), the source signal can be recovered by using the
observation data of each array element, which has strong adaptability.

2 Research on Non-stationary Blind Separation Method for Unfixed Signal Based on Extended Joint Diagonalization

2.1 Using Extended Joint Diagonalization Method to Calculate Signal Tensor

When the signal tensor is established by the extended joint diagonalization method, the nonstationary signal is in the transient mixed form, and the mixed matrix is nonnegative. A model of the signal separation of the underfixed blind source is established, as follows:

\[
\begin{bmatrix}
    x_1(t) \\
    x_2(t) \\
    \vdots \\
    x_m(t)
\end{bmatrix} =
\begin{bmatrix}
    a_{11} & a_{12} & \cdots & a_{1n} \\
    a_{21} & a_{22} & \cdots & a_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix}
\begin{bmatrix}
    s_1(t) \\
    s_2(t) \\
    \vdots \\
    s_n(t)
\end{bmatrix} +
\begin{bmatrix}
    e_1(t) \\
    e_2(t) \\
    \vdots \\
    e_n(t)
\end{bmatrix}
\]

(1)

In the formula, \( \mathbf{X}(t) = [x_1(t), \ldots, x_m(t)]^T \) is the observation signal vector, its signal number is \( m \), \( \mathbf{S}(t) = [s_1(t), \ldots, s_n(t)]^T \) is an unknown source signal vector, the number of signals is \( n \), and \( m < n \). \( \mathbf{A} = [a_1, \ldots, a_n] \) is the unknown instantaneous mixing matrix of \( m \times n \). \( \mathbf{E}(t) = [e_1(t), \ldots, e_m(t)]^T \) denotes Gaussian white noise. The white noise has little effect on the source signal when the higher order accumulator of the signal is obtained, so the formula (1) can be reduced to:

\[
\mathbf{X}(t) = \mathbf{A}\mathbf{S}(t)
\]

(2)

The second-order correlation matrix of the nonstationary signal obtained by constant transformation upper formula (2) is as follows:

\[
R_{\mathbf{X}} = E[\mathbf{X}(t)\mathbf{X}(t)^T] = E[\mathbf{A}\mathbf{S}(t)\mathbf{S}(t)^T\mathbf{A}^T] = AE[\mathbf{S}(t)\mathbf{S}(t)^T]\mathbf{A}^T
\]

(3)

In the formula, \( E \) means mathematical expectation. For the second-order correlation matrix of the upper (3) nonstationary signal, the implied semaphore is less than the fourth-order accumulation, and the signal estimation is easily disturbed by Gaussian noise. The interference signal is shown below:
To avoid the problem in the figure above, the second-order correlation matrix is extended to fourth-order cumulant.

$$Q^4 = cum_{[X^r, X^s, X^t, X^u]} \in \mathbb{R}^{m^4 \times 1}$$

(4)

In the formula, $Q^4$ indicates the fourth order of the signal. The signal tensor is a nonnegative value when the mixed matrix is nonnegative in a certain physical sense according to the fourth-order cumulant property. Let the nonnegative tensor be:

$$Y = BC^T$$

(5)

In the formula, $B = [b_1, b_2, \ldots, b_j]$ and $C = [c_1, c_2, \ldots, c_j]$ are non-negative matrices. Calculate the Euclidean distance of formula (5):

$$J(b_1, b_2, \ldots, b_j, c_1, c_2, \ldots, c_j) = \frac{1}{2} \|Y - BC^T\|^2$$

(6)

The residuals in the non-negative matrix are defined and calculated to:

$$Y^{(i)} = Y - \sum_{p \neq i} b_p c_p^T$$

(7)
The formula for calculating the tensor of a nonstationary signal is as follows:

\[ Q^r = \frac{R_{\text{com}^r}}{Y - \sum_{j} b_j c_j^r} \]  

(8)

The tensor of the nonstationary signal calculated from formula (8), the linear time-frequency variation of the nonstationary signal is analyzed, and the blind separation method of the nonstationary signal is studied [6].

2.2 Analysis of linear time-frequency variation of nonstationary signals

When analyzing the linear time-frequency variation of nonstationary signal, the linear superposition of Frequency hopping source signal \( S_i(t) \) and \( S_j(t) \) is adopted. Using the time-frequency analysis method to process the hopping source signal, the time-frequency representation of the hopping source signal can be obtained as follows:

\[ TFR_i(t, f) \text{  } TFR_j(t, f) \text{  } TFR_{ij}(t, f) \]

The processed frequency hopping signal still conforms to the linear superposition principle. The formula is as follows:

\[ TFR_{ij}(t, f) = \alpha TFR_i(t, f) + \beta TFR_j(t, f) \]  

(9)

On the basis of the above formula, the window function is translated on the time axis, and the source signal is divided into several short-time signals, Then Fourier transform each short-time signal so that the frequency of communication signal can be obtained in different time periods. Take Observation signals \( \chi(t) \) as an example, the continuous STFT and its inversion are expressed as:

\[ STFT\chi(t, f) = \int_{-\infty}^{\infty} \chi(\tau) \hat{h}^*(\tau-t)e^{-j2\pi ft}\,d\tau \]  

(10)

In the formula, \( \hat{h}^*(\cdot) \) means window function. By multiplying the observation signal with the window function, the Fourier transform is carried out. The time-frequency resolution of short-time Fourier transform (STFT) is therefore related to the nature of the window function, the width of the window function is inversely proportional to the time resolution. And the selection of window function width will also affect the frequency resolution. That is, when the width of the window function is wide, the number of communication signals contained on the window function will be
more, thus increasing the resolution of the frequency. But it will also have an impact on the time resolution, reducing the time resolution \[^7\]. Explain the contradiction between temporal and frequency resolution according to the heisenberg uncertainty criterion. The uncertainty criterion can be expressed as:

\[
\Delta t \Delta \omega \geq \frac{1}{2}
\]

In the formula, \(\Delta t\) indicates the time period of interception, \(\Delta \omega\) indicates frequency hopping bandwidth. As can be seen from the above formula (11), since there is no simultaneous optimal state of temporal and frequency resolution. Therefore, in the practical application, after filtering the different window functions, the appropriate and parameter setting are adopted, and the time resolution and frequency resolution are distributed according to the characteristics of the frequency hopping signal. Through the short-time Fourier transform processing of frequency hopping signal, the uncertain part of non-stationary signal can be analyzed clearly. The time-frequency diagram of signal transformation is shown in the figure below:

![Fig.2 Time-frequency diagram of nonstationary signal short-time Fourier transform](image)

As can be seen from the above figure, the frequency hopping signal is processed by hamming window through short-time Fourier transform. When the width of the
window function is short, there are fewer points in the spectrum analysis, resulting in lower frequency resolution and higher time resolution. Thus, if there is a frequency-hopping signal, the window function uses a shorter width to locate the frequency-hopping time accurately, however, the frequency resolution is low, and the frequency of the hopping signal will be distributed in a wide band. When the width of the window function is longer, the frequency component of the two time periods before and after the jump will exist simultaneously for a certain period of time, but the frequency distribution of the frequency hopping signal needs to be on a narrower band to make the separation and positioning of the frequency more accurate [8].

2.3 Blind Separation of Nonstationary Signals

To realize the underfixed blind separation of nonstationary signal, the underfixed blind separation model is first established. Using array element method, it is assumed that the array element number of nonstationary signal is $M$, A uniform linear array antenna model with a distance of $\lambda_f$ between two adjacent elements is constructed. In the formula, $\lambda_{\text{null}}$ is expressed as the shortest value of the received signal wavelength, $c$ is the speed of light, $f_{\text{max}}$ is the maximum frequency of the received signal. As shown below:
As shown above, the incident angle of the source signal $S_n(t)$ is $\theta$. The distance between the two adjacent elements is $d$. Assuming that the transmitter of the nonstationary signal transmits $N$ source signals and that the array number of the receiver is $M$, Select the analytical expression for the nth source signal as follows:

$$S_n(t) = a_n(t)e^{i(\omega_n(t)t + \phi_n(t))} \quad (12)$$

In the formula, $\omega_n(t)$ is the source signal frequency function and $\phi_n(t)$ is the source signal phase function. The curves of the two functions are shown below:
As shown in the figure above, for fully sparse nonstationary signals, it is assumed that at some point $f$ has a component $S_m(t)$ to work, the underdetermined blind separation model can be expressed as:

$$
\begin{bmatrix}
    x_1(t) \\
    \vdots \\
    x_J(t)
\end{bmatrix} =
\begin{bmatrix}
    \alpha_{1m} \\
    \vdots \\
    \alpha_{Jm}
\end{bmatrix}
\begin{bmatrix}
    S_m(t)
\end{bmatrix}
$$

The estimation signal is obtained by solving the underdeterministic equation jointly by the upper model [9]. There are a large number of non-strictly sparse signals in the actual situation, such as speech signals, which generally adopt sparse transformation, for example, STFT, wavelet transform, Gabo: transform, etc., making it show good sparsity in the transform domain. Two nonstationary signals with a duration of 3s and a sampling frequency of 16kHz are known. The two time domain signals are linearly mixed, iteratively processed, and transformed into frequency domain by STFT. The frequency domain scatter plot is shown in the following figure:
It can be seen from the graph that the distribution of the time domain scattered point map of speech signal is scattered. The frequency-domain scatter plot is then obtained using iterative STFT processing, as shown in the following figure:

The distribution of frequency domain scatter plot is obviously concentrated on two straight lines, showing obvious clustering characteristics, which indicates that it
shows good sparsity in frequency domain at this time. The nonstationary signals obtained from the final separation model are shown in the following figure (a) and (b).

As can be seen from the above figure, the spectrum corresponding to S1, S2 and S3 is in a regular and stable state, and the non-stationary signal underfixed blind separation is realized [19].

3 Simulation experiment
3.1 Experimental Data Set Preparation

The nonstationary signal is linearly mixed by a hybrid filter matrix of 3x3 randomly generated by matlab. Following the parameters of the short-time Fourier transform, the mixed signal is processed by means, albinism, pre-weighting and Fourier transform, and the dimension of the mixed signal matrix is 129×3446×2(M × N × I). Filter it through the filter, filter processing, where the number of filters I =2, the relevant parameters are: In the case of linear convolution mixing, the initialized mixing matrix A is 129×2×3(M × I × J), where J =3 indicates the number of source signals. The filter length P is 8, the impulse response matrix of Nonnegative Matrix Factorization (NMF) decomposition is 1×3, and the 3D nonnegative convolution mixing matrix is initialized. The dimension of randomly initialized W is M × K and the dimension is K × N. In the case of linear instantaneous mixing, the initialized mixing matrix is 2×3(I × J). the remaining parameter value settings are consistent with the linear convolution case.

The same processing algorithm corresponding to the traditional linear
instantaneous mixing model is used as the contrast algorithm. The two methods adopt the same data set and performance evaluation criteria. In addition, the improved method in the NMF process, after several test adjustments, considering the feasibility and dimensionality reduction purposes, the dimensionality reduction parameter $k$ value is taken as 12.

### 3.2 Experimental result

In order to compare the similarity degree of the signal before and after separation, set the iterative parameter $Z$ to 100-1500 to take 100, each run 10 times, take the average value of the correlation coefficient as the reference data. The correlation coefficients between the nonstationary signals of the three separation methods and the separated signals vary with the number of iterations as follows:

![Graph](attachment:image.png)

(a) Experimental results of traditional separation method 1
As shown in the above three experimental results, the average value of S1, S2 and S3 is 0.873, 0.867 and 0.847 respectively. The average iterative coefficient value of S1 of traditional separation method 2 is about 0.875, and the average iterative coefficient value of S2 is about 0.858. The S1 iteration coefficient is 0.878, the S2 iteration coefficient is 0.874 and the S3 iteration coefficient is 0.858. The resulting
iterative coefficients are larger than those obtained by the two separation methods. The results show that the separation method has better performance and improved the smoothness of the correlation coefficient of diagonalization.

4 Conclusion

In complex electromagnetic environment, the hybrid signals are overlapped in time domain, frequency domain and even airspace. Moreover, because the number of potential source signals is unknown and the number of array elements is limited, the number of sources in the mixed signal is often larger than the number of array elements. At this point, it is very difficult to estimate the parameters and extract information of the mixed signal directly. Usually first need to separate the mixed signal under underfixed conditions. Making full use of the sparsity of the source signal in the time-frequency domain, as well as other statistical characteristics such as independence, cyclic stationary, periodicity and so on, the problem of underfixed blind separation is transformed into a suitable or overfixed problem to complete the separation of the source signal.

5 Fund project

Project of Jiangxi Provincial Department of Education, Item number: JXJG-14-81-2

Reference

[6] ZHANG Hong-mei, CHU Yuan-feng, XU Guang-yan. UAV Formation Extension
Consensus Control Based on Joint Errors[J]. Electronics Optics & Control, 2019,26(08):17-23.


FPGA Design and Optimization Implementation of GPS Positioning Algorithm

Weihong He, Zhi Jiang, Zengshan Tian, Mu Zhou, Shuai Lu
{hewei@cqupt.edu.cn, 1455830611@qq.com, tiansz@cqupt.edu.cn, zhoumu@cqupt.edu.cn, lushuai.139@163.com}

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract. Propose a parallel data processing framework method based on field programmable gate array (FPGA). In this paper, the complexity analysis of GPS classic localization algorithm is carried out. Within the scope allowed by hardware logic resources, the FPGA hardware implementation framework is designed with the idea of area exchange rate, that is, a large number of pipeline operations are used in Verilog language programming to improve system processing speed. The experimental results show that FPGA achieves high-precision positioning, in the WGS-84 coordinate system, the error of the X-axis about 5m, and the error of the Y-axis and the Z-axis is about 2m.

Keywords: field programmable gate array; parallel data processing; pipeline

1 Introduction

GPS pseudo-range single-point positioning is widely used in various vehicles, ships for navigation and monitoring, and field surveys due to its advantages such as fast speed, no ambiguity in the whole cycle, and cheap signal receivers [1, 2, 3]. Especially when the accuracy is not high and the geographical environment is more complicated and dangerous, the pseudo-point single-point positioning can greatly reduce the difficulty of the operation and exert its maximum effect.

At present, a large number of algorithms have been studied for GPS pseudo-range single-point positioning. Reference [4] quoted an iterative extended Kalman filter algorithm to analyze GPS ephemeris data, and simulation results show that the method achieves the effect of improving positioning accuracy. Reference [5] adopted two different improvement strategies for the random noise of the TEC time series in the ionospheric delay and the limitation of the Klobuchar model parameter setting. Reference [6] performed spherical fitting with multiple positioning coordinates solved by multiple sets of satellites in three-dimensional space, which improved the utilization rate of satellite data and the accuracy of satellite positioning. Reference [7] established a dual-mode system through the combination of Beidou system and GPS system to achieve multi-system positioning, which improved the superiority and reliability of the system. At the same time, the realization of GPS pseudo-range single-point positioning algorithm is also one of the main contents of current research. Reference [8] realized the GPS pseudo-range single-point positioning by programming on the eclipse platform using Java language. Although this is only implemented on the PC side, it has certain reference value for the hardware implementation of the algorithm.
Reference [9] built a system from GPS data acquisition to positioning processing through the DSP processor. The reliability and stability of the system are verified through actual measurement. Reference [10] based on ARM plus FPGA platform realized phase smooth pseudo-range differential positioning, which is beneficial to improve the positioning accuracy.

At present, the research on GPS pseudo-range single-point positioning algorithm has matured. Most of the researches are mainly focused on the hardware implementation of the algorithm. The mainstream hardware implementation platforms mainly include FPGA and DSP, FPGA and ARM, or directly through the PC. DSP has unique advantages in processing digital signals, but the serial processing mode of DSP cannot meet the requirements of high real-time performance. The same problem still exists in ARM. Considering that the parallel processing method adopted in FPGA can realize high-speed signal processing, and the repeatable programming feature can ensure that the system can achieve online debugging. Therefore, this article designs a set of FPGA-based parallel pipeline architecture system scheme. Compared with the traditional DSP or ARM serial processing mode, the FPGA’s operating mode reduces the delay of data processing and improves the speed of the entire positioning system. Makes the system highly scalable and better versatile.

2 Analysis of pseudorange positioning algorithm

2.1 Pseudo-range positioning principle

GPS measurement errors can be roughly divided into three aspects according to different sources: (1) errors related to satellites; (2) errors related to signal propagation; and (3) errors related to receivers. Considering that the magnitude of the deviation has a certain stability, the deviation value can be predicted by direct measurement or mathematical models. In this paper, three major error sources are corrected through classic modeling: ionospheric error, tropospheric error, and satellite clock error.

2.2 Pseudo-range positioning principle

As shown in figure 1, the geometric distance between the receiver and the satellite in a certain epoch is \( r^{(n)} \). Where \( x = [x, y, z]^T \) is the unknown receiver position coordinate vector and \( x^{(n)} = [x^{(n)}, y^{(n)}, z^{(n)}]^T \) is the position coordinate vector of satellite n.
The geometric distance from the receiver to the satellite can be expressed as:

$$ r^{(n)} = \|x^{(n)} - x\| = \sqrt{(x^{(n)} - x)^2 + (y^{(n)} - y)^2 + (z^{(n)} - z)^2} $$

(1)

The pseudo-range observation equation of the receiver and the satellite at time $t$ is:

$$ p^{(n)} = r^{(n)} + \delta t_u - \delta t^{(n)} + T^{(n)} + e_p^{(n)} $$

(2)

Where $n = 1, 2, \ldots, N$ is the temporary number of the satellite or satellite-measurement, $p^{(n)}$ is the pseudo-range observation, and $r^{(n)}$ is the geometric distance from the receiver to the satellite. $\delta t_u$, $\delta t^{(n)}$, $T^{(n)}$ and $e_p^{(n)}$ are the receiver clock difference, satellite clock difference, ionospheric error, and tropospheric error, respectively, and $e_p^{(n)}$ are noise measurements for pseudorange.

### 3 FPGA processing of positioning algorithm

Based on the analysis of the characteristics of the pseudo-range single-point positioning algorithm, this paper designs a hardware architecture for multi-floating-point arithmetic based on the pipeline structure for parallel computing. This makes full use of the characteristics of parallel processing of data within the FPGA. A trade-off between speed and area is achieved on hardware resources.

#### 3.1 Cordic algorithm for complex operations

The pseudorange positioning algorithm involves not only simple arithmetic operations, but also complex operations such as trigonometric functions, inverse trigonometric functions, exponential functions, and logarithmic functions. From the perspective of hardware implementation, this article chooses the CORDIC (Coordinate Rotation Digital Computer) algorithm that only requires addition and subtraction to implement these complex operations.
As shown in Fig. 3, the vector V1 is obtained by rotating the angle V2.

The mathematical relationship from vector V1 to vector V2 is:

\[
\begin{align*}
    x_2 &= \cos \theta (x_1 - y_1 \tan \theta) \\
    y_2 &= \cos \theta (y_1 + x_1 \tan \theta)
\end{align*}
\]  

(3)

The core of the CORDIC algorithm is the rotation angle \( \theta \). The algorithm divides a lot of different angle values \( \theta \) that have been specified. So for each value \( \theta \) that needs to be rotated, it can be obtained by \( n \) iterations of the value \( \theta \), among which there are \( \tan \theta = 2^{-i} \), so get the CORDIC algorithm formula:

\[
\begin{align*}
    x(n) &= \prod_{i=1}^{N} (\sqrt{1 + 2^{-2i}} (x_0 - d_i x_0 \times 2^{-i})) \\
    y(n) &= \prod_{i=1}^{N} (\sqrt{1 + 2^{-2i}} (y_0 + d_i x_0 \times 2^{-i}))
\end{align*}
\]  

(4)

In the formula, \([x(n), y(n)]\) is an arbitrary point in the circumferential coordinate system, and \(d_i\) is a judgment factor for determining the rotation direction. In the circular rotation mode, the values of \(\sin(x)\) and \(\cos(x)\) can be directly calculated using the CORDIC algorithm, and more functions can be calculated by rotating in different coordinate systems (hyperbolic coordinate system, linear coordinate system, etc.).

### 3.1 Cholesky factorization for matrix operations

Matrix operations include transpose, multiplication, and inversion. The FPGA implementation of matrix inversion is mainly introduced here. In this paper, matrix inversion is implemented by matrix decomposition. This method can not only reduce the amount of calculations, but also is easy to implement in hardware. Commonly used matrix factorization methods include Cholesky factorization, LU factorization, and QR factorization. Based on Cholesky factorization, it is suitable for conjugate symmetric positive definite matrices. The
matrix inversion module mainly includes three main parts: Cholesky decomposition, triangular matrix inversion and matrix multiplication. The specific processing flow is shown in Figure 5.

The matrix $A$ is first decomposed by Cholesky to decompose the matrix into the product of the lower triangular matrix $L$ and its transpose $L^T$, that is $A = L \times L^T$, then the inverse matrix $L^{-1}$ of the lower triangular matrix $L^{-1}$ is obtained by backward transpose; finally the product of the inverse matrix of the lower triangular matrix and its transpose is realized, which is $(L^{-1})^T \times L^{-1} = A^{-1}$. In the FPGA design, the Cholesky decomposition function module includes two memory blocks and two processing modules. One of the memory blocks is used to store the data of the input matrix. The data of the input matrix is loaded into the module line by line, and the other memory block is used to store the matrix data being processed. For the triangular matrix $L$ obtained by using the Cholesky decomposition module, the inverse matrix $L^{-1}$ is calculated using the triangular matrix inversion algorithm. Finally, the multiplier of FPGA is used to design the product of the inverse matrix of the lower triangular matrix and its transpose, and the data of the resulting matrix is output row by row. Considering that the positioning algorithm requires high time sensitivity, this paper uses more multipliers and dividers in the implementation of Cholesky decomposition. It is necessary to consume more hardware resources on the premise of greatly improving the operation speed.

4 Algorithm verification and error analysis

4.1 Procedure analysis steps
The analysis steps of the pseudo-range single-point positioning algorithm program are shown in Figure 6. The first step is to determine the number of satellites that have received information after preprocessing the input data. Positioning can only be performed when the number of satellites is greater than or equal to four. The second step is to calculate the position and operating speed of each satellite observed, as well as the ionospheric error, tropospheric error, and clock error of each observation satellite. The third step is to use the least squares method to iteratively solve the nonlinear equations and judge the operation results. If it is less than the set convergence value, it is judged to be convergence. The updated value is used as the position value of the receiver. The steps are complete.

4.2 Positioning error analysis

In order to ensure the correctness of the FPGA processing data, a new GNSS hardware platform OEM617D newly developed by Canada's NovAtel Corporation is selected for verification. The board can simultaneously use GPS, GLONASS and BDS dual-frequency signals for measurement and positioning. The positioning results can be directly configured and output, and it also provides carrier phase measurement accuracy of up to 0.5mm.

This paper makes a static measurement of the fixed position on the roof of the Yifu Building of Chongqing University of Posts and Telecommunications, and uses the precision positioning results of the OEM617D board as a reference to analyze the errors of the FPGA calculation results. Analysis of one of the received satellite data, where Figures 7, 8, and 9 are the Verilog calculations for ionospheric error correction, tropospheric error correction, and clock clock error correction, respectively, and Table 1 also shows the FPGA and Comparison of correction error results calculated by OEM617D boards.

![Fig. 5. Actual calculated value of ionospheric correction term in FPGA](image-url)
The error result calculated by FPGA is a single-precision floating-point number expressed by the IEEE754 standard. According to the calculation results given in Table 1, the ionospheric correction value, tropospheric correction value, and clock clock correction value calculated by FPGA and the OEM617D board error value are small, and the positioning requirements are not high, not much impact.

The least square method is a commonly used method to solve the nonlinear equations in the pseudo-range positioning algorithm, and judging the convergence of the final result is one of the important indicators to improve the precision of the pseudo-range positioning. This article determines whether the value x is as small as a preset value. Threshold to determine convergence. In section 4.2, the error correction items of the errors of the ephemeris data are analyzed and compared. It is found that the error of the result value calculated by the FPGA is small. Because this section will use the least squares solution, which involves mathematical iterations and matrix operations, this is the main cause of errors in FPGA calculation results. Figure 8 shows the FPGA implementation results of the Cholesky algorithm, and Figure 9 shows the actual calculation of the positioning results in the FPGA.

Fig. 6. Actual calculated value of tropospheric correction term in FPGA

Fig. 7. Actual calculated value of clock clock offset correction term in FPGA

Fig. 8. FPGA implementation of Cholesky algorithm

Fig. 9. The actual calculated value of the single-point positioning result in FPGA
Table 1 shows the comparison of the positioning results. In this paper, through static data analysis of 200 epochs, it can be found from the table that in the WGS-84 coordinate system, the error of the X axis is about 5m, and the error of the Y axis and the Z axis is about 2m. The main reasons for the error analysis are divided into two points. The first is because of the errors caused by the correction model used in the ionospheric error correction, tropospheric error correction, and clock correction. The second is that the NovAtel board output is the result of carrier phase double difference positioning. In this paper, pseudo-range data is used for positioning. Although the tedious solution process of the ambiguity is subtracted, there is a lack of accuracy.

<table>
<thead>
<tr>
<th></th>
<th>OEM617D board output value</th>
<th>FPGA actual calculated value</th>
<th>Absolute error</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-axis of WGS-84</td>
<td>-1587095.8437</td>
<td>-1587090.3475</td>
<td>5.4962</td>
</tr>
<tr>
<td>coordinate system(m)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y-axis of WGS-84</td>
<td>5322424.3856</td>
<td>5322427.0051</td>
<td>2.6195</td>
</tr>
<tr>
<td>coordinate system(m)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z-axis of WGS-84</td>
<td>3126226.1455</td>
<td>3126228.9598</td>
<td>2.8143</td>
</tr>
<tr>
<td>coordinate system(m)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5 Concluding remarks

Pseudo-range single-point positioning technology has simple calculation steps and can achieve fast positioning, which has great application prospects in civil applications. This paper analyzes the basic principle of the pseudo-range single-point positioning algorithm, and selects an error correction model suitable for hardware implementation. When hardware resources allow, choose and improve algorithms such as CORDIC and Cholesky decomposition that are easy to implement in hardware to achieve complex arithmetic operations. Through Verilog programming and modelsin simulation, comparing FPGA and OEM617D board results shows that in the WGS-84 coordinate system, the error of the X axis is about 5m, and the error of the Y axis and the Z axis is about 2m.

6 Acknowledgement

The authors wish to thank the reviewers for the careful review and valuable suggestions. This work was supported in part by National Natural Science Foundation of China (61771083, 61704015).
7 References


Heading estimation algorithm for complex environment based on GPS single baseline

Zengshan Tian, Shuai Lu, Wei He, Mu Zhou, and Zhi Jiang
{tianzs@cqupt.edu.cn, lushuai.139@163.com, hewei@cqupt.edu.cn, zhoumu@cqupt.edu.cn, 1455830611@qq.com}

School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract. With the rapid development of unmanned technology, the high-precision heading angle determines the accuracy of this unmanned automatic navigation. However, traditional least square method is used to solve the vehicle’s heading angle will jump in the complex terrain environment. Therefore, we propose an unmanned vehicle heading estimation algorithm based on single GPS baseline. First, we establish GPS single residual and double residual observation models to eliminate measurement errors, and combine code phase and carrier phase to form ambiguity combination. Second, in order to eliminate the effects of instability and noise that may be brought by complex environments, this paper proposes model equations based on the prediction and update equations of Kalman filter, the Kalman filtering performs real-time status update on the single residual ambiguity. Finally, the integer ambiguity is searched to find the heading angle of the vehicle. In addition, we performed an algorithm performance test in the actual unmanned vehicle operating environment. The test results shown that the estimated error of the heading when the vehicle is traveling straight and turning is within 1.5°.

Keywords: GPS, Heading estimation, Kalman filtering, Carrier phase.

1 Introduction

In recent years, unmanned and automatic navigation technologies have been greatly developed, the vehicle’s attitude information is more important, in the automatic navigation system, vehicle course angle determines its direction, only accurately measure the position of the vehicle itself and can accurately control the movement direction of the vehicle yaw information, realizes the high precision automatic navigation.

GPS is a satellite navigation and positioning system authorized by the United States Department of Defense, which means a navigation system for timing and ranging[1]. The system was originally used to provide equipment with global three-dimensional precise positioning, timing, speed and attitude information[2]. Due to the relatively slow development of GPS hardware, the high price and lack of flexibility, the promotion of GPS attitude measurement technology has
been limited. In recent years, the development of the chip industry and chip manufacturing processes has made GPS more and more miniaturized\cite{3}. At the same time, the rapid development of GPS carrier phase difference technology, the use of GPS carrier phase to measure carrier attitude information has become another important branch of GPS application research.

The rest of the paper is organized as follows. Some related works are surveyed in Section 2. In Section 3, we introduce the proposed algorithm framework, and introduce the residual error elimination method and Kalman filter to solve the ambiguity principle. Section 4 shows the experimental results and finally the conclusion of this paper is given in Section 5.

2 Related Works

GPS aims to provide high-precision, high-reliability positioning, navigation, and timing services for various users around the clock and around the world. GPS has a wide range of uses, high positioning accuracy and full global ground coverage. Using GPS can not only perform navigation and positioning on the carrier, but also perform speed measurement, attitude measurement, and time service on the carrier. Cars, ships, airplanes, and other carriers can be equipped with GPS terminals for positioning and navigation; GPS can continuously provide three-dimensional position, three-dimensional speed, and time information of dynamic targets for various types of users; there are many GPS satellites And the distribution is reasonable, so any place on the earth can continuously observe at least 4 satellites synchronously, thereby ensuring global, all-weather continuous 3D positioning.

In recent years, more and more researchers have done a lot of research in the field of self-driving heading estimation. The literature\cite{4} uses an adaptive Kalman filter algorithm with unit quaternion as the state vector to fuse three-axis gyroscope and three-axis accelerometer. And the triaxial magnetic sensor measures the signal to achieve the vehicle heading estimation. The literature\cite{5} proposes an attitude angle estimation algorithm based on position correction and Kalman filtering. The literature\cite{6} addresses the problem of cumulative error of attitude angle calculation using inertial navigation in a low-cost vehicle-fixing attitude system. A three-axis gyroscope, three-axis accelerometer, and magnetic force are used The design implements the attitude heading system, which effectively suppresses the influence of errors on the gyro attitude. Literature\cite{7} the combined attitude measurement using GPS and MEMS sensors achieves high-precision vehicle heading angle measurement, but it is expensive and not suitable for large-scale promotion.

In order to solve the problem of vehicle heading angle instability in an unmanned system, we propose a GPS single-baseline heading estimation algorithm. It uses the GPS single-baseline carrier phase measurement model to solve the vehicle heading angle and uses Kalman filtering to solve the problem of instability of the heading angle under the dynamic environment of the unmanned vehicle, and provides the vehicle heading angle with high accuracy and stability.
3 Algorithm Description

The GPS single baseline model is to install the GPS antenna on the vehicle according to a certain structure, and calculate the vehicle’s heading Angle by solving the GPS satellite data through two GPS receivers. This system adopts a single baseline course estimation system composed of two GPS receivers. The two antennas can be connected to form a baseline vector, and the vehicle’s heading angle can be obtained through coordinate transformation.

Fig. 1. Algorithm model.

Aiming at the characteristics of vehicle heading estimation in the unmanned system, we proposed the algorithm model based on GPS single baseline as shown in Fig.1. It consists of three parts: single difference processing, kalman filter status update and double difference residual processing.

3.1 Single baseline residual model

The original observation equation of pseudo-range is:

\[ P^p = \rho^p + c(\delta_i - \delta_p) + d^p_{trop} + d^p_{iono} + \varepsilon^p \]  

\[
\rho^p = \sqrt{(X^p - X_i)^2 + (Y^p - Y_i)^2 + (Z^p - Z_i)^2}
\] is the geometric distance from the satellite to the reference station, \(\delta_i\) is the receiver clock error, \(\delta_p\) is the satellite clock error, \(d^p_{trop}\) is the tropospheric delay, \(d^p_{iono}\) is for ionospheric delay, \(\varepsilon^p\) is for observation noise.

The original carrier phase observation equation is:

\[ \lambda \cdot \delta \Phi^p = \rho^p + c \cdot \delta - \lambda \cdot N^p - c \cdot \delta^p - c \cdot \delta^p_{\text{tr}} + c \cdot \delta^p_{\text{trop}} \]  

\(\delta \Phi^p\) is the actual measured carrier phase value, \(\rho^p\) is the actual distance between satellite \(p\) and receiver, \(N^p\) is the ambiguity of the satellite \(p\).

In the original pseudo-range and carrier phase observation models, there are error terms such as receiver clock error, satellite clock error, tropospheric error, and ionospheric error. In order to eliminate the errors, single-difference residual processing is performed on the data observed by the two antennas.
Residual term of antenna 1:
\[ \begin{cases} y^1_i = \phi^1_i \times \lambda - (\rho^1_i - c \times \delta^j + d^j_{trop,b}) \\ y^2_i = P^1_i - (\rho^2_i - c \times \delta^j + d^j_{trop,b}) \end{cases} \] (3)

Residual term of antenna 2:
\[ \begin{cases} y^1_j = \phi^1_j \times \lambda - (\rho^1_j - c \times \delta^j + d^j_{trop,r}) \\ y^2_j = P^1_j - (\rho^2_j - c \times \delta^j + d^j_{trop,r}) \end{cases} \] (4)

After the difference residuals, the code phase combination is used to solve the integer ambiguity floating point solution.
\[ \begin{cases} P_1 - P_2 = [\rho_1^j + c \delta^j] - [\rho_2^j + c \delta^j] \\ \phi_1^j - \phi_2^j = [\rho_1^j + f \delta^j + N^j_1] - [\rho_2^j + f \delta^j + N^j_2] \end{cases} \] (5)

it can eliminate atmospheric errors and get single difference ambiguity \( N^1_i - N^1_j = (\phi^1_i - \phi^1_j) - (P_1 - P_2) \).

After performing single-difference residuals on antenna 1 and antenna 2 and performing station-to-station difference, the receiver clock error is eliminated, and obtained the double-difference equation \( \lambda \phi^i_{b,r} (t) = \rho^i_{b,r} (t) + \lambda N^i_{b,r} (t) \), after linearization
\[ \begin{bmatrix} \Delta X_r \\ \Delta Y_r \\ \Delta Z_r \end{bmatrix} T = \begin{bmatrix} \Delta X_r \\ \Delta Y_r \\ \Delta Z_r \end{bmatrix} = \begin{bmatrix} \Delta X_r \\ \Delta Y_r \\ \Delta Z_r \end{bmatrix} \] (6)

\[ \begin{bmatrix} \Delta X_r \\ \Delta Y_r \\ \Delta Z_r \end{bmatrix} = \begin{bmatrix} [\Delta X_r, \Delta Y_r, \Delta Z_r] \end{bmatrix} \]

Based on the single-difference ambiguity floating-point solution, the double-difference residual can be obtained
\[ \begin{cases} v_1^j = [(y^1_i - y^1_j) - (y^1_i - y^1_j)] - (y^1_i - y^1_j) \times (N^i_r - N^j_r) - \lambda \times (N^i_r - N^j_r) \\ v_2^j = [(y^2_i - y^2_j) - (y^2_i - y^2_j)] \end{cases} \] (7)

\( i \) is the reference satellite, \( b \) is the main antenna, \( v_1^j \) is the double difference residual corresponding to the carrier phase, and \( v_2^j \) is the double difference residual corresponding to the pseudo range.

### 3.2 Kalman filtering

In order to eliminate the influence of instability and noise caused by complex terrain environment, this paper proposes the model equation based on the prediction and update equation of kalman filter.

Set state estimator is \( x = (r^T, B_1^T, B_2^T)^T \), \( B_i = (B_{r,h,i}, B_{r,v,i}, B_{r,b,i}, \ldots B_{r,k,i})^T \) is the carrier phase \( L \), single difference phase value between stations, \( r_r = [x, y, z] \) is
the position of antenna 2. Single residual observation vector \( y = [\phi_1^T, \phi_2^T, P_1^T, P_2^T] \) based on pseudorange and carrier phase, set the observation equation \( h(x) = (h_{\phi,1}^T, h_{\phi,2}^T, h_{P,1}^T, h_{P,2}^T)^T \), the observation error matrix is

\[
R = \begin{pmatrix}
DR_{\phi,1}D^T & & & \\
& DR_{\phi,2}D^T & & \\
& & DR_{P,1}D^T & \\
& & & DR_{P,2}D^T
\end{pmatrix}
\]

(8)

Single difference factor

\[
D = \begin{pmatrix}
1 & -1 & 0 & \cdots & 0 \\
1 & 0 & -1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 0 & \cdots & -1
\end{pmatrix}
\]

(9)

Carrier phase and pseudorange observation error matrix

\[
\begin{cases}
R_{\phi,i} = \text{diag}(2\sigma_{\phi,1,i}^2, 2\sigma_{\phi,2,i}^2, \ldots, 2\sigma_{\phi,m,i}^2) \\
R_{P,i} = \text{diag}(2\sigma_{P,1,i}^2, 2\sigma_{P,2,i}^2, \ldots, 2\sigma_{P,m,i}^2)
\end{cases}
\]

(10)

where \( \sigma_{\phi,i} \) is the carrier phase standard deviation, \( \sigma_{P,i} \) is the pseudorange standard deviation.

According to the state vector and observation vector set above, and \( P \) is the covariance matrix corresponding to the state quantity, the observation value \( x \) and covariance \( P \) can be updated according to the observation \( y \) at time \( t \) according to the following steps:

\[
\begin{align*}
\hat{x}_k(+) &= \hat{x}_k(-) + K_k(y_k - h(\hat{x}_k(-))) \\
P_k(+) &= (I - K_kH(\hat{x}_k(-)))P_k(-) \\
K_k &= P_k(-)H(\hat{x}_k(-))(H(\hat{x}_k(-))P_k(-)H(\hat{x}_k(-))^T + R_k)^{-1}
\end{align*}
\]

(11)

\( \hat{x}_k \) and \( P_k \) are the state estimation vector and covariance matrix at the observation epoch \( t_k \), \((-)\) and \((+)\) represent the observation value before and after the update, \( h(x) \) represents the observation model vector, \( H(x) \) is the partial derivative coefficient matrix, \( R(k) \) is the observation error.

Combined with the above Kalman filter equation to perform the observation value filtering update, a floating-point solution can be obtained from the antenna position and carrier phase ambiguity at observation epoch.

### 3.3 Heading solution

The ambiguity single-difference floating-point solution obtained after Kalman filtering is converted into double-difference floating-point solutions between stations, and the true ambiguity solution is searched using the least square ambiguity reduction correlation adjustment method.
The double-difference integer ambiguity vector $\tilde{a}$ is obtained through the ambiguity search, and the integer characteristic of the ambiguity is used to further improve the baseline vector estimation accuracy. $\tilde{b} = \tilde{b} - Q_\tilde{b} \cdot Q^{-1} \cdot (\hat{a} - \tilde{a})$ obtained by LAMBDA search can get $\tilde{a}$, $\tilde{b}$ is a fixed solution of the baseline vector correction number.

Since $\tilde{b}$ is the correction number of the baseline vector, let the coordinate of the main reference antenna of the dual receiver in the WGS84 coordinate system be $[X_m \ Y_m \ Z_m]^T$, then the coordinates of antenna 2 relative to antenna 1 is

$$
\begin{bmatrix}
X_s \\
Y_s \\
Z_s
\end{bmatrix} = \begin{bmatrix}
\delta X \\
\delta Y \\
\delta Z
\end{bmatrix} - \begin{bmatrix}
X_m \\
Y_m \\
Z_m
\end{bmatrix} \quad (12)
$$

Assuming that the antenna 1 is the coordinate origin, the baseline vector is equivalent to the baseline vector correction number

$$\tilde{b} = \begin{bmatrix}
X_s \\
Y_s \\
Z_s
\end{bmatrix} = \begin{bmatrix}
\delta X \\
\delta Y \\
\delta Z
\end{bmatrix} \quad (13)$$

Since the baseline vector $\tilde{b}$ is calculated in the WGS84 coordinate system, it is represented by $\tilde{b}_e$, and it needs to be coordinate-converted to obtain the corresponding vector in the ENU coordinate system, coordinate transformation equation:

$$
\begin{bmatrix}
x_n \\
y_n \\
z_n
\end{bmatrix} = 
\begin{bmatrix}
-\sin \lambda & \cos \lambda & 0 \\
-\cos \lambda \sin \phi & -\sin \lambda \sin \phi & \cos \phi \\
\cos \lambda \cos \phi & \sin \lambda \cos \phi & \sin \phi
\end{bmatrix} 
\begin{bmatrix}
x_e \\
y_e \\
z_e
\end{bmatrix} \quad (14)
$$

where $\tilde{b}_n = [x_n \ y_n \ z_n]^T$, $\tilde{b}_e = [x_e \ y_e \ z_e]^T$, subscript $n$ indicates the representation of the corresponding parameter in the navigation coordinate system, and $\lambda$ is the longitude value of the position of the receiver antenna, and $\phi$ is the latitude value of the antenna position.

According to coordinate transformation, the heading can be directly calculated as:

$$\varphi = \arctan \left( \frac{x_n}{y_n} \right) \quad (15)$$

4 Experimental Results

In order to verify the effectiveness of the method proposed in this paper, we set up a test platform as shown in Fig. 1 on a golf unmanned golf cart. A single baseline heading angle solution model was composed of two GPS antennas.

For the performance of the algorithm on unmanned vehicles, we conducted straight-line and turning-path driving tests on a golf unmanned cart. The test environment was a golf course with more slopes and a larger slope.
4.1 Straight path test

In straight path test, we set the start and end points on the map, plan the vehicle’s predefined straight-line path, receive GPS signals during the automatic driving of the vehicle, the processor performs algorithm calculation to solve the heading angle, and performs automatic navigation control of the vehicle.

First set the start and end points on the map, plan the vehicle’s predefined straight-line path, receive GPS signals during the automatic driving of the ve-
vehicle, the processor performs algorithm calculation to solve the heading angle, and performs automatic navigation control of the vehicle.

The track of the straight test path is shown in Fig.3. The unmanned vehicle travels straight from south to north on the golf course. Fig.4 shows the vehicle heading angle calculated by the algorithm. There is a singularity in the heading angle for the ambiguity calculation using the least squares method, and there is no singularity in the heading angle for the ambiguity update using the Kalman filter.

4.2 Turning path test

The turn test path is shown in Fig. 5. A 90° turn path is planned on the golf course. The algorithm solves the heading angle in real time during the driving of the unmanned vehicle.

Fig. 5. Turning test track. Fig. 6. Heading of turning path test.

Fig. 6 shows the solution of the heading angle of the turning test of the unmanned vehicle. Combined with the driving path analysis, the vehicle first travels straight in the direction of 180° and then turns 90°. During the turn, the output of the heading angle is smooth and stable. Steady heading angle output during cornering.

4.3 Test Result Analysis

At last, we counted several experimental tests results and calculated the standard deviation respectively. The comparison of the standard deviation between Kalman filtering and least squares is shown as Tab 1.
Table 1. Heading angle error statistics for straight line test

<table>
<thead>
<tr>
<th>ID</th>
<th>Testing Time (s)</th>
<th>Mean (°)</th>
<th>Variance of ILS (°)</th>
<th>Variance of KF (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>118</td>
<td>90.2</td>
<td>3.552</td>
<td>0.856</td>
</tr>
<tr>
<td>2</td>
<td>141</td>
<td>181.6</td>
<td>4.224</td>
<td>1.076</td>
</tr>
<tr>
<td>3</td>
<td>145</td>
<td>1.432</td>
<td>4.958</td>
<td>0.721</td>
</tr>
<tr>
<td>4</td>
<td>110</td>
<td>1.469</td>
<td>6.881</td>
<td>1.326</td>
</tr>
</tbody>
</table>

From the above results, it can be seen that the standard deviation calculated by the Kalman filtering is smaller than the least squares whether the vehicle is stationary or in motion. Therefore, in a complex terrain environment, using Kalman filter to update the integer ambiguity can solve the problem of singular points in the vehicle’s heading angle.

5 Conclusion

In order to improve the accuracy of the heading estimation of the unmanned vehicle and reduce the equipment cost, the GPS single-baseline heading estimation algorithm proposed in this paper uses the GPS single-baseline model and Kalman filtering to provide real-time status updates to provide high-precision and stable heading angles. In addition, the algorithm test was verified in the actual application environment. The test results of the unmanned vehicle driving straight and turning show that the algorithm can calculate the heading angle for high accuracy and high stability.

6 Acknowledgement

The authors wish to thank the reviewers for the careful review and valuable suggestions. This work was supported in part by National Natural Science Foundation of China (61771083, 61704015).

References


Radio Frequency Fingerprinting Driven Drone Identification Based on Complex-valued CNN

Hao Gu¹, Yun Lin², Yu Wang¹, Hao Huang¹, Sheng Hong¹, Jie Yang¹, Miao Liu¹, Jinlong Sun¹, and Guan Gui¹*

1. College of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, 210003 China
2. College of Information and Communication Engineering, Harbin Engineering University, Harbin, 150001, China
E-mails: {1218012510, guiguan}@njupt.edu.cn

Abstract. Drone detection and identification technique is of great significance both in the military and civilian fields. Radio frequency (RF) fingerprinting of drone is considered as one of promising techniques due to its uniqueness. Deep learning based RF fingerprinting identification technique can extract hidden features in RF data and then achieve excellent performance. Motivated by this idea, this paper proposes a drone identification method using complex-valued convolutional neural network (CNN) algorithm with higher classification accuracy and faster equipment running time. The complex-valued CNN method convolves the complex convolutional kernel and the real and imaginary parts of the data features separately. In order to verify the proposed method, five state-of-the-art recognition algorithms are adopted to compare their recognition performance and equipment efficiency. Simulation results show that our proposed drone identification method can efficiently recognize the signal of various drones within less computation time.

Keywords: Drone identification, complex-valued CNN, intelligent recognition, RF fingerprinting, deep learning.

1 Introduction

In recent years, drones have quietly infiltrated into people's daily lives and have brought drastic changes to people's lives [1]. In rural areas, drones have greatly improved transportation capacity in remote areas and agricultural production efficiency. In cities, drones not only play an important role in solving the problem of logistics, but also provide detailed land use information and law enforcement evidence for urban planning, construction and management, such as road construction, traffic patrols and city law enforcement. However, the widespread application of drones is bound therefore raising a series of technical and public safety issues [2]. Relevant regulators must be able to adopt effective technologies to detect and identify various drones [3].

RF fingerprinting can be used to detect and identify drones. Due to the different locations and configuration parameters of each smart device, the signal sent by smart devices all has their specific fingerprint [4]. The RF fingerprinting identification technology collects signals from various devices
with their own fingerprint through the receiver and distinguishes them. Specifically, RF fingerprinting-based identification method usually includes two parts: training and identification [5]. The training part means that after receiving devices such as antennas collect wireless signals from various smart devices, they collect their inherent signal characteristics (fingerprint). The identification part means that when an unknown signal is received, we can complete the identification of the unknown signal according to the characteristics of signal that have been collected above [6]. Generally, the above-mentioned RF fingerprinting identification process does not require complicated calculations and can be directly embedded in the host of the receiver. This technology is very suitable for various types of internet of things (IoT) devices [7].

Additionally, deep learning algorithms are widely used in wireless communications [8][9][10] and achieve good performance in physical layer [11][12]. Y. Wang et al. [13] used two layers of convolutional neural network (CNN) networks to identify the modulation signals under unknown channels. What’s more, RF fingerprinting technology also uses deep learning algorithms. J. Yu et al. [14] proposed a multisampling convolutional neural network (MSCNN) to identify ZigBee devices efficiently with low cost. Therefore, the existing technologies RF fingerprinting recognition technologies based on deep learning can also be applied to drone detection and recognition methods.

Based on the RF data collected from true drone devices, this paper proposes a drone identification method via complex-valued CNN in order to identify genuine drone devices. Two RF receivers are used to receive the high-frequency and low-frequency parts of the drone RF data, which are 2.4GHz wifi signals. Then discrete fourier transform (DFT) is performed on the RF data from the two receivers and they are connected together to form the entire RF spectrum of drones. After processing the drone signal, we convolve the complex convolutional kernel and the real and imaginary parts of the data features separately and compare the performance of five different algorithms who are trained on independent datasets. Simulation results are given to confirm equipment running time and identification performance of our proposed method.

The remainder of this paper is organized as follows. In Section 2, we introduce the drone identification system and our specific datasets. Section 3 presents the analysis of two drone identification methods and their neural network structure. Section 4 introduces the simulation results and we make conclusion in section 5.

2 System Model and Datasets

2.1 System model

In our proposed drone identification system, we use complex-valued CNN algorithm to detect and recognize the signals of different drones. This system is shown in Fig.1, which consists of three parts: signal processing part, classification part and evaluation part. The first part is to receive and pre-process the complex RF data of drones. Further, the second part is to train the complex-valued CNN algorithm based on those RF signals in order to classify unknown received drone signals. Last but not least, we could analyze the identification accuracy and other evaluation criteria based on the output results of our proposed drone identification system in the evaluation part.

2.2 RF data based drone datasets

Based on the system model, drone datasets are composed of the complex RF data collected from real drones. These real drone signals are provided by a large open source drone database. This database has
collected many very valuable real drone RF data through the following three modules: drones under analysis module, fight control module and RF sensing module. What’s more, we used four types of drone signals in this database, as is shown in Fig. 1. We can see from the figure that the drone signal datasets contain one type of Background activities (collecting RF background activities data when drones are off) and three types of drone activities (collecting RF background activities data when drones are on). More importantly, three types of drone activities represent three different brands of drones, which means different brands have different prices, protocols, and technologies. After IQ sampling the RF data of each drone activity, we can get drone signal datasets.

Additionally, we used a total of 4,400 drone signal samples in our experiment, with an average of 1100 signal samples per drone activity. These samples are randomly assigned to a proportion of 7: 3 for the training and testing of the Complex-valued CNN. The length of each sample is 2048. What’s more, in order to better extract the features in the datasets, we split each sample into in-phase and quadrature component, whose data form is a real-valued matrix of $2 \times 2048$. These datasets will be used as the input to the drone identification via complex-valued CNN system so that it can identify different drone signals. Last but not least, Table 1 shows the software and hardware configuration used in our data preprocessing, neural networks training and algorithm model verification.

**Table 1. Software and hardware configuration details.**

<table>
<thead>
<tr>
<th>Running environment</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Software</td>
<td>Spyder, MATLABR2019a</td>
</tr>
<tr>
<td>Language</td>
<td>Python3.7.1</td>
</tr>
<tr>
<td>Software library</td>
<td>Keras 2.2.2</td>
</tr>
<tr>
<td>CPUs</td>
<td>8 Intel Xeon E3 (x86_64)</td>
</tr>
<tr>
<td>GPUs</td>
<td>4 NVIDIA GTX1080Ti</td>
</tr>
<tr>
<td>Operating system</td>
<td>Ubuntu 16.04.1-Linux</td>
</tr>
</tbody>
</table>
3 Drone Identification Methods

In this section, two different drone identification methods based on deep learning is introduced. One of them is traditional drone detection and identification method based on fully connected deep neural network (FCN), the another one is drone identification via complex-valued CNN. Each layer of neural network and specific parameters of them will be described in details.

3.1 Traditional drone detection and identification method based on FCN

The architecture of traditional drone detection and identification method based on FCN can be divided into three parts: input layer, hidden layer and output layer. Obviously, we need to focus on the hidden layer who contains 4 fully connected (FC) layers. The number of neurons in the first, second, and third FC layers is 256, 128, and 64, respectively. After every fully connected layer, batch normalization (BN) layer and activation function layer that uses ReLU function is added. Moreover, adding the Dropout layer after each layer is also essential, who can reduce network parameters and avoid overfitting. The number of neurons in the fourth FC layer is the same as the number of drone signals we want to identify. In addition, by comparing the true labels of the input layer with the predicted labels of the output layer, we can know the recognition performance of this traditional FCN method.

3.2 Drone identification via complex-valued CNN

Our proposed drone identification method via complex-valued CNN is introduced in this section. Complex-valued CNN is an extension of the traditional real-value CNN in the complex field. Since the drone signal data comes in the form of IQ data, we have reasons to believe that complex-valued CNN will have a stronger ability to extract features. Hence, we add real-valued CNN models in our simulation experiments in order to compare the identification performance of drone signals with complex-valued CNNs.

As we all know, convolutional operation has a very important position in CNN, which can sparse network parameters and extract data features within all directions. Therefore, the significance of complex convolutional operation is self-evident in complex-valued CNN, and it can be achieved by multiple real-valued convolutional operations. First, we introduce a complex vector $\mathbf{h}$ and a complex filter matrix $\mathbf{W}$, whose definition is as follows:

$$ h = x + iy $$

$$ W = A + iB $$

where $x$, $y$ are real vectors, and $A$, $B$ are real matrices. The convolution process of $\mathbf{h}$ and $\mathbf{W}$ is as follows:

$$ W \ast h = (A \ast x - B \ast y) + i(B \ast x + A \ast y) $$

from this formula we can see that the complex convolutional operation can be split into the sum of the convolution of the real and imaginary parts of one vector $\mathbf{h}$ and the same two parts of another matrix $\mathbf{W}$, respectively. According to this formula, we can infer the complex convolution operation between the complex feature map $\mathbf{M}$ and the complex convolution kernel $\mathbf{K}$:

$$ M \ast K = (M_R K_R - M_I K_I) + i(M_R K_I + M_I K_R) $$

where the subscript $R$ represents the real part and the subscript $I$ represents the imaginary part.
After introducing the compress convolutional operation, the structure of complex-valued CNN in our proposed drone identification system is shown in Fig. 2. It consists of two parts: complex convolutional layer and complex FC layer. On the one hand, the first complex convolutional layer contains 128 filters whose convolutional kernel is 16 and the second complex convolutional layer is composed of 64 filters, whose convolutional layer is 8. After the complex convolutional layers, there are also three complex FC layers. The number of neurons in the first, second and third complex FC layer is 256, 128 and 2 respectively. In order to speed up the training speed of the deep network and avoid overfitting, we add a complex BN layer and a Dropout layer after each complex convolution layer or FC layer. Specifically, the number of neurons in the third complex FC layer is determined by the number of drone activities that we want to identify, and the activation function of this layer is Softmax, which can output the predicted probability of each category.

**Fig. 2.** The structure of complex-valued CNN in our proposed drone identification system.

### 4 Simulation results

In this section, we analyze the classification accuracy of five drone identification methods for our drone signal datasets contain one type of background activities and three types of drone activities. These algorithm models are trained on independent datasets, which 70% of them are used for training and 30% of them are used for testing. We comprehensively evaluate the complex-valued CNN-based algorithm proposed in this paper in terms of model classification accuracy, prediction time of a single sample, and confusion matrix.

#### 4.1 Classification performance of drone identification model

Table 2 shows the classification accuracy and GPT-Time of single sample in different drone identification models. These five methods including complex-valued CNN (our proposed), real-valued CNN, traditional drone identification method based on FCN, long short-term memory (LSTM), decision tree (DT) and support vector machine (SVM). First, we can clearly see from the table that deep learning methods get higher recognition accuracy than machine learning algorithms (DT, SVM). The traditional FCN algorithm is 30% higher than LSTM algorithm and reaches 85%. Additionally, two CNN-based recognition algorithms both get accuracy of more than 90%. Our proposed complex-valued CNN-based
identification algorithm has the highest accuracy rate, even reaching 99.5%. Because of lacking the ability to extract features in the complex domain, real-valued CNN algorithm’s accuracy rate is 7% lower.

Next, we focus on the running time of a single sample within all algorithms, who are all running on the same machine. We can see from the third column of the table that the running GPU-time of SVM algorithm is the longest, reaching 63ms, who is dozen times of other algorithms. The reason is that SVM algorithm is not suitable for classification models with too long data length. The running GPU-time of DT is the shortest, only 0.068ms. However, the GPU-time of neural networks is basically similar. Complex-valued CNN algorithm runs slightly efficient than other deep learning algorithms because it has fewer neurons. Taken together, complex-valued CNN algorithm not only has higher classification accuracy, but also has less running GPU-time.

Table 2. Accuracy and GPU-Time of each drone identification model.

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Accuracy (%)</th>
<th>GPU-Time (ms/sample)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex-valued CNN (proposed)</td>
<td>99.50</td>
<td>1.3448</td>
</tr>
<tr>
<td>Real-valued CNN</td>
<td>92.93</td>
<td>5.0233</td>
</tr>
<tr>
<td>FCN</td>
<td>85.05</td>
<td>1.6006</td>
</tr>
<tr>
<td>LSTM</td>
<td>55.63</td>
<td>2.0082</td>
</tr>
<tr>
<td>DT</td>
<td>53.93</td>
<td>0.0685</td>
</tr>
<tr>
<td>SVM</td>
<td>23.48</td>
<td>63.5852</td>
</tr>
</tbody>
</table>

4.2 Confusion matrices of deep learning algorithm models

Finally, we provide four confusion matrix of deep learning algorithm models in Fig. 3, which facilitate some details of this four algorithm performance. As is shown in this figure, BG represents the background activities label and D1~D3 represent the drone1~drone3 activities label. By observing the squares on the diagonal of all confusion matrices, who represent correctly predicted labels, we can distinguish the two CNN-based algorithms from other deep learning algorithms. LSTM's confusion matrix is very confusing, and it is almost impossible to accurately predict all drone activities signals. FCN algorithm just performs well in D3 labels. In contrast, two CNN algorithms perform well in predicting three types of drone activities, and the difference in the accuracy of predicting background activities signals has led to their overall gap. The above shows that the complex-valued CNN can accurately distinguish every drone signals so that our proposed method has better recognition ability and robustness.
Conclusions

In this paper, we have proposed a drone identification method via complex-valued CNN driven by RF fingerprinting. The proposed complex-valued CNN algorithm is based on real-valued CNN by convolve the real and imaginary parts of the convolution kernel separately. Then, we compare complex-valued CNN with other popular algorithms by training and verifying the classification accuracy and recording running time of them. Simulation results show that the proposed drone identification method via complex-valued CNN achieve excellent performance in identify drone signals and need less equipment loss. In the future work, we will consider expanding the drone datasets and trying to identify RF data of different running modes in the same drone. More importantly, we expect to prune some neurons in the complex-valued CNN proposed in this paper in order to ensure the identification accuracy and improve the network operation efficiency.

References


Jeju, South Korea, Oct. 15–18 2019, pp. 87–90.


Clustering-Based Edge Compression Method with Application to Electromagnetic Object Recognition

Ziyan Yan¹, Qibin Zheng², Yun Lin³, Qingjiang Shi⁴
{1831577@tongji.edu.cn¹, zhengqb_2018@tongji.edu.cn², linyun@hrbeu.edu.cn³}

School of Software Engineering, Tongji University, Shanghai, 201804, China¹,²,⁴
School of Information and Communication Engineering, Harbin Engineering University, Harbin, 150001, China³

Abstract. Collaborative intelligence has attracted more and more attention. By properly portioning deep neural networks (DNN) and distributing the DNN calculation to the edge and cloud, we could reduce the prediction delay and power consumption to meet the actual application requirements. Toward this direction, this paper proposes an edge compression method based on clustering to address the issue of high data communication cost and time delay between the edge and cloud. Specifically, by using K-means clustering algorithm, this method compresses the output layer of the edge DNN, reducing the amount of transmission data and thus the delay and energy consumption. Based on the compression-based edge-cloud collaboration paradigm, we propose a distributed inference scheme for electromagnetic object recognition. The simulation results show that the proposed method can greatly reduce communication cost while maintaining the prediction performance.

Keywords: deep learning, object detection, edge compression, clustering.

1 Introduction

Deep learning technology has been widely used in the field of image / video object detection and recognition, and achieved good prediction results [1][2] since 2012. Traditional DNN-based prediction methods are generally performed in two ways: transmitting the original data to the cloud for prediction, or directly do the prediction on the edge device. However, the former brings great pressure on the communication bandwidth, which will result in serious delay, while in the latter the prediction capability can be constrained by the device performance and power consumption.

1.1 Collaborative intelligence

Considering the problems and limitations of the above two methods, a "collaborative intelligence" method [3] was proposed in related research works, which can be used to optimize the delay and energy consumption of prediction tasks. This work first studies the significant differences in computing time and output data size of each network layer of AlexNet [4], as shown in Figure 1. It can be seen that in computing time, fc6 and fc7 of full connection layer are significantly more than those of other layers, while pool1~pool5 of pooling layer can significantly reduce the data size, and the size after pool5 layer is smaller.
than the original input. In addition, this work also compares the total delay when a prediction task adopts pure cloud mode (only in the cloud center) and pure edge device mode (only in the edge device). The results show that the communication time accounts for more than 94.1% of the total time in pure cloud mode, and the computing time in pure edge device mode is much longer than that in pure cloud mode.

![Fig. 1. Calculation time and output data size of each layer of AlexNet [3].](image)

Based on the above results, as shown in Figure 2, collaborative intelligence divides the neural network into two parts, so as to share the computing to the edge device and cloud center. This method has two main advantages: 1) the edge device only needs to upload the output data of the hidden layer of the neural network, so that when the output of the hidden layer is less than the original data, it can greatly reduce the traffic; 2) selectively put the heavy calculation in the cloud center, reducing the calculation of the edge device. Through the idea of neural network partition, collaborative intelligence integrates the advantages of pure cloud mode and pure edge device mode, so as to achieve more outstanding performance indicators.

![Fig. 2. Three ways to perform prediction tasks [3].](image)

However, the cooperative intelligence [3] does not consider the compression operation of the edge transmission data, which can result in an increase in the network communication delay (also energy consumption) caused by the output of the hidden layer of the neural network when the output of the hidden layer is larger than the original data. Therefore, it can be expected that DNN-based collaborative intelligence calls for some methods to achieve good accuracy but with less overhead.

### 1.2 Edge compression

Edge compression refers to the use of data compression method to compress the output of edge partition neural network on the basis of collaborative intelligence, so as to effectively reduce the traffic, communication delay and energy consumption.
The authors in [5] researched the output data characteristics of CNN's first convolution layer, chose to use 8 bits to quantize the floating-point number of each unit of the output data, effectively reducing the number of bits of the transmitted data, and then use the lossless image coding technology PNG to compress the quantized data. The experimental results show that the average compression rate of the method is 28.57%, and the performance improvement in delay and energy consumption is 4.9 times and 4.6 times respectively. The work in [6] studies the impact of lossless compression / lossy compression on the hidden layer, and conducts experiments based on the object detection task of the yolo9000 model [7]. Experimental results show that the influence of quantization combined with lossless compression on object detection accuracy can be ignored, while lossy compression can get higher compression rate, but it also affects detection accuracy. The detection accuracy can be effectively improved by using compression enhanced training method. The authors in [8] investigate the difference between deep feature data and natural image data, and propose a simple and effective near lossless deep feature compressor. Compared with HEVCIntra, the bit rate of this method is reduced by 5%, which is much lower than other commonly used image codecs. In [9], a computing architecture based on collaborative intelligence is designed. The architecture introduces a unit, which is composed of a separate convolution layer. It can reduce the dimension of hidden layer output of CNN dividing points, so as to achieve the compression effect. Based on resnet-50 model [10], the experimental results show that the performance of this method in terms of delay and energy consumption is improved by 53 times and 68 times respectively.

1.3 Contribution of this paper

Based on the above research, this paper studies the edge compression of CNN model, and proposes a cluster-based edge compression method (CECM). This method can use clustering algorithm to compress the traffic, and then reduce the delay and energy consumption during the transmission process. In addition, we customize this method with a particular application to the electromagnetic object recognition, and the experiment shows that the method can effectively reduce the transmission traffic and energy consumption on the premise that the accuracy of the original neural network is almost not significantly deteriorated.

In the following content of this paper, Chapter 2 describes the detailed steps of edge compression method based on clustering. Chapter 3 describes the application scenario of the method in electromagnetic object recognition. Chapter 4 evaluates and verifies the performance of the method through simulation experiments. Finally, Chapter 5 concludes this paper.

2 Edge compression method based on clustering

As previously mentioned, the neural network can be divided into two parts, which are deployed separately on the cloud and edge devices, named cloud DNN and edge DNN for easy of exposition. The forward calculation of the DNN is accomplished by relaying the output of the edge DNN to the cloud followed by the rest forward calculation of the cloud NN. This method can take advantage of the cloud and the edge device. However, it may incur large communication overhead if the output of the edge DNN has high dimension. Hence, we should properly design the interface between the edge and cloud and special attention should
be paid to how to downsize the amount of data transmission. Obviously, if the data transmitted by the edge is considered to be compressed, even if the output data of the hidden layer of the neural network is larger than the original data, it can effectively reduce the traffic and the delay. Thus, our method is proposed to reduce the amount of data communication process by edge compression based on clustering. We perform K-means clustering on the output of the DNN, and let the edge send clustering results to the cloud, so that the amount of data during the communication process can be greatly reduced. The detailed steps of Cluster-based Edge Compression Method (CECM) are shown as follows:

**Step1.** Firstly, a high-precision deep neural network model is trained, and divided into two sub neural networks $N_e$ and $N_c$ according to the network layer, where we deploy $N_e$ in the edge device and $N_c$ in the cloud center.

**Step2.** After the edge device processes the input data through $N_e$, it generates the m-dimensional output vector $\mathbf{v}_e = [x_1, x_2, ..., x_m]^T$, where $x_i (i = 1, 2, ..., m)$ are floating-point number. We apply K-means clustering [11] to the set $\{x_i | i = 1, 2, ..., m\}$ which all elements $x_i (i = 1, 2, ..., m)$ from the output vector $\mathbf{v}_e$, compose the collection, generate $k$ clusters $S = \{S_1, S_2, ..., S_k\} (1 \leq k \leq m)$ and obtain the k-dimensional vector $\mathbf{v}_1 = [\mu_1, \mu_2, ..., \mu_k]^T$ composed of the center points $\mu_j (j = 1, 2, ..., k)$ from each cluster, where

$$
\mu_j = \frac{1}{|S_j|} \sum_{y \in S_j} y, \quad j = 1, 2, ..., k .
$$

(1)

Where $|S_j|$ is the number of elements in the cluster $S_j$, $y$ is the element of cluster $S_j$. Then, each cluster is labeled with a different label $1 \sim k$, so that the output vector of edge device $\mathbf{v}_e$ can be mapped to m-dimension label vector $\mathbf{v}_2 = [l_1, l_2, ..., l_m]^T$, where

$$
l_i = \arg \min_j ||x_i - \mu_j||, \quad i = 1, 2, ..., m, \quad j = 1, 2, ..., k .
$$

(2)

It can be seen from (2) that $l_i \in \{1, 2, ..., k\}$, we encode $1 \sim k$ with binary, thus $\lfloor \log_2 k \rfloor$ bits are needed at most to represent $l_i$. Then the transmission capacity can be reduced compared with the floating-point form of direct transmission.

**Step3.** The edge device transmits the vector $\mathbf{v}_1$ composed of center point coordinates in the vector $\mathbf{v}_e$ to the cloud center.

**Step4.** After the cloud center receives the vectors $\mathbf{v}_1$ and $\mathbf{v}_2$, replace the label in the vector $\mathbf{v}_2$ with the center point coordinate in the vector $\mathbf{v}_1$ to get the m-dimension vector $\mathbf{v}_c = [\mu_1', \mu_2', ..., \mu_m']^T$, which is the input of the cloud center sub neural network, Where

$$
\mu_i' = \mu_{l_i}, \quad i = 1, 2, ..., m .
$$

(3)

The cloud center will process according to the input, and finally get the corresponding recognition results.

**Step5.** Repeat step 4 for $p$ times to get the average performance (recognition speed and power consumption) of the current overall depth neural network model, $p \in [3, 20]$. Then adjust the partition point of the neural network and $k$, and repeat steps 1 to 4 $t$ times ($t \in [50, 150]$), finally select the deep neural network model with the best performance.
Figure 3 shows the network structure schematic diagram of the distributed neural network. In Figure 3, the content of the red box indicates that the output vector of the specified layer of neural network is clustered (the number of clusters $k$ is 3) to obtain the vector $\mathbf{v}_1 = (0.3, 1.5, 2.7)$ which is composed of three cluster center points, then we use the center point to approximate the original output vector as the input of the next layer, so as to reduce the amount of data in the communication process.

Fig. 3. Clustering and approximate representation of hidden layer output.

3 Application in electromagnetic object recognition

In the era of information-based war, with the development of new electronic and information technology, the electromagnetic object (such as airplane, warship, etc.) often carries many kinds of electronic equipment. However, due to their characteristics of stealth and high speed, these kinds of electromagnetic objects make the recognition system unable to easily and quickly recognize them and react in time. Such situation greatly increases the risk of attack. Therefore, it is required that the recognition of electromagnetic object not only needs to distinguish accurately but also needs to distinguish quickly, so as to make timely defense and attack the target object effectively.
As shown in Figure 4, we give an example of the workflow of using this CECM method to identify electromagnetic objects in the hypothetical battlefield. Firstly, the reconnaissance equipment (edge equipment) is used as the acquisition node to collect the signal from the electromagnetic object. After collection, the data is processed preliminarily and the intermediate results are output through the neural network deployed on the edge device. Then the edge devices compress the intermediate results by K-meaning clustering, and forward the compressed data to the cloud center through a communication link. After receiving the data from the edge node, the cloud center decompresses it, and takes the decompressed content as the input of the remaining neural network. Then the final recognition result can be obtained after some processing.

![Fig. 4. Schematic diagram of distributed identification of electromagnetic object.](image)

**4 Experiment section**

In this section, some simulations are carried out to validate our proposed method by focusing on a communication modulation recognition task. Note that modulation classification can be viewed as a part of object recognition task. The used dataset and the simulation setup is firstly introduced, followed by some simulation results.

**4.1 Dataset and Simulation Setup**

RadioML2016.10a [12] is a synthetic dataset generated by GNU radio software, which consists of 11 modulation modes with different SNR (8 digital modulation and 3 analog modulation). The dataset can be used for modulation pattern recognition. We divide the dataset into two parts, 80% of which is the training set and 20% of which is the test set. The specific information is shown in Table 1:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of samples</td>
<td>220000</td>
</tr>
<tr>
<td>Number of training set samples</td>
<td>176000</td>
</tr>
<tr>
<td>Number of test set samples</td>
<td>4400</td>
</tr>
<tr>
<td>Feature dimension</td>
<td>(2,128)</td>
</tr>
</tbody>
</table>

**Table 1. Dataset Description**
<table>
<thead>
<tr>
<th>Category (modulation mode)</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modulation mode value</td>
<td>8PSK, AM-DSB, AM-SSB, BPSK, CPFSK, GFSK, PAM4, QAM16, QAM64, QPSK, WBFM</td>
</tr>
<tr>
<td>Number of Signal to Noise Ratio (SNR)</td>
<td>20</td>
</tr>
<tr>
<td>SNR Value</td>
<td>-20, -18, -16, -14, -12, -10, -8, -6, -4, -2, 0, 2, 4, 6, 8, 10, 12, 14, 16, 18</td>
</tr>
</tbody>
</table>

We use convolutional neural network VT-CNN2 [13] as a performance baseline. At the same time, on the basis of VT-CNN2, the proposed method CECM is obtained by clustering compression. Next, we compare and verify the performance of the two methods by experiments.

4.2 Prediction Accuracy

As shown in Figure 5, we compare VT-CNN2 with CECM, and draw the prediction accuracy curve of each cluster when $k$ is 2–10. Among them, dotted curve and solid curve represent the prediction accuracy of VT-CNN2 and CECM under different Signal to Noise Ratio (SNR) respectively. It can be seen from the Figure 5 that when SNR increases, the accuracy of VT-CNN2 gradually increases, and the accuracy changes insignificantly after SNR ≥ 0, maintaining at 70%~75%; in general, CECM keeps the similar change trend with VT-CNN2; with the increase of $k$, the two curves gradually coincide, and the accuracy loss between CECM and VT-CNN2 decreases. When $k$ is greater than 2, the change of recognition accuracy is almost stable.
**Fig. 5.** Prediction accuracy comparison between CECM and VT-CNN2.

**Figure 6** compares the confusion matrix of VT-CNN2, CECM ($k = 2$) and CECM ($k = 10$). In diagram, the vertical coordinate in the **Figure 6** represents the real category of test samples, and the horizontal coordinate represents the category of predicted results of test samples. The element in row $i$ and column $j$ of confusion matrix indicates the proportion of samples of class $i$ predicted as class $j$ in all samples of class $i$. The larger $k$ is, the higher similarity the two methods have.

**Fig. 6.** Confusion matrix of VT-CNN2, CECM ($k = 2$) and CECM ($k = 10$).

It can be seen from the above experiments that, when the total number of classes $k$ increases, CECM has recognition performance close to VT-CNN2. Particularly, when $k$ is greater than 2, their performance is almost identical.
4.3 Compression ratio

Now let us look at the compression ratio of these two methods. From the specific implementation steps in Chapter 2, it can be seen that the dimensions of the output vector $v_e$ of the edge device and the label vector $v_2$ after clustering conversion are both $m$. The dimensions of the center vector $v_1$ after clustering conversion are $k$. It can be concluded that without cluster conversion, the data transmission between the edge device and the cloud center in the original model is: $m \cdot B$, where $B$ denotes the number of data bits occupied by a floating number. While after clustering transformation in our scheme, the amount of data transmission is: $k \cdot B + m \lceil \log_2 k \rceil$. Therefore, the compression ratio of CECM method is:

$$\frac{\text{data transmission in our scheme}}{\text{data transmission without compression}} = \frac{k \cdot B + m \lceil \log_2 k \rceil}{m \cdot B}$$

To be specific, let us consider the case when $B=32$ and $m=256$, and outline the compression ratio for $k=2$ to $10$ in Table 2.

<table>
<thead>
<tr>
<th>Cluster number $k$</th>
<th>Compression ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>3.91%</td>
</tr>
<tr>
<td>3</td>
<td>7.42%</td>
</tr>
<tr>
<td>4</td>
<td>7.81%</td>
</tr>
<tr>
<td>5</td>
<td>11.33%</td>
</tr>
<tr>
<td>6</td>
<td>11.72%</td>
</tr>
<tr>
<td>7</td>
<td>12.11%</td>
</tr>
<tr>
<td>8</td>
<td>12.50%</td>
</tr>
<tr>
<td>9</td>
<td>16.02%</td>
</tr>
<tr>
<td>10</td>
<td>16.41%</td>
</tr>
</tbody>
</table>

4.4 Tradeoff between accuracy and compression

As shown in Figure 7, we can see that, with the increase of cluster number $k$, the recognition accuracy increases, but the compression ratio will increase accordingly. Particularly, when $k=4$, we obtain the best tradeoff between accuracy and compression.

5 Conclusion

Referring to the idea of collaborative intelligence and edge compression, this paper proposes an edge compression method based on clustering, which can reduce the amount of transmission data. The experimental results show that this method can greatly reduce the communication overhead while maintaining the prediction performance. In the future, we will consider a possible extension of this method to the case of multiple edge nodes.
Fig. 7. The maximum precision loss and compression ratio changes with number of cluster $k$.

Acknowledgments. This work was supported in part by the National Key Research and Development Project under grant 2017YFE0119300, the NSFC under Grants 61671411, and by the Fundamental Research Funds for the Central Universities under grant 22120180113.

References

Efficient Resource Allocation For UAV Swarm Communication Systems With Awareness of Environmental Interference

Yinfei Qian, Linyun Zhou, Ziyan Yan, Qingjiang Shi

School of Info. Sci. & Tech., Zhejiang Sci-Tech University, Hangzhou, 310018, China
School of Software Engineering, Tongji University, Shanghai, 201804, China

Abstract. Unmanned aerial vehicle (UAV) communication has drawn significant interests from the industry and academic due to its low consumption, high maneuverability, and flexible mobility. This paper studies a multi-UAV wireless communication system where UAVs are operated by a ground control center to perform certain tasks with a planned trajectory. In order to ensure the reliable communication of multiple UAVs, we formulate a mixed-integer problem aiming to maximize the minimum signal-to-noise ratio (SINR) of UAV swarm by jointly optimizing multiuser spectrum access and transmission power. To tackle this problem efficiently, we propose an iterative algorithm based on majorization-minimization method. Extensive simulation results demonstrate that the proposed algorithm can yield substantial gains in terms of communication performance as compared with a baseline scheme.

Keywords: UAV communication, wireless communication, spectrum access, transmission power.

1 Introduction

With many benefits such as low consumption, high maneuverability, and on-demand deployment, multi-UAV formation has received increasing attention in military field. The multi-UAV collaborative operation makes up for the limited capability of single UAV, improves the fault tolerance ability of the system, and makes tasks more efficient. This scheme improves the success probability of single mission, enables its application in military reconnaissance, target strike, electronic countermeasure, battlefield evaluation, etc. Under this background, UAV-aided wireless communication technique is gradually rising [1], [2]. On one hand, leveraging the existing network architecture, UAVs could act as temporary mobile base stations (BSs) to realize the rapid movement of wireless network coverage; on the other hand, the UAV-aided mobile relay system forms virtual multi-antenna array for the communication between remote nodes, which ensures the reliability of emergency communication.

To complete tasks efficiently, the UAV formation needs to interact quickly and safely, especially in highly dynamic, ultra dense wireless network environments. Nonetheless, multi-UAV collaboration brings several challenges. First, UAVs should correctly receive the control signals and quickly follow the instructions from BS. Therefore, the stability and efficiency of communication link will be a critical technical challenge. Furthermore, due to the limited
available network resources, the reasonable use of network resources can effectively improve communication quality. Finally, in most applications, the external interferences of radiation sources have negative influence on communication system, which we can’t ignore. Those above challenges inspire us to study the resource allocation for multi-UAV wireless communication system with awareness of environmental interference in this paper.

Recently, several studies have been devoted to the UAV-based wireless communication network. The authors in [3] proposed a time-frequency resource blocks allocation and power optimization algorithm to promote the reliability of control signals among UAVs. In [4], the authors studied a path planning scheme to minimize the overall inspection time and energy in a specific scenario. The work in [5] addressed the physical-layer security problem in a UAV communication system by jointly optimizing the UAV’s trajectory and transmit power under stringent energy constraints. Meanwhile, the authors in [6] jointly considered the multiuser scheduling, power allocation and trajectory optimization problem, which aimed to realize the fair performance of UAV swarm. Moreover, [7] investigated the placement problem in static-UAV enabled networks. Each UAV serves as a static BS to maximize the communication coverage for ground users.

Unlike the existing works [3]-[7], this paper considers resource coordination for efficient control and reliable communication in a BS-controlled-UAV network. In this paper, we study a scenario where a ground BS controls UAV formation, and UAVs should receive the control signals and quickly execute the instructions. The quality of wireless link is mainly affected by the interferences from external radiation sources. Therefore, our goal is to ease the impact of interference and improve the quality of control signals by jointly optimizing multiuser spectrum access and transmission power. To tackle this problem efficiently, we propose an iterative algorithm based on majorization-minimization method. Finally, extensive simulations validate the efficacy of the proposed algorithm.

The remainder of this paper is organized as follows: Section 2 presents the system model and problem formulation. Section 3 introduces our iterative algorithm to solve this problem. The simulation results and performance evaluation are provided in Section 4. In Section 5, we make a conclusion to this paper.

2 System Model and Problem Formulation

2.1 System Model

As shown in Figure 1, we consider an uplink communication scenario where multiple UAVs controlled by a ground BS performing some tasks with a planned trajectory. In this scenario, the BS sends control signals through a limited spectrum to multiple UAVs at each time slot. We assume that many external radiation sources spread across the region and cause interference. In addition, we assume that the central processing unit at the ground BS can not only acquire the flight dynamics of UAV swarm, but also perceive channel state information (CSI). Due to the shortage of spectrum resources, the available frequency spectrum is usually limited. Moreover, we assume that the external interference can be sensed by enabling cognitive radio function and the received interferences of each UAV on different channels are very different. Under the above assumptions, the goal of this paper is to jointly design spectrum access and power control to combat the external interference and improve the reliability of control signals.
We use $\mathcal{M} \equiv \{1, ..., M\}$ to denote the set of UAVs, $\mathcal{N} \equiv \{1, ..., N\}$ the set of communication channels, and $\mathcal{S} \equiv \{1, ..., S\}$ the set of time slots. We assume that the length of time slot is set to be sufficiently small such that the UAV’s location seems to be unchange within every time slot. In the considered scenario, the ground-to-air link is mainly affected by the probability of line-of-sight (LoS) and non-line-of-sight (NLoS). According to [8], at time slot $s$ the channel gain between the BS and UAV $k$ on channel $i$ can be expressed as

$$ g_{s,k}^i = c_{s,k} (F + f_{s,k}^i)^{-2} $$

with

$$ c_{s,k} = \begin{cases} 10^{-32.4 - \eta_{\text{LoS}}} \times d_{s,k}^{-2}, & \text{LoS link} \\ 10^{-32.4 - \eta_{\text{NLoS}}} \times d_{s,k}^{-2}, & \text{NLoS link} \end{cases} $$

where $F$ denotes the baseline carrier frequency, $f_{s,k}^i \in \{\Delta f_1, ..., \Delta f_N\}$ denotes the frequency interval between the $F$ and the channel $i$ that UAV $k$ uses in time slot $s$, $d_{s,k}$ is the distance between the UAV swarm and BS. What’s more, $\eta_{\text{LoS}}$ and $\eta_{\text{NLoS}}$ indicate the additional attenuation index for LoS and NLoS links, respectively.

Let us denote by $a_{s,k}^i$ a binary variable, which indicates channel $i$ is occupied by UAV $k$ at time slot $s$ if $a_{s,k}^i = 1$ and $a_{s,k}^i = 0$ otherwise. Since the BS needs to transmit control signals to the UAV formation in each time slot, sufficient frequency bands is necessary for UAVs. Due to the shortage of available frequency spectrum, we assume that each UAV only assigns one
channel, and different UAVs must select different channels, which yields the following constraints

\[
\begin{align*}
M & \leq N \\
\sum_{i=1}^{N} d_{i,k}^j & = 1, \quad \forall k \in \mathcal{M}, \quad \forall s \in \mathcal{S}, \\
f_{i,k}^j & \neq f_{i,m}^j, \quad \forall k \neq m, \quad \forall i \neq j,
\end{align*}
\]

Now we are ready to write down the expression of SINR of each UAV at time slot \( s \). Specifically, the corresponding received SINR of UAV \( k \) can be expressed as follows

\[
r_{s,k} = \frac{\sum_{i=1}^{N} d_{i,k}^j p_{s,k} c_{i,k} (F + f_{i,k}^j)^2}{\sum_{i=1}^{N} d_{i,k}^j (\sigma_{i,k}^j)^2}
\]

Here, the uplink transmission power of the BS for UAV \( k \) is denoted by \( p_{s,k} \), which is subject to the constraint \( \sum_{i=1}^{M} p_{s,k} \leq P_{\text{max}} \), with \( P_{\text{max}} \) denoting the maximum transmission power of ground BS. Furthermore, the numerator of (3) represents the useful signal, and the denominator is the external interferences from radiation sources, where \((\sigma_{i,k}^j)^2\) denotes the power of interference (plus noise) on channel \( i \) received by UAV \( k \) at time slot \( s \) and it is assumed to be sensed through enabling cognitive radio function [9].

### 2.2 Problem Formulation

We denote \( \mathbf{A} = \{ a_{i,k} | s \in \mathcal{S}, k \in \mathcal{M}, i \in \mathcal{N} \} \in \mathbb{R}^{S \times M \times N} \). For time slot \( s \), the matrix \( \mathbf{A}_s = [a_{i,k}^1, \ldots, a_{i,k}^M]^T \) denotes the channel selection scheme of multi-UAV formation, where \( a_{i,k} = [a_{i,k}^1, \ldots, a_{i,k}^M]^T \). Obviously, only one element in each column of matrix \( \mathbf{A}_s \) is 1 and others are 0. Similarly, we define \( \mathbf{P} = \{ p_{s,k} | s \in \mathcal{S}, k \in \mathcal{M} \} \in \mathbb{R}^{S \times M} \) as the power allocation matrix. \( \mathbf{p}_s = [p_{s,k}^1, \ldots, p_{s,k}^M]^T \) represents transmission power at time slot \( s \). Next, we consider control instructions and interference signals. Let \( \mathbf{\Phi} = \{ \Phi_{s,k} | s \in \mathcal{S}, k \in \mathcal{M}, i \in \mathcal{N} \} \in \mathbb{R}^{S \times M \times N} \), where \( \Phi_{s,k} = (F + f_{s,k}^j)^2 \). The matrix \( \mathbf{\Phi}_s = [\Phi_{s,k}^1, \ldots, \Phi_{s,k}^M]^T \) denotes the channel frequency coefficient of multi-UAV swarm and \( \mathbf{\Phi}_k = [\Phi_{s,k}^1, \ldots, \Phi_{s,k}^M]^T \). In addition, we also define \( \mathbf{C} = \{ c_{i,k} | s \in \mathcal{S}, k \in \mathcal{M} \} \in \mathbb{R}^{S \times M} \) and \( \mathbf{c}_s = [c_{s,k}^1, \ldots, c_{s,k}^M]^T \) to represent distance factor of channel gain. Then, we define \( \mathbf{\Sigma} = \{ (\sigma_{i,k}^j)^2 | s \in \mathcal{S}, k \in \mathcal{M}, i \in \mathcal{N} \} \in \mathbb{R}^{S \times M \times N} \) as the external interference matrix. For time slot \( s \), the matrix \( \mathbf{\Sigma}_s = [\sigma_{s,k}^1, \ldots, \sigma_{s,k}^M]^T \) represents the external interference of UAVs, where \( \sigma_{s,k}^j = [(\sigma_{i,k}^j)^2, (\sigma_{i,k}^j)^2]^T \). Define \( \mathbf{e}_k \) as a unit column vector.
with the $k$-th element being 1. Therefore, for expression (3), we have the following transformations:

$$p_{s,k} = p^T_s e_k$$
$$c_{s,k} = c^T_s e_k$$

$$\sum_{i=1}^{N} a_{s,k}^i (F + f_{s,k}^i)^{-2} = e_k^T \Phi^T_s A e_k$$

$$\sum_{i=1}^{N} a_{s,k}^i (\sigma_{s,k}^i)^2 = e_k^T \Sigma A e_k$$

In general, the higher SINR indicates better communication quality of the UAV network, we hope to improve the corresponding SINR level of whole wireless communication system. Hence, the paper introduces the max-min-fairness index of the SINR in order to maintain the reliable communication for each UAV. The optimization problem of the whole flight process is formulated as follows:

$$\max_{\{A, p\}} \sum_{s=1}^{S} \min_{k} \left( \frac{p^T_s e_k e_k^T e_k^T \Phi^T_s A e_k}{e_k^T \Sigma A e_k} \right)$$

s.t. $1^T A e_k = 1, \ \forall k \in M, \ \forall s \in S,$
$$1^T A e_i \leq 1, \ \forall i \in N, \ \forall s \in S,$$
$$a_{s,k}^i \in \{0,1\}, \ \forall k \in M, \ \forall i \in N, \ \forall s \in S,$$
$$1^T p_s \leq P_{max}, \ \forall s \in S,$$
$$0 \leq p_{s,k}, \ \forall k \in M, \ \forall s \in S.$$

The purpose of our optimal model (8) is to maximize the minimum SINR of UAV swarm by jointly optimizing multiuser channel access and transmission power at each time slot. (8b-8f) are the channel assignment constraints, i.e., each UAV is only assigned with one channel and any two UAVs access different channels. Note that the inequality in (8c) is due to that the number of UAVs is less than the number of available channels. Obviously, problem (8) is a Mixed-Integer Nonlinear Program (MINLP), which belongs to NP-Hard problem.

3 Proposed Solution

In this section, we propose a simple and efficient iterative algorithm to solve problem (8) by adopting majorization-minimization (MM) method.
First, we can see that problem (8) is separable across $s$. As a result, we only need to focus on the problem of each time slot, i.e.,

$$\max_{(A, P)} \min_k \left( \frac{p^T_{e_k} e^T_{s} e^T_{s} \Phi^T_{e_k} A e_k}{e^T_{s} \Sigma^T_{s} A e_k} \right)$$

(9a)

s.t. $1^T A e_k = 1$, $\forall k \in \mathcal{M}$,

(9b)

$1^T A e_i \leq 1$, $\forall i \in \mathcal{N}$,

(9c)

$d^i_{s,k} \in \{0,1\}$, $\forall k \in \mathcal{M}$, $\forall i \in \mathcal{N}$,

(9d)

$1^T p_s \leq P_{max}$,

(9e)

$0 \leq p_{s,k}$, $\forall k \in \mathcal{M}$.

(9f)

We rewrite (9) as

$\max_{(A, P)} \min_k \left( \frac{p^T_{e_k} e^T_{s} e^T_{s} \Phi^T_{e_k} A e_k}{B \sum_{m=1, m \neq k}^M e^T_{s} A^T A e_m + e^T_{s} \Sigma^T_{s} A e_k} \right)$

(10a)

s.t. $1^T A e_k = 1$, $\forall k \in \mathcal{M}$,

(10b)

$d^i_{s,k} \in \{0,1\}$, $\forall k \in \mathcal{M}$, $\forall i \in \mathcal{N}$,

(10c)

$1^T p_s \leq P_{max}$,

(10d)

$0 \leq p_{s,k}$, $\forall k \in \mathcal{M}$.

(10e)

As compared to (9), we have cancel the constraint (9c) in (10) but introduce an extra term $B \sum_{m=1, m \neq k}^M e^T_{s} A^T A e_m$ in the objective of (10). It is readily known that when the parameter $B$ is sufficiently large, we have $B \sum_{m=1, m \neq k}^M e^T_{s} A^T A e_m = 0$, $\forall k \in \mathcal{M}$ to maximize the objective. As a result, problems (9) and (10) are equivalent when a sufficiently large $B$ is used.

Furthermore, it is seen that, fixing $A_{s,k}$, the power control can ensure that all the SINR values are the same. Let $\gamma$ be the optimal objective value. Then we have

$$p^*_{s,k} = \frac{B \sum_{m=1, m \neq k}^M e^T_{s} A^T A e_m + e^T_{s} \Sigma^T_{s} A e_k}{e^T_{s} e^T_{s} \Phi^T_{e_k} A e_k}, \forall k$$

(11)

Considering that the power constraint (10d) must be satisfied with equality at the optimality, (11) leads to
Therefore, problem (10) is equivalent to

\[
\min_{\lambda_i} \sum_{k=1}^{M} \frac{B \sum_{m=1}^{M} e_i^T \Lambda e_m + e_i^T \Sigma e_k}{c_i^T e_k f_i \Lambda e_k} \tag{13a}
\]

s.t. \( \mathbf{1}^T \lambda_i e_k = 1, \ \forall k \in \mathcal{M}, \tag{13b} \)

\( a_{s,k}^i \in \{0,1\}, \ \forall k \in \mathcal{M}, \ \forall i \in \mathcal{N}. \tag{13c} \)

Note that for a binary variable \( a \), we always have \((a - 0.5)^2 = 0.25\) which is a constant. Moreover, it holds true that \( 0.25 = \max_{0 \leq a \leq 1} (a - 0.5)^2 \). Hence, we can relax the binary constraint \( (13c) \) to \( 0 \leq a_{s,k}^i \leq 1 \) with a penalty term \(-\mu \| \lambda_i - 0.5 \mathbf{E} \|_F^2\), where \( \mathbf{E} \) denotes the matrix of all one. That is, we consider the following problem

\[
\min_{\lambda_i} f_\mu (\lambda_i) \tag{14a}
\]

s.t. \( \mathbf{1}^T \lambda_i e_k = 1, \ \forall k \in \mathcal{M}, \tag{14b} \)

\( 0 \leq a_{s,k}^i \leq 1, \ \forall k \in \mathcal{M}, \ \forall i \in \mathcal{N}. \tag{14c} \)

where

\[
f_\mu (\lambda_i) = \sum_{k=1}^{M} \frac{B \sum_{m=1}^{M} e_i^T \Lambda e_m + e_i^T \Sigma e_k}{c_i^T e_k f_i \Lambda e_k} - \mu \| \lambda_i - 0.5 \mathbf{E} \|_F^2 \tag{15}\]

Two observations can be made as follows. First, problem (14) is equivalent to (13) when \( \mu \) is sufficiently large. Second, if \( \mu \) is large enough, the objective function (15) can be shown to be concave. Hence, we can use MM method [10] to solve problem (14). In the \( r \)-th iteration of MM, we find a upper bound of (15) by using Taylor approximation and instead minimize the upper bound, i.e., equivalently solve the following problem

\[
\min_{\lambda_i} \text{Trace}(\mathbf{W} \lambda_i) \tag{16a}
\]

s.t. \( \mathbf{1}^T \lambda_i e_k = 1, \ \forall k \in \mathcal{M}, \tag{16b} \)

\( 0 \leq a_{s,k}^i \leq 1, \ \forall k \in \mathcal{M}, \ \forall i \in \mathcal{N}. \tag{16c} \)
where $W \triangleq \nabla_{A_{i}} f_{\mu}(A_{i}^{-1})$. Note that (16) can be decomposed into $M$ independent subproblems in the form of

$$
\min_{a_{s,k}} \sum_{i=1}^{N} a_{s,k} w_{i,k} \\
\text{s.t.} \sum_{i=1}^{N} a_{s,k} = 1, \quad \forall i \in \mathcal{N}. 
$$

where $w_{i,k}$ is the $(i,k)$-th element of $W$. Obviously (17) has a closed-form solution, i.e.

$$
a_{s,k}^h = 1 \text{ and } a_{s,k}^j = 0, \forall i \neq i_k \text{ with } i_k = \arg\min_{i} w_{i,k}.
$$

**Algorithm 1**: The proposed algorithm for problem (14)

1. Initialize $A_{s,0}$, $T_{\max}$ and $\Delta \mu$
2. Set $r = 0$
3. Repeat
   4. Solve for $A_{s}^{r+1}$ problem (16) with $W \triangleq \nabla_{A_{i}} f_{\mu}(A_{i}^{-1})$
   5. If $f_{\mu}(A_{s}^{r+1}) - f_{\mu}(A_{s}^{r}) > 0$
      \[ \mu = \mu + \Delta \mu \]
   6. Else
      \[ r = r + 1 \]
   7. End
8. Until $f_{\mu}(A_{s})$ converges, or the maximum iteration number $T_{\max}$ is reached.
9. Compute $\gamma = P_{\max} \left( \sum_{i=1}^{M} \frac{e_{i}^{T} \Sigma_{i}^{T} A_{i} e_{i}}{c_{i}^{T} e_{i} e_{i}^{T} \Phi_{i}^{T} A_{i} e_{i}} \right)$
10. Compute $P_{s,k} = \gamma \frac{e_{k}^{T} \Sigma_{k}^{T} A_{k} e_{k}}{c_{k}^{T} e_{k} e_{k}^{T} \Phi_{k}^{T} A_{k} e_{k}}, \forall k$

We summarize the proposed algorithm in Algorithm 1. By increasing $\mu$, we can ensure that $\mu$ could be large enough so that the objective (15) becomes concave constantly and the algorithm goes into the phase of MM method with guaranteed convergence. Our algorithm can always keep the objective nonincreasing and guarantee outputting binary variable in each iteration. What’s more, when $B$ is large enough, we can also ensure the terms $B \sum_{m=1}^{M} c_{m}^{T} A_{m} e_{m} = 0, \forall k \in M$. Hence, once $A_{s}$ is obtained, we can find compute $P_{s}$ as in the last step of Algorithm 1.
4 Numerical Results

We consider an scene of $5 \times 5$ km$^2$ with the BS located at $x_0 = (0, 0, 0)$ and the UAV formation flies from the BS to a certain destination at the altitude of $H = 1.0$ km. $L$ radiation sources randomly distributed in the network and the power is assumed about -90 dBm. Similarly, the maximum transmission power of the control BS is set to be 30 dBm. The baseline carrier frequency is set to be $F = 500$ MHz and the channel interval is $\Delta f = 5$ MHz. For comparison, we introduce the simple algorithm, called Random method as a baseline. The algorithm includes two steps, the first step is to distribute channels to UAVs randomly, then allocate power to ensure that all the SINR values of UAVs are the same.

First, we consider the convergence behaviors of the proposed algorithm. We randomly select one time slot of the MM method, at the scenario of $M = 12$, $N = 25$ and $L = 3$. To avoid the algorithm converging to the undesired local points prematurely, we set the maximum iteration number in Algorithm 1 as $T_{\text{max}} = 500$. From the result of Figure 2, we can observe that the MM method makes the objective function descend rapidly, until it converges. As a result, the simulation results verify the effectiveness of the proposed algorithm.

![Graph](image-url)
Figure 3 shows the performance of the UAV formation \((M = 12, N = 21)\) for the proposed algorithm with the Random algorithm, for the case of \(L = 3\) and \(5\). It is observed that the communication quality of UAVs declines gradually with the distance from the BS increases. Moreover, with the increasing number of radiation sources, the interference in this system is correspondingly to rise sharply. We can come to the conclusion that the proposed MM method always achieves the higher SINR, compared to the Random algorithm. The main reason is that the MM method makes the channel assignment scheme more reasonable, which can reduce the interference in communication system.

![Fig. 3. UAV swarm performance during the flight phase.](image)

In Figure 4, we illustrate the system performance obtained by two algorithms under different scenario. For a fair comparison, we assume that there are 5 radiation sources with fixed locations. It is seen that the result of MM algorithm offers superior performance over that of Random algorithm. As a result, for the formation with the same number of UAVs, allocating more bandwidth will get better effect on communication system. Similarly, when the available frequency bands of system is fixed, the network performance will gradually deteriorate with the increase of users.
Fig. 4. UAV system performance under different scenario.

5 Conclusion

In this paper, we have studied the reliable communication in a multi-UAV enabled wireless network. The spectrum access and power allocation are jointly optimized by using an iterative algorithm based on majorization-minimization method. The simulation results show that the proposed algorithm substantially outperforms the baseline algorithm in terms of system performance.

Acknowledgments. This work was supported in part by the National Key Research and Development Project under grant 2017YFE0119300, the NSFC under Grants 61671411, and by the Fundamental Research Funds for the Central Universities under grant 22120180113.
References

Research on statistical method of channel information transmission risk probability in large-scale wireless communication system

MAfeng¹, WU Hong-e²

¹{mafeng3213@163.com¹, wuhonghe321313@163.com²}
²(1. School of Mathematics and Statistics, Nanyang Institute of Technology, Nanyang 473000 China;
2. School of Mathematics and Statistics, Nanyang Institute of Technology, Nanyang 473000 China)

Abstract: In order to improve the communication security of the large-scale wireless communication system channel during the information transmission process, the probability transmission method is used to analyze the information transmission risk and provide data reference for the operation of the wireless communication system. Follow the operating principle of a large-scale wireless communication system to build a system's information transmission channel model, and install information acquisition equipment to obtain channel information. Based on this, the optimal risk decision rule is established, and the channel transmission interference is predicted. At the same time, the channel fading characteristics are analyzed. The statistical value of the risk probability is obtained by calculating the interference prediction results and the channel fading characteristics of the system information transmission channel. Through comparative experiments, it is concluded that compared with the traditional statistical methods of risk analysis, the statistical error of the risk probability obtained by the design method is reduced by about 56%. Applying this method to the daily work of the system can effectively maintain the stability of wireless communication and safety.

Keywords: Wireless communication system; communication channel; information transmission; risk probability statistics;

1 Introduction

A wireless communication system, also called a radio communication system, is a system consisting of three parts: a transmitting device, a receiving device, and a wireless channel. It uses wireless electromagnetic waves to implement information and data transmission. According to the classification of working frequency band or transmission means, it can be divided into medium wave communication, short wave communication, ultra short wave communication, microwave communication and satellite communication [1]. The new generation mobile communication system is a stage and goal in the evolution of mobile communication systems. It not only uses new wireless transmission technologies to improve the performance of communication systems, but also integrates with existing various wired
and wireless networks to ensure system communication. At the same time, the communication range of the system is expanded, and a large-scale wireless communication system is obtained.

The use of a large-scale wireless communication system for transmission of related data requires the use of the system's wireless communication channel. The wireless communication channel is a physical channel for transmitting information, that is, a wireless frequency band, and it is a data signal transmission channel using wireless signals as transmission media. At present, the mainstream wireless protocols are all formulated by the IEEE. In the three wireless standards recognized by the IEEE, IEEE802.11b, IEEE802.11g, and IEEE802.11a, the number of channels is different. In the case of 802.11b / g, the available channels will overlap and interleave in frequency, resulting in only three non-overlapping channels in the service area covered by the network. As a result, users in this service area can only share the data bandwidth of these three channels. These three channels will also be interfered by other radio signal sources, because the 802.11b / gWLAN standard uses the most commonly used 2.4GHz radio frequency band.

However, the channels of large-scale wireless communication systems are susceptible to the influence of radio waves and other channels in the external environment during the transmission of information, leading to risks such as signal interruption, poor communication, and signal loss in the communication system. In order to ensure the stable reliability and security of the system operation, it is necessary to use corresponding methods to carry out risk statistics and assessment on the channels of large-scale wireless communication systems before information transmission, and refer to the statistical results to select a more secure channel for transmission to avoid transmission signal interruption. And loss, thereby ensuring system communication stability.

After a long period of research and application, it has been found that using the risk probability as the evaluation index of the system transmission channel can directly reflect the risk value of the current communication channel, thereby quickly determining the feasibility of channel information transmission. The risk probability assessment method can generally be divided into two steps, one is the identification and analysis of the risk, and the other is the calculation of the risk value. Among them, risk identification refers to identifying potential problems that may lead to cost overruns, delayed schedules, or reduced performance, and a qualitative analysis of their consequences [2]. Risk assessment refers to quantifying the risks and consequences of potential problems and determining their severity. This may involve the comprehensive application of multiple models, and finally get a comprehensive impression of system risk. Risk analysis is the evaluation and estimation of risk impacts and consequences, including qualitative and quantitative analysis. By integrating the risk analysis results and
quantifying them, the risk probability value of the corresponding communication channel can be obtained. However, the single risk assessment is accidental and its application value in actual work is low. In order to solve the above problems, a probability statistics method is introduced to realize the risk assessment of channel information transmission in a large-scale wireless communication system, and the statistical results of the risk probability are obtained.

Probability statistics is a mathematical method to study the statistical laws of random phenomena in nature. Probability theory, as a branch of mathematics, generally includes the probability of random events, statistical independence, and deeper regularity. Probability is a quantitative indicator of the likelihood of a random event. In an independent random event, if the frequency of an event in all events, it is relatively stable in a larger range around a fixed constant. It can be considered that the probability of this event occurring is this constant. The probability value for any event must be between 0 and 1. Combining the probability statistics method in the mathematical category, and merging it with the channel risk estimation of large-scale wireless communication systems, the transmission communication risk is quantified, and the statistical analysis results are used as a reference. Indirectly guarantee the stable and safe operation of the communication system.

2 Design of statistical methods for information transmission risk probability

In order to quantify the risk of channel information transmission in a large-scale wireless communication system, the risk is converted into a probability to be expressed in combination with probability statistics, and the statistical results of comprehensive risk probability are obtained by following statistical theory. The ultimate purpose of designing the system channel information transmission risk probability statistical method is to provide data references for information transmission in wireless communication systems, thereby ensuring the communication security of information data. The implementation of the channel information transmission risk probability statistics method for large-scale wireless communication systems is shown in Figure 1.
2.1 Construction of a large-scale wireless communication system transmission channel model

There are two main channel modeling methods based on large-scale wireless communication systems: statistical channel models and deterministic channel models. The construction of the statistical model needs to be based on a large amount of actual measurement data and obtain a large number of time-domain frequency-domain channel responses through multiple measurements in different scenarios. Related channel large-scale and small-scale parameters are extracted in further statistical analysis [4]. The method used for deterministic modeling is to physically reconstruct the propagated scene through computer simulation. The impulse response of the analog signal in the actual channel is calculated based on the electromagnetic wave propagation characteristics. In order to improve the application performance of the channel model, a transmission channel model of a large-scale wireless communication system is established in accordance with the flow of Figure 2 in combination with two model construction methods.

![Fig. 1 Statistical block diagram of channel information transmission risk probability](image1)

2.1.1 Establish large-scale wireless communication system

The large-scale wireless communication system must provide many important messages or services to the target node through the common channel, including synchronization signal, reference signal, control signal and multimedia broadcast multicast service. For the actual spatial correlation channel, a reasonable method is to make the transmitted signal of the system have the spatial omnidirectional characteristics, and then complete the comprehensive
coverage. Within a certain range, the omnidirectional transmission of the system common signal is shown in Figure 3.

![Figure 3. Signal transmission diagram of large-scale wireless communication system](image)

### 2.1.2 Connecting communication transmission channels

Suppose $h \in C^{M \times 1}$ represents the channel vector between the antenna array and a single antenna target node in a large-scale antenna communication system, $h \sim CN(0, R)$. For the single ring scattering model under the far-field assumption, the channel covariance matrix can be expressed as:

$$R = \frac{2}{\pi} \int_{\frac{-\pi}{2}}^{\frac{\pi}{2}} v(\theta) s(\theta)^H p(\theta) d\theta$$  \hspace{1cm} (1)

In the formula (1), variables $v(\theta)$ and $p(\theta)$ represent the response vector and angular power spectrum of the uniform linear array, respectively. Angle power spectrum $p(\theta)$ may obey different distribution according to different terrain scenes [5].

### 2.1.3 Analog wireless communication transmission

Considering the transmission scenario of large-scale wireless communication system, we focus on a subcarrier of a symbol to be transmitted. The common signal $s \in C^{M \times 1}$ is transmitted from M antennas of the system to all communication nodes. The signal obtained from the perspective of a communication receiving node can be expressed as:

$$y = \sqrt{\rho} h^H s + z$$  \hspace{1cm} (2)

Formula 2 $\rho$ is the total average transmission power of the communication system. The specific information transmission process is shown in Figure 4.
2.2 Obtain system statistics channel information

In the channel model of the large-scale wireless communication system, the channel information collection equipment is installed on the corresponding channel location, and the channel information of the cargo system is collected according to the method shown in Figure 5.

Considering the uplink transmission link in beam domain of large-scale wireless communication system, \( \mathbf{y}_m \in \mathbb{C}^{T \times 1} \) represents the received signal of the m-th beam on the server side of the system, and \( \mathbf{h}_m \in \mathbb{C}^{K \times 1} \) represents the channel from all users to the m-th beam on the server side of the system, then the received signal of the m-th beam on the server side can be expressed as:

\[
\mathbf{y}_m = \mathbf{x}^H \mathbf{h}_m + \mathbf{z} \quad (3)
\]

Then the received signals of all beams on the server side of the system can be expressed as:

\[
\mathbf{y} = \mathbf{x}^H \mathbf{H} + \mathbf{Z} \quad (4)
\]

The matrix corresponding to \( \mathbf{y} \), \( \mathbf{H} \) and \( \mathbf{Z} \) is:
Taking the communication information received by the server as the original, the channel information of the system can be obtained by using the mapping principle.

2.3 Establish the optimal risk decision rules

The purpose of establishing the optimal risk decision rule is to determine whether there is transmission risk in the channel based on the acquired channel transmission information of wireless communication system [6]. Firstly, the input scan results and matching rules are processed to determine whether they meet the format requirements. If the format requirements are met, the scan results and matching rules are split according to the match sub result separator.

If the format requirements are not met, the corresponding error handling is called [7]. Because there is a corresponding relationship between the scanning results and the matching rules, the number of disassembly results of the scanning results and the matching rules is the same, and they correspond to each other. Further processing is carried out for each scan sub result and matching item pair. The matching item type field in the matching item is used to determine the category of the matching item. The number type matching item and character type matching item correspond to different response results respectively.

2.4 Prediction channel transmission interference

The transmission interference prediction is carried out for the channel with transmission risk. The types and modes of channel interference are shown in Table 1:

<table>
<thead>
<tr>
<th>Jamming type</th>
<th>Missing part</th>
<th>Reason for frame loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synchronous</td>
<td>PHY head</td>
<td>Transmission failure due to frame collision of nodes in the listening range</td>
</tr>
<tr>
<td>jamming</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asynchronous</td>
<td>PHY head, MPDU subframe</td>
<td>Data transmission outside the monitoring range interferes with frame transmission</td>
</tr>
<tr>
<td>interference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical</td>
<td>PHY head, MPDU subframe</td>
<td>Frame damage caused by channel physical factors</td>
</tr>
<tr>
<td>interference</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For any node i, the probability that it is affected by synchronous interference, asynchronous interference, and physical interference and cause frame transmission failure is denoted as $P_{i}^{sym}$, $P_{i}^{asy}$, and $P_{i}^{phy}$ respectively. Considering that physical interference has little and stable influence on frame transmission, in order to simplify interference prediction...
method, quantitative value $\phi$ is used to represent probability $P_i^{\text{phy}}$ [8]. Then the prediction of channel interference can be realized according to the flow shown in Fig. 6.

Assume that there are $N_i$ other sending nodes in the monitoring range of node $i$, and there may be other sending nodes outside its monitoring range, and use frame transmission status information to derive the prediction results of synchronous interference and asynchronous interference.

### 2.5 Analyze channel fading characteristics

In a complex environment, the propagation path of a signal is diverse. One path may be a simple line-of-sight propagation, or it may pass through various non-visual paths through various modes such as diffuse reflection and diffraction. Distance spread. The wireless communication channel has a strong openness. At the same time, the complex receiving environment and the mobility of the receiver make the channel analysis extremely complex and time-varying [9]. Based on the prediction results of channel transmission interference, the characteristics of channel fading risk are analyzed. In general, the channel fading is in a composite form, as shown in Figure 7.

**Fig. 6.** Channel interference prediction flow chart

Assume that there are $N_i$ other sending nodes in the monitoring range of node $i$, and there may be other sending nodes outside its monitoring range, and use frame transmission status information to derive the prediction results of synchronous interference and asynchronous interference.

**2.5 Analyze channel fading characteristics**

In a complex environment, the propagation path of a signal is diverse. One path may be a simple line-of-sight propagation, or it may pass through various non-visual paths through various modes such as diffuse reflection and diffraction. Distance spread. The wireless communication channel has a strong openness. At the same time, the complex receiving environment and the mobility of the receiver make the channel analysis extremely complex and time-varying [9]. Based on the prediction results of channel transmission interference, the characteristics of channel fading risk are analyzed. In general, the channel fading is in a composite form, as shown in Figure 7.

**Fig. 7.** Principle diagram of composite fading

It can be seen from the figure that the composite fading consists of two parts: small-scale fading and large-scale fading. Among them, path loss and shadow fading describe the intensity change of the average receiving power within a certain moving distance when the distance between the transceiver is long, which usually refers to the situation of more than dozens of wavelengths, so they are collectively called large-scale fading. However, multipath fading is
caused by the different transmission paths, different transmission directions and distances, and different transmission modes such as reflection, diffraction and transmission. The final received signal is synthesized by the superposition of multipath signals with different phase and amplitude, and the phenomenon of fast fading will appear. Sharp changes can be observed when the relative distance of transceiver changes very little, so the fast changes in short distance are called small-scale fading. The propagation characteristics of wireless channel can be expressed by formula 6

\[ g = W \cdot Q \cdot j \]  

(6)

Where \( W \) is path loss, \( q \) is shadow fading and \( j \) is multipath fading. The specific value of each variable parameter can be calculated by formula 7.

\[
W = 10 \log \left( \frac{W_r}{W_t} \right)
\]

\[
Q = W(d_0) + 10n \log \left( \frac{d}{d_0} \right) + X_a(t)
\]

\[
f(t, \tau) = \sum_{i=0}^{L} a_i \left( t \right) e^{-2\pi f_i \delta(t - \tau)}
\]

(7)

Where \( W_t \) and \( W_r \) represent transmitting signal power and receiving signal power respectively. \( d_0 \) is the reference distance, \( W(d_0) \) is the path loss at \( d_0 \), \( n \) is the path loss index, which is determined by the propagation environment. \( X_a(t) \) is the Gaussian random variable with the mean value of zero and the standard deviation of \( \sigma \). In addition, \( t \) and \( \tau \) represent time and delay respectively, and \( L \) is the number of resolvable multipaths. \( a_i \) is the attenuation of each path, \( \{ f_i \} \) is the propagation delay of each path, and \( f \left( t_i \right) \) is the Doppler shift. By combining formula 6 with formula 7, we can get the result of the fading characteristic of the wireless communication channel with risk.

2.6 Statistical information transmission risk probability

Based on the channel fading characteristics, the risk value of wireless communication system channel information transmission is calculated, and the corresponding statistical results are obtained by means of probability statistics. The calculation process of channel transmission risk value is shown in Figure 8.
Call the data collection module to collect the original data, the intrusion detection system collects the external threat data in real time, the configuration verification system checks the internal configuration of the target host regularly, and returns the configuration results. The external threat data is the model observation value, the internal configuration is the internal threat data, and the internal and external cause correlation matrix is calculated with the external threat data. After the channel information data is obtained, the internal and external correlation matrix, observation value output matrix and initial state probability are called to calculate a new real-time state transition matrix according to the system changes caused by internal and external threats. Finally, the real-time risk total value of the target host is calculated based on the model algorithm combined with four probability matrices [10]. Taking the interruption risk of large-scale wireless communication system channel as an example, if the channel remains constant during the transmission of the whole codeword, the interruption probability can be used to evaluate the system performance, and the corresponding interruption risk probability can be expressed as:

$$P = P\left\{ \log \left( 1 + \frac{\rho}{N} h^H V V^H h \right) < R \right\} \quad (8)$$

By integrating all possible transmission risks, the statistical results of transmission risk probability can be obtained. The statistical method of channel information transmission risk probability in large-scale wireless communication system is realized.

**3 Experimental results and analysis**

In order to verify the statistical method of channel information transmission risk in large-scale wireless communication system, a performance comparison experiment is designed in the experimental environment. The experimental environment is a large-scale wireless communication system, which encodes the channel information in the wireless communication system. In this experiment, 12 channels are taken as the research object, and the statistical method of information transmission risk and the traditional statistical method are introduced.
into the background program of the system in the form of program code, which ensures the independent operation of the three statistical methods. The object of study is No.1-4 and No.5-8, respectively. The communication transmission channel is controlled by two traditional risk statistical methods, and the designed risk is controlled by two traditional risk statistical methods. The statistical object of probability statistical method is the communication transmission channel numbered 8-12. The purpose of this experimental verification is divided into two parts. One is to verify the risk statistical error of the three methods by comparison, the other is to verify the application results of the three methods. The comparison results of risk statistical error of the three methods are shown in Figure 9.

![Risk statistical error comparison curve](image)

**Fig. 9.** Risk statistical error comparison curve

Figure 9 shows three cycles with 10 slots per cycle. The average error values of the two traditional risk statistical methods are 16.4 and 15.8 respectively, while the average error values of the channel information transmission risk statistical methods of the large-scale wireless communication system are 7.2, which is reduced by about 56%. Then, according to the results of the three risk statistical methods, the low risk value probability is selected to transmit the actual data information, and the stability of the system is observed according to the artificial adjustment of the external environment signal-to-noise ratio Sexual change, as shown in Table 2.

**Table 2** Comparison results of wireless communication system stability

<table>
<thead>
<tr>
<th>Signal-to-noise ratio</th>
<th>Information transmission size/MB</th>
<th>Probability statistics of transmission risk based on quantum secure</th>
<th>Probability statistics of transmission risk based on intuitive multiplication</th>
<th>Data reception under the control of design risk probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coherence block No.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>communication</td>
<td>preference</td>
<td>/MB</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---------------</td>
<td>------------</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.31</td>
<td>3.58</td>
<td>3.42</td>
<td>3.88</td>
</tr>
<tr>
<td>2</td>
<td>14.23</td>
<td>11.26</td>
<td>12.55</td>
<td>13.54</td>
</tr>
<tr>
<td>3</td>
<td>4.08</td>
<td>2.34</td>
<td>2.78</td>
<td>3.27</td>
</tr>
<tr>
<td>4</td>
<td>5.25</td>
<td>3.22</td>
<td>3.05</td>
<td>4.09</td>
</tr>
<tr>
<td>5</td>
<td>8.70</td>
<td>7.25</td>
<td>7.12</td>
<td>8.10</td>
</tr>
<tr>
<td>6</td>
<td>9.32</td>
<td>8.74</td>
<td>8.58</td>
<td>9.14</td>
</tr>
<tr>
<td>7</td>
<td>3.79</td>
<td>2.08</td>
<td>1.97</td>
<td>3.31</td>
</tr>
<tr>
<td>8</td>
<td>5.28</td>
<td>3.99</td>
<td>3.66</td>
<td>5.01</td>
</tr>
<tr>
<td>9</td>
<td>11.23</td>
<td>10.81</td>
<td>10.67</td>
<td>11.08</td>
</tr>
<tr>
<td>10</td>
<td>18.34</td>
<td>16.53</td>
<td>16.85</td>
<td>17.94</td>
</tr>
</tbody>
</table>

In the calculation experiment, the actual data transmission volume is 84.53mb, among which the actual data reception volume based on the transmission risk probability of quantum secure communication is 69.8mb, the packet loss rate is 17.42%, and the actual data reception volume based on the transmission risk probability of intuitive multiplication is 70.65mb. The packet loss rate is 18.65%, but in the design of large-scale wireless communication system, under the constraint of channel information transmission risk statistical method, the actual received data is 79.36mb, that is to say, compared with the traditional method, the risk probability statistical method can effectively reduce the packet loss rate by about 11.31%. Even in the case of serious noise interference, it can accurately select the channel with low risk probability for data transmission, which has high application value.

4 Conclusion

By collecting and counting the transmission state information of large-scale wireless communication system channel, this paper puts forward the statistical method of information transmission risk probability from the perspective of interference probability. The application value of the risk probability statistical method in the actual system operation is proved by the experimental verification. However, when the environment interference is large, there will still be some packet loss in the application of the method, which needs to be further optimized in the future research work.

References


Radio Frequency Fingerprint Identification Based on Multi-Intervals Differential Constellation Trace Figures

Yang Yang1,3, Aiqun Hu2, Jiaobao Yu2
{17601544638@163.com, aqhu@seu.edu.cn, yujiaobao@seu.edu.cn}
School of Cyber Science and Engineering, Southeast University, 210096 Nanjing, China1,
School of Information Science and Engineering, Southeast University, 210096 Nanjing, China2,
School of Power Supply and Engineering, Nanjing Institute of Railway Technology, 210031
Nanjing, China3

Abstract. Differential constellation trace figure (DCTF) has been demonstrated good performance on radio frequency fingerprint (RFF) identification. However, DCTF easily blurs at low SNRs. This paper proposes two novel RFF identification methods for zigbee devices based on multi-intervals DCTFs. First, a low-complexity motion features extraction method is devised based on high-density regions. Besides, an improved 3D-2D CNN model is utilized to extract motion features and spatial features. We collected 54 different ZigBee devices for experiments and classified them by these two methods. The experimental results show that compared with using single DCTF, which is generated by a single differential interval, these two methods can effectively improve the identification accuracy at different SNR levels. The classification accuracy rate of the 3D-2D CNN classifier is over 92% even under the SNR level of 5 dB.

Keywords: radio frequency fingerprint, differential constellation trace figures, 3D-2D CNN, motion features, ZigBee.

1 Introduction

The Internet of Things (IoT) combines the networking of various information sensing devices to achieve the function of interconnection of people, machines and things at any place and at any time. The development of IoT has significantly improved the intelligence of various industries. ZigBee is a new type of wireless communication technology. As an important IoT device, it has the characteristics of low speed, low power consumption and low cost. It is suitable for communication between electronic devices with short transmission distance and low speed [1]. However, due to its cost limitation, ZigBee access authentication is generally based on MAC addresses, thus it is easily attacked by rogue devices [2, 3].

Radio frequency fingerprint (RFF) has been successfully used to identify ZigBee devices [4, 5, 6]. Compared with software or protocol layer-based authentication methods, RFF identification technology uses the inherent characteristics of the physical layer analog transmission circuit hardware, and has the same advantages of anti-spoofing and anti-cloning as biological fingerprints.

Based on data samples, the identification methods of RFF can be divided into transient and steady states [7, 8, 9]. DCTF is a new type of image-based identification method [10]. After differential processing, it distorts the constellation figure by expanding the characteristics of the
fingerprint. It can reflect the characteristics of the device fingerprint through statistical information such as the contour and point distribution of DCTF, which can more intuitively identify ZigBee devices.

There are many machine learning methods suitable for RFF identification. But when faced with a large number of devices, such as large-scale wireless sensor networks, machine learning methods usually cannot achieve a high identification accuracy. As a very effective end-to-end learning method, deep learning has also been widely used in the RFF field recently [11, 12, 13]. Convolutional neural networks (CNN) is translation invariant and can learn spatial hierarchies of patterns [14, 15]. Thanks to their effective learning of complex and abstract visual concepts, they are widely used in computer vision. DCTF is an effective method to reflect the RFF of the device on the figure in the form of statistical information. The fingerprint features of the device can be clearly and prominently displayed on the figures. It is appropriate to identify DCTFs with CNN models. A DCTF classification model based on 2D CNN deep learning was proposed in [16]. Besides, the effects of parameters such as the size and the number of sample points of DCTF on the classification results were explored. The results show that DCTF can be effectively used to identify device fingerprints at high SNRs.

However, DCTF scattered points are directly distributed on a limited image size. Scattered points at different timing intervals overlap due to noise, which has a great impact on fingerprint characteristics. Compared with the methods based on baseband signal data, single-frame DCTF identification method from each ZigBee device tends to perform poorly at low SNR scenarios. Hence, an efficient approach for identifying DCTFs at low SNRs is still urgently required.

In this paper, we use multiple differential intervals to generate multi-frame DCTFs from each device, and extract motion features to identify different devices by two different methods. The main contributions of this paper are as follows:

- We propose two methods for extracting motion features from DCTFs with different differential intervals, which can be used for RFF identification. The multi-intervals DCTFs can be regarded as a continuous motion process of RFF. In order to verify our methods, we conducted a lot of experiments, and collected up to 54 zigbee devices with a total of 2430 frames of signals.
- We propose a unique method by extracting the high-density regions from DCTFs to generate trajectory figures. The generated trajectory figures can be directly used to classify the ZigBee devices. The experimental results show that this method have better classification performance under 5 dB and 10 dB than DCTF on 2D CNN.
- We also propose a 3D-2D fusion model for RFF identification by extracting motion features from 10 consecutive intervals DCTFs at the same time through 3D convolution kernels, and extracting the spatial features through deeper 2D convolution kernels. Compared with the classification method based on DCTF, this model can get a much higher accuracy rate of 92% even at 5 dB.

2 Differential Constellation Trace Figures

ZigBee uses the 2.4 GHz direct sequence spread spectrum (DSSS) O-QPSK modulation format defined by IEEE 802.15.4 [17]. Each byte is divided into two symbols, and each symbol is 4 bits. Each symbol is mapped to one of 16 pseudo-random sequences, and each sequence has 32 chips, thus completing the mapping from symbols to chips. In the end, In-phase and quadrature channels both transmit data at a speed of 1 M Chips/s. In order to complete the drawing of the
constellation trace, we use an oversampling rate of 10MHz to collect the data, that is, 10 points are used on each half-sine wave, and a total of 40K sampling points are collected as the DCTF sample points.

The signal transmitted by the ZigBee transmitter can be expressed by the following expression:

\[ X(t) = (a_r x_r(t) + b_r + j(a_q b_q(t) + b_q))e^{-j2\pi f_{ct}t} \]  

(1)

Where \( a_r \) and \( a_q \) are the I/Q gain imbalance factor, respectively. \( b_r \) and \( b_q \) are the DC offset of I/Q, respectively. \( f_{ct} \) is the carrier frequency of the transmitter.

After the signal is transmitted, assuming that the channel is ideal, the received signal on the receiver is

\[ Y(t) = X(t)e^{j2\pi f_{ct}t} \]  

(2)

Since the transmitter and receiver have different carrier frequencies \( f_{ct} \neq f_{scr} \), carrier frequency offset \( \theta = f_{scr} - f_{ct} \) can be used as a fingerprint feature. In order to obtain a stable differential constellation trace figure, we need to perform the following key steps:

- After the receiver receives the baseband signal and performs energy normalization, first perform the shift operation with shift value of \( \varphi \) on the I/Q channel for feature highlighting:

\[ Y'(t) = (a_r x_r(t) + b_r + j(a_q x_q(t) + \varphi) + b_q))e^{-j2\pi f_{ct}t} \]  

(3)

Since the signal of O-QPSK is a constant envelope, it is in the shape of a circle on the constellation figure. After this, it shows a distorted ring shape. This process can enlarge the morphological characteristics of the fingerprint, but it will not change the expected value of the cluster centers of the signal points.

- The rotation factor \( e^{-j2\pi \theta t} \) is a function of time \( t \), which will cause the constellation figure to change with time and appear as a rotation. By introducing a differential processing to compensate this rotation effect, the resulted constellation figure can be fixed.

\[ S(t) = Y'(t) \cdot Y^*(t + \lambda T_s) \]  

(4)

where \( T_s \) is the sampling interval.

- After differential processing, the statistical characteristics of I and Q values can be reflected in the gray value of each pixel, that is, the higher the frequency of occurrence of the same position, the higher the brightness of the position. It is finally presented in the form of a heat figure or a gray scale figure.
Different differential time intervals $\lambda$ will affect the rotation factor, so that the constellation figure has different point position distributions. When $\lambda = 0$, $S(t) = Y'(t) \cdot Y''(t)$, it is a positive real number, and the trajectory points are concentrated on the positive half axis of the I-axis. When $\lambda$ gradually increases from 1, the aggregation points on the constellation figure gradually spread outward in the form of a distorted ring. Some DCTFs under different $\lambda$ are shown in Fig. 1.

3 RFF Identification Based on motion features

DCTF under low SNRs will severely blur due to the influence of noise, resulting in a low recognition rate. RFF identification using multi-intervals DCTFs is a very feasible method. In Section A, we propose a low-complexity motion features extraction method based on high-density region (HDR).

In Section B, we propose a 3D-2D CNN structure to identify different devices.

3.1 Proposed RFF Identification Based on HDR Method

For the different ZigBee devices $I$, the DCTF can generate 4 centers $I_{p}^{i}$, $p = 1, 2, 3, 4$. The mathematical expectation of the center position value is determined by the fingerprint characteristic values and the differential interval $\lambda$ [10], which can be expressed as:

$$I_{p,\lambda}^{i} = RFF_{p,\lambda}^{i} \cdot e^{-j2\pi \lambda \theta}$$  \hspace{1cm} (5)
Where $RFF_{p, \lambda}$ are fingerprint feature values for different cluster center $p$ under different devices $l$ and $\lambda$. High-density regions will be formed around the centers. In low SNR scenarios, the contour features of DCTF will be destroyed. The high-density area has better anti-noise performance due to a large number of points distribution.

The matrix $\Phi^l$ of DCTF is defined as the statistical results of the signal points in each subzone ($N \times N$). In order to obtain motion trajectory figures under $\lambda = \{\lambda_1, \lambda_2, ..., \lambda_k\}$, we set thresholds $\alpha$ to extract high-density regions and set the values of matrix to 255 or 0.

$$
\Phi^l_{i,j}(m,n) = \begin{cases} 
255 & \text{when } \Phi^l_{i,j}(m,n) \geq \alpha \\
0 & \text{ otherwise (}0 \leq m,n < N) 
\end{cases}
$$  \hfill (6)

Then we generate $k$ matrices $\{\Phi^l_{i,a1}, \Phi^l_{i,a2}, ..., \Phi^l_{i,ak}\}$ under $\lambda$. Add these high density matrices to get the trajectory matrix $\Phi^l_{i,\lambda}$.

$$
\Phi^l_{i,\lambda} = \sum_{a=1}^{k} \Phi^l_{i,a} 
$$  \hfill (7)

Normalize the maximum and minimum values of the trajectory matrix $\Phi^l_{i,\lambda}$ to (0, 255) and generate a grayscale image of the normalized matrix $\Phi^l_{i,\lambda}$.

$$
\Phi^l_{i,\lambda} = \frac{\Phi^l_{i,\lambda} - \min(\Phi^l_{i,\lambda})}{\max(\Phi^l_{i,\lambda}) - \min(\Phi^l_{i,\lambda})} \times 255 
$$  \hfill (8)

For different devices, the generated trajectory figures can be used to identify different devices. The process can be found in Fig. 2. The trajectory figures (in RGB) generated by different devices are shown in the Fig. 3.
3.2 3D-2D CNN model for RFF Identification

3D-CNN model can extract motion features from the multi-frame images, it is widely used in the field of human action behavior recognition. 3D convnet is well-suited for temporal feature learning, while 2D convnet only learn spatial information [18, 19, 20, 21].

In order to obtain more accurate classification performance at low SNRs, we use a 2-layer 3D convolution kernels to extract the motion features between DCTFs of 10 consecutive differential time intervals \(\{\lambda_0, \lambda_0 + 1, ..., \lambda_0 + 9\}\), and add a 2-layer 2D convolution kernels after the 3D convolution kernels, which is used to extract more abstract spatial features. The 3D-2D CNN structure in Fig. 4 is as follows:

- **Input layer:** In order to reduce the complexity of the model, we use a 61x61 grayscale image instead of a color heat map at the input end to classify, that is, the size of the input end is 10x61x61x1.

- **3D convolution layer:** We use 2 consecutive 3D convolution layers to extract the motion track information between multiple frames. The convolution kernel sizes are 8x3x3x5 (8 3x3x5 3D kernels) and 16x3x3x3 (16 3x3x3 3D kernels).

- **2D convolution layer:** After the output of the second 3D convolution layer, the dimensions of the feature maps are reshaped, transformed from (57,57,16) to (57,57,64), and then sent to the 2D CNN to extract more abstract spatial features. The 2D kernels are 32x5x5 (32 5x5 2D kernels) and 64x5x5 (64 5x5 2D kernels). In order to reduce the dimension and size of the features, we add a layer of Max Pooling (2x2) behind each 2D kernel.
Fig. 4. 3D-2D CNN model structure.

- **Fully connected layer**: The feature maps finally input to the fully connected layer. After flattening, an one-dimensional tensor is formed. After vertical connection, it is sent to the fully connected layers consisting of weights and biases. To prevent overfitting, we use the dropout strategy and set its value to 0.5. Finally, it is sent to the softmax classifier to get the predicted device number.

We divide all data sets into a training set and a test set according to 4:1. We choose K-fold cross-validation to divide all training sets into 5 partitions, one of which is the validation set and the remaining four partitions are used for training. When the classification result of the validation set is the highest, the network model is selected as the best model. Our training chooses cross entropy as the classification loss function, the SGD optimizer as the optimizer. The learning rate is 0.01, the batch size is 128. The epoch is 200, and the model parameters are total 4,005,890. The process of identification with 3D-2D CNN can be seen in Fig. 5.

4 EXPERIMENTS AND DISCUSSION

4.1 Experimental System and Processing

To verify our proposed methods (HDR and 3D-2D CNN), we collected 54 Ti CC2530 ZigBee devices for experiments. USRP N210 is used as the receiving device, and each ZigBee collects 45 frames of signals at close range for a total of 2430 frames of signals. Due to the good signal acquisition environment, the SNR of the scene is about 30 dB, and the influence of multipath is ignored. The collected signal is considered to be an ideal signal in the LOS scene.

On the PC side, we use Matlab 2016a to perform energy normalization preprocessing on the collected data. In order to evaluate the classification performance of the model under different SNRs, tested from 0-30 dB (AWGN) every 5 dB. Subsequent I/Q shift and differential processing operations are performed to finally generate multi-intervals DCTFs.

For the collected 2430 frames, we choose continuous differential intervals to generate trajectory figures and multi-frame DCTFs, then send them to 2D CNN and 3D-2D CNN for classification to verify the proposed methods separately. These models were built
under the Tensorflow 1.14 framework with the python 3 environment. The generation and classification process after signal acquisition are all completed on the same computer, configured as CPU I7-9700F, 32GB of RAM, GPU RTX 2060s.

4.2 Identification Accuracy with Different $\lambda$

To test the accuracy of RFF recognition based on single-frame DCTF method under different SNRs, we collected 54 devices and tested the RFF recognition performance at 5-20 dB for $\lambda$ is 1-15. 2430 DCTFs are generated for each differential interval. We use the generated DCTFs to train and test on 2D CNN model. For SNRs above 20 dB, a classification accuracy of more than 98% can be obtained. When SNR is at 10 dB, the performance of RFF recognition decreases rapidly.

It can be seen from Fig. 6 that the classification performance is low and unstable when $\lambda$ is 1-4. This is because when $\lambda$ is small, the point distribution is more clustered, and the difference of DCTFs is small. At the same time, when $\lambda$ is higher than 5, the classification performance is of slight differences and remains stable above 92%. Below 10 dB, there is a large difference in classification performance under different $\lambda$. This is because the noise destroys the characteristics of DCTF seriously, and the 2D CNN model classification cannot obtain stable classification results.

4.3 Comparison of Different Classification Methods

In order to evaluate the performance of different classification methods (3D-2D CNN versus HDR versus DCTF) under different SNRs, we divide the SNRs into two levels,
Fig. 6. Accuracy for DCTF on 2D CNN with different $\chi$ under different SNRs.

Fig. 7. Accuracy for different methods with different $\lambda$ under different SNRs.

We select multi-intervals DCTFs with different $\lambda$ as the input of 3D-2D CNN for classification under continuous intervals $\lambda$ within 80. For the HDR method, we choose the continuous intervals $\lambda$ of 10-15 and 15-20 for verification. The classification results are shown in Fig. 8 and the fusion matrices of 3D-2D CNN, HDR, DCTF are shown in Fig. 8-10.

At low SNRs, the 3D-2D CNN classification performance is very sensitive to $\lambda$, and the classification variation range exceeds 60%. When $\lambda$ is 1-10, the classification performance is poor, and the classification accuracy is less than 20% at 0 dB, which is
lower than HDR and DCTF classification methods. When $\lambda$ is small, the points of DCTF are too clustered.

When SNR is between 5 dB and 10 dB, HDR is slightly improved over the DCTF method, which proves that motion features have better anti-noise performance at low SNRs.

The classification results of the 3D-2D CNN model is rapidly improved over 5 dB, and the performance of most differential intervals is above 90%, and the highest even reaches
95.49%. It is much higher than HDR and DCTF. When SNR reaches 10 dB, all the 3D-2D models with different $\lambda$ are above 94%, and the highest is 98.03%.

In the mid-to-high SNRs, when the SNR level is 15 dB, all the models of 3D-2D CNNs reach more than 98% except $\lambda$ is 1-10. However, HDR and DCTF also achieved a classification accuracy of 92.1% ($\lambda = 15 - 20$) and 91.79%. When SNR is higher than 20 dB, all methods achieve a classification accuracy close to 100%.

4 CONCLUSION

This paper proposes two RFF identification methods based on continuous multi-intervals DCTFs. We first proposed a low-complexity motion features extraction method, and successfully used these features for ZigBee device identification. In order to improve the low identification rate under a low SNR level, a multi-intervals DCTFs based on DL RFF identification method is proposed. The 3D convolution kernels are used to extract more complex multi-frame motion features. Using 2D convolution kernels to extract static spatial features, classification accuracy of 92% and 98% were obtained at 5 dB and 10 dB. The classification results at low SNRs are significantly improved. In the future, we will consider ways to improve the accuracy of RFF identification based on DCTF in a time-varying multipath environment.

5 ACKNOWLEDGMENT

This work was supported in part by the Research Fund of Jiangsu Provincial Key Research and Development Program under Grant BE2019109, National Mobile Communications Research Laboratory 2020B05, Southeast University.

References


Study on the optimization model of rural spatial main function area distribution based on 5g Technology

WEN Wen¹,LI Yan-dong ²
{wenwen8779@163.com1,sdfds8574@163.com2}
1.Yaha School of Built Environment, Haikou University of Economics,Haikou 571127,China;
2.Haikou University of Economics,Haikou 571127,China)

Abstract: In view of the problem that the traditional main function area allocation optimization model leads to the small rural profit, the optimization model of rural space main function allocation based on 5G technology is studied. First of all, the distribution index system of the main function area of the village is established, and the weight of each index system is determined. The quadrant method is used to determine the main function area, and the 5G technology is used to schedule the communication resources of the main function area. Combining with the established main function area allocation index system, the construction of the main function area allocation optimization model is completed. By comparing with the traditional allocation optimization model, it is proved that the functional area allocation optimization model based on 5G technology can improve the rural profit with better application prospect.

Keywords:5G Technology; Rural Space; Main Function Area Distribution; Optimization Model;

1 Introduction

The main function area is the area that can represent the core function of the region. Each region because of the different main functions, division of labor and cooperation, common prosperity, common development. The main function is determined by its own resources, environmental conditions, social and economic basis, and is also given by a higher level of region [1]. The main function is different, the region type will have the difference. The natural environment and resource conditions of various regions of China are quite different, so that all regions cannot develop in accordance with a unified development model. The main function area is an effective way to promote the coordinated development of the region and realize the rational distribution of population and economy, is the urgent need to realize sustainable development and improve the utilization ratio of resources, is the inevitable requirement to adhere to the people-oriented and realize the equalization of public services, and is an important measure to improve the level of regional regulation and enhance the effectiveness of regional macro-control [2]. Therefore, the spatial main function area can better promote the regional development. For the rural space, the rational allocation of the main functional areas
can better develop and protect the rural resources. The traditional main function area allocation optimization model, because the overall grasp of the village is not comprehensive in the distribution, the allocated area cannot maximize the advantages of rural space, there are some limitations [3].

The fifth generation of mobile communication technology is the latest generation of cellular mobile communication technology, and it is also an extension after 4G, 3G and 2G systems. The performance goals of 5G are high data rate, reduced delay, energy saving, reduced cost, improved system capacity and large-scale device connectivity [4]. The first phase of the 5G specification in Release-15 is to adapt to early commercial deployment. Phase II of the release-16 will be completed in April 2020 and submitted to ITU as a candidate for IMT-2020 technology. The ITU IMT-2020 specification requires speeds up to 20 Gbit/s, enabling wide channel bandwidth and large capacity MIMO. Based on the above analysis, this paper will study the optimization model of the main function area of rural space based on 5G technology.

2 Constructing the Optimization Model of the Functional Area of Rural Space Based on 5G Technology

2.1 Establishing the Distribution Index System of the Rural Main Function Area

The distribution of main functional areas should mainly consider natural ecological conditions, carrying capacity of soil and water resources, location characteristics, environmental capacity, existing development density, economic structure characteristics, population agglomeration, participation in the international division of labor and other factors. As shown in the following table, this paper first establishes the distribution index system of the main function area of rural space.

<table>
<thead>
<tr>
<th>Primary indexes</th>
<th>Secondly indexes</th>
<th>Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource utilization</td>
<td>Land resources are available per person</td>
<td>Assess the amount or carrying capacity of available land resources in an area</td>
</tr>
<tr>
<td></td>
<td>Available water resources</td>
<td>The amount of water available in an area in the future</td>
</tr>
<tr>
<td></td>
<td></td>
<td>The amount or potential to support socio-economic development</td>
</tr>
<tr>
<td></td>
<td>Environmental capacity</td>
<td>To assess the absorption capacity of an</td>
</tr>
</tbody>
</table>

Table 1 Distribution index system of main function area of rural space
After establishing the distribution index system of the main function area of the rural space shown in the above table, the weight of each index of the index system is determined.

### 2.2 Determine the weight of each index in the index system

For the environmental level indicators in Table 1, experts from relevant fields are invited to rate the ecological environment of the target main functional areas of rural space. Combined with the results of the following formula, the index weight of the first-level index environment in the distribution index system is obtained.

\[
E = \frac{\sum_{i=1}^{n} R_i B_i}{\sum_{j=1}^{n} B_j} \quad (1)
\]

In formula (1), \( E \) is the ecological carrying capacity of regional unit area. \( R_i \) is a global equilibrium factor for regional productivity of the same land type; \( B_i \) represents productive land area in the region. If the average score of the expert is divided into \( \bar{x} \), then the index weight calculation formula of the first-level index environment in the main function area allocation index system is as follows.

\[
W_E = \frac{E + \bar{x}}{2} \quad (2)
\]

For the traffic index in the first class index, traffic accessibility to the response area can
be used. The formula for calculating accessibility is as follows:

\[
A_i = \frac{\sum_{j=1}^{n} T_{ij} M_j}{\sum_{j=1}^{n} M_j} \quad (3)
\]

In formula (3), \( A_i \) For the accessibility of \( i \) roads in regional traffic networks, \( T_{ij} \) is the shortest distance from Road \( i \) to Road \( j \), time is used in this paper to measure the length of the distance, and \( M_j \) is the attraction user of the \( j \) road using the gravitational size parameter [5].

The resource index in the distribution index system of the main function area of rural space can first compare the per capita available land resources with the per capita available water resources, and take the small value as the molecule in the calculation formula. Then, compare the risk of natural disasters and environmental capacity, take the large value as the denominator of the formula, that is:

\[
W_r = \frac{\min(x_1, x_2)}{\max(x_3, x_4)} \quad (4)
\]

In formula (4), \( x_1 \) Land resources available per capita, \( x_2 \) for available water resources, \( x_3 \) Risk of natural disasters, \( x_4 \) for environmental capacity. After calculating the weight of the primary index in the system, the weight of the secondary index in the distribution system is calculated [6]. According to the following formula, the weight of the first-level index in the distribution index system of the main function area of rural space can be calculated.

\[
W_i = \frac{\sum_{j=0}^{n} S_i \cdot \mu_j}{\sum_{j=0}^{n} S_j} \quad (5)
\]

In formula (5), \( W_i \) is the weight of \( i \) first-level indicator, \( S_i \) The degree of
importance set in accordance with the criteria for distribution of \( i \) first-level indicator, \( \mu_i \) Relevant index for the allocation of \( i \) first-level indicator, \( n \) is the number of secondary indicators included in the first-order indicator. After determining the weight of the first-level index, the weight of the second-level index can be determined according to the following formula.

\[
W_{ij} = W_i \cdot \mu_{ij} \quad (6)
\]

In formula (6), \( W_{ij} \) is the weight of the secondary indicator under the \( i \) -first-level indicator, \( \mu_{ij} \) the allocation index for secondary indicators \[7\].

After determining the weight of each index in the distribution index system of the main function area of the rural space, the index weight in the distribution index system is used to locate the main function area of the rural space.

2.3 Location of main function area of rural space

According to the distribution index system of the main function area of rural space, the image limit method is adopted to determine the main function area and establish the number axis. The x-axis represents the economic development support (including the carrying capacity of resources and the carrying capacity of the environment), and the y-axis represents the natural ecological binding force (including the development density and the development constraint potential).

As shown in the following figure, in area I, the x value is high and the y value is low. There are two situations in Area II: The x value is high, the y value is moderate and the x value is moderate. There are also two cases in area III: low x value, moderate y value and moderate x value, and low y value. In Area IV, the x value is low and the y value is high \[8\].
Among them, development zones are forbidden to be important ecological service functional areas and environmentally sensitive areas, including nature reserves and drinking water source protection areas.

Restricting the development zone as the ecological fragile zone and the environment sensitive zone around the development zone, its resources and environment carrying capacity is weak, the large-scale agglomeration economy and population conditions are not good enough, and it is related to the ecological security of the larger area. Including the experimental areas of nature reserves, important water conservation areas, serious areas of soil erosion, scenic spots and so on.

Optimal development area refers to the area with high density of land development and weakening of carrying capacity of resources and environment, which is the core of economy. The key development area refers to the areas with strong carrying capacity of resources and environment, great potential for development and good conditions for economic and population agglomeration.

The entropy value method is used to divide the main function area, assuming that there are $m$ items to be evaluated, $n$ items evaluation index, thus forming an evaluation of the original data matrix $X = \{x_{ij}\}_{i,n}, x_{ij}$ represents the value of the evaluation index of item $j$ of the $i$ sample. The information entropy is defined as follows:

$$H(x) = -\sum_{j=1}^{n} g(x_j) \ln(x_j)$$  \hspace{1cm} (7)

By means of standardized data processing, the information entropy values $e$ of the
evaluation index and \( j \) of the information utility value are calculated as follows:

\[
e_j = -K \sum_{i=1}^{n} y_{ij} \ln y_{ij} \quad (8)
\]

In formula (8), \( K \) is constant [9]. When \( m \) samples are in a completely disordered distribution state, \( y_{ij} = 1 \) and \( K = \frac{1}{\ln m} \) at this time. After calculating the entropy value, combined with the weight of each index in the distribution index system, the location of the main function area in the rural space is located. After locating the main function area, in order to make better use of the resources of the main function area, use 5G technology to schedule the communication resources of the main function area.

### 2.4 Using 5G Technology to Schedule Main Function Area Communication Resources

When the main function area of rural space is allocated, the main function area resources are allocated according to the communication transmission in rural space. The following figure is a schematic diagram of using 5G technology to schedule the main function area resources.

![Fig. 2 Schematic diagram of the resources of the main function area of the scheduler](image)

Set a total of \( t \) 5G communication users assigned to \( Q \) main function areas \( c_1, c_2, \ldots, c_Q \), each main functional area is provided with a C-RAN system RRHs, several adjacent main functional areas are divided into a local area, a total of \( U \) regions, and each
region is provided with a microcloud server. On this basis, we define $b_{k,j,d}$ and $p_{k,j,d}$ as the frequency band and transmit power corresponding to the service $s_j$ required for user $y_j$ in the main functional area $k$, the downlink rate $R_i$ corresponding to this service.

$$
\begin{align*}
R_i &= \sum_{y_i \in s_k} x_{k,j,d} b_{k,j,d} l_b (1 + \gamma_i) \\
\gamma_i &= \frac{p_{k,j,d} G_{k,j}}{\sum P_d G_{d,j} + N_0 b_{k,j,d}} 
\end{align*}$$

(9)

In formula (9), $x_{k,j,d} = 1$ indicates that user $y_j$ in the main function area $k$ is using the service $s_j$; $\gamma_i$ denotes the signal-to-noise ratio; $G_{k,j}$ is the transmission loss of user $y_j$ in the main function area $k$; $G_{d,j}$ is the transmission loss of user $y_j$ that interferes with the base station to the main function area $k$; $P_d G_{d,j}$ interference for other base stations; $N_0 b_{k,j,d}$ Gaussian white noise interference.

For the resource scheduling scheme of network C-RAN, assuming that the frequency band of a certain time point system distributed to the main functional area $k$ is $B_k$, then the constraints are as follows:

$$
\sum_{y_i \in s_k} \sum_{x_{k,j,d}} x_{k,j,d} b_{k,j,d} \leq B_k
$$

(10)

According to the above conditions, the communication resources in the main function area are scheduled [10]. In the main function area after positioning, combined with the optimal communication resource scheduling model and the allocation index system, the allocation of the main function area is optimized to complete the construction of the model.

### 2.5 Complete the construction of allocation optimization model

The optimal use of 5G technology to schedule the main function area communication resource is optimized. Get the assignment optimization model objective function shown below.
\min f(x) \\
\text{s.t. } \begin{cases} 
Ax \geq b \\
l \leq x \leq h
\end{cases} \quad (11)

In the above formula, $A$ allocates the expense parameter for the main function area, $b$ is the constraint vector, and $l$ and $h$ are the cost function constraints for the rural space main function area. There are more than one optimization parameters for the distribution of the main function area of the rural space satisfying the above objective function, the parameters of the optimization model are determined, and the model is cross verified.

Use the training data set and use the established allocation index system to assign weight to the data set. Multiple optimization model parameters are developed by training. These model parameters are used to calculate the cross validation error for the cross validation set separately, and the model parameters with the minimum cost function are selected. Final error checking of the model using actual data. If the error between the output and the actual value of the final model is within the expected error range, then the error of the model is determined to meet the requirements. If the error between the output of the model and the actual value is not within the expected error range, continue to train the model until the model output meets the requirements. At this point, the construction of the allocation optimization model of rural main function area based on 5G technology is completed.

3 Experiment

This paper constructs the rural main function area allocation optimization model based on 5G technology. This section will design contrast experiments to test the model performance constructed in this paper.

3.1 Experiment content

The experimental group is the traditional allocation optimization model based on improved PSO, and the experimental group is the rural main function area allocation optimization model based on 5G technology studied in this paper. The contrast index of the experiment is the rural profit growth ratio of the experimental area after two optimization models.

The experimental data are processed and analyzed on the computer platform configured in the following table.

<table>
<thead>
<tr>
<th>Table 2 Parameters of Computer Virtual Simulation Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>project</td>
</tr>
<tr>
<td>CPU</td>
</tr>
</tbody>
</table>
3.2 Experimental data

The experimental objects selected in this paper are A and B two rural spaces. The experimental group was rural A, and the contrast group was rural B. The specific parameters of the experimental object are shown in the table below.

<table>
<thead>
<tr>
<th>project</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geological landforms</td>
<td>Mountains and plains</td>
<td>Plains and hills</td>
</tr>
<tr>
<td>Climate</td>
<td>Monsoon climate of medium latitudes</td>
<td>Monsoon climate of medium latitudes</td>
</tr>
<tr>
<td>The water resources</td>
<td>18 rivers and streams</td>
<td>16 rivers and streams</td>
</tr>
<tr>
<td>Soil resources</td>
<td>White pulp soil, sand ginger black soil</td>
<td>Meadow soil, paddy soil, mulch soil,</td>
</tr>
<tr>
<td>Vegetation resources</td>
<td>Shrub and broadleaf forests</td>
<td>Meadows, shrubs, cultural vegetation</td>
</tr>
<tr>
<td>Cultivated land area</td>
<td>1,605,230</td>
<td>1,724,010</td>
</tr>
<tr>
<td>Woodland area</td>
<td>75,369</td>
<td>52,147</td>
</tr>
<tr>
<td>The number of residents</td>
<td>35,000</td>
<td>42,700</td>
</tr>
<tr>
<td>Number of industries</td>
<td>8</td>
<td>11</td>
</tr>
</tbody>
</table>

Use the experimental data shown in the above table to complete the experimental verification process, record and analyze the experimental data, and draw the corresponding conclusions.

3.3 Experimental result

The experimental results are shown in the following figure, analyze the information in the diagram, and draw the experimental conclusion.
Fig.3 Experimental results

As can be seen from the above figure, the P/E growth ratio of village A using the experimental group model is much higher than that of village B using the contrast group model. Analyzing the trend of the curve, the profit growth ratio curve of rural A fluctuates in a certain range, and finally remains stable. The rural B-earnings ratio curve showed a downward trend as a whole, and then the downward trend slowed down, gradually approaching zero. indicating that the rural profitability of the applied experimental group model is better. To sum up, the allocation optimization model based on 5G technology can improve the rural profit and have a better application prospect.

4 Conclusion

The main function zoning involves all aspects of population, economy, society, resources and environment, which needs a lot of coordination and overall planning. This is a very challenging innovation work different from any previous zoning. The main function orientation is a tangible hand for the government to regulate and control the economic and social development under the market economy system. It is of great value and innovative significance for the implementation of the scientific development concept, the establishment and improvement of the socialist market economy system, the promotion of resource conservation and environmental friendliness, and the overall planning of the efficiency and fairness of urban and rural regional development.

This paper constructs the allocation optimization model of the rural main function area based on 5g technology, through the comparison experiment with the traditional distribution optimization model, it is proved that the village with the allocation optimization model constructed in this paper is more profitable and can promote the use of resources in the rural main function area.

5 Fund project
supported by the philosophy and Social Sciences Planning of Hainan Province, Project No.: HNSK(YB)18-52

Reference


Research on 5G Millimeter Wave Phased Array Antenna for Broadband Communication

Li Jihao1, Yang Zhenfeng2, Li Yibing3, Sun Xiaohang, Lv Ji, Zhong Wei, Dong Junfeng
ljh@hd-mw.com1, hengdian@hd-mw.com2,3
Nanjing Heng Dian Electronics Co. Ltd, Nanjing, China1,2,3

Abstract. A highly integrated millimeter wave communication active phased array antenna has been developed to meet the demand of a new generation of 5G millimeter wave communication. It has a high gain for compensation of millimeter wave transmission loss. It can generate multi-beams for multi-user by beamforming to multiplex time-frequency resources with an interference suppression between multiple users. The highly integrated millimeter wave communication active phased array antenna has a wide scanning range, a high emission EIRP value, a good cross-polarization characteristics, a high-speed real-time beam control, and a good interactive performance of the host computer software. It has a system verification to meet the need for a new generation of 5G millimeter wave broadband communications. 5G millimeter wave phased array antenna test result for broadband communication shows that data transmission rate is more than 2.8 Gbps with a spectrum width of 800 MHz and a modulation of 64 QAM. At the same time, highly integrated millimeter wave communication active phased array antenna can be applied to satellite communications and tactical mobile hotspots, etc.

Keywords: phased array antenna, phased array, antenna array, millimeter wave communication

1 Introduction

The traditional active phased array antenna is usually realized by tile and brick structure transceiver module based on Multi Chip Module technology[1,2]. This causes a complicated process, a high cost and a big size of the traditional active phased array antenna. It is difficult to meet the demand of mass production of future 5G millimeter wave communication base stations.

With the development of new generation of 5G mobile communication technology, a lot of research has been conducted on the millimeter-wave phased array antenna technology that may be applied to communication and detection. It is foreseeable that in the next 3 to 5 years, more and more technological breakthroughs will appear in millimeter wave devices, packaging, integration and even system architecture design. Low-cost and other aspects relying on the civil technology, the millimeter wave phased array technology that meets the needs of future military applications will also show a rapid and even leapfrog development [3]

5G communication requires high data transmission rate. But it is difficult to find continuous broadband frequency band below 6 GHz, and it is difficult to greatly improve spectrum utilization efficiency by simply relying on coding technology. While millimeter wave frequency
band can provide continuous large bandwidth, which can meet the requirement of high data transmission rate of 5G communication. Therefore, it is necessary to conduct research on millimeter wave antenna technology in order to realize millimeter wave broadband communication.

The transmission loss of the millimeter wave signal in the atmosphere is larger than that of the low-band microwave signal. The millimeter wave active phased array antenna technology brings a higher gain based on the multi-antenna beamforming technology, which can compensate the large transmission loss of the millimeter wave. In addition, the millimeter-wave active phased array antenna technology realizes multi-user multiplexed time-frequency resources by time-division multi-beam or simultaneous multi-beam. It can suppress interference between multiple users through beamforming, which improves the capacity and frequency effectiveness of the 5G communication system. Therefore, millimeter wave active phased array antenna technology has become one of the key technologies of 5G communication.

2 Highly integrated millimeter wave phased array antenna technology

The research on millimeter-wave phased array antenna technology will inevitably need to combine specific requirements such as military early warning detection, wireless communication, and multi-functional integration, and keep up with the latest advances in millimeter-wave frontier technology in the industrial world. The possible technical approaches in the field of phased array causes military electronic equipment to achieve a leapfrog in performance, and the cost will continue to fall. This not only meets the objective needs of equipment development, but also provides a good reference for the subsequent application of millimeter wave technology to the low-band microwave band application [3].

The prototype of the 5G millimeter wave phased array antenna developed is shown in Figure 1. A multifunctional antenna board is used to integrate the antenna unit, feed network, power network and wave control network. The multifunctional transceiver chip is used to integrate the transceiver channel to simplify the phase array antenna topology. It reduces the thickness of the phased array antenna, which can meet the requirements of thin, light and modular [4,5,6].
5G millimeter wave phased array antenna structure schematic diagram refers to Figure 2. It is composed of multi-function antenna board, multi-function transceiver chip, wave control board, power module and host computer control program.

When the millimeter-wave active phased array antenna operates in the transmitting state, the transmission switch state, the numerically controlled phase shifter and the numerically controlled attenuator control code of the transceiver channel are set to complete the power allocation according to the azimuth, elevation angle and other information from the terminal. After phase, attenuation setting and power amplification, signal is sent to the antenna front to realize the space power synthesis, so that the radiation direction of the phased array antenna is directed to the target direction.

When the millimeter wave active phased array antenna operates in the receiving state, parameters such as the switch state, the numerically controlled phase shifter and the numerically controlled attenuator control code are set according to the azimuth angle and elevation angle information returned by the terminal. At the same time, the received high-frequency signals are amplified and phase-shifted and attenuated by the respective receiver channels after passing through the antenna array. And the received high-frequency signals enter the receiving feed network. They form the sum signal, the azimuth difference signal and the elevation difference signal. And they output to the receiver for related signal processing.

2.1 Millimeter wave phased array broadband array antenna technology

In order to achieve a thin and light structure, the millimeter wave phased array antenna uses a form of magnetoelectric dipole antenna. The array antenna is divided into three layers as a whole, which is fed by a slot through a strip line. The structure of the antenna cell is shown in...
Figure 4. The antenna cell size is 4*4 mm. The simulation antenna gain is more than 6.4 dB between the frequency range of 26-31GHz as Figure 5 shows. Figure 6 shows that 3dB beamwidth of the antenna is about 90 degrees.

The standing wave characteristic of the antenna is shown in Fig. 7 and S11 of antenna is maintained below -18 dB.

Fig. 4. Antenna cell structure

Fig. 5. Gain of antenna cell
2.2 Millimeter wave phased array antenna wave control technology

The beam control system controls the phase shifter of the transceiver channel according to the control instruction of the terminal. And it performs the beam scanning and pointing function of the phased array antenna. At the same time, it controls the transmission and reception state of the multi-function chip of the transceiver channel according to the control instruction from the terminal. The phased array working mode is switched. In order to ensure the stable and reliable operation of the system, the beam control system also controls the power modulation driver.

The beam control board performs the timing control of the beam controller and the calculation of the wave control digital code based on FPGA. FPGA has the characteristics of fast speed and high flexibility. In order to improve the response speed of the beam control system, the main function of the beam control is realized in the FPGA.
2.3 Host computer control software for millimeter wave phased array antenna

The host computer control software is responsible for setting the beam scanning azimuth and elevation angles, setting the transmitting state and receiving state, and monitoring the phase shift code and pattern.

The communication between the host computer and the beam control board is completed through Ethernet, SPI interface or serial port RS-232. The host computer for beam control test based on VC6.0 is used to control and observe the results of the beam control system. Its operation interface is shown in the figure 11. Its functions include interface communication, angle data transmission, transmission status control and monitoring, line phase shift base code, and column-wise shift base code monitoring.
3 Test and verification of millimeter wave phased array antenna

Scanning pattern test results of tiled millimeter-wave phased array antennas are shown in Figure 12 [7].

The normalized emission scan pattern of the millimeter-wave phased array antenna has good pattern characteristics in the range of ± 45 degrees. And the side lobe level is lower than -12dB.
Figure 13 shows the equivalent isotropically radiated power (EIRP) test results of a tiled millimeter-wave phased array antenna at 28 GHz. The EIRP value is greater than 51 dBm.

Figure 14 shows the cross polarization test results of the tile millimeter wave phased array antenna in the range of ± 45 degrees. The cross polarization isolation is better than -30 dB in the range of ± 45 degrees.
Figure 15 shows the 5G millimeter wave phased array antenna test results for broadband communication. Data transmission rate is more than 2.8 Gbps with a spectrum width of 800 MHz and a modulation of 64 QAM.

Conclusions

A highly integrated millimeter-wave communication active phased array antenna has been developed for the needs of next-generation 5G millimeter-wave communication. It has a high gain to compensate for millimeter-wave transmission loss. It forms a multi-beam to realize multi-user multiplexing of time-frequency resources. And it uses beamforming to suppress multi-user Interference. Highly integrated millimeter-wave communication active phased array antenna has a wide scanning range, a high transmit EIRP value, a good cross-polarization characteristics, a high-speed real-time beam control performance, and a good computer software for human-computer interaction. It meets the needs of the new generation of 5G millimeter wave broadband communications. 5G millimeter wave phased array antenna test result for broadband communication shows that data transmission rate is more than 2.8 Gbps with a spectrum width of 800 MHz and a modulation of 64 QAM. At the same time, highly integrated millimeter wave communication active phased array antenna can be applied to satellite communications and tactical mobile hotspots, etc.

References

Research on Access Rate Optimization Algorithm in D2D Communication System

Yanliang Ge$^1$, Qiannan Zhao$^1$, Guanghua Zhang$^{1,*}$, Weidang Lu$^2$

$^1$ Northeast Petroleum University, Daqing 163318, China;
$^2$ Zhejiang University of Technology, Hangzhou 310014, China;
* dqzgh@nepu.edu.cn

Abstract. In the uplink mode based on cellular network, device to device (D2D) multi-user shared channel can effectively improve the spectrum utilization and overall communication capacity of the system, but it will also bring serious interference problems, which will affect the end-user experience. To solve this problem, an access rate optimization algorithm based on interference minimum allocation multiplexing resources is proposed in this paper. In the communication system where D2D users and cellular users coexist, the objective function of minimum interference is established firstly, and then the Hungarian algorithm is used to optimize the objective function, so that the total interference of D2D users in the system is the minimum, and the maximum number of D2D users accessing the system is achieved. The simulation results show that the algorithm can reduce the interference of D2D users to cellular users in a certain range, and also increase the number of D2D users accessing the system to a certain extent.

Keywords: D2D communication, Cellular network, Resource allocation, Interference control.

1 Introduction

With the rapid development of wireless communication technology and the rapid popularization of intelligent terminal equipment, with the exponential growth of network traffic, it puts tremendous pressure on operators' core networks and spectrum, and mobile networks are facing severe challenges [1]. Device to device (D2D) communication technology is also called terminal pass-through technology. It directly implements data exchange and service provision without the need for base station transfer. The introduction of D2D communication in the cellular network can reduce the burden of the base station, reduce mobile terminal power consumption, improve transmission efficiency, and increase spectrum Utilization, improve system performance.

Under the control of the base station, D2D communication technology can use cell resources in orthogonal mode or multiplex cell resources in non-orthogonal mode (multiplexing mode) for communication, at present, most researches on D2D communication in multiplex mode [2]. In the multiplexing mode, the introduction of D2D
communication technology into the cellular networks will bring serious interference problems [3]. How to allocate resources to reduce interference becomes the research focus of D2D communication.

In [4], an algorithm based on joint mode selection and resource allocation is proposed to achieve the maximum capacity and minimum interference of the communication system. In [5], a greedy heuristic algorithm is proposed, which takes the channel allocation problem as a mixed-integer nonlinear programming, and then improves the total throughput as much as possible under the requirement of ensuring the entire user's signal to interference ratio (SINR). In [6], according to the threshold of the signal-to-noise ratio of cellular users and D2D users, as well as the maximum transmitting power of user equipment, the transmitting power of D2D users is calculated, to ensure the quality of service of users. In [5, 6], although the quality of communication service is guaranteed, the improvement of system throughput is limited. In [7], under the condition of ensuring the quality of service (QoS) cellular users and D2D users, the access set of D2D users is selected by linear programming to ensure that the interference of D2D communication will not affect cellular communication; secondly, a channel allocation scheme is proposed to maximize system throughput. In [8], a scheme of resource allocation and power control is proposed to ensure the interference of cellular link is controllable. In [9], a simplified bipartite graph is proposed, which uses the Hungarian algorithm for resource allocation to maximizing the rate. In [10], the authors propose a priority heuristic algorithm of D2D for multiplexing channel resources between users and cellular users, which can dynamically adjust the multiplexing channel resources, effectively increase the number of users allowed for D2D communication, and reduce the interference to cellular users. In [11], the resource allocation problem of D2D users is transformed into a parameter planning problem that can be solved by the Dinkelbach algorithm, and joint power control, relay selection and resource allocation scheme is obtained by the Hungarian algorithm. In [12], it guarantees the quality of service (QoS) requirements of users and puts forward a resource allocation method of grouping based on the greatly improved access rate. However, the system throughput of this scheme is relatively low, and once the number of D2D users increases, it cannot guarantee that the distance between D2D users in the group is within a reasonable range.

In the allocation of D2D multi-user shared channel, reduce the interference between cellular users and D2D users and maximize the number of D2D users accessing the system, to allocate communication resources reasonably. In the communication system where D2D users and cellular users coexist, a scheme is proposed to allocate multiplexing resources according to the minimum interference.

2 System Model

For the convenience of research, this paper only considers the scenario of D2D users multiplexing cellular communication resources in the uplink. Suppose the base station (BS) is in the center of the cellular system, there are also M cellular users (CU) in the system, and the set is expressed as $C = \{CU_1, CU_2, \cdots, CU_m, \cdots, CU_M\}, K \text{ D2D}$
users (DU), the set is expressed as $D \{DU_1, DU_2, \cdots, DU_k, \cdots, DU_K\}$ and $N$ uplink channel, CU and DU are randomly assigned in the system. DU exists in pairs, each D2D user pair includes D2D transmitting user (DUT) and D2D receiving user (DUR). In this paper, it is referred to as T and R for short. In Figure 1, the system model is shown.

Compared with the downlink, the spectrum utilization of uplink communication links in the cellular networks is lower. and the interference of the receiver of D2D users is relatively less when multiplexing the uplink communication. At the same time, the base station can also control the spectrum resources in the system according to its interference. Therefore, this paper only considers the sharing scenario of uplink resources in a single cell. To ensure the communication quality, at most one DU can be reused for one CU, and at most one CU can be reused for a pair of DU users.

3 Problem Description

When DU reuses the resources of CU, it will be interfered with by CU and other DU. In addition, DU will also interfere with the base station. In order to reduce the interference between cellular users and D2D users and allocate the communication resources reasonably, this paper establishes a system optimization model. In the system model of Figure 1, in order to improve the communication efficiency, when the kth
D2D pair multiplexes the communication link of the mth cellular user, the receiver of the D2D user will be interfered by the uplink cellular channel, so the signal interference ratio of the D2D receiver meets the following requirements:

$$SINR_{k,m} = \frac{G_k P_{k,m}}{N_0 + A_{m,k}} \geq T_0$$  \hspace{1cm} (1)

Where $G_k$ is the channel gain between $R_k$ and $T_k$ of the D2D user, $P_{k,m}$ s the transmission power of $T_k$ multiplexing the mth uplink cellular user, $N_0$ is the white noise received by $R_k$, $A_{m,k}$ is the interference value of DU link caused by the interference of CU link. In order to ensure the normal and effective communication of DU, it is necessary to meet the requirement that the signal to interference ratio is at least $T_0$. Then transmit power $T_k$ needs to meet:

$$P_{k,m} \geq P_{k,m}^{(0)}$$  \hspace{1cm} (2)

When $DU_k$ multiplexes $CU_m$, $T_k$ needs to achieve the minimum transmit power.

$$P_{k,m}^{(0)} = \frac{T_0 (N_0 + A_{m,k})}{G_k}$$  \hspace{1cm} (3)

According to the system model in Figure 1, it is known that the D2D user may also cause interference to the base station when multiplexing the cellular user, and the interference to the base station when multiplexing the mth cellular user channel for the kth D2D user:

$$I_{k,m} = P_{k,m} H_k$$  \hspace{1cm} (4)

Where $H_k$ is the channel gain between $T_k$ and base station.

According to Shannon formula, the throughput of D2D users when multiplexing cellular channel as:

$$C_{k,m} = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{n=1}^{N} \log_2 (1 + SINR_{k,m}^n)$$  \hspace{1cm} (5)

Among them, $C_{k,m}$ is the throughput when multiplexing the mth channel for the kth D2D user, and $SINR_{k,m}$ is the maximum signal to interference ratio obtained under multiplexing cellular user communication.
In order to maximize the capacity of the system, the essence is to find the maximum SNIR. As formula (1) (4) shows, formula (5) of maximum capacity problem can be further transformed into the problem of minimum interference.

\[ \min \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{n=1}^{N} \beta_{k,m}^{n} I_{k,m}^{(n)} \]  

\[ (6.1) \beta_{k,m}^{n} = \{0,1\}, \forall k, m, n \]  
\[ (6.2) \sum_{k=1}^{K+1} \sum_{m=1}^{M+1} \beta_{k,m}^{n} = 1, \forall n \]  
\[ (6.3) \sum_{k=1}^{K+1} \sum_{n=1}^{N} \beta_{k,m}^{n} = 1, \forall m \]  
\[ (6.4) \sum_{m=1}^{M+1} \sum_{n=1}^{N} \beta_{k,m}^{n} = 1, \forall n \]  

Where \( \beta_{k,m}^{n} \) means that the kth pair of DU can reuse the nth channel resource shared by the CU, and \( 1 \leq n \leq N, 1 \leq m \leq M + 1, 1 \leq k \leq K + 1 \). When allocating resources in the actual communication system, \( N, M, K \) need to meet \( K < M < N < K + M \). Part of the DU is allocated as a dedicated channel resource, so the DU will not be interfered with by other users, and does not interfere with the CU. The other part of the DU needs to reuse CU resources, and reasonable resource allocation needs to be made to make the access system with as little interference as possible and as many D2D users as possible. If \( n \leq N, m = M, k = K \), \( \beta_{k,m}^{n} = 1 \) means that DU and CU may reuse channel n; if \( n \leq N, m = m + 1, k \leq K \), \( \beta_{k,m}^{n} = 1 \) means that the DU uses a dedicated channel n; if \( n \leq N, m = M, k = K + 1 \), \( \beta_{k,m}^{n} = 1 \) means that the CU uses a dedicated channel n; if \( k = K + 1, m = M + 1, n \leq N \), \( \beta_{k,m}^{n} = 1 \) means that dedicated channel n is idle.

4 Resource Allocation Algorithm

4.1 Algorithm Analysis

The algorithm proposed in this paper is to ensure the communication quality of the system and maximize the communication capacity of the system, and then carry out a reasonable allocation of user communication resources so that the performance of the communication system reaches the optimal and D2D users access the system most.
The algorithm steps are as follows: First, construct an interference matrix, and again based on the combination of cellular users, D2D users, and channels in the matrix, and find the smallest interference, finally, resource allocation is based on this data. Because there are \( m \) free channels left in the system, when allocating the free channels to users, only specific cellular users need to be determined to be allocated to specific D2D users for multiplexing, while the free channels are allocated to the remaining D2D users. Suppose all communication channels are the same. In order to ensure the normal communication quality, the interference threshold value of the received information of the base station is preset as \( I_0 \).

(1) If \( I_{k,m}^{(n)} \leq I_0 \), means that when the \( k \)th pair of D2D users can reuse the \( n \)th channel resource of the \( m \)th cellular user and the interference value is less than the interference threshold received by the base station, access to the D2D pair of users is allowed. The access admission parameter is \( \theta_{k,m}^{(n)} = 1 \); if \( I_{k,m}^{(n)} > I_0 \), means that the \( k \)th pair of D2D users cannot reuse the \( n \)th channel of the \( m \)th cellular user, then set the admission parameter as \( \theta_{k,m}^{(n)} = 0 \).

(2) If D2D uses a dedicated channel \( n \), it can be considered that the D2D user reuses the \( n \) channel used by a cellular user with zero interference, so we assume \( I_{k,m+1}^{(n)} = 0 \). If the dedicated channel is used by a cellular user, it is considered that the cellular user and a D2D user with zero interference to the base station multiplex the \( n \)th channel, so we assume \( I_{k,m+1}^{(n)} = 0 \). In this way, we can calculate an admission matrix of \( N \times (m + 1) \times (k + 1) \), which is \( \theta_{k+1,m+1}^{(n)} \), and an interference matrix of \( N \times (m + 1) \times (k + 1) \) is \( I_{m+1,k+1}^{(n)} \).

(3) From equation (4), we can know that the problem of maximum capacity is further transformed into the problem of minimum interference, and then it is transformed into the 0-1 assignment problem of the interference matrix for resource reuse. Therefore, in the channel resource reuse allocation, the obtained interference matrix is applied to the Hungarian algorithm to find the optimal resource reuse strategy\(^{[13]} \).

(4) In the dedicated channel allocation, it is only necessary to find the minimum interference value in the interference matrix \( I_{K\times M} \), and then directly assign an uplink channel to the user for D2D.

4.2 Simulation Parameters

In the simulation, the communication cell is a circular cell with a radius of 500 meters. A base station exists in the center. D2D user pairs and cellular users are randomly and uniformly distributed in the cell. Each cellular user can only occupy one cellular channel, and each cellular channel is only allocated to one cellular user or one
D2D pair user or one D2D pair user and one cellular user. Compared with the heuristic algorithm and the random algorithm, the algorithm proposed in this paper takes the interference threshold of the base station and the distance between the D2D transceiver users as variables, and its system allows the number of D2D users to access to increase significantly. Due to the random algorithm resource allocation, the D2D users in the system select the channel resource multiplexing randomly, without considering the interference between users, so the communication quality of D2D users cannot be guaranteed. But the algorithm improved in this paper is to establish the objective function of minimum interference, and use the Hungarian algorithm to achieve the optimization of this objective function, so that the D2D receiving end of the system is subject to the best pairing strategy with the least total interference, thus bringing the number of D2D users increased access systems. By using MATLAB for simulation comparison, data is randomly generated each time and the average value is obtained. The main system simulation parameters are shown in Table 1.

### Table 1. System simulation parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell radius/m</td>
<td>500</td>
</tr>
<tr>
<td>Total system bandwidth/MHz</td>
<td>10</td>
</tr>
<tr>
<td>Path loss model for D2D link/dB</td>
<td>$148+40\log d$</td>
</tr>
<tr>
<td>Path loss model for cellular link/dB</td>
<td>$128.1+37.6\log d$</td>
</tr>
<tr>
<td>D2D link shadow fading standard deviation/dB</td>
<td>12</td>
</tr>
<tr>
<td>Cellular link shadow fading standard deviation/dB</td>
<td>10</td>
</tr>
<tr>
<td>Noise spectral density/(dBm/Hz)</td>
<td>-174</td>
</tr>
<tr>
<td>D2D receiver SINR threshold/dB</td>
<td>-10</td>
</tr>
</tbody>
</table>

### 4.3 Analysis of Simulation Results

Figure 2 shows the relationship between the interference threshold of the base station and the logarithm of system access D2D under different algorithms. It can be seen from the figure that when the interference threshold of the base station is increasing, the D2D allowed by the system to access user communication is also increasing. When the interference threshold of the base station is about -70dBm, the D2D logarithm of the three algorithms can reach the maximum allowable access value of the system. It can be seen from Figure 2 that the algorithm and the heuristic algorithm have a smaller difference between the number of D2D users allowed by the system when the interference threshold of the base station increases. So the system allows fewer D2D users to access. However, compared with the comparison algorithm, when the interference threshold of the base station is between -150dbm to 120dbm, the number of users allowed to access D2D increases faster with the increase of the threshold.
Fig. 2. The relationship between the interference threshold of the base station and the log of D2D

Fig. 3. The relationship between the distance between D2D users and the logarithm of users accessing D2D.

Figure 3 shows the maximum distance between the users of D2D receiving and sending end between 20m ~ 140m and the logarithm relationship between the users of D2D allowed to access the system. It can be seen that in the process of increasing the maximum distance, the D2D allowed to access keeps more access to the number of users. As can be seen from Figure 3, as the distance between the users of the D2D transceiver increases, the number of D2D users allowed to access the system gradually decreases, which is caused by the more interference the more D2D users. Moreover, the system of the algorithm in this paper allows the number of D2D users to ac-
cess to decrease significantly less than heuristic and random algorithms. When the maximum distance between D2D transceiver users is 20m ~ 100m, the algorithm and heuristic algorithm in this paper can maintain a higher number of D2D users accessing the system, but as the maximum distance between D2D transceiver users increases, the interference gradually increases. When the maximum distance between D2D transceiver users is between 100m and 140m, the resource allocation algorithm based on the objective function of minimum interference in this paper makes the number of D2D users allowed to be accessed by the system significantly higher than its comparison algorithm.

![Graph](image)

**Fig. 4.** The relationship between the distance between users of D2D and the average interference caused by each pair of D2D.

Figure 4 shows the relationship between the maximum distance between the users of D2D receiving and sending end and the average interference value caused by each pair of D2D users. The interference value of each pair of D2D users increases with the increase of the maximum distance of D2D receiver and transmitter users. With the increase of the distance between the receiver and the transmitter, the transmitting power of D2D to the user's transmitter is also increasing, so the interference will also increase. When the maximum distance between the transmitter and receiver of D2D is 30m ~ 60m, the interference caused by the improved algorithm is obviously lower than other algorithms.

5 Conclusion

This paper studies the resource allocation problem of multiple D2D and multiple uplink cellular users in a single cellular system by establishing the communication mod-
el of D2D communication technology introduced into the cellular network, and carries out theoretical design and simulation analysis, so as to obtain the resource allocation scheme with the minimum interference as the target. This model takes the interference threshold of the base station and the distance between D2D communication pairs as variables to analyze, and optimizes the resource allocation process of D2D communication. The simulation results show that under the condition of limited communication resources and guaranteed user service quality, the algorithm maximizes the number of users of D2D access to the maximum extent allowed by the system, and reduces the interference of D2D to users of cellular users within a certain range. However, this system model only studies the interference of a single cell and does not consider the interference of neighboring cells, so the resource allocation of the multi-cell system model can be considered on this basis.

References


Research on the performance of multi-user detection algorithm based on serial and threshold in the IoT

Jingqiu Ren, Jianpeng Zhang, Zonglin Gu, Guanghua Zhang, Weidang Lu
1 Northeast Petroleum University, Daqing 163318, China; 2 Zhejiang University of Technology, Hangzhou 310014, China; dqzgh@epu.edu.cn

Abstract. In the 5G system for the Internet of Things (IoT), the detection and identification of a large number of IoT users or devices has become a key issue in the research of IoT technology. In the multi-user detection algorithm based on 5G technology, due to the slow convergence speed and high complexity of the traditional algorithm, people began to consider the conditions of adding serial and threshold to study, in order to improve the performance of multi-user detection in the IoT based on 5G. In this paper, based on the research of Maximum Logarithm Message Passing Algorithm (Max-log-MPA) based on threshold, the serial condition setting is added to analyze the performance of Max-log-MPA algorithm based on serial and threshold (S-T-Max-log-MPA) and Max-log-MPA algorithm based on threshold (T-Max-log-MPA), so as to obtain a better multi-user detection algorithm based on 5G in the Internet of things. The simulation results show that the S-T-Max-log-MPA algorithm has lower bit error rate and better performance in the experiment, which is more suitable for multi-user detection in the Internet of things based on 5G.

Keywords: Internet of things, sparse code multiple access technology, MPA algorithm, bit error rate, serial

1 Introduction

In recent years, with the rapid development of big data, cloud computing, smart city, etc., the network is facing the growing demand of high capacity, high transmission rate and low delay, and the Internet of Things (IoT) industry has entered a stage of rapid development. Especially with the rapid development of 5G technologies, the application of the IoT breaks the established concept of people to people interconnection of the existing mobile network, and realizes the real IoT, people to people and people to things, bringing great convenience to people's production and life. In the application of IoT based on 5G, the number of users is usually large. How to use multi-user detection to solve such problems as reducing the cost of machine terminals, resource allocation and low-cost IoT terminal coverage has become an important research issue.
Sparse Code Multiple Access (SCMA) is one of the core technologies of 5G. It is a new orthogonal multiple access technology using sparse codebook, and it is the extension of Low Density Signal (LDS) \[4\] \[5\]. When detecting the accuracy of data and dealing with multi-user access problems, SCMA multiple access technology realizes the capacity of more users under the same number of resources, increases the overall throughput of the network, and realizes massive connections through code field sparse expansion and non-orthogonal superposition. In addition, LDS integrates the modulation idea of sparse spread spectrum and high-dimensional modulation technology, mapping the bit data stream in the link to the multi-dimensional codeword in the preset codebook, to solve the system overload problem caused by a large number of data connections \[6\]. However, in the existing SCMA system, when the Message Passing Algorithm (MPA) performs multi-user detection of the received signal \[7\], the calculation complexity of the algorithm is too high and the transmission accuracy is poor, so it is difficult to apply it to the actual decoding process \[8\]. Therefore, the accuracy of multi-user detection algorithm in the transmission of information is still one of the important challenges in the multi-user detection of the IoT \[9\]. In reference \[10\], it is described that in the context of IoT communication, compressed sensing is used for multi-user detection. Many studies assume that the channel information is known, and in practical applications, channel estimation is also needed which improves the complexity of the system. The Maximum logarithm Message Passing Algorithm (Max-log-MPA) based on log domain and the algorithm based on serial MPA (S-MPA) are proposed in reference\[11\] \[12\], it reduces the computational complexity, but the system loss is large. In reference \[13\], an improved multi-user detection algorithm based on partial edge is proposed. Although the complexity of the algorithm is reduced, the BER performance is also reduced. In reference \[14\], an improved serial scheduling based MPA (ISS-MPA) detection scheme is proposed. Although it can maintain good bit error rate performance, the complexity effect of the detection algorithm is not particularly good. In reference \[15\], a threshold based MPA algorithm is proposed, which uses a threshold to control the algorithm, so as to reduce the complexity of detection algorithm. In reference \[16\], an algorithm is proposed to reduce the decoding complexity of SCMA by introducing a weighting factor instead of the iterative process of MPA algorithm. In reference \[17\], a threshold-based Max-log-MPA low complexity multiuser detection algorithm is proposed. The influence of the change of signal-to-noise ratio on complexity is analyzed to reduce the complexity of multiuser detection better.

Among the above algorithms, a new method is proposed to compare with the traditional MPA algorithm, so as to prove the superior performance of the proposed algorithm and analyze the performance of each algorithm. Therefore, based on the research of Max-log-MPA algorithm based on threshold, this paper adds serial condition setting to analyze the performance of the Max-log-MPA algorithm based on serial and threshold (S-T-Max-log-MPA) and the Max-log MPA algorithm based on threshold (T-Max-log-MPA), so as to select an optimal multi-user detection algorithm to solve the problem of multi-user detection in the application of IoT based on 5G.
2 SCMA system model

Sparse code multiple access (SCMA) is a non-orthogonal access method based on code domain [18]. At the transmitter, each user has its own codebook [19], and the transmitter consists of two parts: multidimensional modulation and sparse spread spectrum [20-22]. Sparse spread spectrum plays an important role in transmission. The information bits of different users are mapped into multi-dimensional codewords after channel coding, and then are non-orthogonal stacked on the same resource block by sparse spread spectrum.

The SCMA uplink model consists of transmitter, transmission channel and receiver, as shown in Figure 1. It can be seen that in the multi-user SCMA system, different binary bitstreams are mapped to the codewords of SCMA multi-dimensional codebook by different users through different codebooks. Assuming that the number of users and codebook of IoT is J, and the codebook length is K (J > k), then the users of IoT $\{u_1, u_2, \ldots, u_J\}$ get binary bit data stream $b_j(b_1, b_2, \ldots, b_J)$ after source coding and channel coding. When entering the SCMA encoder transmission, the j-th user’s bit stream is mapped to the orthogonal subcarrier through the codebook. The mapping process can be expressed as $f: B^{J\log_2M} \rightarrow \chi^{[23][24]}$, B represents the set of binary numbers, and $\chi$ is the user’s codebook. Because the channels of each layer of the upper link are different, the channel factor is different. Suppose the channel factor is $h_j(h_1, h_2, \ldots, h_J)$.

- $b_1$: SCMA encoder
- $b_2$: SCMA encoder
- $b_j$: SCMA encoder
- AWGN
- MPA Multiuser detection
- $b_1(1)$
- $b_2(1)$
- $b_j(1)$

Figure 1 uplink SCMA communication system model

$M$ in $B^{J\log_2M}$ represents the size of codebook. Assuming that the user node data is divided into several groups according to a-bit, the size M of codebook is:

$$M = 2^a$$  \hspace{1cm} (1)

First, define an overload factor $\lambda = \frac{J}{K}$, assume that there are 6 user nodes and 4 time-frequency resource blocks, as shown in Figure 2. It can be seen that at this time, the overload factor $\lambda = 1.5$, in other words, the system has 150% overload capacity. As shown in Figure 3, user1, …, user6 represents six different user nodes. The bit information of these six user nodes is mapped to the codewords of different codebooks, and each user node has only one codebook corresponding to it.
From figure 2 and Figure 3, it can be seen that the area where element 0 is located is a white area, and the area where element 1 is located is a colorful area, it means that in user1, orthogonal time-frequency 1, 3 resources transmit signals, 2, 4 resources do not transmit signals, and so on, the following matrix can be obtained:

\[
F_{4 \times 6} = \begin{bmatrix}
0 & 1 & 1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 \\
1 & 0 & 0 & 1 & 1 & 0
\end{bmatrix}
\] (2)

Assuming that the time of each IoT user is synchronous, the signal received by the base station is the weighting of all signals:

\[
y = \sum_{j=1}^{6} diag(h_j)x_j + n
\] (3)

Where, \(x_j = [x_{1,j}, x_{2,j}, \ldots, x_{K,j}]^T\) is the codeword sent by user j. \(h_j=[h_1, h_2, \ldots, h_K]^T\) represents the receiver channel vector, which is the channel gain matrix. \(n\) represents the Gaussian white noise in the transmission channel \(n \sim CN(0, \sigma^2 I)\). In this case, the signal received by the time-frequency resource at K can be expressed as:
Because of its sparsity, the code word conflict at time-frequency resource $K$ will be greatly reduced [25].

3 T-Max-log-MPA algorithm

The T-Max-log-MPA algorithm is based on the Threshold based MPA algorithm (T-MPA), adding the judgment of the necessary conditions for the stability of user nodes. Before updating the message, first judge whether the user information nodes meet the necessary conditions for the stability of user nodes, and then judge whether they pass the threshold conditions. Only when users who meet the threshold conditions and pass the necessary condition of user node stability can be decoded in advance.

In the SCMA iteration process, if the user node $u_j$ in the factor graph is in the same position in the $i$-th cycle iteration process, that is, the $i$-th iteration is equal to the result of the $i'$-th iteration [26]:

$$
\arg \max_{1 \leq m \leq M} q'(x_{i, m}) = \arg \max_{1 \leq m \leq M} q'(x_{i', m}), i < i' < I_{\text{max}}.
$$

This indicates that the user $u_j$ node is stable. Therefore, in the $i$-th iteration and the $(i+1)$-th iteration, the same position of the largest element in the codeword credibility vector is a necessary condition for the stability of the user node,

$$
\arg \max_{1 \leq m \leq M} q'(x_{i, m}) = \arg \max_{1 \leq m \leq M} q'^{i+1}(x_{i', m}).
$$

The algorithm can be divided into three steps:

Step 1: receive condition initialization:

$$
I_{\xi_k \to u_j}^0(x_j) = \frac{1}{M}
$$

(4)

$$
I_{\xi_k \to u_j}^i(x_j) = \sum_{x_{\xi_k}} \left\{ \frac{1}{\sqrt{2\pi}\delta} \exp \left( -\frac{1}{2\delta^2} \left\| y_k - \sum_{k':i} \sum_{x_{k'}} \frac{i}{x_{\xi_k}} \prod_{m \neq i} I_{\xi_m \to u_k}^{i-1} (x_{j}) \right\| \right) \right\}
$$

(5)

Where, $i$ is the number of iterations, $\xi_k$ and $\zeta_k$ represents the non-zero position set of row $k$ and column $j$ in the $F$ matrix respectively.

Step 2: update the asset node:

$$
I_{u_j \to \xi_k}^i(x_j) = \prod_{m \neq j} I_{u_j \to \xi_k}^i(x_j)
$$

(6)

Step 3: when formula (5) and formula (6) reach the maximum number of iterations, the output probability after MPA decoding $Q(x_j)$ is:

$$
Q(x_j) = \prod_{k \in \xi_j} I_{\xi_k \to u_j}^{I_{\text{max}}} (x_j)
$$

(7)

By combining formula (4) - (7), we can get:
\[ I'_{c_j \rightarrow a_{u_j}}(x_j) = \frac{1}{\sqrt{2\pi\delta}} \times \max_{t=1,2,\ldots,N} \left\{ \frac{1}{2\delta^2} \left\| y_k - \sum_{v \in c_j} h_{k,v} x_k,v \right\|^2 + \sum_{v \in c_j} I'_{a_{u_j} \rightarrow v}(x_j) \right\} \quad (8) \]

\[ I'_{a_{u_j} \rightarrow v}(x_j) = \sum_{v \in c_j \cap m} I'_{a_{u_j} \rightarrow v}(x_j) \quad (9) \]

Where, \( t \) still represents the number of iterations. When the algorithm passes the maximum number of iterations, the output probability of each user's codeword is:

\[ Q(x_j) = \sum_{v \in c_j} I'_{a_{u_j} \rightarrow v}(x_j) \quad (10) \]

### 4 S-T-Max-log-MPA algorithm

Based on the Max-log-MPA algorithm, the S-T-Max-log-MPA algorithm introduces the serial update algorithm and threshold MPA algorithm, in the Max-log-MPA algorithm, index (EXP) algorithm is changed into sum algorithm and maximum value. In the serial update algorithm, user node message update is integrated into resource node information update to reduce the complexity of information storage. In the threshold MPA algorithm, hard decision is used to effectively reduce the user node information that needs to be updated in each cycle. This algorithm combines the advantages of the three, can effectively reduce the complexity of the detection algorithm while maintaining a good bit error rate.

Because the S-T-Max-log-MPA algorithm first judges the stability of IoT users in the iterative update process, the formula (5) of the resource node update process of the algorithm is modified as follows:

\[ I'_{a_{u_j} \rightarrow v}(x_j) = 2 \times \frac{1}{\sqrt{2\pi\delta}} \times \max_{s_j} \left\{ \frac{1}{2\delta^2} \left\| y_k - \sum_{v \in c_j} h_{k,v} x_k,v \right\|^2 \times \prod_{m \in c_j \cap j} I'_{a_{u_j} \rightarrow m}(x_j) \right\} \quad (11) \]

After that, the log likelihood ratio (LLR) of each user's coding bit determines the user:

\[ Q(x_j) = ap_v(x_j) \times \prod_{m \in c_j \cup j} I_{a_{u_j} \rightarrow m}(x_j) \quad (12) \]

Where, \( ap_v(x_j) \) represents the prior probability of the user J code word.

\[ \text{LLR}_{j,s} = \log \left( \frac{P(b_i = 0)}{P(b_i = 1)} \right) = \log \left( \frac{\sum_{m \in c_j \cap s} Q(x_j)}{\sum_{m \in c_j \cap s} Q(x_j)} \right) \quad (13) \]

Where, \( \text{LLR}_{j,s} \) is the log likelihood ratio, \( \sum_{m \in c_j \cap s} Q(x_j) \) represents the output probability of the decoded variable node, \( \sum_{m \in c_j \cap s} Q(x_j) \) represents the output probability of the variable node to be decoded. \( P(b_i = 0) \) represents the probability of the decoded
variable node VN, and \( P(b_i = 1) \) represents the probability of the variable node VN to be decoded.

5 Analysis of BER performance

In the SCMA system, the T-Max-log-MPA algorithm combines the codeword credibility and user node stability, eliminates the exponential operation, and increases the judgment of the necessary conditions for the user node stability. Before the message is updated, the stability and threshold conditions of the user node are judged successively, which can not only increase the reliability of the decision codeword, but also reduce the loss of posterior soft information during the transmission process, which improves the accuracy of message delivery and BER performance. The S-T-Max-log-MPA algorithm adds the serial mode to the T-Max-log-MPA algorithm. Because of the serial mode, the asynchronous mechanism is also added to the algorithm. Because of the existence of the asynchronous mechanism, in each iteration, all resource nodes can process and deliver messages at the same time, and the received messages can be delivered in a timely manner, which accelerates the convergence performance of the algorithm, and achieves the desired effect when the number of iterations is small, and further improve BER performance.

6 Analysis of simulation results

![Figure 4 BER performance comparison between T-Max-log-MPA algorithm and S-T-Max-log-MPA algorithm when \( T_{\text{max}} = 2 \)](image)

Figure 4 shows the BER performance comparison between T-Max-log-MPA algorithm and S-T-Max-log-MPA algorithm when \( T_{\text{max}} = 2 \). It can be seen from Figure 4 that when \( T_{\text{max}} = 2 \), the overall BER performance of S-T-Max-log-MPA algorithm is...
higher than that of T-Max-log-MPA algorithm within the threshold value range of $0 \leq \frac{E_b}{N_0} \leq 14\text{dB}$, and the smaller the threshold, the more obvious the change, and the greater the BER performance difference. Comparing the BER performance of S-T-Max-log-MPA and T-Max-log-MPA with the change of signal-to-noise ratio, we can see that when $\frac{E_b}{N_0} = 0$, if the threshold value $th = 0.01$, the BER of them are 20.55% and 26.83% respectively, and the BER of S-T-Max-log-MPA is 6.28% lower than that of T-Max-log-MPA. If the threshold value $th = 0.10$, the BER performance of the two algorithms are 21.08%, 28.13% respectively, and the BER performance of S-T-Max-log-MPA is 7.05% lower than that of T-Max-log-MPA. When the threshold value $th = 0.6$, the BER performance of the two algorithms is 24.7%, 29.48% respectively, which is lower than 4.78% of T-Max-log-MPA algorithm. When $\frac{E_b}{N_0} = 14$, calculate the BER when the thresholds are $th = 0.01$, $th = 0.1$, $th = 0.6$ respectively. The BER performance of the S-T-Max-log-MPA algorithm under the above three thresholds are: 2.033%, 2.177%, 5.417%, T-Max-log-MPA algorithm are 5.85%, 5.85%, 6.367%, and their bit error rates differ by 3.817%, 3.673%, 0.959%. It can be seen from the above that under the same threshold, with the increase of $\frac{E_b}{N_0}$, the smaller the bit error rate is, the better the optimization effect is. Therefore, when $T_{\text{max}}=2$, the performance of S-T-Max-log-MPA is better than that of T-Max-log-MPA.

Figure 5 shows the performance comparison between S-T-Max-log-MPA algorithm and T-Max-log-MPA algorithm when $T_{\text{max}}=3$. When $\frac{E_b}{N_0} = 0$, S-T-Max-log-MPA algorithm at the threshold value $th = 0.01$, 0.1, 0.6, the corresponding BER performance is 18.02%, 20.03%, 24.58%, T-Max-log-MPA algorithm corresponding values are 24.38%, 25.32%, 25.78%, and the BER performance differs by 6.36%, 5.29%, and 1.2%. The performance of S-T-Max-log-MPA algorithm is best. When $\frac{E_b}{N_0} = 14$, S-T-Max-log-MPA algorithm at the threshold value $th = 0.01$, 0.1, 0.6, the corresponding BER performance is 1.383%, 2.133%, 3.9%, and T-Max-log-MPA algorithm corresponding values are 2.917%, 3.367%, 3.683%. It can be seen from the
results that when the threshold is less than 0.6, the performance of S-T-Max-log-MPA is better than that of T-Max-log-MPA. Therefore, when $T_{\text{max}}=3$, the threshold value is less than 0.6, S-T-Max-log-MPA performance is better than T-Max-log-MPA.

Figure 6 shows that when $T_{\text{max}}=5$ and the BER performance of S-T-Max-log-MPA algorithm and T-Max-log-MPA algorithm under the same threshold. As can be seen from the Figure 6, the BER performance of the S-T-Max-log-MPA algorithm is the best when the $E_b/N_0$ changes from 1 to 14 and $t_h$ is less than 0.6. However, when $t_h=0.6$, the BER of S-T-Max-log-MPA algorithm is worse than T-Max-log-MPA algorithm, which indirectly shows that the best threshold value is less than 0.6 in the calculation process of the algorithm. Therefore, when selecting the optimal result, the part with threshold value $t_h<0.6$ is taken.

7 Conclusion

Based on the research of 5G based multi-user detection method in the IoT, this paper in-depth the S-T-Max-log-MPA algorithm and the T-Max-log-MPA algorithm. Through the comparison of experimental simulation, when the maximum number of iterations is 2, 3 and 5, when the threshold value is less than 0.6, with the change of $E_b/N_0$ of signal-to-noise ratio, the BER performance of S-T-Max-log-MPA algorithm is better than that of T-Max-log-MPA algorithm, and the smaller the threshold value, the more obvious the curve change, the more significant the BER performance improvement. It can be concluded that BER performance of S-T-Max-log-MPA algorithm is better than that of T-Max-log-MPA algorithm. The former algorithm can better solve the problem of poor BER performance of threshold messaging algorithm at low threshold, and can be better applied to multi-user detection based on 5G IoT.
Funding: This work was supported by the Youth Science Foundation of Northeast Petroleum University under Grant No. 2019QNL-34 and Graduate Innovative Research Work and the National Natural Science Foundation of China under Grant (61871348).

Conflicts of Interest: The authors declare no conflict of interest.

References


11. Song Chunxue, Wen Ping, Zhang Xuechen. FPGA implementation of sparse code multiple access system based on 5g wireless communication [J]. Application of electronic technology, 2016, 42 (7): 812.


23. Donghong Cai; Pingzhi Fan; Xianfu Lei; Yingjie Liu; Dageng Chen. Multi-Dimensional SCMA Codebook Design Based on Constellation Rotation and Interleaving, 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), Nanjing, China, 15-18 May 2016.


Support Precise Latency for Network Based AR/VR Applications with New IP

Lijun Dong, Lin Han, Richard Li
{liju.dong, lin.han, richard.li}@futurewei.com

Futurewei Technologies Inc., 2330 Central Expressway, Santa Clara, CA, USA, 95050

Abstract. The emerging Augmented Reality (AR), Virtual Reality (VR) and Holographic applications have brought up a lot of challenges to technologies such as information display, image processing, fast computing and networking. This paper gives a quantified analysis on the latency requirement that AR and VR impose to networking. Most importantly, the paper discusses on how to use New IP, an advanced data packet framework to support the precise low latency requirement.

Keywords: AR, VR, networking, latency budget, precise latency constraint, New IP, contract spec, metadata, holographic application, hologram.

1 Introduction

The multimedia has been evolved from the 2-D audio/video in 4K/8K resolution to AR/VR video. Augmented Reality (AR) is a live direct or indirect view of a physical, real-world environment whose elements are augmented by computer-generated sensory input such as sound, video, graphics or GPS data. Virtual Reality (VR) is a computer technology that uses software-generated realistic images, sounds and other sensations to replicate a real environment or an imaginary setting, and simulates a user's physical presence in this environment to enable the user to interact with this space. A VR viewport delimits the scene horizontally from the viewport center, whose angle can normally range 120 degrees. In order to ensure good immersion, a displayed viewpoint’s pixels need to at least have 4K resolution. Therefore, the resolution of the full 360-degree scene is at least 12K.

AR/VR technologies have enormous potential in many different fields, such as entertainment, remote diagnosis, and remote maintenance, etc. Holograms, haptics, and other sensory data will further provide immersive and feel-like-real user experience, with which the real and virtual world to the users will be extensively blurred. In a hologram, the same object or scene is captured from different angles. A different image depicting the same object or scene will be seen by a viewer from different viewpoints, depending on the relative position of the reviewer’s eyes.

Powerful cloud capabilities have improved the VR user experience and reduced device cost, promoting the evolution of VR from local to cloud-based VR. Cloud VR can make full use of the distributed computing capabilities of many-core servers, Graphics Processing Unit (GPU) clusters, as well as latest rendering and artificial intelligence (AI) technologies. Figure 1. shows an example of cloud VR - virtual concert. In the virtual concert, the musicians could be performing in different places in the world, while the audience could be sitting on the beach enjoying the concert as if he is present in front of the stage with those musicians’ 3D projections. The holographic data is represented through use of point clouds consisting of volumetric data in
a conceptual three-dimensional box. The large volumetric data needs to be streamed through the Internet to the end user such that the rendering of the interested object/scene’s image from any 360-degree viewing angle can be achieved. We call the AR/VR applications that involve large volumetric data streaming as network-based AR/VR in the following sections of the paper.

"VR sickness" [1] already shows in AR/VR applications if the information about the virtual environment received by the human brain is not always consistent due to data transmission lagging. In the network-based AR/VR, sufficient realism inevitably requires both extremely low latency and high bit rate.

The paper focuses on providing a quantified analysis on transport latency budget in supporting network-based AR/VR applications. Most importantly, in the paper the authors discuss on how to achieve precise latency budget by leveraging a novel and thriving data packet framework called New IP.

2 Latency Budget

Latency is the most important performance metric for AR/VR, holographic streaming. Motion to Photon (MTP) is defined as the time needed for user movement to be fully reflected on a display screen. A high MTP latency causes sickness and nausea. When a user wearing VR headset makes a movement, the mind expects the display on the screen is also updated promptly and appropriately to reflect the movement. When MTP latency is high, the display also fails to show the user’s movement, then the user can be disoriented and feel sick, resulting in very poor and intolerable VR experience. It often considers that the MTP latency less than 20 ms is necessary to convince your brain that you are presented in an augmented or simulated world. Some research even shows that the MTP latency must be smaller than 17ms [2] for sensitive users. Latency greater than 20 ms not only degrades the visual experience, but also tends to result in VR sickness, which is also known as cyber-sickness. It is caused by a sensory mismatch or
conflict to the signals the balance system is sending to the brain. Taking the analogy to riding in a car, the vestibular system tells the brain that you are moving, but the proprioceptive system indicates that you are sitting still. You may experience car sickness if your visual system may be getting different signals depending on what you are looking at, and confuses your brain about your movement since you cannot predict exactly when you are going to slow down, speed up or turn as a passenger. VR sickness can be minimized by keeping MTP latency below the threshold, above which humans can detect the lag between the visual input and self-movement.

Table 1. Current and projected latency in network based AR/VR

<table>
<thead>
<tr>
<th>Latency</th>
<th>Current value (ms)</th>
<th>Projected value (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>T2</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>T3</td>
<td>110 to 1000</td>
<td>5</td>
</tr>
<tr>
<td>T4</td>
<td>0.2 to 100</td>
<td>?</td>
</tr>
<tr>
<td>T5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>T6</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>MTP</td>
<td>130 to 1180</td>
<td>13+?</td>
</tr>
</tbody>
</table>

The network-based AR/VR involves many more factors from one end to another end. The major components of E2E delay in network-based AR/VR include: T1: Sensor detection and action capture; T2: Computation for region of interest (ROI) processing, rendering and encoding; T3: Group of Pictures (GOP) [3] framing and streaming; T4: Network transport; T5: Terminal decoding; T6: Screen refresh.

Table 1. shows the current and projected values for each component of end-to-end latency. If we anticipate that the technology development and advancement would bring down the latency of some components, such as reducing the latency caused by ROI processing, rendering and encoding (T2) to 2 ms, GOP framing and streaming (T3) to 5ms by using improved parallel hardware processing, and screen refresh latency (T6) to 0.01ms by using OLED, etc., then the budget for the round trip network transport delay (T4) will be around 5 to 7ms.

We can see that MTP latency is currently much greater than 7ms. The network transport latency is comprised of physical propagation delay and switching/forwarding delay at each network device.

- The physical propagation delay: This is the delay caused by the speed limit of signal transmitting in physical media. Taking the fiber as an example, the optical transmission rate cannot exceed the light speed, i.e. 300km/ms in free space. However, light travels slower the fiber optic core because the refractive indexes of light are different in free space and in the glass. In normal optical fiber, the light speed is about 200km/ms [4]. In order to reduce the physical propagation delay, the physical distance between user and AR/VR server needs to be limited. The deployment of AR/VR server should be close to user as much as possible.

- The switching/forwarding delay: This delay normally is much more than the physical propagation delay, which can vary from 200us to 200ms at each hop.

We cannot expand the physical scope of an AR/VR application beyond the speed-of-light limit. However, we can ensure that application processing and transport related latencies do not significantly reduce this limited scope. As a rule of thumb, they should consume no more than 5-10% (1-2ms) of this 20ms budget, and preferably less.
3 Support Precise Latency Budget by New IP

3.1 What is New IP?

New Internet Protocol (New IP) [5][6][8] has been proposed to overcome the three major issues that are caused by the fixed structure of the IP packets in the current Internet, e.g., statistical multiplexing, best-effort paradigm, and an IP address-based reachability. New IP is a data plane technology that defines a new network datagram format, its specification, and corresponding capabilities in the network nodes. The New IP datagram format is shown in Figure 2, which includes 3 components, namely, a) addressing evolution, b) the contract inclusion, and c) the payload extension.

![Diagram of New IP datagram format]

The New IP Address (Shipping Spec) evolution aims to replace the current fixed type of addressing in order to provide flexibility to include all types of addresses and fit different reachability scenarios. The New IP shipping specification is backward compatible with the existing address schemes (e.g., IPv4 and IPv6).

The New IP Contract (Contract Spec) inclusion provides a series of apparatus to enable a large selection of network capabilities, their functioning and regulative control at the finest packet-level granularity. Contracts carry specific requirements and parameters associated with time-engineered services for media services as we discussed in the previous section of this paper. Figure 3, shows the contract structure. Basically contract is composed of multiple contract clauses, each of which is a combination of Event, Condition, Action and Metadata. Action describes how New IP nodes should treat the packet when certain designated event/condition is met. The metadata is a set of parameters that are associated with the actions or applications.

The New IP Payload (Payload Spec) associates semantics to the data payload. New IP payload provides options to the receiver to consume any residual information in the payload while allowing the network to drop portions of the payload when congestion occurs. This type of communication is named as Qualitative Communication [7], which helps to mitigate retransmission overheads and delays when faced with slow or congested conditions.
3.2 How New IP Supports the Precise Latency Requirement?

The Contract Spec in New IP format is able to carry latency requirement to precisely support computational multiplexing approaches on the switches or routers, which provides much finer granularity of time assurance at packet level compared to the current statistical multiplexing.

For AR/VR or holographic data (later is called multimedia data in general), by inserting the precise latency requirement in the New IP contract clause (Metadata), the residual latency can be evaluated in a hop-by-hop basis by intermediate New IP node (switches or routers). We regard the packets with time constraints as latency-sensitive packets. A computational multiplexing scheduler at each New IP node is capable of determining the precise position in the outgoing queue of an output port for a latency-sensitive packet based on its latency constraint compared to other packets. With computational multiplexing, simultaneously arriving packets with latency budget that are intended to be forwarded on the same output port are scheduled based on their respective latency budgets. This capability is only possible when datagrams have flexible structure as designed in New IP to carry their latency budget because control plane methods are not designed to handle per packet requirements.

In order to ensure the extreme low latency with precise budget, two contract clauses are designed for the multimedia data packets. The first contract clause is shown in Figure 3, the action is Latencyguarantee, which instructs the intermediate New IP nodes to perform computational multiplexed scheduling of the packet. The metadata includes: (1) total budget is the end-to-end latency constraint between the time when the sender sends out the packet and the time when the packet reaches the receiver, which is set to be at most 2ms for cloud-based AR/VR applications; (2) residual budget is computed as the total budget subtracted by the elapsed time when the packet arrives at a New IP node; (3) residual hop indicates the number of hops between the current node to the receiver. One straightforward
scheduling algorithm is to place the latency-sensitive packets in one prioritized queue per output port, and schedule them according to average per-hop residual budget, which is defined as residual budget divided by residual hop. This scheduling algorithm makes sure the packet with the smallest latency budget gets transmitted ahead of the packets with larger budgets. However, it has a fundamental fairness issue, that the packets with larger per-hop residual budget are being “starved” at the end of the queue. The researches on this topic are highly encouraged.

**Contract Clause 2**

![Contract Clause 2](image)

The second contract clause is designed to instruct the intermediate New IP nodes to drop portions of packet payload when encountering network congestions instead of dropping the packet completely. The action is set to be PartialDrop. It is proposed in [10][11], the packet payload could be divided into multiple equally sized chunks, over which the random linear network coding is applied. When network congestion happens that requires dropping the latency-sensitive packet entirely (according to the current implementation in the Internet), the New IP node could remove/trim chunks from the tail of the packet payload as many as needed until the outgoing queue is able to retain the packet. In this way, the receiver is still able to obtain some parts of the packet payload, such that only the lost portions are retransmitted. In the contract clause 2, some metadata is also configured to assist PartialDrop execution, which includes: (1) NetworkCoded is used to indicate that the packet payload chunks are applied with random linear network coding; (2) Full DoF indicates the complete number of degrees of freedom in order to decode the payload data chunks; (3) Current DoF indicates the degree of freedom of the remaining chunks in the payload. When a chunk is removed from the payload, the current DoF is subtracted by 1; (4) Coefficients contains the coefficients for the remaining coded chunks in the payload. When a chunk is removed from the payload, the corresponding coefficient is also deleted from the Coefficients metadata.

4 Conclusion

The network-based AR/VR and future holographic applications impose many new challenges to the networking technologies, especially the transport layer. The paper analyzes the maximum value that the network transport latency could have. The paper discusses one possible solution to address the precise latency requirement, which leverages the flexible, programmable New IP data packet framework. The paper also outlines the research directions and opportunities when leveraging New IP framework.

References

A Novel Sample-Enhanced Dataset based on MFF for Large-Angle Face Recognition

He Wang¹, Yan Wang¹,*, Jie Liu¹, Guisheng Ying¹
{wanghe11@hrbeu.edu.cn, wangyanj@hrbeu.edu.cn*, liujie@hrbeu.edu.cn, yinguisheng@hrbeu.edu.cn }¹
¹ College of Computer Science and Technology, Harbin Engineering University, China

Abstract. Large-angle face recognition has always been a huge challenge due to the scarcity of large-angle dataset. In this paper, a novel sample-enhanced dataset is constructed, which is composed of various angle face picture samples from -90° to 90° relative to the front face. The constructed dataset is obtained by enhancing large-angle face samples of the CASIA-WebFace dataset. The large-angle face samples are generated from small-angle face samples of the CASIA-WebFace dataset, which is based on the multi-task feature framework (MFF). By employing these sample datasets, four trained FaceNets are achieved for face recognition. Finally, to test the effectiveness of the four face recognition networks for the large-angle face, 300 large-angle face pictures of different basketball players are selected as the samples of the experiment. The results demonstrate that the accuracy of large-angle face recognition has been greatly improved when utilizing the FaceNet that is trained by the novel enhanced dataset.

Keywords: Enhanced-dataset, large angle, MFF, face recognition.

1 Introduction

With the progress of technology, face recognition is greatly developed, which is applied everywhere in our daily life, such as the ticket checking system of railway station, the mobile phone face unlock system and so on. Most them are focused on small-angle face pictures. However, the angle of the face is various in different scenes. For the large-angle face pictures, the face recognition system may not work well. Yet, large-angle face recognition has always been a huge challenge due to the scarcity of large-angle dataset. In fact, large-angle face recognition plays a vital role in our daily life, such as identifying criminals in video surveillance. Therefore, constructing new large-angle face samples is particular important and essential, which will drive the progress of large-angle face.

With the development of neural networks, small-angle face recognition has been overcome. However, because it is difficult for neural networks to extract enough features for large-angle faces, high-angle face recognition has always been a huge challenge. At the same time, the published data sets such as CASIA-WebFace [1] and LFW [2] have rela-tively small angular deflection ranges, and there have been few reports of high-angle face recognition data sets in recent years, which has limited high-angle people Progress in face recognition. Although many people such as DR-GAN proposed by Liu et al. [3] and chu at al. [4] have proposed solutions for multi-pose face recognition, so far few people have tried solutions for large-angle face recognition from the perspective of datasets.

In this paper, a 3D face reconstruction method based on the MFF [5] framework is used to
generate a large-angle face sample set. It is the first time to use the MFF sample generation of 3D face reconstruction for the large-angle dataset application. MFF is a framework that can strongly constrain the pixel intensity and the 3DMM prior parameters to the BFM [6] face model. The overall of architecture of the paper is shown in figure 1. Firstly, we align the face samples to obtain Face landmarks, and then fit the Face landmarks to the 3DMM model to solve the shape, expression, pose, and lighting parameters required by the MFF algorithm. Finally, the face model reconstructed based on the MFF algorithm is subjected to depth estimation to complete three-dimensional reconstruction, and further generates face samples with larger angles. At last, based on the generated high-angle face sample set, the recognition of high-angle face images under the real environment of 15 basketball players is completed, which verifies the effectiveness of our sample generation method.

2 Method

In this section, the procedure of generating a multi-pose face dataset is introduced in detail, which includes face alignment, pose estimation, 3DMM coefficient solution and MFF frame generation of multi-pose faces.

2.1 Face alignment

Face alignment is one of essential and crucial work for face recognition. However, the work has become more difficult with the increasing of complexity of the background and the pose of face.

To imitate more realistic face recognition environment in our daily life, the CASIA-WebFace dataset [1] is employed in the experiment, which is obtained from a web crawler. The pictures from the CASIA-WebFace have more complicated lighting background to make it closer to the environment of face recognition in daily life. Based on the CASIA-WebFace dataset, the 3D face recognition method is utilized to generate a face recognition dataset with the performance of a large angle. Thus, the generated dataset can achieve the requirement of large angle and wild environment. For the face alignment of the pictures from the generated dataset that has more face poses and complicated environment, the face alignment network (FAN) method proposed by Adrian Bulat [7] is adopted, which is shown in figure 2. The adopted FAN method is constructed by stacking four Hour-Glass (HG) networks. The Hour-
Fig. 2. The Face Alignment Network (FAN) constructed by stacking four HGs in which all bottleneck blocks. This network uses RGB images as input and face landmarks as output.

Fig. 3. Face alignment results of CASIA-WebFace using FAN method

Glass (HG) networks [8] are based on one of the state-of-the-art architectures for human pose estimation. By doing the experiment, the final results are presented in figure 3. From figure 3, it can be concluded that FAN method is more robust for the complex background and lighting of the face in CASIA-WebFace and there are fewer failures.

2.2 Pose estimation

Pose estimation is an important part in the generation of large-angle face samples. The use of large-angle face samples for projection to retain more identity information is one of the characteristics of 3D face reconstruction. We use 5 landmarks (nose tip, chin, left eye, left eye, right eye, (Left mouth corner, right mouth corner) as 2D face key points, 3D face key points, which are set based on the the work in [9]. The pose estimation of face images in CASIA-WebFace is completed. Due to the good effect of face alignment in the section 2.1, the accuracy of the pose estimation is improved. The results are shown in figure 4.

From figure 4, we performed pose estimation on the face samples in CASIA-WebFace. Where figure 4(a), (b), (c), and (d) are the cases of 45°, -45°, 56°, and -57° around the Y axis. In most cases, pose estimation can be performed accurately, but a lot of face occlusions will reduce the accuracy of the pose estimation method.
2.3 3D face model fitting

We use the 3DMM fitting method to obtain the 3DMM prior parameters that are needed for the MFF algorithm. The 3DMM model is a universal 3D face model [10]. It can approach any real face shape under the constraints of shape and texture conditions. We refer to the work of Chu et al. [4] to solve 3DMM parameters.

Each 3D face shape (with \( n_{vert} \)) is represented by a \( 3n_{vert} \times 1 \) dimensional vector, which stands for (0-6) seven expressions, \( i \in (0, n_{id} - 1) \) for the index value of identity.

\[
S_{i,e} = (X_1, Y_1, Z_1, X_2, \ldots, Y_{n_{vert}}, Z_{n_{vert}})
\]  

(1)

\[
S = \bar{S} + A_{id} \alpha_{id} + A_{exp} \alpha_{exp}
\]

(2)

Where \( S \) is the 3D face, \( \bar{S} \) is the average shape. We use BFM, which is the most widely used 3D spatial face model. \( A_{id} \) is an identity morphable model whose principal axis of variations are in the matrix. \( \alpha_{id} \) is the coefficient. \( A_{exp} \) is the principle axes trained on the offset between expression scans and neutral scans. \( \alpha_{exp} \) is the weight.
Finally, formula (3) was used for shape constraint, requiring the key points of the target 3D model to coincide with the key points of the 2D image after being projected.

\[
\arg_{f,R,3d.atid.atexp} = \min ||S_{2dt} - S_{2d}||
\]  

(3)

Where \( f \) is the scale factor, and \( R \) is the rotation matrix. Finally, we follow the work of Liu et al [9]. The PCA dimension reduction method is used to solve the 3DMM prior parameters required by the MFF framework. The effect of BFM model and parameter fitting is shown in figure 5.

2.4 MFF framework and generate large angle face

MFF is a multi-feature task framework that can reconstruct the 3D face with identity information under the constraints of 3DMM prior parameters and texture conditions, which is presented in figure 6.

Fig. 6. Schematic diagram of multi-tasking framework MFF (Pixel intensity feature, Edge feature, Specular highlight feature, Gaussian Prior feature, Texture constraint feature)

The superiority of the MFF framework is reflected in the fact that it constrains 5 features (Pixel intensity feature, Edge feature, Specular highlight feature, Gaussian Prior feature, Texture constraint feature) at the same time, and there are few cases of failure. The purpose of the fitting algorithm using multiple features is to maximize the posterior probability of the model parameters, not only as done in SNO [11], but also to consider the different characteristics of the model parameters. Its fusion algorithm is shown in formula four:

\[
\min_\theta \tau^C C^C + \tau^E C^E + \tau^S C^S + \tau^P C^P + \tau^T C^T
\]  

(4)

In this function, the \( C^C \) is the pixel intensity of the picture, \( C^E \) is the edge feature, \( C^S \) is the specular highlight feature, \( C^P \) is the 3DMM prior parameters, finally, \( C^T \) is the texture constraints. \( \tau^S \) is the feature weighting factors. Among them, Pixel intensity feature, Edge feature, Specular highlight feature, three features are obtained from the input RGB image, and
Gaussian Prior feature is obtained from the 3DMM prior parameter of 2.2. It is minimized using a Levenberg-Marquardt optimization algorithm [12].

![Fig. 7. Large-angle face generation results (a) 30° (b) 60° (c) 75° (d) 90°](image)

After using MFF multitasking to form a strong constraint on facial identity information, we follow the method of Liu et al. To mark some anchor points outside the surface area and estimate their depth. After the depth information is obtained, a large-angle face image can be generated. As shown in Figure 7. Among them, (a) is the input picture, and the yaw angle is estimated to be 30 degrees after attitude estimation. (b), (c), and (d) are the results after the large angle is generated, and the identity information is better retained.

3 Experiments

In this section, the experiment is implemented to verify the effectiveness of the enhanced large angle face datasets, which is divided into two parts. In part 1, the generation of the large angle face dataset is introduced in detail. The results of the test of the generated large angle face dataset are analyzed and discussed in part 2, which is a good validation for the generated large angle face dataset.

3.1 Large-angle face dataset generation

3.1.1 Datasets

To generate and test the large angle face dataset, another two datasets, as the fundamental samples, are employed in the experiment. The first one is the CASIA-WebFace dataset, which is composed of 10575 subjects and 494414 images. The pictures from the CASIA-WebFace dataset are produced from the Internet. Thus, the sample pictures of the CASIA-WebFace dataset have more complicated background, which can be used to be closer to the environment in our daily life.

The second dataset we used is composed of several real pictures, which is to verify the effectiveness of the method generating the sample set. In this dataset, face pictures of 15 basketball players are extracted. These face pictures have various angles, most of which have yaw angle greater than 60° around the y-axis. The backgrounds of these face pictures are from different basketball game, which can keep these face pictures in a wild state. The sample dataset is constructed by 300 face pictures in total and 20 face samples are extracted from every basketball player in average. Each image of these face pictures is cropped to 160*160 px. The constructed dataset is named as validation set and some of the face picture samples are presented in figure 8.
3.1.2 Sample set generation

To generate high angle face samples, 100 subjects are extracted from the CASIA-WebFace dataset, which is called CASIA-WebFace (100). The face alignment of these selected 100 subjects is finished by using the method of FAN. Then the selected dataset is divided into two intervals (-90°-0° and 0°-90°) by employing the method of pose estimation using five key points mentioned in the section 2.2. To more accurately estimate the rotation angle of faces pictures, the divided two big intervals are further partitioned into smaller intervals by the step of 15°. Finally 12 sample intervals are obtained, which is shown in table 1. From table 1, it can be concluded that most of the samples in the CASIA-WebFace (100) are focused on the interval from -60°-60° and a few samples are distributed outside the interval from -60°-60°. Therefore, some additional work is conducted to enhance the samples especially in the interval from -90°-60° and 60°-90°.

The detail procedure of sample enhancement has been mentioned in the section 2.3 and 2.4. Firstly, 3DMM Fitting method is adopted to solve the prior parameters of the former CASIA-WebFace (100). These prior parameters are required by the MFF framework. Based on the MFF 3D face reconstruction, the lower angle face pictures in CASIA-WebFace (100) can be well projected to higher angle face pictures. The corresponding intervals are detail depicted as flowing:

- Projecting samples in (30°, 45°) (-45°, -30°) in CASIA-WebFace(100) to (46°, 60°) (61°, 75°) (76°, 90°) (-60°, -46°) (-75°, -61°) (-90°, -76°) within six intervals.
- Projecting images in (46°, 60°) and (-60°, -46°) sections of CASIA-WebFace (100) to (61°, 75°), (76°, 90°), (-75°, -61°), (-90°, -76°) within 4 intervals.

Finally, the projected samples are added into the original CASIA-WebFace (100) and a new sample dataset can be obtained. The newly generated sample (it is called as Samples-Data) have many face sample no matter any intervals from -90° to 90°. The detail comparison of both samples between the original CASIA-WebFace (100) and Samples Data is presented in table 1.

<table>
<thead>
<tr>
<th>Angle</th>
<th>CASIA-WebFace(100)</th>
<th>Samples-Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>[-90°~75°)</td>
<td>0</td>
<td>818</td>
</tr>
<tr>
<td>[-75°~60°)</td>
<td>16</td>
<td>834</td>
</tr>
<tr>
<td>[-60°~30°)</td>
<td>229</td>
<td>818</td>
</tr>
<tr>
<td>[-30°~15°)</td>
<td>806</td>
<td>806</td>
</tr>
<tr>
<td>[15°~0°)</td>
<td>1089</td>
<td>1089</td>
</tr>
<tr>
<td>[0°~15°)</td>
<td>996</td>
<td>996</td>
</tr>
<tr>
<td>[30°~45°)</td>
<td>865</td>
<td>865</td>
</tr>
<tr>
<td>[45°~60°)</td>
<td>608</td>
<td>608</td>
</tr>
<tr>
<td>[60°~75°)</td>
<td>12</td>
<td>845</td>
</tr>
<tr>
<td>[75°~90°)</td>
<td>1</td>
<td>834</td>
</tr>
</tbody>
</table>

By the comparison in the table 1, it can be obviously deduced that the number of the samples from the CASIA-WebFace (100) dataset has greatly increased in the intervals from -90°--45° and 45°-90°. Especially in the higher intervals from -90°--60° and 60°-90°, the growth is the most pronounced and the number of the samples have increased at last 818, which has greatly expanded the CASIA-WebFace (100) dataset in the aspect of large angle face dataset. Also, the large angle face recognition is driven by the generated dataset Samples-Data.

We generated a large-angle face recognition sample set of 12 angle intervals, as shown in Figure 8. The first line is -90°-0°, and the second line is 0°-90°.

### 3.2 Validation of large-angle face generation method

Most of the face recognition methods are relied on the deep learning. Transfer learning [13] is a hot research topic in the field of deep learning, which can increase the efficiency of training. The expanded dataset Samples-Data is developed by the CASIA-WebFace (100), which mainly focuses on the large angles of the CASIA-WebFace (100). The FaceNet pre-trained models [14][15] for CASIA-WebFace transfer learning training, and only trained the last softmax layer to extract high-level features.

The Samples-Data sample set is divided into 3 intervals, which are 0°<=|Y|<=60°, 0°<=|Y|<=75°, 0°<=|Y|<=90° (Y is around the Y-axis Rotational yaw angle) respectively. The FaceNet is used to train the sample sets in three intervals and CASIA-WebFace (100) to investigate the influence of different face angles on the extracted features of the neural network, which can be used to further verify the effectiveness of the sample enhancement method. FaceNet is trained using four types of sample sets and the four neural networks after fine-tuning training are named A, B, C, and D, respectively. The validation set consisting of 300 basketball
player face pictures are utilized to verify the effect of large angle face recognition. The results are shown in table 2.

Table 2. FaceNet recognition accuracy of the validation set under four different training sets.

<table>
<thead>
<tr>
<th>Training</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.8026</td>
<td>0.8438</td>
<td>0.8672</td>
<td>0.8012</td>
</tr>
</tbody>
</table>

From table 2, the minimum recognition rate of the validation set using A network is 0.8026 and D network is 0.8012. Because the training sets of A and D lack large angle face picture samples. The recognition rate of B network is 0.8438, which is greatly improved compared to A and D. Because the training set of $60^\circ \leq \mid Y \mid \leq 70^\circ$ is added to make the neural network learn more deep features. The recognition rate of the verification set under the C network reaches 0.8672, which is the highest accuracy of the four networks. The main reason is that the training set of the D network includes faces in extreme states (the face has a yaw angle of 90 around the Y axis). The samples allow the neural network to learn enough features.

4 Conclusion

In this paper, a novel sample-enhanced dataset is constructed, which is composed of various angle face picture samples from -90° to 90° relative to the front face. To generate the constructed dataset, multi-task feature (MFF) framework is employed in this paper. The generated dataset is developed by the CASIA-WebFace dataset, which greatly expands the samples of the CASIA-WebFace dataset especially for the large angle samples. The FaceNet is employed and trained by using the generated dataset. To test the effectiveness of large-angle face recognition, 300 large-angle face pictures of different basketball players are selected as another sample by using four different trained networks. The results demonstrate that the accuracy of face recognition has been greatly improved when the larger angle face picture samples are used to the network.

References


A Three-party Repeated Game Model for Data Privacy in Mobile Edge Crowdsensing of IoT

Mingfeng Zhao[1][0000–0002–1629–5793], Lei Chen[2][0000–0002–3919–8056], Jinbo Xiong[1][0000–0001–9985–1953], and Youliang Tian[3][0000–0002–5974–1570]

1 Fujian Provincial Key Laboratory of Network Security and Cryptology, College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350117, China, zmf1900953654@163.com, jbxiong@fjnu.edu.cn
2 College of Engineering and Computing, Georgia Southern University, GA 30458, USA, lchen@georgiasouthern.edu
3 State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University, Guiyang 550025, China youliangtian@163.com

Corresponding author: Jinbo Xiong (jbxiong@fjnu.edu.cn)

Abstract. The low request response delay of mobile edge crowdsensing (MECS) paradigm allows quick interactions among entities in practical scenarios. However, there often exist dishonest behaviors in such interactions, and the personal information leakage involved seriously threatens the privacy and security of sensing users. To tackle this problem, previously we had proposed a non-repeated three-party game model, without the consideration of multiple interactions in the actual scenario. Based on game theory, this research therefore proposes a three-party repeated game model. Specifically, we propose the corresponding social norms for different phases of sensing data. It analyzes all possible behaviors deviating from rationality, calculates the change of corresponding payoff function, and explores the influencing factors and constraints of players’ honest behaviors based on the premise of maximizing interests. Finally, a significant number of simulations and numerical analyze indicate that the proposed model is feasible and effective in maximizing the benefits of game participants.

Keywords: Privacy protection · Mobile edge crowdsensing · Game Theory · Three-party repeated game model · Nash Equilibrium.

1 Introduction

In recent years, the popularity of wireless mobile and 5G terminal devices is growing explosively [21, 18, 30]. Smart phones, wireless Bluetooth wearable devices and other integrated sensors are becoming more abundant, making them capable of powerful sensing and computing [35, 32]. Based on this background, a paradigm MCS, mobile crowdsensing, has emerged [42, 22]. In MCS, a large number of mobile users who carry the afore mentioned IoT intelligent terminal equipment as the basic sensing unit interact with the sensing platform with the
help of wireless sensor network to realize the collaborative work of task distribution and data collection [35, 28], so as to complete the large-scale and complex social sensing tasks under various scenarios [33]. Meanwhile, with the development of 5G base stations, the throughput of real-time data communication and link bandwidth would be greatly improved. Nowadays, we are committed to building a smart city paying more attention to strong interactions with high real-time requirements, such as intelligent transportation systems [12], crowdsourced bus service [6], connected autonomous vehicle service [4, 5, 39], and mobile health system [40, 31]. This promotes the emergence of the MECS paradigm [36], which includes sensing users, edge nodes and cloud service providers. By introducing the concept of mobile edge computing (MEC) [2, 34], it solves the communication bottleneck of traditional cloud computing. In an MEC, a large number of scattered edge nodes with different application services use their edge attributes closer to the client, combined with certain computing and storage capacity, to carry part of the functions of the remote cloud server [17].

On the other hand, the explosive growth of the number of intelligent IoT devices has also brought massive multi-dimensional and heterogeneous source data, and the sensing activities of multi-user collaboration may also expose their social association attributes and other privacy information. Specifically, on the one hand, the sensing user will choose the task published by the edge node, and upload the data collected from the real-time sensing activity to the edge node to obtain the task reward. In this process, there exist risks of user privacy leakage, such as sensing data ownership migration and adversary attacks that may be encountered in the process of data transmission. At the same time, as semi trusted and resource constrained entities, edge nodes may also have potential active disclosure behaviors and node attacks. On the other hand, in the process of the transaction between the edge node and the service provider, the data ownership has been migrated again, and the cloud service provider, as an untrusted platform, may actively disclose the user’s private data to adversaries exchange benefits. Therefore, it is urgent to solve the problem of user data privacy leakage in MECS network. The center of big data privacy protection is privacy protection technology.

The current work can be divided into: data distortion (focusing on differential privacy technology), data encryption (such as homomorphic encryption, secure multi-party computing, functional encryption, etc.) and restricted publishing (focusing on data anonymity). As another hot field in the background of big data, artificial intelligence can effectively drive the level of privacy protection, while reducing the risk of privacy leakage in the application process with the help of privacy protection technology. [38] is a good example as it integrates the concept of game theory. However, it lacks the consideration of possible multiple interactions between entities in the actual scene over a period of time. In game theory, there are important differences in the nature of players in dealing with short-term and long-term relationships. A tacit or cooperative relationship that is difficult to form in the short term can constrain each other’s behaviors through long-term potential retaliation, sanctions and other threatening behaviors, as
shown in the work of [24], [26]. However, most of the existing work focuses on
the long-term behavior relationship between the two entities. Therefore, in order
to deal with data privacy of users in MECS network, this paper aims to build
a repeated game model for three-party entities, in order to find the influencing
factors and constraints that regulate the benign behaviors of multi-party entities
in MECS. The main contributions of this paper are summarized as follows:

– Build a repeated game model based on three-party entities, and analyze the
deviation behavior and payoff change among players in multiple phases of
the sensing data life cycle.
– On the premise of maximizing benefits, identify the influencing factors and
constraints of honest behaviors of players in different phases of MECS.
– Through a large number of simulation experiments and numerical analyse,
the proposed three-party repeated game model is proofed feasible and effec-
tive, suitable for MECS paradigm.

The rest of this paper is organized as follows. Section 2 introduces the related
works. Section 3 gives the preliminaries. Section 4 describe the proposed repeated
game model in detail. Section 5 discusses the experimental results and theoretical
analysis of the proposed model. Section 6 gives a summary of the research.

2 Related Works

As mentioned above, in the mobile edge group intelligence perception network,
user privacy threats caused by perceived data leakage have attracted many schol-
ars to conduct relevant research.

On one hand, the work mainly focuses on the application of cryptography
and block chain technology. The method based on cryptography, the scheme de-
sign and construction of data encryption, anonymity, disturbance, aggregation
and other aspects can be carried out. Blockchain technology provides verifica-
tion support and portable management for data security sharing, high reliability,
tamper resistance and so on [23], [19], [29]. In the intelligent perception paradigm
of mobile edge group based on the background of Internet of things, mobile edge
computing is one of the core technologies supporting the architecture, which
meets the needs of low latency and fast corresponding service requests of In-
ternet of things applications. Li et al. [16] proposed a privacy data aggregation
scheme for mobile edge computing to assist Internet of things applications, which
not only guarantees the data privacy of terminal equipment, but also provides
source authentication and integrity check, saving half of the communication cost
compared to the traditional schemes. Considering the problem of data privacy
protection in the process of collecting personal information, a protection algo-
rithm based on differential privacy model is proposed in [15], and a time window
partition and a dynamic network community discovery algorithm is designed to
reduce the differential privacy noise. With the help of layered sampling, the time
cost and cumulative errors are reduced. Experiments show that the algorithm
can keep the important structural features of the original network graph on the
premise of satisfying the differential privacy protection model. In addition, users mainly collect sensor data through intelligent Internet of things devices equipped with sensors. In view of this, many methods to protect the privacy of intelligent terminal devices have been proposed. Blasco et al. [3] put forward a three-layer method to protect the privacy of citizens in order to solve the problem that personal privacy is easy to be mined and attacked due to the need of smart city services to access sensitive data of users [13, 14]. By combining the first layer and the second layer of homomorphic public key encryption, local data collection is safe, and the third layer adds differential privacy to control the spread of public information.

When users enjoy personalized services provided by various context aware applications, their sensitive information hidden in the context is exposed. Zhang et al. [41] designed a privacy protection deception strategy based on the passive defense strategy for most of the current mechanisms. They proposed a new technology: FakeMask, essentially a privacy check algorithm that can adaptively release a fake context according to the current context of the user, greatly limiting the adversary to infer the actual context. Experimental evaluation and scheme comparison in real smartphone environment show that FakeMask has outstanding performance.

On the other hand, from the perspective of game theory, through modeling and analyzing the behavior game between players, we can solve the prisoner’s dilemma and other problems. In MECS network, many application scenarios involve different entities with multiple target conflicts. The process of conflicts is actually the choice and game of the best strategy, and the ultimate goal is to maximize their own interests. At present, game theory has been successfully applied in many representative communication and network scenarios, such as defenses against DoS attack in wireless network [1], data privacy protection in social networks [20] and privacy protection model in transportation systems of IoT [27]. In [25], Moura et al. investigated and studied the main challenges of mobile edge computing services to wireless resources based on game theory. They discussed the specific game strategies, model evaluation and balance constraints in the edge network scenarios by classical game and evolutionary game, and emphasized the application trend and research direction of game theory model in mobile edge computing services in the future. Jin et al. [8] considered two groups of service providers with different request strategies to obtain the reward matrix. This work aims at the problem of excessive permission request in the current smart phone terminal, thus established two groups of evolutionary game model for user privacy protection, and analyzed the stability strategy of the model. Kim [9] proposed an MCS control scheme based on multi-level game model and differential privacy concept in view of the serious loss of personal privacy in mobile group intelligence perception. From the perspective of an MCS server differential privacy(DP) controller and mobile devices, the dynamics of their interactions are captured and analyzed, and the game process is repeated step by step to explore effective solutions for promoting interaction among players. At present, many real-world application scenarios can be simulated as prisoners’
dilemma, and the relevant research literature also provides a variety of strategies; however, it rarely conforms to the design objectives of the intelligent agent: reactivity and initiative. In [11], the risk attitude and reputation factors are combined into infinite repeated games, and the original game theory matrix is transformed into a new matrix with cooperative equilibrium. By analyzing the repeated prisoner’s dilemma and the results of simulation experiments, it is verified that the performance of agents considering the above two factors in the decision-making process, in both active and passive manner is improved. Xiong et al. [38] propose an AI (Artificial Intelligence)-enabled three-party game framework by combining machine learning and game theory, discuss the privacy leakage problem of entity interaction in typical application scenarios of MECS, and provide an effective and efficient scheme for ensuring data privacy in the MECS of IoT.

In MECS, there are entities such as users, edge nodes, cloud service providers, attackers and so on. From the above literature, mobile edge computing, as an important component technology of MECS, has many scenarios based on game theory. [9] provides an example of a scheme combining game theory and cryptography to solve the problem of privacy protection in MCS paradigm. The repeated prisoner’s dilemma game under multiple factors is considered in [11]. [38] combined the AI algorithm on the basis of game theory and cryptography, provided an effective solution to the problem of privacy leakage risk in MECS paradigm, without considering the impact of repeated interaction between entities over a period of time. In the practical application scenario of MECS, entities often interact multiple times over a period of time, and therefore it is necessary to consider the subjective and objective factors in the interaction process of multiple entities. In addition, most of the above works are repeated games between two parties, without the extension to multiple parties, which stimulates the development of our work in this paper.

3 Preliminaries

In this section, we formally define our system model, threat models and assumptions. We then introduce the problem description and design goals.

3.1 System Model

As shown in Fig.1, in MECS paradigm, there exist entities such as sensing users (SUs), edge nodes (ENs) and cloud service providers (CSPs), which are described as follows.

(1) Sensing users (SUs): a large number of ordinary people who apply smartphones, tablets, wearable devices and other mobile devices as basic perceptual units. After selecting the sensing task independently, they utilize various sensors integrated on the device to carry out sensing activities, and upload the data to ENs with the help of wireless network.
(2) Edge nodes (ENs): With the help of mobile Internet, sensing tasks are distributed to SUs and sensing data are collected and processed by ENs. A large number of nodes with different functions and dispersions cooperate to complete large-scale and complex social sensing tasks, and employ the processed data to trade with CSPs.

(3) Cloud service providers (CSPs): provide real-time services to SUs to meet their personalized needs by using sensing data or service models obtained from transactions with ENs.

In our system model, the life cycle of sensing data in an MECS network is divided into sensing data uploading phase and sensing data trading phase. In practical application scenarios, there may exist adversaries who attempt to obtain the user private data in each phase.

3.2 Infinitely Repeated Game

An infinite number of repeated games is a game process in which players repeat the same structure for many times and there is no fixed time to end the game. The behavior process of repeated implementation is called stage game. The specific definition is as follows:

Definition 1: An infinite number of repeated games can be expressed as a tuple $< N, S, P, H, \delta, T >$, where,
- $N$: a finite set of $n$ players
- $S_i$: action strategy sets of $n$ players, where $i \in N$, and action profile could be denoted as $S = \times_{i \in N} S_i$
- $P_i$: the payoff function of $i \in N$ at every stage
- $H$: $S \rightarrow R^n$, the set of players' payoffs at the end of each stage game
- $\delta$: the discount factor of players to evaluate the payoffs, $0 \leq \delta \leq 1$
- $T$: the number of the stages

Definition 2: The stage game $G$ is a strategic game. Combined with its repetition times $T$, we can determine a "$T$-repeated game" process, and denoted as $G(T)$.

$$G = \{ S_i, \pi_i; i = 1, ..., n \}.$$

Fig. 1: System model of MECS
In (1), $G$ is the original game of $G(T)$, and each repetition in $G(T)$ is called a stage of $G(T)$. It can be seen from the above that $S_i$ is the strategy set of player $i$; $\pi_i$ is his/her payoff at each stage, and it depends on $(S_1, S_2, ..., S_n)$.

According to the system model architecture shown in Fig.1, it is obvious that there is a chronological order for the players’ strategy selection. We assume that player 1 has the priority to choose strategies, and there is no limit on player 2’s strategy choice. This is also true between player 2 and player 3. In view of this situation, the expansion form of the game is commonly used to analyze [10]. The possible behaviors of the players are represented in the form of the behavioral game tree, and the payment of each stage of the game process is given at the leaf node of each branch, as shown in Fig.2. Here, $t \in T$, obviously, when $t = T = 1$, it means that each player has played a single game.

![Extended behavioral game tree](image)

**Fig. 2: Extended behavioral game tree**

### 3.3 Single-stage Game

Here, we will describe the behavior of the initial stage game based on the work of [38]. For the sensing data uploading phase and the sensing data trading phase, there are two different player sets $N'$ and $N''$: SUs, adversary $A_1$, ENs, and ENs, CSPs and adversary $A_2$. The player’s strategy set of two phases $S_i(i \in (N', N''))$ correspond to $(M, M')$ in [38]. Based on the structure of payoff function of players in different phases, we respectively obtain the player’s payoff set $(h, h')$ at the end of the first stage game from [38]. It is worth noting that we all abide by two hypothesis: ① SUs will not carry out sensing activities at any cost, and the adversary will not blindly launch unprofitable attacks; ② the accuracy of data largely determines the privacy of user information, so the payoff function of each player will be constructed based on this. Thus, it is easy to know that the adversary $A_1$ will not launch an attack when SUs do not upload their sensing data, so play2 in Fig.2 has only three behavior branches, and affected play3 has six behavior choices.
4 Repeated-stages Game Model

In a single-stage game, each participant only pays attention to the current payoff. However, the interactions between these participants are long-term and repetitive in practice. Under these circumstances, participants will consider the impact of current behavioral strategies on future payoffs. Therefore, based on the situation of single-stage strategy game [38], we model their multiple interaction processes as a repeated game model.

4.1 Sensing Data Uploading Phase

When $\mu_5 < \mu_3$, the optimal strategy profile $\ell = 1, \sigma = 0, \tau = 0$ is the only pure strategy Nash Equilibrium solution. If there is only one pure strategy NE solution in the original game, each participant would adopt the NE strategy profile of the original game in the next stage [7]. Therefore, the result of repeated game in this case is that SU uploads sensing data, $A_1$ does not launch an attack and EN does not leak privacy, so the user’s private information is well protected.

When $\mu_5 > \mu_3$, the optimal strategy profile $\ell = 1, \sigma = G(d_2) - \mu_1, \tau = 1$ is the mixed strategy NE solution. We introduce a discount factor $\delta$ [37] due to the payoff of infinitely repeated games being endless. We can use the same discount factor $\delta$ to discount the future payoff of each stage, so that the total payoff can be limited and comparable.

Given $\delta = e^{-r\Delta}$, where $r$ is the preference rate for time and $\Delta$ is the length of a period. For one path of infinitely repeated games, we assume that the payoff of participants at each stage are $\pi_1, \pi_2, \pi_3, \ldots$, respectively. Also, we use $\delta_{SU}, \delta_{A_1}, \delta_{EN}$ to denote the discount factor of SU, $A_1$, EN, respectively. Therefore, the total payoff of participants is shown in Formula (2).

$$\pi = \pi_1 + \delta \pi_2 + \delta^2 \pi_3 + \cdots = \sum_{i=1}^{\infty} \delta^{i-1} \pi_i.$$ (2)

For the case of $\mu_5 > \mu_3$, we propose social norms that are in line with the actual situation. Usually, an SU always uploads sensing data, and refuses to upload until $n_1$ times leakage behaviors are discovered. For $A_1$, the initial strategy is not to attack. It would switch to attack strategy and sustain $n_3$ times until the SU uploads data $n_2$ times continuously. Additionally, if the SU does not upload data, $A_1$ would not launch an attack. EN chooses not to leak at the moment. Once EN finds that $A_1$ launch $n_3$ consecutive attacks, it would switch to leak strategy. Because the SU could not determine who committed the dishonest act at this time, and EN refuses to admit. Also, when $A_1$ does not launch an attack for $n_4$ consecutive times, EN switches to the behavior of not leaking. The more the participants deviated from the social norms, the more their later payoffs decreased. In order to facilitate the analysis of problems without losing generality, we assume that $n_1 = 1, n_2 = 1, n_3 = 2, n_4 = 2$.

If all participants abide by the initial behavior of the above social norms, the strategic path in infinitely repeated games would be $(\ell = 1, \sigma = 0, \tau = 0) \rightarrow$
(ℓ = 1, σ = 0, τ = 0) → (ℓ = 1, σ = 0, τ = 0) → ⋯ (ℓ, σ, τ are omitted in the following text). Therefore, we can calculate the total payoffs of SU, A₁ and EN as follows:

\[
\begin{align*}
&u_{SU}(d) = G(d)(1 + \delta_{SU} + \delta_{SU}^2 + \delta_{SU}^3 + \cdots) = \frac{G(d)}{1-\delta_{SU}} \tag{3} \\
u_{EN}(d) = G(d)(1 + \delta_{EN} + \delta_{EN}^2 + \delta_{EN}^3 + \cdots) = \frac{G(d)}{1-\delta_{EN}} \\
u_{A_1}(d) = 0(1 + \delta_{A_1} + \delta_{A_1}^2 + \delta_{A_1}^3 + \cdots) = 0
\end{align*}
\]

1. Considering the deviation of A₁’s behavior

Assume that starting from the first round of the game, the new strategy path of the game would be described as follows:

\[
(1, 1, 0) \rightarrow (1, 1, 0) \rightarrow (1, 1, 1) \rightarrow (0, 0, 1) \rightarrow (0, 0, 1) \rightarrow (0, 0, 0) \cdots
\]

Now, the total payoffs \(u_{A_1}(d)\) after A₁’s behavior deviates would be:

\[
u_{A_1}(d) = (\mu_6 \text{Sens}(d) - R)(1 + \delta_{A_1} + \delta_{A_1}^2).
\]

2. Considering the deviation of EN’s behavior

Similarly, starting from the first round of the game, the new strategy path of the game would be described as follows:

\[
(1, 0, 1) \rightarrow (0, 0, 1) \rightarrow (0, 0, 1) \rightarrow (0, 0, 0) \rightarrow (0, 0, 0) \cdots
\]

Now, the total payoffs \(u_{EN}(d)\) after EN’s behavior deviates would be:

\[
u_{EN}(d) = G(d) + (\mu_5 - \mu_3) \text{Sens}(d).
\]

3. Considering the simultaneous deviation of A₁ and EN’s behaviors

The behavior of both deviates simultaneously in the first round, and the strategy profile is (1, 1, 1); the new game strategy path would be described as follows:

\[
(1, 1, 1) \rightarrow (0, 0, 1) \rightarrow (0, 0, 1) \rightarrow (0, 0, 0) \rightarrow (0, 0, 0) \cdots
\]

Now, the total payoffs of \(A_1\) and EN would be described as follows respectively:

\[
\begin{align*}
u_{A_1'}(d) = \mu_6 \text{Sens}(d) - R \\
u_{EN'}(d) = G(d) + (\mu_5 - \mu_3) \text{Sens}(d)
\end{align*}
\]

In order to restrain the game participants from choosing deviation behaviors, their payoff function should satisfy the following inequalities:

\[
\begin{align*}
u_{A_1'}(d) < \nu_{A_1'}(d) \\
u_{EN'}(d) < \nu_{EN'}(d) \\
u_{A_1'}(d) < \nu_{A_1'}(d) \\
u_{EN'}(d) < \nu_{EN'}(d)
\end{align*}
\]

We can have the following inequalities:

\[
\begin{align*}
\mu_6 < \frac{R}{\text{Sens}(d)} \\
\delta_{EN} > \frac{G(d) + (\mu_5 - \mu_3) \text{Sens}(d)}{G(d) + (\mu_5 - \mu_3) \text{Sens}(d)}
\end{align*}
\]

Obviously, When the discount factor \(\delta_{EN}\) of EN and \(\mu_6\) satisfy the relevant range, the game participants will not choose to deviate from the social norms, so as to protect the sensing user’s personal private information.
4.2 Sensing Data Trading Phase

The analysis method of this phase is similar to that of the uploading phase. We use $\delta_{\text{CSP}}$ and $\delta_{A_2}$ to denote the discount factor of CSP and $A_2$ respectively. For the case of $k_3 > k_2$, we also propose social norms that are in line with the actual situation. Usually, an EN always chooses to trade data, and refuses to trade until $m_1$ disclosures are found. A CSP would firstly chooses not to leak any data, and switches to leakage behavior once finds out that $A_2$ has launched $m_2$ times of attacks. Additionally, $A_2$‘s initial strategy is not to attack. And they would launch $m_3$ times of attacks continuously once found that CSP have leaked data. Meanwhile, when $A_2$ finds that the CSP no longer leak data $m_4$ times, if will choose not to attack for maximized payoffs. In order to support a complete strategy transformation process, we assume that $m_1 = 1, m_2 = 1, m_3 = 1, m_4 = 2$.

Similarly, we can have that when $k_4 < \frac{3C}{Sens(d)}$, the range of discount factor is:

$$\delta_{\text{CSP}} > \frac{(k_3 - k_2)Sens(d)}{G(d) + (k_3 - k_2)Sens(d)}$$

(9)

Whereas, in the case of $k_4 > \frac{3C}{Sens(d)}$, the range of discount factors are:

$$\begin{cases} 
\delta_{\text{ASP}} > \frac{(k_3 - k_2)Sens(d)}{G(d) + (k_3 - k_2)Sens(d)}, \\
\delta_{A_2} < \frac{k_3Sens(d) - C - \sqrt{k_3^2 Sens(d)^2 - 2k_4 Sens(d)C - 3C^2}}{2C}
\end{cases}$$

(10)

As the range of $k_4$ changes, when the discount factors range of relevant game participants are satisfied, they would not have any reason to deviate from social norms. In this way, it prevents the private information of users from being leaked.

5 Model Analysis and Validation

In this section, we conduct numerical analysis on the proposed three-party repeated game model, including sensing data uploading phase and sensing data trading phase. The experiments were carried out on a desktop computer running Windows 7 system, which was configured with Intel core i5-5200U, 2.20 GHz CPU and 8 GB RAM, and the software used is MATLAB R2016a.

5.1 Sensing Data Uploading Phase

In this section, the parameter settings are the same as above. In the sensing data uploading phase, Fig.3(a) illustrates that when $\delta_{EN}$ exceeds a certain value, EN would change from the leaking strategy to the non-leaking strategy, and this value depends on $\mu_5 - \mu_3$. With the increase of $\mu_5 - \mu_3$, the critical value also increases, which is consistent with formula (8). $\mu_5$ is the positive influence coefficient reflecting the payoff obtained by EN’s leakage behavior, and $\mu_3$ is the reputation punishment coefficient of EN’s leakage behavior. $\mu_5$ is a constant value for the sensing data of certain accuracy. By increasing $\mu_3$, that is, increasing the punishment for EN on users’ side, the critical value goes down, which makes the
Fig. 3: Performance evaluation results: (a) shows the influence of $\mu_5 - \mu_3$ and $\delta_{EN}$ on strategy selection of EN; (b) and (c) show the influence of $\mu_5 - \mu_3$ and $\delta_{EN}$ on payoffs of EN and SU; (d) shows the influence of $k_3 - k_2$ and $\delta_{CSP}$ on strategy selection of CSP; (e) and (f) show the influence of $k_3 - k_2$ and $\delta_{CSP}$ on payoffs of CSP and EN; (g) shows the influence of $k_4$ and $\delta_{A_2}$ on strategy selection of $A_2$; (h) and (i) show the influence of $k_4$ and $\delta_{A_2}$ on payoffs of $A_2$ and EN.
constraint conditions in equation (8) easier to be reached, forcing EN to choose the non-leaking strategy. Fig.3(b) and Fig.3(c) show that the payoff of EN and SU increases with the increase of $\delta_{EN}$. At the critical value, the payoff of SU will increase greatly with the change of strategy choice of EN. It is verified that EN would be more inclined to comply with social norms within the range of constraint conditions.

5.2 Sensing Data Trading Phase

In the sensing data trading phase, Fig.3(d) shows that CSP will be more prone to non-leakage behavior with the increase of $\delta_{CSP}$, and its critical value depends on $k_3 - k_2$. $k_2$ is the punishment coefficient for CSP’s leakage behavior, and $k_3$ is the positive influence coefficient reflecting the payoff of CSP obtained through the leakage behavior. Similarly, we can adjust $k_2$ on the side of EN to reduce the critical value according to formula (9), forcing CSP to comply with social norms and choose not to leak. As can be seen from Fig.3(e) and Fig.5(f), the payoff of CSP and EN are increasing with the rise of $\delta_{CSP}$. In detail, Fig.3(e) shows in the case that the value of $k_3 - k_2$ increases, the payoff of CSP increases with the rise of $\delta_{CSP}$, which indicates that the smaller $k_2$ is, the lighter the punishment on CSP is, and CSP is more likely to choose the leakage strategy. It is verified that $k_2$ is the critical factor influencing CSP’s choice of leakage and non-leakage strategy. Additionally, when $\delta_{CSP}$ is at the critical value, the payoff remains the same regardless of whether CSP chooses the leakage behavior or not. When $\delta_{CSP}$ exceeds the critical value, CSP would gain more from choosing non-leakage behavior over leakage behavior, which verifies CSP’s corresponding constraint conditions on compliance with social norms in formula (10). Fig.3(f) shows that EN’s payoff has a significant change at the critical value, which also reveals the key influence of CSP’s leakage and non-leakage behavior on EN’s payoff.

When $k_4 > \frac{\Delta C_{Sens(\delta)}}{S_{Sens(\delta)}}$, formula (10) indicates that a smaller discount factor $\delta_{A_2}$ is required to satisfy the constraint. Fig.3(g) manifests that $A_2$’s strategy is more inclined to leak with the increase of $k_4$, forcing us to choose smaller $\delta_{A_2}$. Additionally, Fig.3(h) shows that when $k_4$ is at a small value and $\delta_{A_2}$ does not meet the constraint conditions, the payoff from $A_2$’s choice of leakage is much less than the choice of non-leakage. Therefore, on the premise of meeting the constraint conditions, the adversary $A_2$ after repeated games would not deviate from the social norms. In this case, EN’s payoff would also increase significantly in this range, which is consistent with the results in Fig.3(i).

6 Conclusion

In view of the user data privacy problem of two-party interactions discussed in most of the existing work, based on game theory, we constructed a three-party repeated game model for sensing data uploading phase and the sensing data trading phase in MECS. By considering the possible interaction strategies among
participants in different phases, analyzing their potential deviation strategies, and calculating the change of payoff, this paper further explores the influencing factors and constraints that regulate participants’ honest behaviors in the game. Finally, through simulation experiments and numerical analysis, our proposed model shown feasible and has a certain guiding significance for user data privacy protection in MECS applications.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China (61872088, 61872090, U1905211, 61772008); the Natural Science Foundation of Fujian Province (2019J01276); the Science and Technology Major Support Program of Guizhou Province (20183001); and the Guizhou Provincial Key Laboratory of Public Big Data Research Fund (2019BDKFJJ004).

References


The design of wireless communication and positioning system of electric vehicle based on CAN bus

MA Xi-jie1, XU tian2, Li yu-chao3, WU chen-yue4, BAI yan-hui5
{tiantai0200@tom.com1, yanghong1910@163.com2, jiangyuan1923@163.com3, pengrenjie235@tom.com4, gaozhenyu101@2980.com5}
1, 2, 3, 4 Nanjing Institute of Technology, Nanjing 211167, China
5 School of Mathematics and Computer Science, Chifeng University, Chifeng 024000, China

Abstract: Because the throughput of the traditional wireless communication and positioning system of EVs is reduced, which can not meet the needs of users, the wireless communication and positioning system of EVs based on CAN bus is designed. Using the central control board to collect the real-time status, exchange information data, select the tps5430 switching power conversion chip to realize the two-way transmission with MCU address and data; under the support of hardware structure, optimize the initialization procedure, use the toa positioning principle to complete the positioning and ranging, extract the effective information, and realize the design of the wireless communication positioning system of the programmable electric vehicle based on CAN bus. Using computer simulation software, the performance of the system is tested. The results show that when the traffic load increases to 45mbps or higher, the throughput curve of the traditional system gradually shows a downward trend, while the throughput of the designed system does not change significantly after the traffic load increases, which can meet the needs of users.

Keyword: CAN Bus; Additional range electric vehicle; Wireless communication location

1 Introduction

Automobile manufacturing industry is a strategic industry related to the national economy and people's livelihood, social stability and economic development, and it is an important part of China's national economy. In recent years, the severe energy crisis, environmental degradation and national strategy put forward an urgent demand for the research and development of electric vehicles. Countries around the world have formulated national policies for the research and development of electric drive related technologies and the development of electric vehicle industry [1-3]. Electric
control system is one of the most core technologies of electric vehicles. Because of the complexity of system structure and working conditions, it is difficult to break through the key technologies of electric control system, which seriously restricts the promotion and application of electric vehicles. Electric vehicles are not only related to automobile itself and infrastructure construction. According to the relevant definition of the Ministry of science and technology, China's strategic emerging industries include energy conservation and environmental protection industry, new generation information technology industry, biological industry, high-end equipment manufacturing industry, new energy industry, new material industry, new energy automobile industry and other seven major industries[4-6]. Most of the industrial strategic emerging industries are related to the development of electric vehicles. The production, manufacturing and sales of electric vehicles, infrastructure supporting links, after-sales service links, battery rental / sales / service links, electric vehicle warranty and financial services and other links will have a profound impact on the industrial structure. The incremental electric vehicle is equipped with vehicle charging power unit and auxiliary power unit to drive the vehicle with electric energy. When the electric energy is sufficient, the battery provides all the energy needed for the vehicle to run. When the electric energy is insufficient, the auxiliary power unit works to provide the energy needed for the vehicle to run. The matching internal combustion engine has small power, and its efficiency and emission can be controlled in a better state. The fuel consumption rate and harmful emissions can be significantly reduced. The battery capacity is large, which can provide a long driving range of pure electric vehicle, and there is no driving range anxiety problem due to the existence of auxiliary power system. The battery capacity is significantly smaller than the battery capacity of pure electric vehicle. The manufacturing cost is greatly reduced, the available types are diversified, and the use cost is relatively low; the internal combustion engine can provide the heat required by the air conditioning power and vehicle heating, reduce the battery consumption, and increase the pure electric driving mileage of the vehicle; there is no need to build large charging facilities, and night charging can be used, which is conducive to the peak valley regulation of the power grid, and improve the operation quality of the power grid[7]. This paper studies the wireless communication and positioning system based on CAN bus, optimizes the system performance, and meets the user's requirements for the use of EVs.
2 Hardware Design of Wireless Communication Positioning System for Range-increasing Electric Vehicle Based on CAN Bus

2.1 Central control panel

The central control board is the core of the whole control system. It is the brain of the whole charger and exchanges data for each unit. The central control board communicates with human-computer interaction unit and integrated measurement unit through 485 bus; communication with card reading unit through 232 bus; the central control board and the integrated measurement unit are respectively connected with the S+ and s-contact points of the charging gun, and communicate with the electric vehicle BMS through the CAN bus; the central control board drives the indicator light, emergency stop switch and drain circuit through the IO port; the sampling circuit is driven through the AD port, and the insulation detection circuit is driven through the serial port[8-10].

![Communication block diagram of integrated measurement unit](Image)

**Figure 1** Communication block diagram of integrated measurement unit

The integrated measurement module is used to collect the necessary data in the charging process: the measurement of the output voltage and current of the charging unit, the judgment of the connection state of the charging gun, the measurement of
the ambient temperature and humidity, the real-time data measurement of the electric energy meter, the data measurement of the lightning protection module, and the communication with the BMS to collect the real-time state data of the battery. The data is uploaded to the controller through the RS485 bus interface and the central control board.

2.2 Circuit design

Tps5430 is a DC-DC power supply chip with outstanding comprehensive performance, which has the advantages of high conversion efficiency, high output current and over-current protection function. The 12V DC supplied by the auxiliary power supply is input from VIN pin, and after the chip conversion, the pH pin outputs 5V DC. The output voltage of 5V is determined by R7 of 10K Ω and r10240 of 3K Ω. Inductance Li and capacitance C2 filter the output circuit, making the output voltage more stable; diode d11349 provides protection for the circuit. circuit diagram as follows:

![Figure 2 5V power circuit diagram](image)

The power chip reg1117-3.3 is selected to convert 5V voltage into 3.3V voltage. After the 5V voltage input is filtered by low frequency and high frequency, the 3.3V voltage is output at 2 or 4 parallel terminals, and then the 3.3V voltage output is more stable by low frequency and high frequency filtering.

The circuit uses the clock chip ds3231 to provide timing for MCU. Ds3231 is a high-precision clock chip produced by the American credit semiconductor company,
As shown in figure 3, Can provide accurate information from year to second; with C bus interface, it can realize two-way transmission with MCU address and data; with low power consumption, it can use its own pin to connect dry battery as standby power supply, and continue to provide timing function when the main power supply is disconnected, etc.

As shown above, the standby power supply uses one-way filter power supply to supply power to RTC. After the main power supply is cut off, the RTC standby power supply will supply power to the RTC to ensure the continuous timing function. Reset is realized by connecting the external circuit. On the premise of normal operation of the clock circuit, LCP1788 is connected to the low level through the corresponding pin, so as to realize the RESET of the single chip microcomputer. That is to say, the reset can be realized by inputting the low level at its reset pin and keeping it for more than several machine cycles. JTAG interface is used as the test and debugging interface for the central control board, which is convenient for debugging in the programming process, as shown in the figure below:

Figure 3 DS3231 External interface circuit diagram
JTAG interface provides 5 signals to connect with lpc1788, namely TMS, TDI, TDO, TCK and trst. TMS / swdio is the pin that j-link outputs debugging mode signal to lpc1788; TDI is the data input pin; TDO is the pin that receives LCP1788 data; TCK is the clock signal sent by the emulator to the target board; TRST is used to reset JTAG simulation mode. Pin 3, pin 5, pin 7, pin 9 and pin 13 are connected with pull-down resistance and grounded, and pin 1 and pin 2 are connected with 3.3V power supply. At this point, the optimization of the hardware part of the system is finished.

3 Software Design of Wireless Communication Positioning System for Electric Vehicle Based on CAN Bus

3.1 Initialization design

After the system initialization, for the positioning base station, it has been waiting for the UWB signal. After receiving the UWB signal sent by the tag, it immediately enters the interrupt processing, runs the sds-tw algorithm, sends the distance data to the tag and continues to return to the status of waiting for the tag. For the tag, it circulates the UWB signal to the four base stations in turn, and after
receiving the information returned by the base station, it passes the SDS-TWR algorithm calculates the distance from four base stations, calculates the arrival time difference according to the TDOA model, and sends it to the main base station through UWB communication, and the main base station sends the data to the upper computer through WiFi. Manually place each base station on each coordinate axis and origin of the space rectangular coordinate system, and the system can start to run the self-built coordinate system software. In the positioning system, the label needs to calculate the coordinates of its own location according to the coordinates of the base station. Therefore, before starting the positioning, the position coordinates of the base station are required to be known. Because the system adopts the method of self-building coordinate system by software, it is time-consuming and laborious to choose the manual fixed-point measurement method in advance, and may cause large errors. In the process of system initialization, each base station is manually placed on the coordinate axis of the spatial rectangular coordinate system. Base station 1, that is, the location of the main base station, is the origin of the coordinate. Other base stations can get the distance between the main base station and the main base station through the UWB communication module, and calculate the location coordinates. After the location coordinates of each base station are known, the positioning system can work normally. The specific process of system initialization is as follows: for the base station, hardware initialization is the first step, and then all the base stations are initialized to wait for the tag signal; for the tag, after the hardware initialization, enter the coordinate initialization process, send the coordinate initialization flag to the primary base station, and then wait for the return information of the primary base station until the primary base station sends the coordinate initialization information back. Label begins the first positioning process; after receiving the coordinate initialization flag sent by the label, the primary base station communicates with other base stations for ranging, and measures the distance between these base stations and the primary base station, and returns the distance information to the label as the coordinate value of each base station in the coordinates, completing the initialization of coordinates.

3.2 Location and ranging

When radio waves propagate in different environments, the degree of signal loss is also different. The ranging technique based on signal intensity calculates the distance of the transceiver based on the loss of radio waves in propagation. The loss
of signal can be divided into loss caused by distance, loss caused by complex terrain and multipath loss. The ranging technique based on signal strength mainly estimates the distance between the receiver and the receiver according to the relationship between the signal intensity and the distance. The distance between the two ends is obtained by calculating the flight time of the radio waves at both ends. The main idea is that when the radio waves emitted by the transmitter reach the receiver, the receiver also sends a radio wave, and the transmitter calculates the distance by the delay of the transmission and reception time. As shown below:

![Diagram showing TOA Positioning principle](image)

**Figure 5** TOA Positioning principle

As shown above, you can get the following formula:

$$d = \frac{\left[ (T_3 - T_0) - (T_2 - T_1) \times v \right]}{2}$$  \hspace{1cm} (1)

In formula, $v$ is the speed at which radio waves travel, $T_0, T_3$ is the time when the transmitter transmits and receives the signal, $T_1, T_2$ is the time when the receiver receives and transmits the signal respectively.

The distance between the transceivers is estimated by calculating the time difference between the transmitter and the receiver, let the velocity of the signal be
$v_1, v_2$ is the distance between the transmitter and the receiver is $d$, The time difference between the signal arriving at the receiving end is $T_2 - T_1$, there are:

$$
\begin{align*}
L &= (T_2 - T_1) \times s \\
d &= \frac{v_1 \times v_2}{v_1 - v_2}
\end{align*}$$

In order to meet the positioning requirements, different positioning algorithms should be selected according to different applications. The location accuracy of the algorithm based on distance measurement is high, but it is easy to be affected by the outside. In order to ensure the accuracy of the distance, the results need to be modified, which will lead to the complexity of the algorithm and the high cost of hardware. So choose the appropriate location algorithm according to the situation.

If the absolute time of device A and device B is not strictly synchronized, Measuring distance will cause error, To avoid this error, SDS-TWR Algorithm, Its principle is similar to that of two-range ranging, The transmitter transmits the signal after encountering the object, The signal immediately returned, The transmitter detects the reflected signal, and the time interval between the transmitting and receiving can be obtained, This time interval is the time when the signal goes through the two-way distance. When the time is multiplied by the speed of the signal, the distance information can be calculated. Algorithm flow chart:
Figure 6 SDS-TWR Schematic diagram of ranging principle

Equipment A sends a signal to equipment B and records the value of the moment. B receives the signal and records the value of the moment. Then it sends a message to A and records the value of the moment. A receives the signal and records the value of the moment. A sends the signal to B again and records the value of the moment. B receives the signal and records the value of the moment. A and B record the value of the moment three times respectively.

3.3 Location Information Extraction

GPS chip communicates with MCU through serial port 2. As long as the GPS chip is powered on, the GPS chip will continuously transmit data to MCU in NMEA-0183 format, and the amount of information is very large. The main task of GPS software design is to analyze NMEA-0183 protocol, and then extract the required data according to the needs. First, judge whether the GPS information is available, that is to say, judge whether the GPS chip has searched more than four satellites and can accurately locate them; then write the corresponding program algorithm to extract the current longitude, latitude and speed information of the module, process these information, then extract and process them again, so that the vehicle status can be monitored all the time.

GSM / GPRS chip communicates with SCM through serial port 1. SCM can control GSM / GPRS chip by following the format of at command.
through serial port. So far, the design of the wireless communication and positioning system of the electric vehicle based on CAN bus is completed. The following simulation experiments are designed to verify the performance of the designed system.

4 System Function Testing

In order to verify the performance of the wireless communication positioning system, a comparative experiment is proposed to simulate the operation process of the designed system, compare its throughput under different traffic loads with the original system, get the test results, and complete the function test.

4.1 Performance test preparation process

During the experiment, the throughput of the system needs to be detected, which is completed by using Tamosoft Throughput test performance test software. Tamosoft throughput test can simulate TCP and UDP data flow in real time, calculate important indicators, and generate test results in the form of numbers or charts. The specific parameters of the experimental platform are shown in the table 1 below:

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Experimental Platform Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Specific parameters</td>
</tr>
<tr>
<td>subject</td>
<td>Model: DT-610L-JH61MAI</td>
</tr>
<tr>
<td>structure</td>
<td>Rack</td>
</tr>
<tr>
<td>Memory type</td>
<td>Non-ECC</td>
</tr>
<tr>
<td>Memory capacity</td>
<td>64 G</td>
</tr>
<tr>
<td>Number of memory slots</td>
<td>4</td>
</tr>
<tr>
<td>Memory size</td>
<td>4 G</td>
</tr>
<tr>
<td>Display chip</td>
<td>C236</td>
</tr>
<tr>
<td>network controller</td>
<td>Gigabit Ethernet</td>
</tr>
<tr>
<td>Expansion slot</td>
<td>PCIE</td>
</tr>
<tr>
<td>mainboard</td>
<td>Chipset: C236</td>
</tr>
<tr>
<td>Embedded network controller</td>
<td>Gigabit Ethernet</td>
</tr>
<tr>
<td>storage</td>
<td>Number of internal hard disks: 4</td>
</tr>
<tr>
<td>Disk array card</td>
<td>R121i supports raid0, 1, 5, 10</td>
</tr>
</tbody>
</table>

647
<table>
<thead>
<tr>
<th></th>
<th>CD drive</th>
<th>DVD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard disk type</td>
<td>SATA; SSD; hybrid hard disk</td>
<td>SATA; SSD; hybrid hard disk</td>
</tr>
</tbody>
</table>

From the data in the above table, it can be seen that the experimental platform can meet the test requirements, input the original values in the operation interface, test the system performance, in order to ensure the accuracy and preciseness of the test results, other experimental variables are the same except for the experimental objects.

4.2 Performance Test Results Analysis

After completing the preparation of the above experiments, the results obtained are compared with the original system, and the specific control results are shown in the following figure:

![Figure 7 Test results diagram](image)

It can be seen from the above results that when the traffic load is kept at 0–45 Mbps, the throughput of the original system is almost the same as that of the designed system, and the network utilization increases with the increase of network traffic. However, when the traffic load increases to 45mbps or higher, the network congestion is caused by too many packets in the network, resulting in the gradual change of the network throughput curve of the original system, showing a downward trend, which proves that when the network traffic load is too large, the original system can not maintain normal operation and can not meet the user's resource demand. The designed system throughput curve changes smoothly, does not cause network congestion.
congestion due to the increase of traffic load, and can meet the needs of users.

5 Concluding remarks

Because the throughput of the traditional wireless communication positioning system of EVs is reduced, which cannot meet the needs of users, this paper designs the wireless communication positioning system of EVs based on CAN bus. Through the design of hardware and software, the design of wireless communication positioning system is completed, and the performance of wireless communication positioning system is optimized. It is hoped that this research can contribute to the development of the electric vehicle. In the research process of the wireless communication positioning system of the electric vehicle based on the CAN bus, the long positioning time leads to the reduction of the efficiency of the wireless communication positioning of the electric vehicle based on the CAN bus. Therefore, in the future research, the positioning time is the research focus to further improve the effectiveness of this system.

6 Fund projects

"Challenge Cup" support project of science and technology innovation fund for college students of Nanjing University of Engineering in 2020 (tz20200007)

References


Remote monitoring system for embedded network equipment faults based on Internet of Things

Wei-xue LIU
{tiantiankaixin255@tom.com}
(Collage of Mathematics and Compute Science, Chifeng University, Chifeng 024000, china)

Abstract: The current remote monitoring system for network equipment faults mainly uses three technologies: web, Internert, and CAN. Although these three systems meet the needs of the monitoring scope. However, when collecting data, due to insufficient information and simple analysis of the data, the accuracy of system monitoring is poor. To this end, a remote monitoring system for embedded network equipment faults based on the Internet of things is studied. Based on the requirements analysis and combined with the industrial Internet of Things technology, the overall scheme of the system is given, and the system hardware design is carried out from the three aspects of the collector, communication equipment and host, and the corresponding software is designed in detail based on the three main system hardware. The results show that compared with the three remote monitoring systems for network equipment faults based on web, Internert, and CAN, the system has higher monitoring accuracy while ensuring the monitoring range, which indicates that the system has stronger comprehensive monitoring performance.

Key words: Internet of things; Internet equipment; Fault monitoring; system design

1 Introduction

With the rapid development of social science and technology, research and innovation in various fields are constantly emerging, and knowledge has grown rapidly in the form of explosion. Some of the main characteristics of the information age are networking, digitalization, and informationization. Because the characteristics of the network can transmit information very quickly, the realization of informatization must rely on a perfect network [1]. The Internet has now developed into the lifeblood of today's information society and a key foundation for the development of the knowledge economy, and has had a significant impact on all aspects of social life, especially social and economic development. Since the 1990s, computer networks, typified by the Internet, have developed rapidly. From the initial scientific research and education network to a commercial network, it has gradually changed all aspects of our work and life and accelerated global informationization. The process of the revolution,
now people's work, life, study and communication can not leave the Internet. In this context, more and more devices are connected to the network, and the types of network devices are increasing. The current basic network devices are: computers, hubs, switches, bridges, routers, gateways, network interface cards (NICs), wireless Access points (WAP, printers and modems, fiber optic transceivers, fiber optic cables, etc., so failures are unavoidable, which brings a lot of inconvenience to people's lives and work [2]. Network equipment failure monitoring is of great significance to the long-term normal operation of communication systems, and to social production and life, as well as national defense security. In some special environments such as military weapon systems, the reliability of network systems is particularly important. Especially during the execution of certain critical tasks, the occurrence of network failures may lead to the loss of important information and even cause mission failure. Fault prediction for network equipment is of great significance for improving the reliability of network systems, ensuring the smooth execution of key tasks, and reducing the burden of network management and maintenance.

Currently, network equipment fault monitoring systems mainly include three types of network equipment fault remote monitoring systems based on web, Internet, and CAN. However, although the monitoring scope of these three systems meets requirements, the monitoring performance is insufficient. Aiming at the above problems, a technical solution of the remote monitoring system for embedded network equipment faults based on the Internet of things is proposed. In the absence of monitoring, notify the network administrator of the faulty device name and other information in the form of a mobile phone message in real time, so that the network administrator can detect the network fault in a timely manner and repair the fault in a short time to ensure network smoothness and data security [3]. The test results show that: compared with the fault remote monitoring system based on web, Internet and CAN, the system has better monitoring performance, a wider monitoring range, and has reached the expected goals of the research.

2 Design of remote monitoring system for embedded network equipment faults

With the rapid development of the current Internet of Things technology, related equipment has entered the era of the Internet of things, making the application rate of different intelligent IoT machinery and equipment continue to increase. The normal operation of the network center equipment is the basic condition to ensure the unit's Internet of things access. Sudden failures in the network equipment during operation will cause great losses if it is not repaired in time. Therefore, effective methods are sought to ensure the normality of the equipment Operation has become a hotspot for related personnel analysis [4]. For this reason,
designing a reasonable intelligent IoT equipment fault monitoring system has high application significance.

2.1 System framework

Industrial Internet of Things technology can help enterprises to establish a complete network architecture and realize the monitoring and management of the entire production process of equipment. At present, the industrial Internet of Things technology is relatively mature, and there are successful cases in other industries. Therefore, it is feasible to apply the Industrial Internet of Things technology to the equipment digital monitoring industry [5]. The Industrial Internet of things model is divided into four levels. This section divides the hierarchical structure of the equipment monitoring management system based on the industrial Internet of things model and system requirements. It is divided into physical layer, interconnection layer, network layer, data layer, service layer, application layer, as shown in Figure 1.

**Fig. 1 IoT technology hierarchy**

Physical layer: Mainly refers to the underlying physical equipment and perception system. During the operation of network equipment, its operating data will be stored by the equipment's PLC, so the physical equipment and sensing system here mainly refers to the PLC of the machine tool equipment [6].

Interconnection layer: It mainly refers to the interconnection communication system of

---

Physical layer:

- Equipment monitoring management information system
- Equipment monitoring database
- Equipment interconnection communication system
- Physical layer: equipment1, equipment2, equipment3

Interconnection layer:

- The server
- Equipment monitoring service system

Network layer:

- Internet, fieldbus, industrial Ethernet, wireless network

Data layer:

- Equipment monitoring database

Service layer:

- Application layer

Application layer:

- Equipment monitoring management information system
- Equipment monitoring service system
network equipment. It is responsible for the interconnection between equipment and the interconnection between the equipment and the network layer, and collects the equipment operation data in the equipment PLC, and sends these data to the network layer.

Network layer: It refers to the network equipment monitoring network system, which is a communication bridge between the systems, corresponding to various network systems in the production environment. In the factory, the production data collected by the interconnection layer needs to be passed upwards, and at the same time, the operation instructions of each device need to be issued to the interconnection layer. Due to the strong functionality of industrial Ethernet, it is suitable for various complex environments. Industrial Ethernet is mainly used as a device monitoring network system [7].

Data layer: Mainly refers to the network equipment monitoring database to realize the classified storage of real-time operating data of the equipment.

Service layer: Refers to the network equipment monitoring service system, which summarizes and organizes equipment operation data to meet the requirements of enterprises to implement digital monitoring.

Application layer: Network equipment monitoring and management information system, through the analysis of network equipment operation data, the overall control and optimization of operation.

Based on the analysis of the composition of the Internet of things technology, the architecture of the device monitoring and management system is designed. As shown in Figure 2, it is mainly divided into four levels: the device interconnection acquisition layer, the network transmission layer, the data storage layer, and the application service layer.

2.2 System hardware

Fig.2 System framework hierarchy
In terms of physical layout, the system adopts a split design, which is divided into a host and a collector. This design can prevent the signal from being shielded when the system uses wireless transmission. The host implements the core functions of the front-end monitoring system, which can be installed in network equipment, and the collector is responsible for data collection [8]. The host and the collector can use GPRS communication. The hardware module diagram of the front-end monitoring system is shown in Figure 3.

![Fig.3 Hardware module diagram of the front-end monitoring system](image)

(1) Collector
In this paper, the system data acquisition module is mainly responsible for collecting real-time parameters of the faulty equipment in the IoT environment, and detecting the system status of the faulty equipment.

![Fig.4 Circuit physical diagram of the data acquisition module](image)

(1 is communication line, 2 is LF2407A chip, 3 is converter)

The data acquisition module mainly includes sensors, power adapters, wireless communication modules, equipment fault data storage modules, A / D converters and communication interfaces. The sensor is used to collect the data of the faulty device, the
power adapter provides energy for the overall data acquisition module, and the wireless communication module implements the wireless transmission of data through the IEEE 802.15.4 protocol. The data acquisition module performs operations such as filtering, amplifying, sampling / holding, and A / D conversion on the signals sent from the field of the faulty device through sensors, and stores the data of the faulty device in the device fault data storage module. Device information is passed between modules [9].

In order to ensure the utilization of the server, generally one acquisition device is connected to 3-5 network devices for acquisition, and the devices are mainly distinguished according to the IP number and port number set in the device to achieve the correspondence between the device and the operating data. In order to ensure the reliability of data collection and upload of the equipment in some unexpected situations, a backup server is configured for each server. When the collection server fails to work normally, the system will automatically call the backup server for jobs.

(2) GPRS communication equipment

The system of this paper realizes the efficient transmission of faulty device data in the Internet of things through the communication network module. The communication network module is the bridge of the overall system and the link between the data acquisition module and the detection and diagnosis module. It plays an important role. The communication module is mainly composed of a SIM900B controller, a radio frequency module, a power module, a memory, and an application interface. Its advantages are: providing voice, data, SMS fax and other services; built-in TCP / IP protocol stack; low power consumption, only 2.0ma current in sleep state. The core controller of the GPRS terminal communication module comes from SIMCom's wireless quad-band SIM900B controller, which can send data information with the lowest power consumption, as shown in Figure 5.

![Fig.5 M900B controller](image)

SIM900B controller characteristic parameters are shown in Table 1.
Table 1 SIM900B controller characteristic parameters

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Explain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual frequency</td>
<td>Support both GSM and GPRS, working frequency 850 / 900 / 1800MHz</td>
</tr>
<tr>
<td>GPRS rate</td>
<td>GPRS downlink transmission, maximum 85.6kbps</td>
</tr>
<tr>
<td>Serial port</td>
<td>Standard serial port, transmission rate 1.2k-115.2kbps</td>
</tr>
<tr>
<td>Interface control</td>
<td>Support mobile and Unicom SIM card</td>
</tr>
<tr>
<td>function</td>
<td>At command control</td>
</tr>
<tr>
<td>power supply</td>
<td>3.1-4.8VDC</td>
</tr>
<tr>
<td>working temperature</td>
<td>-40-85 °C</td>
</tr>
</tbody>
</table>

(3) Host

The host is at the center of the entire monitoring system, and maintains a close relationship with the sensor nodes. It can interact with the user through a WEB browser. The host can receive the data collected by the sensor and save the data in the host's database. At the same time, the host can provide a communication interface based on the HTTP protocol. The user can send control instructions to the host through these interfaces. It can query the corresponding data stored in the ECS database according to the instructions, or feed back to the sensor node according to the instructions sent by the web end, so as to control the sensor node of the lower computer.

In this article's remote monitoring system, we adopted the cloud Elastic Compute Service (ECS) and Open Storage Service (OSS) provided by Alibaba Cloud to jointly constitute the host used by the monitoring system. The ECS cloud server handles storing data, and OSS provides auxiliary storage space, which can be customized according to user needs. Figure 6 below shows the structure of the host system.
2.3 System software design

(1) Acquisition program

The specific process of collection is shown in Figure 7.

Fig. 6 Structure of the host system of this system

Fig. 7 Acquisition program flow
First, when the device starts running, the acquisition program will use the host as a client to establish a connection with the collector. After the connection is established, the heartbeat packet of the acquisition program will periodically detect the connection status to ensure the reliability of the connection. Secondly, the collection function in the collection program will be called to collect equipment operation data. There are two main collection functions, which are used to correspond to the device register type and the device coil type. Then, the acquisition program will judge the length of the data collected at the same time. Finally, the collected data will be filled into the message in order and transmitted to the host to realize the collection of equipment operation data.

(2) Communication program

The network access device based on SIM900A establishes a communication connection with the sensor node through the XBEE coordinator. If the XBEE module uses the AT mode, it will lack a lot of information. Here, the API frame mode is used for communication connection. The SIM900A module's GPRS service supports TCP/IP protocol, using TCP/IP protocol to establish network communication [10].

According to the function of sim900a based network access device, the application program of network access device is designed by Arduino software. The specific process is as follows:

Power on the SIM900A-based network access device and perform a series of initialization. Use the XBEE coordinator to establish a GPRS network, search for node devices in the network, and send handshake information # HELLO $ to the node. Determine whether the sensor node is connected to the coordinator by receiving feedback information # RECEIVES. If it is connected, establish a node communication list, otherwise continue testing until a device is connected to the network. Send a AT command to the SIM900A module with a single chip microcomputer to establish a TCP/IP connection with the cloud server. Based on the connection status, determine whether SIM900A has established a good communication connection with the cloud server. If not, apply for a TCP/IP connection again. The single-chip microcomputer waits to receive data. If the serial port connected to SIM900A triggers an interrupt, it means that data is sent from the host server. The single-chip microcomputer analyzes the received command, changes related parameters according to the command, and then looks up the communication list to pass the control command through the XBEE coordinator. Send to the sensor node for feedback control operation, the microcontroller continues to wait to receive data. If the serial port connected to the XBEE coordinator triggers an interrupt, it means that data has been uploaded from the sensor node. The microcontroller first analyzes the received data to determine whether the collected data exceeds the set
threshold. If not, the data is normal. The MCU directly encapsulates the data and uploads it to the cloud server through SIM900A. If the data exceeds the threshold, it indicates that an exception has occurred. The MCU sends an AT instruction to the SIM900A module through the serial port, and sends the exception information to the designated mobile phone. Then send the corresponding feedback instruction to the sensor node through the XBEE coordinator, and then encapsulate the data through the TCP connection of SIM900A, upload it to a specific port on the host server, and realize the data upload.

(3) Host software architecture

The software architecture of this host is shown in Figure 8. Based on the database, the system connects with the network access equipment through the transceiver program on the server, receives the data collected by the sensor node, stores it in the database, and sends the control instruction to the sensor node. The WEB server provides a functional interface to facilitate Users interact with the cloud platform.

![Fig.8 Host software architecture](image)

**3 System simulation test**

This paper designs a remote monitoring system for embedded network equipment faults based on the Internet of things. In order to prove that this system has the characteristics of
simple structure, easy networking, remote monitoring, cross-platformity, and user-friendly operation, experiments are performed below to verify.

### 3.1 Monitoring arrangement

The experiment monitored four networks located in different rooms. The specific monitoring locations are shown in Figure 9 below.

![Monitoring arrangement](image)

**Fig.9 Monitoring arrangement**

Nodes 1, 4 and the gateway are placed in Room A on the north side, node 2 is placed in Room B on the south side, and node 3 is placed in Room C. Each node has the same composition and is used to monitor the operating status of the equipment.

### 3.2 System operating environment

This system test is based on the premise that the device is not installed in the network device, according to the operation of the network device, the normal operation and failure of the network device are simulated, and implemented through a variety of network communication methods to achieve the test. The specific test environment is as follows:

- **Server:** Windows Server 2003 32bit;
- **Client:** Windows XP Professional 32bit, Windows7 Ultimate 32bit;
- **Browser:** IE8.0, IE9.0, Google Chrome;
- **High and low temperature:** YSL-GDW-010 high and low temperature test chamber -50~80 °C;
- **Web environment:** Wired network, China Unicom 3G network, mobile 3G network, telecommunication 3G network;
- **Front-end equipment:** Hi3515 development board, infrared sensor, level sensor, door status sensor, headset, camera.

### 3.3 experiment procedure

The network equipment monitoring experiment mainly includes four parts: WEB-based application, host, gateway, and lower-level sensor nodes. The collector transmits the collected
information to the host, and after the initial processing by the microcontroller, it sends it to the X4 gateway through the XBEE module. The gateway forwards the data to the host and stores it in the SQL database. The collector transmits the collected information to the host, and after the initial processing by the microcontroller, it sends it to the X4 gateway through the XBEE module. The gateway forwards the data to the host and stores it in the SQL database. The control instructions are forwarded to the destination address through the gateway, so as to control the sensor devices and execution devices of the lower computer to achieve the purpose of remote monitoring.

3.4 Test Results

(1) System monitoring performance

In order to diagnose the different fault characteristics of the network equipment, four kinds of faults are set on the network equipment, and each kind of fault is deployed according to FIG. 6. The experiment is repeated 10 times.

<table>
<thead>
<tr>
<th>project</th>
<th>Monitoring accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet of things</td>
<td>96.58</td>
</tr>
<tr>
<td>web</td>
<td>92.55</td>
</tr>
<tr>
<td>Internert</td>
<td>88.14</td>
</tr>
<tr>
<td>CAN</td>
<td>90.35</td>
</tr>
</tbody>
</table>

From the experimental results in Table 2, it can be seen that compared with the remote fault monitoring system based on web, Internert, and CAN, the system designed in this paper has higher monitoring accuracy, which shows that the system has higher fault identification capabilities. And robustness, which greatly enhances the effectiveness and accuracy of equipment fault monitoring.

(2) Monitoring scope

The monitoring range refers to the maximum range that the monitoring system can cover. Because the monitoring range is circular, it is generally judged by the radius. Table 3 below shows the monitoring ranges of the four monitoring systems.

<table>
<thead>
<tr>
<th>project</th>
<th>Monitoring range (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet of things</td>
<td>12.36</td>
</tr>
<tr>
<td>web</td>
<td>8.69</td>
</tr>
<tr>
<td>Internert</td>
<td>9.58</td>
</tr>
<tr>
<td>CAN</td>
<td>10.2</td>
</tr>
</tbody>
</table>

From the experimental results in Table 3, it can be seen that the monitoring radius of the system in this paper is not much different from the monitoring radius of the three network
equipment fault remote monitoring systems based on web, Internert, and CAN. It can be seen that the monitoring scope of this system meets the needs.

4 Concluding remarks

Network environment equipment fault diagnosis and protection technology application decision has become an important research topic for network environment operation and maintenance. However, these fault diagnosis systems continue to improve the diagnostic capabilities and expand the scope of diagnosis, but also exposed in the actual operating environment. Some problems occurred, such as excessively high false alarm rates, alarm floods, poor correlation between alarms, and poor monitoring performance of the system itself. In view of the above problems, this paper designs a remote monitoring system for embedded network equipment faults based on the Internet of things. After testing, the system has higher monitoring accuracy and a wider monitoring range, which effectively ensures the stable operation of network equipment. However, the monitoring time is not considered in the research process, which leads to the decrease of remote monitoring efficiency of embedded network equipment fault. Therefore, in the next research, we will focus on the research of monitoring time to improve the efficiency of the system.

5 Fund projects

Major Project of the Ministry of Education: "Research on the Construction of Ethnic Disciplines and Majors in Ethnic Areas: Cases of Inner Mongolia and Northeast Ethnic Universities" (number: mjzxzd1405)

References


Embedded intelligent vacuum monitoring method of high voltage circuit breaker based on Internet of things

LIU Wei-xue
{tiantiankaixin255@tom.com}
(Collage of Mathematics and Compute Science, Chifeng University, Chifeng 024000, China)

Abstract. A certain degree of vacuum is the premise to ensure that the intelligent high-voltage circuit breaker will extinguish the arc during operation, and the degree of vacuum will drop in operation due to some reasons, so it needs to be monitored frequently. In view of the low accuracy of the traditional embedded intelligent high-voltage circuit breaker vacuum monitoring method, the embedded intelligent high-voltage circuit breaker vacuum monitoring method based on the Internet of things is proposed. Based on the analysis of the principle of embedded intelligent high-voltage circuit breaker vacuum degree monitoring, using the Internet of things sensor technology, the embedded intelligent high-voltage circuit breaker vacuum degree monitoring. The experimental results show that the proposed method has high accuracy and good monitoring effect, which is of great significance to the construction of smart grid.

Keywords: Internet of things technology; Embedded system; Intelligent high voltage circuit breaker; Vacuum degree;

1 Introduction

The high-voltage circuit breaker has dual functions of control and protection in the power system. It is one of the equipment with the largest maintenance workload in the power system equipment. Its overall operation, maintenance and repair are closely related to the whole power system. Its reliable breaking capacity is indispensable for the construction of a safe and reliable smart grid, and plays an important role in the power system [1]. In recent years, because of its unique advantages, vacuum circuit breaker has been gradually applied in distribution network at home and abroad. In Baoding area, more than 90% of the circuit breakers used in 10kV distribution network are vacuum circuit breakers, and vacuum circuit breakers are used in new and reconstructed distribution network projects. The main performance index of vacuum circuit breaker is the vacuum degree in the arc extinguishing chamber. The vacuum value directly affects the arc extinguishing effect, and then affects the breaking capacity of the switch [2]. If the vacuum degree changes to a certain critical value, the normal arc extinguishing function cannot be completed, and the load current cannot be cut off, which will directly damage the protected lines and equipment; it will also cause power grid accidents, such as large-area power outage, affect
people's normal life order, and cause great economic losses. The on-line monitoring system of vacuum degree of high-voltage circuit breaker can know the working state and defect position of circuit breaker in time, reduce the premature or unnecessary power failure test and maintenance, reduce the maintenance workload, reduce the maintenance cost, improve the pertinence of maintenance, and significantly improve the reliability of power system. Therefore, through the on-line monitoring of the vacuum degree parameters of the embedded intelligent high-voltage circuit breaker, further reasonable fault analysis can provide practical guidance for the safe, reliable and economic operation of the power system [3-4].

At present, domestic and foreign research in this area has made some achievements. In reference [5], a vacuum degree detection system based on panning discharge principle is proposed, and a complete on-line vacuum degree detection system of high-voltage vacuum circuit breaker is designed. The system is mainly composed of power supply system and signal acquisition and processing system, which realizes the power supply of vacuum sensor and other equipment. Under the environment of high-voltage and strong electromagnetic field, it can measure the vacuum degree signal with high precision and directly detect the vacuum degree of vacuum circuit breaker online; based on the analysis of on-line monitoring requirements of medium voltage vacuum circuit breaker, a general design framework of on-line monitoring system for medium voltage vacuum circuit breaker is designed in reference [6], and several key technical problems in the process of software and hardware design that need to be solved in the process of on-line monitoring of circuit breaker are introduced in detail. The on-line monitoring system developed according to the technical scheme in this paper has been applied to many products Successful application, perfect function, good performance and high reliability. But the accuracy of the above two methods is low.

In view of the above problems, this paper proposes an embedded intelligent vacuum monitoring method based on the Internet of things.

2 Internet of things technology

The Internet of things refers to the real-time collection of any object or process that needs to be monitored, connected and interacted through sensors, RFID technology, global positioning system and other technologies, as well as the collection of all kinds of required information such as sound, light, heat, electricity, mechanics, chemistry, biology, location, etc. through various possible network access, it can realize the ubiquitous link between things and things, things and people, and realize the connection between objects and people Intelligent perception, recognition and management of process. Internet of things technology is an extension and extension of Internet technology. Through Internet of things technology, the equipment is automatically and real-time identified, located, tracked, monitored and triggered corresponding events. It is important to carry
out wireless transmission of operation data and monitoring instructions to achieve real-time management and control of the equipment. By using the Internet of things technology, through the perception of the outside world, an intelligent vacuum monitoring network of high-voltage circuit breakers is constructed to comprehensively monitor the factors affecting the operation of high-voltage circuit breakers [7-9]. The application of RFID, ZigBee and other Internet of things technology to the design of high-voltage circuit breaker not only makes it have the function of traditional high-voltage circuit breaker, but also realizes identification, monitoring and control. Finally, the integrated monitoring host follows IEC 61850 protocol and transmits data to the integrated information platform through optical fiber. Its main technical features are: low power consumption, large network capacity, short time delay, reliability, security and so on.

3 Monitoring principle

The embedded intelligent high-voltage circuit breaker relies on the vacuum in the arc extinguishing chamber to complete the arc process. The theory of gas discharge shows that whether the collision ionization and recombination of charged particles in gas are balanced determines whether the gas decomposes due to the discharge phenomenon or has the electrical strength characteristics of insulation. In the environment of constant temperature, electric field and polar distance, vacuum degree is the main reason of gas molecular movement in the arc extinguishing chamber. The larger the free path of charged particles is, the more difficult it is to collide in the process of motion. In high vacuum environment, the collision ionization between charged particles is weakened. At this time, the collision ionization and recombination of charged particles are in equilibrium state [10-12]. The structure of vacuum interrupter is shown in Figure 1.

Fig. 1. Vacuum interrupter structure
When the vacuum degree in the arc extinguishing chamber is high enough, only a small voltage value is needed to maintain the electron flow generated by field emission between the contact and the shield. At this time, the free path of the electron is smaller than the distance between the electrode and the shield. The electron reaches the shield without collision, and the potential on the shield increases as the charge accumulates on the shield. The negative charge accumulated on the shield will make it the negative potential reaches a peak [13-14]. When the vacuum inside the shield deteriorates, the pressure of the gas becomes larger and the particles such as gas molecules become more, which leads to the decrease of the insulation characteristics of the gas and the pre breakdown voltage. It is easy to have pre breakdown and pre discharge, and the potential on the shield will decrease. When the real void is close to the critical value, the travel of electrons will be smaller, and fierce collision will occur before reaching the enclosure wall. The internal gas molecules will absorb the electrons produced by field emission and form heavy and slow drifting negative ions. There is almost no charge accumulation on the enclosure wall [15-16]. The above analysis shows that the potential change on the shield can reflect the internal vacuum of the embedded intelligent high voltage circuit breaker. Figure 2 shows the change curve of AC potential and DC potential of shield.

(a)AC potential
At present, there are many on-line monitoring methods for the vacuum degree of embedded intelligent high-voltage circuit breaker, among which the coupling capacitance method is widely used. The coupling capacitance method is developed on the basis of the potential change detection method. According to the change of the surrounding electric field caused by the change of the potential on the shield, the change of the potential measured by the coupling voltage probe outside the shield wall is used to indirectly obtain the change of the gas pressure inside the arc extinguishing chamber [17]. The main idea is to insert a detection contact between the vacuum bubble cover and the earth, and keep a certain distance between the contact and the vacuum bubble cover, so that the detection contact will not affect the safe and stable operation of the circuit breaker. The distance between the detection contact and the vacuum bubble shield can be calculated according to the theory or the test. The schematic diagram is shown in Figure 3.
Fig. 3. Equivalent capacitance diagram of coupling capacitance method

In Figure 3, C1 is the capacitance between the conducting rod and the shield. The capacitance value has been determined when the switch leaves the factory, and it is very stable as long as the vacuum degree of the arc extinguishing chamber remains unchanged; C2 is the capacitance between the shield and the contact. The capacitance value is basically determined by the area of the contact and the distance between the contact and the shield. In fact, as long as the position of the detection contact is fixed, the capacitance value remains unchanged; C3 is the capacitance value between the contact and the earth. Like C2, the capacitance value is basically determined by the area of the contact and the distance between the contact and the shield. In fact, as long as the position of the detection contact is fixed, the capacitance value remains unchanged. In practice, the capacitance of C1 and C2 is much larger than that of C3. In order to increase the capacitance of C2 as much as possible, solid materials with large dielectric constant are filled between the contact and the shield. In normal operation, three equivalent capacitors are connected in series to divide the voltage. Since the capacitance value of C2 is the largest, the voltage of C2 is the smallest. As long as the air pressure of vacuum bubble changes slightly, the relative value of the voltage change on C2 will be large, which greatly improves the sensitivity of monitoring. In addition, the capacitance value of C2 can be adjusted according to the actual needs, which greatly facilitates the design of online monitoring system. In fact, C2 can be designed as adjustable capacitance to cooperate with the commissioning and sensitivity test of online monitoring system [18-20].

4 Embedded intelligent vacuum monitoring of high voltage circuit breaker based on Internet of things

According to the monitoring principle, a low-cost and low-power short-distance wireless network communication technology is applied to the vacuum degree monitoring of embedded intelligent high-voltage circuit breaker. Through online monitoring of the vacuum degree parameters of embedded intelligent high-voltage circuit breaker, the fault analysis of high-voltage circuit breaker is realized [21-22]. The CC2530 chip produced by TI company is selected for wireless communication control chip, which is based on the wireless technology of pressure ee802.15.4 protocol. The chip of ZigBee technology integrates the enhanced 8051 single-chip microcomputer core, which can directly write programs. The chip model used is cc2530f256. The structure size is small, the development board size is reduced, the cost is reduced, and the receiving signal is sensitive, With strong anti-interference ability and few external components, the direct communication distance between hardware nodes can reach more than 100 meters by using a suitable and simple main circuit [23-25]. The Internet of things needs to continuously sense the vacuum degree data of high-voltage circuit breaker according to multiple sensor nodes, because the vacuum degree data of
high-voltage circuit breaker received by sensor nodes has the disadvantages of complex, variable, numerous and poor repeatability, which has great redundancy in the time and space of vacuum degree data detection of high-voltage circuit breaker. Therefore, based on the analysis of the network boundary effect, using the Internet of things technology, the vacuum data flow of high-voltage circuit breakers can fundamentally reduce the redundant vacuum data of high-voltage circuit breakers, and reduce the energy consumption of sensor nodes. In this paper, the short-range wireless network communication technology, according to the basic parameter replication mechanism, can fundamentally reduce the generation of redundant high-voltage circuit breaker vacuum data, and can more effectively reduce the errors in the calculation process.

If a node in a cluster in this sensor network is $S_0, S_1, ..., S_n$, then the sequence instruction received by the $j$ node is $C_{j,vol}$, and the basic form of the sequence is $C_{j,vol} = C_{j,0} + C_{j,1} + ... + C_{j,n-1}$.

In the sequence matrix, each sensor node can identify $n$ high-voltage circuit breaker vacuum data points, in which $C_{j,n}$ represents the high-voltage circuit breaker vacuum data in a node.

According to all sensor nodes containing the vacuum degree data of high-voltage circuit breaker, the basic sequence matrix $n \times N$ of vacuum degree data $C^a$ of high-voltage circuit breaker can be constructed:

$$C^0 = \begin{pmatrix}
C_{0,0} & C_{0,1} & ... & C_{0,n-1} \\
C_{1,0} & C_{1,1} & ... & C_{1,n-1} \\
... & ... & ... & ... \\
C_{n,0} & C_{n,1} & ... & C_{n,n}
\end{pmatrix} = \begin{pmatrix}
C_{0,0} + C_{1,0} + ... + C_{n,0} \\
C_{0,1} + C_{1,1} + ... + C_{n,1} \\
... + ... + ... + ...
\end{pmatrix} \quad (1)
$$

Among them, the matrix presents the instructions received by any sensor node, specifically the vacuum data flow of the $n$-th high-voltage circuit breaker over time; the row of the matrix represents the vacuum data of the $n$-th high-voltage circuit breaker of many nodes at a certain time. On this basis, the column and row transformation for equation (1) refers to the transformation of time $S$ and space $d$, and the transformation formula is as follows:

$$C^t = \begin{pmatrix}
S^t_{0,0} & S^t_{0,1} & ... & S^t_{0,n-1} \\
... & ... & ... & ... \\
S^t_{n-1,0} & S^t_{n-1,1} & ... & S^t_{n-1,n-1} \\
d^t_{0,0} & d^t_{0,1} & ... & d^t_{0,n-1} \\
... & ... & ... & ... \\
d^t_{n-1,0} & d^t_{n-1,1} & ... & d^t_{n-1,n-1}
\end{pmatrix} \quad (2)$$
Where, \( C^0 \) represents the vacuum degree data of the transformed high-voltage circuit breaker. The algorithm monitors the data transmission time point, because there will be a lot of data reduction in the process of data transformation, so this step is also to reduce the energy consumption of node transmission in the sequence matrix row transformation. The row transformation formula is:

\[
C_i^0 = \begin{bmatrix}
S_{i0}^0 & S_{i1}^0 & \ldots & S_{iN}^0 \\
\vdots & \vdots & \ddots & \vdots \\
S_{i0}^N & S_{i1}^N & \ldots & S_{iN}^N \\
S_{i0}^N & S_{i1}^N & \ldots & S_{iN}^N
\end{bmatrix}
\]

(3)

Where, \( C_i^0 \) refers to the vacuum degree data of high-voltage circuit breaker after row transformation. Compared with the data flow of vacuum degree of high-voltage circuit breaker initially received by nodes, the amount of redundant vacuum degree data of high-voltage circuit breaker is much less, and more free time space is also released. After transformation, according to formula (3), we can further get:

\[
C_i = \begin{bmatrix}
S_{i0} & S_{i1} & \ldots & S_{iN} \\
\vdots & \vdots & \ddots & \vdots \\
S_{i0} & S_{i1} & \ldots & S_{iN} \\
S_{i0} & S_{i1} & \ldots & S_{iN}
\end{bmatrix}
\]

(4)

In the formula, \( C_i \) represents the vacuum degree data of high-voltage circuit breaker after time and space release. According to the same way above, the data of vacuum degree of high-voltage circuit breaker is transformed. On the basis of the data change of vacuum degree of high-voltage circuit breaker, the corresponding recursive processing is done, and the wavelet coefficients of different degrees are obtained. Finally, the wavelet coefficient is compared with the original vacuum degree data of high-voltage circuit breaker. According to the comparison results, the transformed vacuum degree data of high-voltage circuit breaker greatly reduces the pressure of time and space to a certain extent, which is very important for the current energy limited Internet of things.

According to the wavelet function, a distributed transform is designed, which is suitable for any transform support length. By removing the spatial-temporal correlation of the vacuum degree data of the high-voltage circuit breaker in each node of the ring, the redundant phenomenon of the vacuum degree data of the high-voltage circuit breaker can be reduced to a certain extent, and then the transmission efficiency of the vacuum degree data of the high-voltage circuit breaker between each node of the ring and the cluster head node can be realized.
Considering that any cluster in the Internet of things uses \( N \) node: \( S_0, S_1, \ldots, S_i, \ldots, S_{N-1} \) to construct a ring, assuming that the vacuum degree data of the high-voltage circuit breaker stored in the \( y \) node \( S_y \) of the ring is represented by \( c_y \), then the vacuum degree data of the high-voltage circuit breaker collected by this cluster can be represented as a whole vector set: \( d = (c_0, c_1, \ldots, c_{N-1}) \). This vector set needs to pay attention to the difference between \( d \) and other ordinary vectors. In the ring model, the 0th element \( c_0 \) of vector \( d \) is fundamentally adjacent to the last element \( c_{N-1} \).

The adjacent nodes in the ring belong to the virtual network adjacent to each other in space, so the vacuum degree data of high-voltage circuit breaker stored in the adjacent nodes in the ring may have higher spatial correlation. On the basis of the distribution transformation of the ring model, the calculation of each low-frequency or high-frequency coefficient is aimed at an adjacent node of the support length, so the spatial correlation can be used to eliminate the redundant vacuum data.

The effective way to solve the boundary effect is to expand the boundary, and link the sensor nodes into a ring, which is equivalent to the periodic expansion of the signal, so as to solve the problem of boundary effect, and realize the efficient monitoring of vacuum degree data of high-voltage circuit breaker.

**5 Simulation experiment**

In order to verify the validity of the embedded intelligent vacuum monitoring method based on the Internet of things, a simulation experiment is carried out. The experimental environment is shown in Table 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating system</td>
<td>Microsoft Windows XP</td>
</tr>
<tr>
<td>Processor</td>
<td>Intel(R)Celeron(R) 2.6GHz</td>
</tr>
<tr>
<td>Internal storage</td>
<td>6.0 GB</td>
</tr>
</tbody>
</table>
Hard disk 4.0 GB

Database management software Microsoft SQL server 2010 R2

JDK 1.6

Mathematical software MATLAB

Take the high-voltage circuit breaker of 126kV electronic equipment as an example, monitor the vacuum degree of the high-voltage circuit breaker, and build the experimental platform as shown in Figure 4.

Fig. 4. Construction of Experimental Platform

Set parameters of high-voltage circuit breaker in the experimental platform, as shown in Table 2.

<table>
<thead>
<tr>
<th>Order number</th>
<th>Name</th>
<th>Unit</th>
<th>Numerical value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rated voltage Power frequency withstand voltage Rated</td>
<td>kV</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>kV</td>
<td>42</td>
</tr>
<tr>
<td>3</td>
<td>lightning impulse</td>
<td>kV</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Unit</td>
<td>Value</td>
</tr>
<tr>
<td>---</td>
<td>------------------------------------------------------------------------------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>4</td>
<td>Rated current</td>
<td>A</td>
<td>1250</td>
</tr>
<tr>
<td>5</td>
<td>Rated short-circuit breaking current</td>
<td>kA</td>
<td>31.5</td>
</tr>
<tr>
<td>6</td>
<td>Rated short-circuit making current (peak value)</td>
<td>kA</td>
<td>80</td>
</tr>
<tr>
<td>7</td>
<td>Rated dynamic stability current (peak value)</td>
<td>kA</td>
<td>80</td>
</tr>
<tr>
<td>8</td>
<td>Rated thermal stability current (effective value)</td>
<td>kA</td>
<td>31.5</td>
</tr>
<tr>
<td>9</td>
<td>Rated short-circuit breaking current breaking times</td>
<td>order</td>
<td>30</td>
</tr>
<tr>
<td>10</td>
<td>Rated thermal stability time</td>
<td>s</td>
<td>4</td>
</tr>
</tbody>
</table>

According to the parameters of the high-voltage circuit breaker, through the vacuum condition in the arc extinguishing chamber of the high-voltage circuit breaker, and by changing the voltage between the breaks of the circuit breaker, the change curve between the breakdown voltage between the breaks and the vacuum degree in the arc extinguishing chamber is obtained, as shown in Figure 5.
It can be seen from the figure that when the pressure in the arc extinguishing chamber is less than $1.33 \times 10^{-2}$Pa, the withstand voltage strength of the arc extinguishing chamber is high and almost unchanged, about 100kV. When the pressure in the arc extinguishing chamber is greater than 0.133Pa, the breakdown voltage of the fracture gradually decreases, and when the voltage is about 0.93pa, the breakdown voltage of the fracture reaches the minimum value, less than 1kV, and then increases with the pressure. The breakdown voltage between ports increases gradually. It shows that this method can effectively monitor the vacuum degree of the embedded intelligent high voltage circuit breaker.

In order to further verify the effectiveness of this method, the monitoring accuracy of this method, literature [5] method and literature [6] method are compared and analyzed, and the comparison results are shown in Figure 6.

According to figure 6, the monitoring accuracy of the embedded intelligent high-voltage circuit breaker vacuum degree in this method can reach up to 100%, which is higher than that of the literature method.

6 Concluding remarks

Due to the low accuracy of the traditional embedded intelligent high-voltage circuit breaker vacuum monitoring method, this paper proposes an embedded intelligent high-voltage circuit...
breaker vacuum monitoring method based on the Internet of things. Based on the principle of embedded intelligent high-voltage circuit breaker vacuum degree monitoring and the Internet of things sensing technology, the embedded intelligent high-voltage circuit breaker vacuum degree is monitored. The experimental results show that this method can accurately monitor the vacuum degree of intelligent high-voltage circuit breaker and promote the construction and development of intelligent power grid.

7 Fund projects

Major Project of the Ministry of Education: "Research on the Construction of Ethnic Disciplines and Majors in Ethnic Areas: Cases of Inner Mongolia and Northeast Ethnic Universities" (number: mjzxzd1405)

References


Parallel Capacitance., no. 99, pp. 1-7 (2017)


Embedded Livestock and Poultry Breeding Environment Data Acquisition Method Based on Internet of Things

XU Xiao-hua¹, CHEN Chang-xi²*
{yg69030@163.com¹,xuxh69850@163.com²*}
(1.Tianjin Agricultural University,Tianjin 300384,China; 2.Tianjin Agricultural University,Tianjin 300384,China)

Abstract: The current method of collecting environmental data of livestock and poultry facility breeding cannot automatically screen out abnormal data, so an embedded method of collecting environmental data of livestock and poultry breeding based on Internet of things was designed. First design topology structure of the Internet of things, make between cluster head nodes can communicate with each other, improve the communication ability of member nodes, introducing the environmental perception technology, including the temperature, humidity, ammonia, hydrogen sulfide and carbon dioxide, thermoelectric effect of relying on the thermocouple temperature sensor acquisition, harmful gas depends on the choice of gas molecules light absorption characteristics, absorb certain wavelengths of light conditions to obtain the corresponding information of harmful gas concentration, the final design of acquisition circuit, data acquisition process, completed based on Internet of embedded data acquisition method of design of livestock and poultry breeding environment. The experimental results show that the designed collection method can summarize abnormal data, and separately list the occurrence time and corresponding items of abnormal data, which provides great convenience for controlling the breeding environment.

Key words: Internet of things; Embedded; Livestock and poultry breeding environment; Data acquisition method;

1 Introduction
China is a large livestock and poultry farming and consumption country. As an important part of agriculture, livestock and poultry farming is of great significance to improving people's diet and living standards, and plays a key role in promoting the rapid development of the national economy. The environment of livestock and poultry breeding is the main factor affecting the growth of livestock and poultry. The environment not only affects the growth of livestock and poultry, but also affects the yield and quality of livestock and poultry [1-2]. With the rise and development of intensive high-density breeding technology, the livestock and
poultry breeding model has gradually changed from a stocking or low-density low-level artificial breeding model to a scientific, large-scale, mechanized breeding facility breeding model. Although China is a large country in the world for livestock and poultry breeding, the informationization of livestock and poultry facility breeding is still weak. At present, the environmental control of livestock and poultry breeding seriously restricts the improvement of livestock and poultry health and production performance. At present, the data collection methods of livestock and poultry breeding facilities can not automatically screen out abnormal data, so an embedded data collection method of livestock and poultry breeding environment based on the Internet of things is designed. First, design the topology of the Internet of Things so that the cluster head nodes can communicate with each other, improve the communication capabilities of member nodes, and introduce environmental sensing technologies, including the sensing of temperature, humidity, ammonia, hydrogen sulfide, and carbon dioxide. The temperature depends on the thermoelectric effect of thermocouple sensor to collect, and the harmful gas depends on the selective absorption characteristics of gas molecules to detect the absorption of a certain wavelength of light to obtain the corresponding concentration information of harmful gas. Finally, a collection circuit is designed to implement the data collection process.

2 Research on Embedded Data Collection Method of Livestock and Poultry Based on Internet of Things

The overall design of the embedded livestock and poultry breeding environment data collection method adopts design concepts such as low power consumption and general equipment ideas. Multi-sensors are added to the breeding environment, and the interaction between the sensors is completed through the Internet of Things. The overall structure of aquaculture environment data collection is shown in the figure:
As can be seen from the above figure, the overall structure of the breeding environment data collection consists of a power supply, a microprocessor, a ZigBee communication module, a temperature and humidity sensor, an illumination sensor, a carbon dioxide sensor, a hydrogen sulfide sensor, and an ammonia sensor. After the composition, the real-time information of the environmental parameters is displayed on the LCD display[3-4]. Next, a comprehensive study was conducted on the collection methods of livestock and poultry environmental data.

2.1 Designing the topology of the Internet of Things

The topology of the IoT network refers to the networking structure of ZigBee nodes, which directly affects the network's working performance. Therefore, it is necessary to select a specific ZigBee network topology for different applications and functional characteristics. The existing ZigBee network topology can be divided into flat network, hierarchical network, hybrid network and Mesh network according to the functions of the nodes [5]. This article mainly uses ZigBee's hierarchical network as the field network structure for breeding environment data collection, as shown in the following figure:
Hierarchical network structure is an extension of the traditional flat network structure, and the peer network nodes are divided into cluster head nodes with data aggregation and network organization capabilities and member nodes with data aggregation and forwarding functions only. In a hierarchical network, only the cluster head nodes have perfect layer and network layer protocols, which determines that only the cluster head nodes can communicate with each other, and the member nodes can only collect data and communicate with their cluster head nodes. Hierarchical network improves the routing structure and management efficiency, but increases the cost of hardware [6]. Another disadvantage that cannot be ignored is the poor communication ability of the member nodes, which reduces the robustness of the network. Based on the above analysis, the data transmission from the sensing data collection site to the remote data center is completed in the transmission mode of ZigBee and GPRS relay. The designed topology is shown in the following figure:

Figure 2 ZigBee's hierarchical network structure
**Figure 3** Topological structure of aquaculture environment data collection method

The topology consists of sensor nodes deployed on pallets, turnover boxes, and ZigBee network coordinators installed in storage vehicles and warehouses. The coordinator and the GPRS data transmission unit are designed to connect hardware. Through the server and the intranet application server, the data is released as content to the public and intranet users.

### 2.2 Introduction of environmental awareness technology

Livestock and poultry facilities breeding environment perception technology is a composite technology that integrates multi-disciplines such as electronic technology, opto-electromechanical technology, computer and information processing, and is comprehensively used. It is widely used in data collection for livestock and poultry breeding environment. The sensitivity and stability of the environmental monitoring sensors for livestock and poultry facilities breeding performance directly determine the quality of the environmental data collected from livestock and poultry [7]. Reliable and accurate environmental perception is the basis for subsequent data collection analysis and processing. Temperature sensing is used to detect the cold and hot degree of the livestock and poultry breeding environment. In actual livestock and poultry breeding environment temperature sensing applications, most of them use negative temperature coefficient thermistor sensors:
In the above formula, \( T \) represents the measured temperature, the unit is K; \( T_0 \) in \( T = 273 + t \) represents the reference temperature, the unit is K; \( T_0 = 273 + t_0 \), \( R_t \) and \( R_0 \) represent the resistance of the thermistor at the temperature of \( T \), \( T_0 \), and \( B \) is the thermistor Material constant. The value of \( B \) can be calculated:

\[
B = \ln \left( \frac{R_t}{R_0} \right) 
\]

According to the definition of the temperature coefficient of the thermistor, the relative change in resistance \( \alpha \) can be expressed as:

\[
\alpha = \frac{1}{R_t} \frac{dR_t}{dt} = -\frac{B}{T^2} \quad (3)
\]

Without the need for an external power supply, the thermistor uses a thermocouple temperature sensor to convert the temperature conversion into a potential change sensor. The thermoelectric effect of the thermocouple sensor is shown in the figure:
Figure 4 Thermoelectric effect of thermocouple sensor

In the closed loop shown in the figure above, there are two different metals A and B. If the temperature of metal A and metal B is different, that is, \( T > T_0 \), then a thermoelectric potential \( E(T, T_0) \) will be generated in this closed loop. By detecting the potential, Capable of sensing temperature changes and completing temperature collection.

Another important aspect of environmental perception is the perception of harmful gases. It is necessary to use carbon dioxide sensors, hydrogen sulfide sensors, and ammonia sensors to perform real-time online monitoring of harmful gases in livestock and poultry facilities. At present, electrochemical gas sensors in China are characterized by high measurement accuracy, fast speed, easy use, good environmental applicability, and low price [8]. There are many applications for sensing harmful gases in livestock and poultry facilities. The performance analysis of the infrared carbon dioxide sensor is shown in the following table:

<table>
<thead>
<tr>
<th>Performance indicators</th>
<th>Electric chemical formula</th>
<th>Infrared type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Stability</td>
<td>low</td>
<td>High</td>
</tr>
<tr>
<td>Volume</td>
<td>Large</td>
<td>Small</td>
</tr>
<tr>
<td>Anti-interference</td>
<td>weak</td>
<td>Strong</td>
</tr>
<tr>
<td>ability</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td>Low</td>
<td>slightly higher</td>
</tr>
<tr>
<td>Service life</td>
<td>Short</td>
<td>Long</td>
</tr>
<tr>
<td>Post maintenance</td>
<td>Large</td>
<td>Small</td>
</tr>
</tbody>
</table>

It can be seen from the above table that the basic detection principle of harmful gases is the selective absorption of light by gas molecules, that is, gas molecules can only absorb photons whose energy is exactly equal to the difference between two energies. When a beam of infrared monochromatic light or composite light passes through the gas to be measured, if the molecule of the gas to be measured selectively absorbs the light in some frequency band of the radiation light, the absorption spectrum will be generated; According to the gas molecular structure, the gas absorption spectrum of different molecular structure is different from each other. By detecting the absorption of a certain wavelength of light, the concentration information of the corresponding harmful gas can be obtained qualitatively and quantitatively.

2.3 Design data acquisition circuit
The data acquisition circuit designed in this paper is mainly completed by TD SIM4 100, and its specific circuit is shown in the figure below:

**Figure 5** Interface circuit of TD SIM4 100

TD SIM4 100 is a multi range, 12 bit DAC chip, its working voltage is only 5V. During operation, the circuit can receive analog signals higher than the power supply voltage and lower than the ground potential; the chip has 8 independent analog input channels; four programmable input ranges are provided for the input analog signals: ±10V, ±5V, 0-5V, 0-10V, which increase the effective dynamic input range to 14 bits [9]. The chip provides a flexible interface for 4-20 mA signal and sensor powered by ±12 V or ±15 V to single 5 V system; the chip has 5 MHz bandwidth, 100 kspS throughput; internal / external clock and internal / external start selection, 8 + 4 parallel data interface, internal 4.096 V or external reference voltage; it has low current shutdown mode controlled by software and hardware. Among them, the input and output of 8-bit tristate data I/O port are compatible with TTL or CMOS logic level. Its data conversion process starts from writing the control word, and after the completion of data conversion of TD SIM4 100, a standard notification signal will be given. The system can control its configuration to be query mode or interrupt mode to realize...
data reading. See the following table for specific control conditions:

<table>
<thead>
<tr>
<th>BIP</th>
<th>RNG</th>
<th>Analog input range/V</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0-5</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0-10</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>±5</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>±10</td>
</tr>
</tbody>
</table>

Firstly, the initialization of sc2410 system resources is carried out in the system initialization program of S3C2410, including the I/O mode, the DMA controller of interrupt system, etc. then, the initialization process of GPRS module is completed, and the UART0 of S3C2410 is configured as FIFO mode through the work mode configuration. Then, according to the GPRS communication protocol, use the serial port to send at instruction to initialize GPRS, connect the remote data service, and establish the data path. After the above initialization process is completed, start the data collection process [10-12]. The specific process of collection is shown in the figure below:

In the process of data acquisition, different analog input range and sampling mode are selected according to the characteristics of different channel signals by programming. At the same time, the conversion is started. After the conversion, the 12 bit data is read into memory twice. Finally, when all the 8-way data are converted, the data sending program is called, and
the sampling data is sent to the Internet of things terminal through GPRS module for data processing. So far, the research on the data collection method of embedded livestock and poultry breeding environment based on the Internet of things has been completed.

3 Experiment

In order to verify the effectiveness of this method, we choose a pig house for experimental test. We use the traditional data collection method and the data collection method designed in this paper to collect the environmental data of the pig house at the same time, and compare the collection results.

3.1 Experiment preparation

The tested pigsty is located to the south, with a length of 100m, a width of 15m, a spacing of 35m, bilateral fences, and about 20 domestic pigs in the shed. There are windows on the north and south walls of the pigsty, and two doors on the East and west sides, which play the role of ventilation and lighting, and three sensor nodes are installed in them, as shown in the following figure:

![Figure 7 Pig house sensing node placement](image)

The outdoor temperature is about 25 °C, as shown in the figure above, three sensor nodes are installed in the livestock house, including two bit routing nodes and one coordinator node. The black dot in the figure represents the position of the sensor. Node 1 is installed on the right side of the livestock house door, with a height of 1.5m from the ground. Node 2 is installed on the wall in the middle of the livestock house, with a height of 2m from the ground. Node 3 is installed on the south side of the livestock house. The height of the wall from the ground is 1.5m. The test period of the system is 15 days, and the time interval of data upload is 10 minutes. The server analyzes the data according to the data uploaded by the central node, and then obtains the real-time environment parameters.
3.2 Experimental results and analysis

According to the above experimental process, the embedded livestock and poultry breeding environment data collection results of this method and the embedded livestock and poultry breeding environment data collection results of the traditional method are respectively obtained, as shown in the following figure:

![Figure 8](image) Data acquisition results of this method

![Figure 9](image) Acquisition results of traditional methods

It can be seen from the comparison of the acquisition results of the above two methods that the traditional acquisition method and the acquisition method designed in this paper can collect the temperature, humidity, ammonia value, hydrogen sulfide value and carbon dioxide value in the pig house. But for the abnormal values beyond the normal range, the traditional methods can not be directly reflected; and the collection method designed in this paper can summarize the abnormal data, list the time of abnormal data and the corresponding items separately, which provides great convenience for controlling the breeding environment.

In order to further verify the effectiveness of the method in this paper, the accuracy of data collection of embedded livestock and poultry breeding environment of the traditional...
method and the method in this paper are compared and analyzed, and the comparison results are shown in Figure 10.

![Figure 10 comparison results of acquisition accuracy](image)

According to the analysis of Figure 10, the accuracy rate of the embedded livestock and poultry breeding environment data collection of this method is higher than that of the traditional method.

## 4 Concluding

In order to better control the breeding environment of livestock and poultry and have an intuitive understanding of the abnormal data generated by the breeding environment, an embedded data collection method based on the Internet of things is designed. Firstly, the topological structure of the Internet of things is designed, and the environmental sensing technology is introduced, including the perception of temperature, humidity, ammonia, hydrogen sulfide and carbon dioxide. Finally, the collection circuit is designed to realize the data collection process. The experimental results show that the designed collection method can summarize the abnormal data, and enumerate the occurrence time and corresponding items of the abnormal data separately, which provides great convenience for controlling the breeding environment. In the research of the embedded livestock and poultry breeding environment data collection method based on the Internet of things, because the embedded livestock and poultry breeding environment monitoring is not in place, there are still some deficiencies in the collected embedded livestock and poultry breeding environment data. In the future, we will focus on the embedded livestock and poultry breeding environment monitoring to improve the monitoring accuracy, so as to ensure the integrity of the embedded livestock and poultry breeding environment data.

## 5 Fund projects

Post Scientist of National Broiler Industry Technology System (CARS-41-G25)
The Science and Technology Planning Program of Tianjin Science and Technology Bureau, China (18ZXRHNC00080)
Information Post of Tianjin Dairy Industry Technology System Innovation Team(ITTCSR2020023)

Reference


Deep Learning based OTDOA Positioning for NB-IoT Communication Systems

Guangjin Pan, Tao Wang, Xiufeng Jiang, Shunqing Zhang
guangjin_pan@shu.edu.cn, twang@shu.edu.cn, XiufengJiang@shu.edu.cn
Shanghai Institute for Advanced Communication and Data Science,
Key laboratory of Specialty Fiber Optics and Optical Access Networks,
Shanghai University, Shanghai, 200444, China

Abstract. Positioning is becoming a key component in many Internet of Things (IoT) applications. The main challenges and limitations are the narrow bandwidth, low power and low cost which reduces the accuracy of the time of arrival (TOA) estimation. In this paper, we consider the positioning scenario of Narrowband IoT (NB-IoT) that can benefit from observed time difference of arrival (OTDOA). By applying the deep learning based technique, we explore the generalization and feature extraction abilities of neural networks to tackle the aforementioned challenges. As demonstrated in the numerical experiments, the proposed algorithm can be used in different inter-site distance situations and results in a 15% and 50% positioning accuracy improvement compared with Gauss-Newton method in line-of-sight (LOS) scenario and non-line-of-sight (NLOS) scenario respectively.

Keywords: Positioning, NB-IoT, observed time difference of arrival, deep neural network.

1 Introduction

Various type of smart terminals, such as self-driving cars [1], smart meters [2], and wearable devices [3], provide a paradigm shift in our daily works and lives. With the urgent requirement for the connected world, the IoT transmission has been identified as one of the most important technologies for future wireless networks, and the massive machine-type communication (mMTC) has been selected as one of the most important scenarios for the coming 5G communication system [4]. In order to provide a smooth evolution to the mMTC transmission, NB-IoT is proposed to offer initial IoT services with wide coverage, low power and spectrum consumption, and large connectivity [5].

Among NB-IoT applications, including smart grid, safety monitoring and vending machines [6], the localization services become a fundamental feature and according to [7], more than 40% IoT connections will be related to location information by 2020. With the limit power and cost budget for NB-IoT terminals, the conventional global navigation satellite system (GNSS) based localization schemes are not available in the current NB-IoT systems, and the most common approach, according to 3GPP standard [8], relies on observing the time difference of downlink reference signals, which is often referred to as observed time difference of arrival (OTDOA).

Although the OTDOA based localization schemes have been successfully utilized in the traditional long-term evolution (LTE) systems [8], the extension to NB-IoT systems is not straight forward. For example, an iterative expectation maximization based successive interference cancellation algorithm has been proposed in [9] to jointly consider the residual
frequency offset, the fading coefficients of different channel taps, and the time-of-arrival information of different cells. In [10], the estimation of phase differences for different frequency hopping reference signals has been utilized to reduce the positioning error. However, the achievable positioning accuracy of the above schemes is quite limited and the key positioning issues in the NB-IoT systems have not been completely investigated as explained below.

Limited Resolution. To maintain a low cost implementation of NB-IoT systems, the baseband sampling frequency has been reduced from the conventional LTE requirement (30.72 MHz) to much lower rates (1.92 MHz). Although this type of scheme greatly reduces the potential energy consumption, the resolution of observed reference signal time difference (RSTD) will be reduced as well, which causes a larger positioning error.

Complexity. Traditional algorithms, such as the famous Gauss-Newton method [11], require several iterations to achieve a satisfied positioning result, and the associated power consumption may not be suitable for some low power NB-IoT applications. Therefore, a low complexity and high accuracy algorithm will be desirable.

On the other hand, some deep learning schemes are proposed to solve positioning problem. [12] and [13] propose to use deep learning algorithm instead of KNN to improve the accuracy of fingerprint-based positioning. [14] uses deep learning to estimate the absolute position directly from the raw channel impulse response (CIR) data. All of these works perform well with multiple antennas. However, high precision positioning is also important in the single input single output (SISO) system.

In this paper, we propose a deep learning inspired positioning framework to address the above challenges in NB-IoT systems. Different from the traditional brute-force application of neural networks, we propose to use calculated RSTD results as commonly adopted in the OTDOA scheme rather than realtime measured signal strength to achieve better positioning performance. Meanwhile, we also exploit the generalization ability of neural networks and evaluate the positioning performance under different inter-site distances. As demonstrated in the numerical experiments, our proposed scheme can achieve as much as 15% and 67% accuracy improvement in LOS scenario and NLOS scenario respectively, if compared with traditional Gauss-Newton methods.

The rest of this paper is organized as follows. Section 2 provides some preliminary information about RSTD measure and Gauss-Newton method. Section 3 presents the deep learning based framework for position estimation, and the numerical examples are given in Section 4. Finally, we conclude this paper in Section 5.

2 Preliminaries

Consider an OTDOA based positioning scheme as shown in Fig. 1, where each user equipment (UE) is assumed to be connected with multiple base stations (BSs). Since the perfect synchronization between UE and BSs is in general difficult to obtain in the practical systems, RSTDs, i.e., the time-of-arrival (TOA) differences of reference signals from different BSs, are commonly adopted in the positioning process. Given the measured RSTDs, we can draw different hyperbolic curves and estimate the UE’s position based on the inter section region formed by different curves as shown in Fig. 1. In this section, we briefly elaborate the composition of the RSTD estimation error and the traditional Gauss-Newton algorithm to compute the estimated positions.
Fig. 1. Illustration of the OTDOA based positioning scheme. The UE can receive the signals of 3 BSs and obtain TOAs, e.g., $\tau_1$, $\tau_2$ and $\tau_3$. Then the three hyperbolic curves can be drawn by the time differences. The intersection of the three hyperbola is the position of the UE.

2.1 Composition of RSTD Error

The OTDOA based scheme is to estimate the two-dimensional (2D) UE position, based on the pre-known positions of BSs, e.g. $\{p_i = [x_i, y_i]^T\}$, and the observed RSTDs, e.g. $\{\Delta \tau^*_i\}$. RSTD is obtained by subtracting the TOA of the reference BS and the neighboring BS, e.g,

$$\Delta \tau^*_i = \tau^*_i - \tau^*_0 = \tau_i - \tau_0$$

where $\tau^*_i$ ($i = 0,1,\ldots,N_{BS} - 1$) indicates the measured TOA, and $N_{BS}$ is the total number of observed BSs. In practical system, the RSTD is often computed by the additive noise as given below, for $i = 1,2,\ldots,N_{BS} - 1$,

$$\Delta \tau^*_i = \frac{1}{c} \left( ||p - p_i|| - ||p - p_0|| + n_{i,0} + e_i \right)$$

where $c$ is the speed of light, $n_{i,0}$ and $e_i$ are the difference of measurement error including the sampling error and NLOS error between the serving BS and the neighbouring BS respectively.

2.2 Gauss-Newton Method

Mathematically, the position estimation process can be obtained by minimizing the overall square errors with respect to the observed TOA results, e.g.,

$$p^* = \arg \min_{p \in \mathbb{R}^2} \sum_{i=1}^{N_{BS}-1} |c\Delta \tau^*_i - h_i(p)|^2$$

$$= \arg \min_{p \in \mathbb{R}^2} \sum_{i=1}^{N_{BS}-1} |u_i(p)|^2$$

$$= \arg \min_{p \in \mathbb{R}^2} \sum_{i=1}^{N_{BS}-1} |u_i(p) + \nabla u_i(p)^*(p^* - p)|^2$$
where \( \mathbf{p}^* = [x^*, y^*]^T \) denotes the estimated position of UE, and \( h_i(p) \) indicates the distance differences with respect to different BSs, which is given by,

\[
h_i(p) = \sqrt{(p - p_i)^T(p - p_i)} - \sqrt{(p - p_0)^T(p - p_0)}
\]

(4)

In the Gauss-Newton based approach \[11\], the estimated position in the \((k + 1)^{th}\) iteration, \( \mathbf{p}^{(k+1)} \), is obtained via the following iterative equation

\[
\mathbf{p}^{(k+1)} = \mathbf{p}^{(k)} + \beta_k \left( \mathbf{H}^T \mathbf{h}(\mathbf{p}^{(k)}) \right)^{-1} \mathbf{H}^T \mathbf{h}(\mathbf{p}^{(k)}) \left( \mathbf{c} \Delta \mathbf{r} - \mathbf{h}(\mathbf{p}^{(k)}) \right)
\]

(5)

where \( \beta_k \) denotes the step size with certain convergence property, \( \Delta \mathbf{r} = [\Delta x_1, \Delta x_2, \ldots, \Delta x_{N_{BS}-1}]^T \), and \( \mathbf{h}(\mathbf{p}) = [h_1(p), h_2(p), \ldots, h_{N_{BS}-1}(p)]^T \). \( \mathbf{H}^T(\mathbf{p}) \) is the Jacobian matrix of \( \mathbf{h}(\mathbf{p}) \), which is defined as,

\[
\mathbf{H}^T(\mathbf{p}) = \begin{bmatrix}
\frac{\partial h_1(p)}{\partial x} & \frac{\partial h_1(p)}{\partial y} \\
\frac{\partial h_2(p)}{\partial x} & \frac{\partial h_2(p)}{\partial y} \\
\vdots & \vdots \\
\frac{\partial h_{N_{BS}-1}(p)}{\partial x} & \frac{\partial h_{N_{BS}-1}(p)}{\partial x}
\end{bmatrix}
\]

(6)

### 3 Deep learning for position estimation

The relationship between RSTD and UE location is highly nonlinear, but the Gauss-Newton algorithm is only a quadratic convergence which will result in an increase in positioning error. On the other hand, when the TOA is NLOS, the optimization algorithm for position estimation is non-convex. Thus, We try to solve these problems with the neural network which is very popular recently. Deep neural networks (DNN) have good capability of classification and regression. DNN's classification ability has been well shown in fingerprint-based positioning [15]. In our algorithm, we use DNN's great regression ability to fit the relationship between RSTD and UE's position.

#### 3.1 TOA Error Distribution

In LTE system, the sampling rate is \( F_s = 30.72 \text{MHz} \) and each sampling period is \( T_s = 1/F_s \). However, in NB-IoT system the sampling-rate could be defined as \( F_s = 1.92 \text{MHz} \) in order to save battery-power. We consider scenarios without inter-cell interference. We did 1000 Monte Carlo simulations, and recorded the TOAs of 7 BSs for each simulation. The error distribution of TOA is shown in Table. 1. What we can learn from the Table. 1 is that the TOA estimation errors are mainly caused by low sampling rate and NLOS. In practice, however, higher sampling rates mean more power consumption, which is not recommended in IoT devices, and if we can mitigate the effect of NLOS, we can greatly improve the accuracy of positioning. We define the NLOS scenario when the TOA error is greater than \( 0 \) Ts. In OFDM systems, some NLOS identification
technologies [16, 17] have already performed well. In this article we assume that the recognition result of NLOS is prior information, and let LOS status \( s = 0, 1 \) where 1 indicates NLOS and 0 indicates LOS.

### Table 1. TOA error distribution under different kinds of channel

<table>
<thead>
<tr>
<th>TOA Error</th>
<th>(-1T_s)</th>
<th>(0T_s)</th>
<th>(1T_s)</th>
<th>(2T_s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWGN</td>
<td>90</td>
<td>6842</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>EPA</td>
<td>121</td>
<td>6737</td>
<td>138</td>
<td>4</td>
</tr>
<tr>
<td>EVA</td>
<td>376</td>
<td>3586</td>
<td>2502</td>
<td>559</td>
</tr>
</tbody>
</table>

#### 3.2 Data Collection and Preprocessing

![Fig. 2. The topology of the simulation. There are a total of 7 BSs, and the intermediate BS is regarded as a serving BS. We randomly generate UEs around the serving BS.](image)

We can regard DNN as a black box \( f(\cdot) \) through a lot of data training. In the problem of OTDOA positioning, the function can be written as \( p = f(\cdot) \).

In the first step, we need a lot of data sets about \( W \) and \( p \) mapping on \( f(\cdot) \) to train the network. We consider the case of \( N_{BS} \), that is, the number of BSs is 7. As shown in Fig. 2, UEs are randomly generated in cells around the intermediate BS. In order to generalize different inter-site distance situations, we normalize the input so that the normalized input parameter can be written as,

\[
w = \frac{[c\Delta \tau_1 x_1 y_1 \cdots c\Delta \tau_6 x_6 y_6 D_{cell} s_1 \cdots D_{cell} s_6]}{D_{cell} (7)}
\]

where \( D_{cell} \) represents the distance of two adjacent cells, and \( s_i \) indicates the \( i^{th} \) BS's LOS status. It is easy to find that the input parameters of BS's coordinates are certain value after normalization, but we still use these parameters as input parameters because they contain the position information even if it is not necessary. We generate 100,000 sets of \( w \) and their corresponding coordinates to train the network.
Fig. 3. Schematic diagram of the DNN structure. The input parameter \( \mathbf{w} \) contains RSTD and BS position information, and the output \( \mathbf{p} \) represents the estimated position.

### 3.3 Neural Network Configuration

**Table 2.** Network structure

<table>
<thead>
<tr>
<th>Layers</th>
<th>Output Shape</th>
<th>Activation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input layer</td>
<td>18</td>
<td>-</td>
</tr>
<tr>
<td>Fully connected layer 1</td>
<td>18</td>
<td>Relu</td>
</tr>
<tr>
<td>Fully connected layer 2</td>
<td>32</td>
<td>Relu</td>
</tr>
<tr>
<td>Fully connected layer 3</td>
<td>16</td>
<td>Relu</td>
</tr>
<tr>
<td>Fully connected layer 4</td>
<td>8</td>
<td>Relu</td>
</tr>
<tr>
<td>Output layer</td>
<td>2</td>
<td>Linear</td>
</tr>
</tbody>
</table>

The network diagram is shown in Fig. 3, and the network structure of the DNN in our experiments is shown in Table 2. An input layer, an output layer and 4 hidden layers form this network, and these four hidden layers have 18, 32, 64, and 8 neural units respectively. The target is to find a function \( f(\cdot) \) and to make the estimated coordinates and real coordinates as close as possible. The optimization function can be written as,

\[
\mathbf{p}^* = \arg \min_{\mathbf{p} \in \mathbb{R}^2} |\mathbf{p} - \mathbf{p}^*|^2
\]

Thus, we choose mean square error (MSE) as loss function and the loss function can be given by,

\[
\mathcal{L} = \frac{1}{N_{\text{batch}}} \arg \min_{\mathbf{p}^* \in \mathbb{R}^2} \sum_{m=1}^{N_{\text{batch}}} |\mathbf{p}^{(m)} - \mathbf{p}^*(m)|^2
\]

where \( N_{\text{batch}} = 32 \) is the batch size for each training step in our experiments. \( \mathbf{p}^{(m)} \) and \( \mathbf{p}^*(m) \) are the \( m \)th training position data. In our experiment, We only use 6 layers. Thus, the training
speed and calculating speed are both faster than the Gauss-Newton algorithm. Use the generated data sets to train the network, and the trained network can be used to estimate UE’s position.

3.4 Model Generalization

Because the BSs’ coordinates are known values, the normalized distance differences are the main parameters affecting network training results and positioning accuracy. It can be proved that the actual distance difference between the serving BS and the neighbouring BS to the UE is in the range of $[-D_{cell}, D_{cell}/3]$. The time difference is converted to RSTD after sampling, and its range of values can be expressed as,

$$\Delta t_i^* \in \left\{ \frac{< F_s D_{cell} >}{c}, \frac{< F_s D_{cell} > + 1}{c}, \ldots, \frac{< F_s D_{cell} >}{c} \right\}$$

(10)

where $\langle \cdot \rangle$ indicates rounding. After normalization, the range of distance difference can be written as,

$$\frac{c\Delta t_i^*}{D_{cell}} \in \left\{ \frac{< F_s D_{cell} >}{c}, \frac{< F_s D_{cell} > + 1}{c}, \ldots, \frac{< F_s D_{cell} >}{c} \right\}$$

(11)

Fig. 4. The CDF of normalized distance difference $\frac{c\Delta t_i^*}{D_{cell}}$.

After the distance difference is normalized, it is distributed in a similar range, and in the case of large inter-site distance, the distribution contains more information. We have carried out experiments when the inter-site distance is 500m, 750m, 1000m, and 10000m. The cumulative distribution function (CDF) of the normalized distance difference $\frac{c\Delta t_i^*}{D_{cell}}$ is shown in Fig. 4. It is like sampling the input space of a neural network. Moreover, the large distance model can carry out more fine-grained sampling. A more fine-grained sampling of input
means that the DNN can fit a more precise function $f(\cdot)$ after training. Therefore, we use 10000m inter-site distance normalized data sets to train the network, and evaluation of the performance of this network by 500m, 750m, 1000m inter-site distance data sets.

4 Numerical Results

In this section, we provide numerical results to compare the performance of the DNN algorithm with the Gauss-Newton one. Gauss-Newton's algorithm. The simulation parameters are summarized in Table 3.

<table>
<thead>
<tr>
<th>Table 3. Scenario parameters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Layers</td>
<td>Output Shape</td>
</tr>
<tr>
<td>Network synchronization</td>
<td>Perfect synchronization</td>
</tr>
<tr>
<td>Cyclic prefix</td>
<td>Normal</td>
</tr>
<tr>
<td>Number of BS/device antennas</td>
<td>1/1</td>
</tr>
<tr>
<td>Macro transmit power</td>
<td>46 dBm for 1.92 MHz</td>
</tr>
<tr>
<td>Terminal noise density</td>
<td>-174 dBm/Hz</td>
</tr>
<tr>
<td>Number of PRS</td>
<td>1</td>
</tr>
<tr>
<td>Consecutive PRS subframes</td>
<td>1</td>
</tr>
<tr>
<td>PRS muting</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Fig. 5. The performance comparison between two algorithms when the inter-site distance is different. The results come from Monte Carlo simulation of 1000 random UE locations. 1), 2), 3) represent the performance of Gauss-Newton algorithm in three different inter-site distance cases. 4), 5), 6) represent the performance of the DNN algorithm which are trained and tested by the same inter-site distance data sets. 7) and 8) represent the performance of the DNN algorithm which are trained by 500m inter-site distance data sets but tested by 750m or 1000m data sets.
4.1 Results of the DNN Algorithm

Fig. 5 illustrates the CDF of the horizontal positioning accuracy of two different algorithms. We consider three different kinds of distance and the performance of the DNN algorithm is better than the performance of Gauss-Newton algorithm when we use the same distance data sets to train and test the network. Lines 1 to 6 indicate that the positioning accuracy of the DNN algorithm can achieve about 25% accuracy improvement.

However, when the inter-site distance of the training and test data sets is different, the performance of the DNN algorithm is very poor even if the data sets are normalized. For example, in the experiment of 750m inter-site distance, the average positioning accuracy is 24.5m when the network is trained by 750m inter-site distance data sets and it is only 43.2m when the network is trained by 500m inter-site distance data sets. That is because when we use the data sets of 500m inter-site distance, the data sets are not general enough to allow the network to learn the characteristics of the inter-site distance of 750m and 1000m.

In order to evaluate the positioning ability of the DNN algorithm under different number of BSs, we set the number of BSs to 4, 5, and 6 respectively. The performance of the proposed algorithm is shown in Fig. 6. It indicates that the more BSs participating in the positioning, the higher the positioning accuracy will be and the proposed the DNN algorithm performs better than the Gauss-Newton algorithm.

4.2 Results of Generalized Networks

After that, we use 10000 m inter-site distance data sets to train the DNN, and use 500m, 750m, 1000m distance data sets to test the network. The result is shown in Fig. 7. In this case, the positioning accuracy of DNN is improved by about 15%. It achieves significant gains under different inter-site distances. Therefore, this method successfully generalizes the network at the cost of some positioning accuracy, enabling high accuracy positioning on the same network even if the inter-site distance is different.

We can also find a trade-off relationship between the inter-site distance of the training sets and the positioning accuracy of the test sets. It indicates that the generalization ability of the
network should be at the expense of positioning accuracy and if we want to cover more inter-site distance situation, the positioning accuracy will be reduced.

**Fig. 7.** The performance comparison between the Gauss-Newton algorithm and the DNN algorithm which is trained by 10000m inter-site distance data sets. 1), 2), 3) represent the performance of Gauss-Newton algorithm in three different inter-site distance cases. 4), 5), 6) represent the performance of the DNN algorithm which are trained by 10000m inter-site distance data sets but tested by 500m, 750m and 1000m data sets.

**4.3 Results of NLOS scenario**

**Fig. 8.** The performance comparison between the Gauss-Newton algorithm and the DNN algorithm under EVA channel. The positioning accuracy will decrease because of the NLOS scenario. Line 4 indicates when the DNN input contains the LOS status $s$, the positioning accuracy can be greatly improved.

We use the AWGN channel to represent the LOS scenario, and the EVA channel to represent the scenario where LOS and NLOS coexist. Fig. 8 illustrates the CDF of the horizontal positioning accuracy of two different algorithms under AWGN and EVA channels. Line 5
indicates that the positioning accuracy will decrease because of the NLOS scenario. However, when the input of DNN contains the LOS status, the positioning accuracy can be greatly improved by about 67% and 50%, if compared with Gauss-Newton algorithm and DNN algorithm without LOS status input, respectively. This figure shows that the LOS status can be beneficial to improve the positioning accuracy.

5 Conclusion

In this paper, we propose to use DNN for location estimation when the TOA estimation accuracy can not be improved in 1.92 MHz sampling rate. In order to solve the DNN generalization problem, we analyze the CDF of the distance difference, and use the large inter-site distance data sets to train the network. We also use the LOS status as input to improve the positioning accuracy in the NLOS scenario. The proposed algorithm can improve positioning accuracy by about 15% in LOS scenario and 67% in NLOS scenario compared with Gauss-Newton method.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) Grants under No. 61671011, 61701293 and No. 61871262, 61401266, the National Science and Technology Major Project Grants under No. 2018ZX03001009, the Huawei Innovation Research Program (HIRP), and research funds from Shanghai Institute for Advanced Communication and Data Science (SICS).

References


Achieving trustworthiness, Transferable, Cross-domain Dynamic Accumulator Authentication in Industrial Internet of Things*

Linjie Wang¹,², Youliang Tian¹, and Jinbo Xiong³
{youliangtian@163.com, wanglinjie_66@hotmail.com, jinbo810@163.com}

¹ State Key Laboratory of Public Big Data, College of Computer Science Technology, Guizhou University, Guiyang 550025, China
² School of Data Science, Tongren University, Tongren 554300, China
³ College of Mathematics and Informatics, Fujian Normal University, Fuzhou 350007, China

Abstract. The authentication problem is one of the most significant challenges in the industrial Internet of Things control system applications. To authenticate the relationships between devices is an effective solution to tackle authentication issue. However, the previous studies of Internet of Things focused on connecting sensors and setting the sharing condition from either private or public in the sense that the relationship authentication issue between devices is not well solved. In this paper, we propose CDAA, a Cross-domain Dynamic Accumulator Authentication with a general undirected graph representing the relationship between devices needed authentication for industrial Internet of Things supporting trustworthiness cross-domain, cryptographic accumulator, and transferability. Specifically, a dynamically updatable cross-domain authentication scheme is proposed based on the cryptographic accumulator and a standard digital signature scheme as the underlay of the CDAA. Finally, we give the analysis and comparison and the results show that the proposed scheme can address the authentication issue. The effectiveness and feasibility of proposed scheme are presented and analyzed through a comparison with traditional systems in practical application.

Keywords: Cross-domain · Authentication · Undirected graph · Accumulator · Bilinear pairings.

* This work is supported by The National Natural Science Foundation of China under Grant Nos. 61662009 and 61772008; Science and Technology Major Support Program of Guizhou Province (No. 20183001); Ministry of Education-China Mobile Research Fund Project under Grant No. MCM20170401; Foundation of Postgraduate of Guizhou Province under Grant No. YJSCX.HJ2019101; Science and Technology Planning Project of Tongren Municipality under Grant No. 202078.
1 Introduction

To date, ubiquitous wireless sensor devices have made the Internet of Things (IoT) attracted enormous attention from academics and industries, which is foreseen one of the most important technologies of this century [1] [2]. However, the industrial Internet of Things (IIoT) offered interconnection and intelligence to industrial systems through sensing devices and actuators has the potential to change the world as the Internet did [3]. In practice IIoT system, for the purpose of protecting the legitimate access each device has its own unique identity and has the capability to sense, compute, authenticate and communicate [4] [5]. Many existing solutions have been introduced into security and privacy solutions in IIoT, e.g., the data security, wireless power transfer, and the deployment and communication problems of devices [6] [7].

In a multi-layer distributed IIoT system, the following situations are often encountered: on the one hand the supervisor entity $A$ wants to command the subordinate entity $B$, $A$ should provide a legal authoritative signature to show the legality of its command [4]; On the other hand, entity $A$ wants to access subordinate entity $C$ of entity $B$ who is in the same regulatory authority as $A$, $A$ also should provide a legal authoritative signature to illustrate the legality of its command. In this paper, vertices and edges are used to represent company members and the authentication relationship between them, respectively. That is, given two vertices $v_i$ and $v_j$ do not have a signature edge, one can generates the legitimate signature edge $(v_i, v_j)$ without the administrative involvement. This property can help the authority to mitigate the pressure of the digital signature authentication.

1.1 Related work

Since the first undirected transitive signature based on the discrete logarithm and RSA assumptions was proposed by Micali and Rivest [8], Shahandashti et al. [9] proposed a short transitive signature based on the bilinear group pairs. A new approach transforming state transitive signature scheme into the stateless ones without loss of security was introduced by Ma et al. [10], and the proposing three concrete protocols were based on hardness of Factoring and RSA problems. The security analysis showed the scheme secure against the adaptive chosen message attack in random oracle model. Subsequent to Micali and Rivest’s undirected graph transitive signature work, Rivest [11] introduced a directed transitive signature scheme and showed it hard to construct based on a sophisticated mathematical group. Like Neven’s [12] researches, Xu [13] also studied the directed trees, but his scheme is more consideration on the property of constant signature size and privacy preserving. Camacho et al. in [14] also introduced a transitive signature protocol for the undirected tree.

In the existing papers, literature [12] all have designed the transitive signature schemes for special directed graphs (e.g. directed tree) and the security proof under adaptive chosen-message attack. In addition to transitive signature studied, a rejectable signatures [13] were used to authenticate general directed
graphs. Followed the accumulator work by Ma et al. [10], Lin et al. [15] presented a novel transitive closed undirected graph authentication scheme to support blockchain-based identity management systems.

1.2 Contributions

In order to address the aforementioned authentication issue between IoT devices in IIOT system, we introduce a cross-domain dynamic accumulator authentication scheme (CDAA). The contributions are outlined as follows:

- In the undirected administrative domain graph, we leverage the transitivity of digital signature authentication to construct the digital authentication relation between the vertices of non-existent edge.
- In our scheme, a signer can directly compute the authentication to the verifier with the transitivity of the signature to prove its legality even when they are not in the same administrative domain.
- After updating the vertices’ witness and signature, the signer only needs to publish and chain the new witness and signature, and the verifier can verify the correctness of the new witness and signature to decide whether to be signature authenticated.
- For the implementation, we give the theoretical analysis and compare CDAA with existing protocols (as shown in Table I and II). We also test the time cost on the public parameters, private key, accumulator update cost of CDAA and so on.

1.3 Organization

The rest of the paper is organized as follow. In Section II, we describes the relevant preliminaries. Section III discusses the system model. In Section IV, we provide the security models of CDAA. The CDAA scheme with its security analysis is proposed in Section V. Section VI presents the comparison of performance and simulation. Conclusion and further discussion are given in Section VII.

2 Preliminaries

In this section, we introduce some fundamental preliminaries required in this paper. \( \mathbb{N} = \{1, 2, \cdots, N\} \) denotes a set of positive integers from 1 to \( N \); \( \mathbb{Z}_N^* \) is a multiplicative group of integers modulo \( N \); \( \| \) represents the concatenation operator on strings; \( f : \mathbb{N} \to \mathbb{R} \) is a negligible function; \( \mathbb{PPT} \) is the abbreviation of probabilistic polynomial time and \( \mathcal{S} \) shows that \( x \) is sampled randomly and uniformly from the set \( \mathcal{S} \).

2.1 Graphs

In this paper, we consider \( G = (V, E) \) as an undirected graph with a nonempty vertices set \( V \subset \mathbb{N} \) and \( E \subset V \times V \), a set of edges. Based on the definition of the partition, we split the vertices set \( V \) into several infinite sets \( V = V_1 \cup V_2 \cup \cdots \cup V_u \) as Fig. 2 shows, where \( u = |V| \).
2.2 Dynamic accumulator

Literature [11] generalized the definition of accumulator as follows. \( X_\lambda (\lambda \in \mathbb{N}) \) denotes the domain of values to be accumulated and \( F_\lambda \) is a set of pairs of functions, as well as \( U_f \) the accumulator domain. For each \((f, g) \in F_\lambda\), let \( f : U_f \times X^\text{ext}_f \rightarrow U_f \) for some \( X^\text{ext}_f \supseteq X_\lambda \) and \( g : U_f \rightarrow U_h \) is a bijective function. For any \( \lambda \in \mathbb{N} \) and \( X = \{x_1, x_2, \ldots, x_q\} \subset X_\lambda \), \( g (f (\ldots f (u, x_1), \ldots, x_1)) \) is called the accumulated value of the set \( X \) over \( u \). Additionally, the following properties are satisfied:

- **Efficient Generation**: Define a sequence of pairs functions \((f, g) \in F_\lambda\) and generate a key pair \((sk_{\text{acc}}, pk_{\text{acc}})\) for the accumulator.
- **Efficient evaluation**: For every \((f, g) \in F_\lambda\), \( u \in U_f \) and \( X \subset X_\lambda \), there is an efficient algorithm to compute \( g (f (u, X)) \).
- **Quasi Commutativity**: For every \( \lambda \in \mathbb{N} \), \((f, g) \in F_\lambda\), \( u \in U_f \) and \( x_1, x_2 \in X_\lambda \), it holds \( f (f (u, x_1), x_2) = f (f (u, x_2), x_1) \).

2.3 Security assumption

In our scheme, several theoretic definitions also have been based, e.g. Bilinear pairing, Diffie-Hellman Exponent assumption \((n\text{-DHE})\), Strong Diffie-Hellman Problem \((n\text{-SDH})\), Hidden Strong Diffie-Hellman Exponent assumption \((n\text{-HSDHE})\) [16] and a standard digital signature scheme \( SDS = (\text{Gen}, \text{Sig}, \text{Ver}) \) [4].

3 System Model

We show an overview of our proposed architecture based on the practical application scenario as well as abstract it for a undirected tree to introduce the authentication.

3.1 Authentication hierarchical architecture

In order to achieve the trusted authentication and ensure valid verification, we establish the architecture graph with three components to enhance the illustration of authentication service and application scenarios in the IIoT. As shown in Fig. 1, the first one is the third party layer (e.g. cloud computation) and then there is user network layer which is followed by the application devices layer.

3.2 Transitive authentication model

In this section, we analyze detailed the relationship between various devices shown in Fig. 1. In the user network layer, we can find that Alice can view the Logistics, Order goods, and E-commerce in the office, as well as Bob can view Storage, Workshop and After-sales service in the office. If Alice’s supervisor Eve wants to monitor the Workshop which belongs to another supervisor’s manage
domain, she needs to go through the legal permission layer by layer, and finally gets Bob’s authentication to permit it, even though she has the legitimate certification. The question is why Alice does not access the Workshop directly since she has legitimate certification.

In order to study this problem, we abstract a graph from above situation, as shown in Fig. 2. It can be clearly seen that all the nodes are valid digital signatures, node $v_{11}$ hence can be authenticated by node $v_{q,h}$ (i.e. the blue dotted line showed in Fig. 2. This authentication problem solved in [4] utilizes the transitive authentication relationship between superior and subordinate for a directed tree. However, if node $v_{km}$ wants to access the node $v_{qh}$, we can see that node $v_{km}$ and node $v_{qh}$ belong to different administrator, that is, in the different administrator domain. In view of the above situation, this paper focuses on studying the digital authentication across the domain between $v_{km}$ node and node $v_{qh}$ (i.e. the red dotted line showed in Fig.2).

4 The security Model of CDAA

In this section, the formal definition of a cross-domain dynamic accumulator authentication scheme is proposed for the general graphs in IIoT.

4.1 Outline of CDAA

A secure cross-domain dynamic accumulator authentication scheme (CDAA) for the general graph $G$ consists of the seven algorithms, i.e.: Setup, EdgeLab,
EdgeVal, AccUpdate, AccWitUpdate and AccVerify. These algorithms consist of the accumulator authority, the signer, an untrusted update entity and a verifier. Among them, the authority is in charge of constructing an accumulator initial algorithm Setup to generate the accumulator key pair \((pk_A, sk_A)\), the initial accumulator status \(Acc_\phi\) and a public state \(State_\phi\). The algorithm EdgeLab is used to add a new edge between node \(v_{ik}\) and \(v_{js}\), and also to obtain a new accumulator state \(Acc_{V \cup \{(v_{ik}, v_{js})\}}\) and state \(State_{U \cup \{(v_{ik}, v_{js})\}}\), as well as a witness \(wit_{(v_{ik}, v_{js})}\).

Two sets of \(V\) and \(V_w\) as a bookkeeping are maintained, where \(V\) contains the values computed by \(AccUpdate\) algorithm in the accumulator and \(V_w\) is the status of accumulator while a witness was \(wit_{(v_{ik}, v_{js})}\) created. The accumulator state \(State_U\) also has a bookkeeping information set \(U\) for adding the elements in it to the accumulator, which is a superset of \(V\) and \(V_w\). The detailed syntax of the CDAA is represented as follows:

- \(\text{Setup}(1^k, N) \rightarrow \{(sk_A, pk_A), Acc_\phi, State_\phi\}\): Initialization Setup algorithm take as input a security parameters \(1^k (k \in \mathbb{N})\), then generates public pa-
rameters set params, the accumulator key pair \((sk_A, pk_A)\), a standard digital signature scheme SDS, the accumulator security key pair \((spk, ssk)\), an empty accumulator value \(Acc_0\) and a state information list \(State_0\).

- **EdgeLabel** \((v_{ik}, v_{js}) \rightarrow \{x_i\}**: Edge Label algorithm is used to generate edge label among nodes needed to construct authentication. Then output a label \(x_i\) for edge \((v_{ik}, v_{js})\).

- **EdgeVal** \((x'_i, ssk) \rightarrow \{s\sigma, w\}**: Edge Value algorithm is used to calculate an value \(w\), the edge representation notation and generate the new edge signature \(\sigma\).

- **AccUpdate** \((v_{ik}, v_{js}) \rightarrow \{Acc, State\}**: Accumulator Update algorithm is used to update \(Acc\) and \(State\) and output the new accumulator \(Acc\) and \(State\).

- **AccWitUpdate** \((x_i) \rightarrow wit'\)**: Accumulator Witness Update algorithm is used to input edge label notation \(x_i\), then outputs the updated witness \(wit'\).

- **AccVerify** \((w'_i, pk_A, g_i, Acc, w) \rightarrow \{0, 1\}**: Accumulator Verify algorithm takes as input parameters and verifies the signature and witness.

### 4.2 Security of CDAA

The security requirements of the CDAA are defined through the following game between the challenger and the attacker. We call the CDAA is secure if no \(\mathcal{PT}\) adversary \(A\) has a non-negligible advantage in the following phase against a challenger \(C\).

- **Phase 1.** The changer \(C\) runs the Setup algorithm to generate the corresponding parameters, e.g. the public parameters params = \((q, G, G_T, e, g, H, H')\), a standard digital signature scheme SDS, the accumulator key pair \((sk_A, pk_A)\), an empty accumulator value \(Acc\) and a state information list \(State\). Then \(C\) initializes \(Acc_0=1\) and \(State_0=\{(\phi, g_1 = g^\gamma, \cdots , g_n = g^{\gamma n}, g_{n+2} = g^{\gamma n+2}, \cdots , g_{2n} = g^{2\gamma n}\}\), then sends \(pk_A, spk\) to the adversary \(A\).

- **Phase 2.** In this phase, number of queries as following to \(C\) is given by the adversary \(A\): EdgeLabel, EdgeVal, AccUpdate, AccWitUpdate and AccVerify.
  - **EdgeLabel** and **EdgeVal** queries. Adversary \(A\) randomly chooses a general graph \(G = (V, E)\). He arbitrarily chooses two nodes \(v_{ik}\) and \(v_{js}\) without edge between them, where \(i < j\). Then he runs EdgeLabel algorithm to obtain edge label \(x_i\) between nodes \(v_{ik}\) and \(v_{js}\), then sends to \(C\).
  - **AccUpdate** queries. \(A\) could change the state of accumulator adaptively, by querying \(C\) to add \(x_i\) to the set. For any \(x_i\), \(A\) queries the witness and \(C\) runs WitAdd algorithm to respond with \(wit'\) to \(C\). In addition, \(C\) runs AccUpdate algorithm to obtain the newly updated state set \(State\) and output the accumulator value \(Acc\).
  - **AccWitUpdate** query. \(A\) can query the witness many times for label notation \(x_i\) (i.e. corresponding with edge \((v_{ik}, v_{js})\)). Each time \(C\) runs AccWitUpdate algorithm to generate and update the witness \(wit'\) to \(A\) as a response.
Phase 3. After all context processes have been finished, $A$ submits a forgery $(x_{j'}, w_{i't'}, Acc_{V'})$ on edge $(v_{ik'}, v_{js'})$ chosen by himself. $A$ wins the game if the following conditions hold:

- $AccVerify(pk_A, x_{j'}, w_{i't'}, Acc_{V'}) = 1$;
- $x_{j'}$ has never been queried during the game.

We use $Adv^{CDAA}_A(k)$ to represent the probability of an adaptive chosen-message adversary $A$ that wins the above game.

**Definition 1.** The CDAA is unforgeable under chosen-message attacks if the following function $Adv^{CDAA}_A(k)$ is negligible for any PPT adversary $A$.

$$Adv^{CDAA}_A(k) = \Pr[\text{params} \leftarrow \text{Setup}(1^k, N) ; \text{Δ EdgeLab}(\text{params}) ; w_{i't'} \leftarrow \text{EdgeVal}(\text{params}, x_{i'}) ; \text{(State}_{U}, Acc_{V}) \leftarrow \text{AccUpdate}(\text{params}, U, V) ; 1 \leftarrow AccVerify(pk_A, x_{i't'}, w_{i't'}) \land (v_{ik'}, v_{js'}) \notin G]$$

### 4.3 Privacy

Let graph $G$ be composed of nodes $i, j, k$ and edges $(i, j), (j, i), (j, k)$. The meaning of privacy is the verifier, not to obtain any other information about the other nodes and edges, receiving a witness for $(i, j)$. That is, the verifier could not get any other information on the graph while he verifies an edges witness.

**Definition 2.** The CDAA is leakage-free, even if an adaptive chosen-message attacker could know any undisclosed information of the group.

### 5 Our Construction

We present our CDAA scheme and the proposed construction as follows. Randomly chosen two nodes $v_{ik}$ and $v_{js}$ for a given group, let $(v_{ik}, v_{js})$ denote an edge.

#### 5.1 Concrete scheme

**Setup:** Take as input a security parameters $1^k$ and $\ell = \lceil k/2 \rceil - 2$, generates an accumulator above and the setup parameters $params = (q, G, \mathbb{G}_T, e, g, H, H')$ of a bilinear map $e : G \times G \to \mathbb{G}_T$, where $H : N \to \mathbb{Z}_N^*$ and $H' : \{0, 1\}^* \to \{0, 1\}^\ell$ are two public hash functions. Let $g$ be the generation of $G$. It does as follows:

- Randomly choose a value $\gamma \in \mathbb{Z}_q$ and run a standard digital signature algorithm $Signa(1^k)$ to obtain a key pair $(spk, ssk)$ for digital signature node.
- Output the accumulator public/private key pair $(pk_A, sk_A)$, where $pk_A = (spk, z = e(g, g)^{\gamma^n})$, $sk_A = (\gamma, ssk)$. 

The public algorithm creates an initial accumulator $Acc_{\phi}=1$ and a public state set $State_{\phi}=(\phi, g_1 = g^{\gamma_1}, \cdots, g_n = g^{\gamma_n}, g_{n+2} = g^{\gamma_{n+2}}, \cdots, g_{2n} = g^{\gamma_{2n}})$.

EdgeLab: This algorithm is used to maintain a state information set $\Delta$ to store the edge labels. It does as follows:

- For a given graph $G$, the label of edge $(v_{ik}, v_{js})$ between node $v_{ik}$ and node $v_{js}$ is denoted by $x_{i'}$, where $x_{i'}=H'(H(i) \parallel H(k) \parallel H(j) \parallel H(s))$ and $i < j$.
- If $x_{i'} \notin \Delta$, then updates $\Delta=\Delta \cup \{\delta_{ik,j}\}$.

EdgeVal: The signer runs this algorithm to calculate a value $w$ and a signature. It does as follows:

- For any $x_{i'} \in \Delta$, compute $w = \prod_{j' \not= i'} g_{n+1-j'+i'}$.
- Run $Signa$ to calculate a signature $\sigma_{i'}=Signa(ssk, g_{i'} \parallel x_{i'})$.
- Output $wit_{i'} = (w, \sigma_{i'}, g_{i'})$.

AccUpdate: The signer runs this algorithm to check whether $(v_{ik}, v_{js}) \notin G$ (i.e. $x_{i'} \notin \Delta$), then does as follows:

- Update the state set $State_{U \cup \{(v_{ik}, v_{js})\}} = (U \cup \{(v_{ik}, v_{js})\}, g_1, \cdots, g_n, g_{n+2}, \cdots, g_{2n})$.
- If $x_{i'} \in \Delta$, calculates the accumulator value $Acc_{V \cup \{(v_{ik}, v_{js})\}} = Acc_{V} \cdot g_{n+1-i'}$, and outputs $Acc_{V} = \prod_{v' \in X} g_{n+1-v'}$; otherwise outputs $\perp$.

AccWitUpdate: The authority runs this procedure to obtain the updated witness after edge $(v_{ik}, v_{js})$ added into the graph $G$. For an edge $(v_{ik}, v_{js})$ with the corresponding label notation $x_{i'}$, does as follows:

- If $(v_{ik}, v_{js}) \in V$, compute $w' = w \cdot \prod_{j \in V \setminus \{v_{ik}, v_{js}\}} g_{n+1-j+i}$.
- Output the updated witness $wit'_{i'} = (w', \sigma_{i'}, g_{i'})$.

AccVerify: For an edge $(v_{ik}, v_{js})$ with the corresponding notation $x_{i'}$, this algorithm inputs $w'_{i'}, pk_A, g_{i'}, Acc_V, w$, and does as follows:

- If $VerA(spk, \sigma_{i'}, g_{i'} \parallel x_{i'}) = 0$, then outputs $\perp$.
- If $AccVerify(pk_A, x_{i'}, wit_{i'}, Acc_{V}) = 0$ (i.e. $\frac{\epsilon(g_{i'}, Acc_{V})}{\epsilon(g, w)} \neq z$), then output reject.
5.2 Correctness

We assume that AccV is an accumulator for $sk_A = (params, \gamma, ssk)$, $pk_A = (params, spk, z = e(g, g)^{\gamma_{n+1}})$, and $state_U = (U, g_1 = g^{\gamma_1}, \ldots, g_n = g^{\gamma_n}, g_{n+2} = g^{\gamma_{n+2}}, \ldots, g_{2n} = g^{2n})$. Before updating the witness, a correct accumulator value $AccV = \prod_{j' \in X} g^{n+1-j'}$ (e.g. $x_{j'} \in \Delta$ corresponding edge $(v_{i'k}, v_{j's}) \in G$) should be calculated. For each $x_{i'} \in \Delta$ (i.e. edge $(v_{ik}, v_{js}) \in G$), the update witness $wit'_{x_{i'}} = (w', \sigma_{i'}, g_{i'})$ is correct while the following equation holds:

$$\frac{e(g_{i'}, AccV)}{e(g, w)} = \frac{e(g, g) \sum_{j' \in X} \gamma^{n+1-j'+i'}}{e(g, g) \sum_{j' \neq i'} \gamma^{n+1-j'+i'}} = z$$

5.3 Edge deletion

The proposed scheme not only supports the addition of edges, but also satisfies the deletion of edges. In this section, we slightly modify our scheme to achieve the deletion of edges. Note that, if the authentication relationships are terminated among the devices, the corresponding edges are also deleted in the graph. Assume that an edge $(v_{i'k}, v_{js}) \in E (x_{i'} \in \Delta)$ is deleted from the graph.

$Delete (AccV, x_{i'}, ssk)$: On input $AccV, x_{i'}$ and $\sigma_{i'}$, this algorithm checks whether $x_{i'} \in \Delta$ and returns $\bot$ otherwise. It computes $\hat{Acc} = AccV / g^{n+1-i'}$, and runs $Signa$ to obtain $\sigma_{\hat{Acc}} = Signa (ssk, \hat{Acc})$. Finally, it returns the updated accumulator $\hat{Acc}$ and its signature $\sigma_{\hat{Acc}}$. Here it should be emphasized that the signature $\sigma_{\hat{Acc}}$ is computed on $\hat{Acc}$ instead of $g_{i'} || x_{i'}$.

$AccWidUpdate (x_{j'}, spk, w, op)$: This algorithm inputs parameters $spk, w, x_{j'}$ and $op$. Here, $\hat{Acc}$ represents the already updated accumulator and Acc the accumulator before the update. Notation $op$ will represent the $x_{j'}$ deleted or added to the accumulator $\hat{Acc}$. It returns the updated witness if $op = 0$ or if $op = 1$.

6 Performance Evaluation

In this section, we give the whole protocol comparison and concentrate on the public parameters, private key, accumulator update cost and so on.

6.1 Computational cost

Since the accumulator add computation, accumulator up-to-date computation and accumulator witness up-to-date operation dominate the computational overhead, we only focus on these operations following estimation. Table I shows computation complexity analysis on the maximum computation cost of one member during the execution between our protocol and literature [4]. In order to easily
illustrate the analysis comparison, an abbreviated notation is given. That is: $C_{h2}$-the cost of computing the prime representative of one element; $C_{hw}, C_H$-the cost of computing one $H'(\cdot)$ and $H(\cdot)$ hash operation respectively; $C_e$ -the cost of computing one exponentiation operation; $C_g$-the cost of computing $e(g, g)$; $C_{mult}$-the cost of computing one modular multiplication; $C_{Sig}$ and $C_{Svf}$-the cost of computing one signing operation and verification operation for SDS respectively. If there is $n$ nodes and $m$ edges (where $m = 1, \ldots, n(n-1)$) in the graph $G = (V, E)$. When adding a new edge $(v_{ik}, v_{js})$ to the graph $G$, the computation cost of all algorithms is light than the literature [4].

6.2 Experimental Test

The pairing-based cryptography (PBC) programs are C programs running on an Intel® Core 3.4 GHz and 8G RAM desktop PC. We used SHA-1 and ECC as the
### Table 1: Computation cost

<table>
<thead>
<tr>
<th></th>
<th>CDAA/PPAG</th>
<th>Ours</th>
<th>Literature [4]</th>
</tr>
</thead>
<tbody>
<tr>
<td>EdgeLab/Elabcost</td>
<td>(C_{hw} + 4C_H)</td>
<td>(C_{hw} + 2C_H)</td>
<td></td>
</tr>
<tr>
<td>EdgeVal/Witcost</td>
<td>((2 + n)C_e + C_{Sig})</td>
<td>(C_{h2} + C_e + C_{Sig}) + ((n^2 - 1)C_{mult})</td>
<td></td>
</tr>
<tr>
<td>AccUpdate/UEvalcost</td>
<td>(C_e + C_{mult})</td>
<td>((n^2 - 1)C_{mult}) + (C_e)</td>
<td></td>
</tr>
<tr>
<td>AccWitUpdate/UWitcost</td>
<td>(C_{hw} + 4C_H + C_e + 2C_{mult} + C_{Sig})</td>
<td>(C_{hw} + 2C_H + C_{h2}) + (C_e + C_{Sig})</td>
<td></td>
</tr>
<tr>
<td>AccVerify/Verfcost</td>
<td>(C_{hw} + 12C_H + C_{Svf}) + ((5 + n)C_e + 2C_{Sig}) + (2C_p + 4C_{mult})</td>
<td>(8C_H + 4C_{hw} + C_{Svf}) + (5C_e + 2C_{Sig} + 4C_{h2}) + ((2n^2 - 2)C_{mult})</td>
<td></td>
</tr>
</tbody>
</table>

hash function and Bilinear Pairing, as well as simulated our experiments on the PBC library (version 0.5.12) and the gnu multiple precision arithmetic library (version 6.0.0a). The Fig.2 (a) shows the execution time cost of comparison in each stage of the accumulator (accumulator, witness, and verification) between our CDAA scheme and PPAG [4]. We assume that there are 9000 elements and the simulation was run more than 100 times. It is worth noting that the bilinear map accumulator is much faster than the RSA accumulator by considerable amount simulations, especially the slopes differ by at least an order of magnitude the total running time of key generation, edge label algorithm, edge value algorithm and accumulator updated algorithm.

The Fig.2 (b) shows the execution time on updating the witness between PPAG [4] and CDAA, depending on the number of elements (from 500 up to 8500). Nevertheless, it is noteworthy that the chosen user nodes are all met the requirements of the PPAG scheme, i.e. all the nodes are subordinates of the signer.

Finally, Fig.2 (c) shows the running time of verification between the PPAG [4] and CDAA as the set number increases. As can be seen, the bilinear-map accumulator is faster than the RSA accumulator. The results clearly reflects that as the number of elements increases, the verification time of both protocols increase. Notably, when there is 9000 elements, the verification time of PPAG is almost 1.5 time than the verification of CDAA.

### 6.3 Comparative evaluation of related schemes

Furthermore, we show an objective comparison with other related ones to measure the security properties in TableII. Take the scheme in [4] as an example, which is another scheme to address the problem solving by this paper. But F. Zhu et al. [4] proposed scheme is based on the directed tree and the relationship between nodes are superiors and subordinates. It therefore could not be suited
Table 2: Computation overhead

<table>
<thead>
<tr>
<th>Scheme</th>
<th>F. Zhu et al.[4]</th>
<th>CDAA Construction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authentication technique</td>
<td>Accumulator standard</td>
<td>Accumulator standard</td>
</tr>
<tr>
<td>Tress</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>General graphs</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Updatable</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Leakage-free</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Based on</td>
<td>Strong RSA</td>
<td>Bilinear Pairing</td>
</tr>
<tr>
<td>Cross domain</td>
<td>NO</td>
<td>Yes</td>
</tr>
<tr>
<td>Authentication</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

for the general tree, in other words, it doesn’t apply to the practical application. We represent a scheme which has one concrete attribute.

Based on above comparisons, we could see that our scheme has desirable properties and it is suited for the practical application. The scheme furthermore is not restricted by the relation between nodes and can be authenticated across different domains.

7 Conclusions

This paper abstracted the relationship of devices as a node in a general graph and took the advantage of a dynamic accumulator based on bilinear pairing to achieve the authentication of equipment. Notably, our authentication scheme not only realizes the authentication of vertical nodes, but also satisfies the authentication transmission of horizontal nodes (i.e. between the same layer). In addition, the merit of dynamic accumulator could provide the feasible solution for edge addition. The security analysis shows that our protocol satisfies secure needs on the accumulator authentication in Industrial Internet of Things. Beyond that, the scheme also satisfies the privacy-preserving authentication. In our future work, we will further design more efficient CDAA scheme to facilitate the authentication speed.

References

Research on target location technology of campus video surveillance system based on WBAN and Internet of things

GAO Jie1, HUANG Yi2
{oijfdisod984@163.com, jahg245@163.com}2

1. Modern Economics & Management College of JXUFE, Gongqingcheng 332020, China;
2. NanChang Business College of JXAU, Gongqingcheng 332020, China

Abstract. In the traditional monitoring process, due to too much subjectivity of people, and the human brain's attention to things is accurate, but the sustainability and diversity are poor. When the monitor observes several surveillance cameras at the same time, the accuracy will be greatly reduced with the increase of time. Based on this, a research on target location technology of campus video surveillance system based on WBAN and Internet of things is proposed. Firstly, a three-dimensional geographical scene was constructed to detect the moving target, then the feature of the target was extracted and the image was preprocessed, and the target positioning was realized through the generation of panorama and perspective. Experimental results show that the proposed method has high positioning accuracy.

Keywords: WBAN; The Internet of things; Campus video surveillance; target.

1 Introduction

With the development of industrial society, information security and social security have become the hottest topics. As an important information technology, video surveillance has been paid more and more attention by the society in the field of security. However, with the popularization and enlargement of the monitoring system, the traditional human-dependent monitoring cannot meet the needs of today's society more and more, and its limitations become more and more prominent [1]. In the traditional monitoring process, due to too much subjectivity of people, and the human brain's attention to things is accurate, but the sustainability and diversity are poor. When the monitor observes several surveillance cameras at the same time, the accuracy will be greatly reduced with the increase of time. On the one hand, the monitor can't observe many things at the same time. On the other hand, the monitor can't pay attention to monotonous things for a long time. Finally, human physiology needs rest and diet, which also affects the reliability and stability of the monitoring system. All of these factors have led to the problems of the traditional human-dependent monitoring system, including high failure rate, poor reliability, and slow corresponding speed. At the same time, due to the huge and complex monitoring system, if all rely on human monitoring, the human cost is quite expensive, it is difficult to maintain such a large investment in the actual monitoring system.

WBAN, Wireless Body Area Network (WBAN, Wireless Boay Area Network) is a communication Network that centers on the human body and aims at collecting various physiological parameters of the human body. Through WBAN, people can communicate and synchronize data with their personal electronic devices (such as pdas and mobile phones). WBAN can communicate with other data communication networks (such as other people's WBAN, wireless/wired access network, mobile communication network, etc.) as part of the entire communication network, and communicate with any terminal on the network (such as PC, mobile phone, telephone, media player, digital camera, game console, etc.). With
the increasing complexity of monitoring scenes, the traditional video surveillance analysis based on two-dimensional images can no longer meet people's needs, and more information needs to be obtained from scenes to add to the analysis factors [2]. However, in today’s monitoring, depth information becomes particularly important. Therefore, the positioning technology of obtaining the three-dimensional space point coordinates corresponding to moving objects from 2d images based on the camera model has become a hot issue in intelligent monitoring technology.

2 Target location technology of campus video monitoring system based on WBAN and Internet of things

2.1 Construction of 3d geographical scenes

The target location technology of campus video surveillance system is realized on the basis of constructing realistic three-dimensional geographical scene, which also provides detection environment parameters for the combination of WBAN and Internet of things. The purpose of this study is to reconstruct the 3d geographical scene in order to simulate the target on the surface of the video surveillance area and provide the virtual object for the virtual camera imaging. In the case that the monitoring area is bare land, high-precision DEM data can be directly used, or 3d laser scanner can be used to scan the area to form point cloud data, and then gis can be used to form the earth surface model (DSM) of regular grid [3]. 3d scene rendering can be done using osgEarth, OSG’s third-party terrain rendering tool. The algorithm idea of osgEarth is to use quadtree data structure to divide terrain data and texture data into layers and blocks, build pyramids of different resolution of digital terrain and texture images, and establish terrain block nodes of different LOD layers [4]. This rendering method is beneficial to improve the efficiency of scene modeling.

2.2 Moving target detection

The moving target is moving and changing, so it needs to be detected to provide a basis for its precise positioning. This step mainly USES the displacement sensor in the wireless sensor node of WBAN to detect the moving target by comparing the input image with the background reference image. Sensor is a kind of high performance sensor used to measure three-dimensional motion attitude. This type of sensor is mainly a measurement System based on Micro Electro Mechanical System (MEMS) technology. Generally, the MEMS system consists of triaxial gyroscope, triaxial accelerometer, triaxial electronic compass, triaxial magnetic sensor, triaxial angular velocity sensor, triaxial acceleration sensor and other motion sensors.

Set the background reference image at time n as \( f_{bn}(x,y,n) \) and the current frame as \( f_{c}(x,y,n) \), then the background difference image is:

\[
f_{d}(x,y,n) = \{f_{c}(x,y,n) - f_{b}(x,y,n)\}
\]  \( (1) \)

Formula (1) shows that the background difference value of each pixel is \( f_{d}(x,y,n) \), so we can get the change detection mask as follows:

\[
e(x,y,n) = \begin{cases} 1, & \text{iff } f_{d}(x,y,n) > T \\ 0, & \text{other} \end{cases}
\]  \( (2) \)

The judgment basis of moving target detection is as follows: if the difference value is greater than \( T \), then \((x,y)\) points belong to the moving target pixel, and vice versa. The \( T \) here is the threshold. The change detection mask set at the time of update interval \( n \)
and \( n-1 \) is \( c(x,y,n) \) and \( c(x,y,n-1) \) respectively, then the motion template \( M(x,y,n) \) at the time of \( n \) is obtained from the images of the two adjacent update intervals, as follows:

\[
M(x,y,n) = \frac{C(x,y,n-1)}{C(x,y,n)}
\]  

\( M(x,y,n) \) is a target factor that determines which pixels in the current frame are used to update the current background. The background obtained according to formula (4) is called the instantaneous background, and then according to formula (5), the weighted average of the instantaneous background and the current background is used to obtain the current new background.

\[
f_b(x,y,n) = \begin{cases} 
  f_b(x,y,n), & \text{if } M(x,y,n) = 0 \\
  f_b(x,y,n-1), & \text{if } M(x,y,n) = 1 
\end{cases}
\]

\[f_b(x,y,n) = \tilde{\varphi}_b(X(X),n) + (1-\tilde{\varphi}_b(X(X),n-1))\]  

Where \( \tilde{\varphi}_b \) is the update factor, \( \tilde{\varphi}_b \) affects the background update speed, which is generally selected by the experiment. After several experiments, \( \tilde{\varphi}_b \) is selected as 0.1.

Hall Monitor sequence was selected for the experimental video stream. The resolution of the image frame was 352×288, and the image was frame 40. The experimental platform is celeron 1.7g, 256mb of memory, compiled with VC++, processing speed of about 15 frames/second. The generation process of the new background in this paper is set as (a), (b), (c), (d), (e), (f), and (a) is the current frame \( f_c(x,y,n) \); (b) is a motion target template; (c) background \( f_i(x,y,n-1) \) of the previous frame; (d) and (e) are the instantaneous background corresponding to the image when \( M(x,y,n) \) is 0 and 1 respectively; (f) new background image for final production.

2.3 Target feature extraction

When object detection is conducted in the video, the purpose is to identify and detect the part of the region (target region) that you are interested in from the complex and changeable background environment, and then calculate the general trajectory of the target to achieve accurate target location.In general, three criteria of real-time, accuracy and robustness are used to judge the target tracking system [5]. Real-time is in the detection of the target, the system is expected to be real-time, continuous detection; Accuracy is in the detection of the target, hope to be able to correctly detect the target as accurate as possible; Robustness means that when the external environment is constantly changing, the detection system is expected to be robust enough to ensure the effectiveness and feasibility of detection. Usually, when detecting and tracking the target, a feature of the target is selected to match with the video image sequence. The selection of features is related to the accuracy of subsequent positioning. The commonly selected target features include color features, edge features, texture features and shape features.

Color features. In the description of color space, RGB color space and HSV color space are usually used. The RGB color space is a color space composed of red (R), green (B), and blue (G). HSV color space is a color space composed of chroma, saturation and brightness. In the description of color features, there are many ways to use color histogram, which specifies the approximate distribution information of color, and the effect is
better. However, only the quantitative distribution information of the color is specified, and the position information of the color body itself is not specified. If similar color distribution information appears near the tracking target, it will lead to tracking error.

Marginal features. Edge feature is mainly the edge distribution information of the image. The edge information characteristics described in this way are better tracked on general targets. If the edge characteristics of the target to be tracked are not so obvious, the tracking effect based on this method is not ideal.

Texture features. Texture features represent a relationship between the pixels of the target region in the image. It is only a statistical feature on the surface of the target, with strong regularity. In this way, the influence of noise can be effectively avoided.

Shape characteristics. Shape feature is to extract some shape information of the known edge region. The shape information includes the shape information of the area and the shape information of the target contour. In this way, the shape information of the whole object can be obtained.

Directional gradient characteristics. The feature information of the direction gradient is obtained by calculating and constructing the histogram of the direction gradient in an area of the image. By dividing the image into several smaller parts, the histogram of the direction gradient of each small region is calculated. The histogram of the direction gradient of each small region can uniquely describe the information of this region.

2.4 Image preprocessing

By high speed CCD camera filmed for the rail fastener continuous video, consecutive video frames is actually made up of many images of (general collection of video is 25 pour/seconds, namely a second 25 images), WBAN and Internet of things through the low-power embedded ARM processor in the system, you can get the three-dimensional human body stance, Angle and azimuth data. In addition, quaternion algorithm and data fusion technology can also be used to obtain special data, and the output data are usually collected in the form of quaternion and euler Angle. Need to be aware of is the use of WBAN continuously and comprehensive data acquisition and the Internet of things will produce a large amount of data, how to filter integration of useful data, eliminating redundant information, which is a key research question, the key is the need of image preprocessing, which must be poured video into images, including smoothing denoising and extract the image edge.

Main purpose of image smoothing can reduce the noise, the noise influence on image signal amplitude and phase is very complex, some noise and image signal are relevant, some have no correlation, also may have a correlation between the noise itself, so if you want to reduce the noise in the image, must according to specific situation adopt corresponding measures, smooth denoising is a kind of can suppress the image noise reduction and image processing technology, from the perspective of the image quality is the main purpose of image enhancement is to improve the image of intelligible degree, can make the enhanced image is more suitable for the human eye observation or is more suitable for machine vision recognition system. When the camera located at the bottom of the monitoring system collects the video of the target coupler, the image quality will inevitably be reduced due to various interferences such as the external environment. In this paper, WBAN enhancement method is used to enhance the image [6], which will obtain good image quality.

The edge of the image is the most basic feature of the bottom layer of the image. The edge information of the object is expressed in the form of the discontinuous local characteristics of the image, such as the change of image gray level, the change of image color and so on. Edge detection can greatly reduce the amount of data, and remove the information that we think is irrelevant and only keep the important attributes of the image structure. In this
paper, edge extraction is divided into two processes: training and testing [7]. In this paper, the rail processing images have rich texture features and the grey value change [8], so this article selected two characteristics to process the image brightness and texture, using machine learning methods, the sleeper edge character and the edge feature extract and obtained the parameter values, we need training process is as follows:

First, 50 images were selected as the training sample set.

Second, the edge segment and non-edge segment of the training sample set are marked. This paper is marked red, so the pixel point can be judged to be on the edge or non-edge point according to the R channel value of the image.

Third, randomly get a certain number of each image pixel, three over ten of them are case also is the point of the edge of the sleeper, seven over ten take negative cases is also edge points, and then calculate the histogram of each operator eigenvalue, because of the brightness, texture feature value has eight directions, and we focus on close to the brightness of the horizontal direction or texture characteristic value, so only select 0, PI / 8, 7 * PI / 8 three directions, computation formula is as follows (6):

\[
x^2(g, h) = \frac{1}{2} \sum_{i} \left( \frac{g_i - h}{g_i + h} \right)^2
\]

(6)

X stands for pixel point, g for brightness, and h for image height.

Finally, the calculated eigenvalues are sent to the classifier for training. The function selected in this paper is:

\[
Y = \frac{1}{1 + \exp(-x^2)}
\]

(7)

Fifth, the trained parameter beta value is calculated.

Secondly, the process of testing. The parameters obtained from training are put back into the original program. Through comparison and research, this paper selects the characteristic values with a calculation radius of 0.025 and a calculation direction of three directions, namely the final test effect diagram, it can be found that the edge of the sleeper can be clearly detected. The edge detected is a discontinuous edge pixel, and the eight connected regions of the pixel are connected to the corresponding length of the line segment, so as to obtain the required rail edge segment and sleeper edge segment.

2.5 Generation of panorama and perspective

Generally, the omni-directional image obtained by the monitoring system is an omni-directional image. Since the omni-directional image is a compressed circular panoramic image, it is necessary to expand the image in order to make it consistent with human observation and understanding habits. At the same time, it is also convenient to use the computer for further analysis and processing. When the panorama is expanded, the 360° ring information can be obtained. When the perspective is expanded, it is more in line with the perspective observed by the human eye, facilitating the processing of details. Meanwhile, relative to the panorama, some spatial distortion phenomena in the generated image are reduced [9].
The development process of the omni-directional image is the inverse process of the omni-directional image generation process. In the process of the panoramic image, the omni-directional image on the CCD is expanded to a cylinder in space. During the process of perspective expansion, some points in the omni-directional image are restored to a specific point in space. When the panorama is expanded, all points in the omni-directional image are mapped to the cylinder with radius L, elevation $\phi_1$, and depression $\phi_2$. Any point $P(X, Y, Z)$ on the cylinder satisfies the following conditions:

$$x^2 + y^2 = L^2$$  \hspace{1cm} (8)

The cylinder is crosscut by $y=0$ and $x=L$, and the cylinder is drawn into a plane from the $xy$ plane in the order of 1, 2, 3 and 4 quadrants. The points mapped to the plane are a panorama generated by the omni-directional image. The coordinate origin of the panorama is the point where the $x$ axis intersects the cylinder in the positive direction. In the spatial coordinate system, the $z$-axis coordinate value is the $Y$-axis coordinate value of the panorama. Here, the intersection line of the original $xy$ plane and cylinder is taken as the positive direction of the $x$-coordinate of the panorama along the stretching direction, and the corresponding relationship between the following coordinates of the panorama $(x_1, y_1)$ and the omnibearing image $(x, y)$ is obtained:

$$x = \frac{\cos(x_1/2L)}{2ac(a^2 + c^2)}$$ \hspace{1cm} (9)

$$y = \frac{\sin(x_1/2L)}{2ac(a^2 + c^2)}$$ \hspace{1cm} (10)

In the viewpoint of the focus of the hyperboloid namely space coordinate origin generated perspective, for mapping plane is G, in the center of the plane as the origin of coordinates, a bit in the plane of the G $(gx, gy)$ and space coordinates, the origin of the lines in the $x, y$ plane projection with the $x$ axis is the direction of the crossing Angle for Ox, and the line in the plane of projection and axis is the direction of Oy, get a G $(gx, gy)$ corresponding relation with the space coordinates:

$$x = (L \cos O_x + g_x \sin O_x)$$ \hspace{1cm} (11)

$$y = (L \cos O_x + g_x \sin O_x) \sin O_y$$ \hspace{1cm} (12)

$$z = L \sin O_y - g_y \cos O_y$$ \hspace{1cm} (13)

Before generating the panorama and perspective, the first thing to determine is the value of the endpoint coordinates in the direction of the height of the required generated image, that is, to determine the $z$ value of the two endpoints on the $Y$-axis of the panorama in the spatial coordinate system. Among them, one method is to calculate the spatial coordinates of the upper and lower endpoints by expanding the length of the axis after calculating elevation $\phi_1$ and depression $\phi_2$ through a and b. This method is simple and does not require prior knowledge of the inner and outer diameters of the omni-directional image. Another way is to take the $xz$
plane in space, and get the points in space for \( r \) and \( r \). The transformation formula is obtained through mathematical analysis as follows:

\[
y = \frac{2ac \sqrt{T^2 + x^2} - fL(a^2 + c^2)}{b_x x}
\]

(14)

The advantage of this method is that when the inner and outer diameters are manually delineated on the omni-directional image, the generating range of panorama and perspective can be changed, and the data deviation caused by the fixing device on the omni-directional lens shell to the first method can be avoided [10]. After the coordinate values of the upper and lower vertices of the height direction of the generated graph are determined by the above method, the respective \( Y \)-axis coordinate values of the top and bottom of the panorama are the coordinate values of the upper and lower vertices. After determining the top and bottom coordinates of the panorama, the value range of the upper and lower endpoints in the perspective can be obtained through trigonometric analysis. The process of generating a panorama and perspective is the process of traversing each point in the generated empty image and starting to find the corresponding point on the omnibearing map.

In the process of panorama generation, it can be found from the corresponding calculation formula that, except for the calculation part of trigonometric function, which is related to the width value \( x \) of the image, most of the calculation is only related to the height value \( y \). In this way, when traversing every point in the panorama, the points in the height direction should be taken as an external loop, so that only the data related to the \( X \)-axis should be calculated in traversing every row, thus greatly reducing the calculation amount of the algorithm. In the process of omni-directional image development, the selection of \( L \) is a key problem. The selected value of \( L \) is directly related to the computational amount of the computer, the quality of the generated image, and the pixel loss and repetition of the image. Of course, the value of \( L \) can be fixed according to the requirement of panoramic display width \( W \) in the project. The advantage of this method is that when the image resolution is large, the calculation amount can be minimized according to the need of display. At this point:

\[
L = \frac{W}{2\pi}
\]

(15)

In general, the generation height of the image can be generated \( h = R - r \), where:

\[
L = \frac{h}{(\phi_1 + \phi_2)}
\]

(16)

At this point, the value of \( L \) varies according to the height of image generation. The advantage is that when the resolution of the original image is different, the quality and speed
of image generation can be effectively guaranteed. In the practical engineering application, we can combine the two and judge and choose according to the actual situation. If the resolution of the original image is within a certain range, and the resolution of the original image is large enough to make the panorama take a long time to generate, then L is considered to be fixed to save the expansion time.

The expansion of the panorama maps the imaging points in space to a cylinder with the z-axis as the axis. When an object in space is not parallel to the z-axis, due to its mapping relationship, it will cause distortion in the z-axis direction after the object is projected. The larger the Angle of intersection with the z-axis, the greater the distortion will be. Figure 2 is an example of image distortion. The line segment L₁=L₂ in the space is not the same length when expanded into a panorama, resulting in distortion in the panorama. Therefore, in order to obtain an accurate positioning in the final positioning link, it is necessary to generate the standard panorama as far as possible to reduce image distortion.

![Fig. 1. Distortion during omni-directional panoramic projection](image)

### 2.6 Target positioning

The final implementation process of the target positioning of campus video monitoring system based on WBAN and Internet of things, namely, the final target positioning process is as follows:

The first step is to scan the image line by line and search for the pixel with a gray value of 255 in the image. Set to get the first gray value is 255 pixels in the image \((x₀, y₀)\), have \((x₀, y₀)\) as the seed pixels onto the stack, and give it marked as a motion object Movingobj₀ (also is a prospect target), and the object pixel count set to 1, the grey value of pixels in the current frame \(f_0(x₀, y₀)\) as seed pixel gray value, \(M(x₀, y₀, n)\) assigned a zero value.

The second step is to take out the seed pixel \(f₀(x₀, y₀)\) from the stack. The gray value of the pixel \(fₙ(x₀, y₀)\) in the current frame is taken as the gray value of the seed pixel. Taking it as the center, the neighborhood pixel is checked. If the domain pixel \((x₁, y₁)\) satisfies \(M(x₀, y₀, n) = 1\), the pixel is considered to be part of Movingobj and pushed onto the stack. Otherwise stop the area from growing further. In order to accelerate the search speed, this paper USES the 5X5 neighborhood window to search. That is:

\[
M(x₁ + k, y₁ + m, n) = \frac{1}{fₙ(x₁, y₁)}
\] (17)
After scanning the 5X5 domain pixel \((x_0, y_0)\) of the pixel, go back to the beginning of this step until there are no seed pixels in the stack.

Third, determine the minimum box containing the moving object Movingobj0. Its target is within range of region \(\{(x, y)\mid x_{\text{min}} \leq x \leq x_{\text{max}}\}\).

Fourth, repeat the above steps until the gray value in the image is zero.

Fifth, according to the number of pixels count of each moving object, remove the moving objects whose count is less than the range, because these are often false targets formed by noise or unprocessed shadows. The following description of the target motion can usually be achieved by describing the trajectory of the target center of mass.

3. Test experiment

In this paper, a target positioning technology of campus video surveillance system based on WBAN and Internet of things is designed. In order to verify the practical application effect of this method, a comparative experiment is designed with the common monitoring system target positioning method. Through comparison, the experimental hypothesis is verified.

3.1 Experimental contents

In the experiment, a school playground was used as the experimental site. Four cameras were installed in the site to shoot a group of video simultaneously. Experiment site pedestrians in 89, as shown in figure 2 and figure 3 is one of the cameras interval of 10 s acquisition of image information, this section is the red box in figure 2 and figure 3 identification of target localization experiment, the blue box identification of target tracking experiment, set a goal according to the given trajectory motion, one goal some shade.

![Fig. 2. Target position 1](image-url)
3.2 Multi-angle target positioning test and analysis in video monitoring

For the target identified by the red box in FIG. 2 and FIG. 3, the proposed method and the traditional method are respectively adopted to locate the target from different perspectives. The positioning results of target position 1 and 2 by the two methods are shown in table 1 and table 2.

<table>
<thead>
<tr>
<th>Angle of view</th>
<th>Actual coordinates/m</th>
<th>The method of this paper</th>
<th>conventional method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(366.27,394.45)</td>
<td>(366.27,394.46)</td>
<td>(369.67,396.59)</td>
</tr>
<tr>
<td>2</td>
<td>(366.27,394.45)</td>
<td>(366.26,394.41)</td>
<td>(370.21,398.32)</td>
</tr>
<tr>
<td>3</td>
<td>(366.27,394.45)</td>
<td>(366.27,394.44)</td>
<td>(371.25,399.12)</td>
</tr>
<tr>
<td>4</td>
<td>(366.27,394.45)</td>
<td>(366.28,394.42)</td>
<td>(369.58,397.89)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Angle of view</th>
<th>Actual coordinates/m</th>
<th>The method of this paper</th>
<th>conventional method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(373.19,394.75)</td>
<td>(373.22,394.72)</td>
<td>(375.86,397.21)</td>
</tr>
<tr>
<td>2</td>
<td>(373.19,394.75)</td>
<td>(373.18,394.72)</td>
<td>(377.68,399.83)</td>
</tr>
<tr>
<td>3</td>
<td>(373.19,394.75)</td>
<td>(373.25,394.78)</td>
<td>(370.12,391.17)</td>
</tr>
<tr>
<td>4</td>
<td>(373.19,394.75)</td>
<td>(373.19,394.28)</td>
<td>(369.26,396.55)</td>
</tr>
</tbody>
</table>

It can be seen from the analysis of table 1 and table 2 that, compared with the traditional method, the positioning coordinates obtained by the method in this paper are closer to the actual coordinates for targets from different perspectives.

In order to verify the positioning accuracy of the proposed method more directly, the positioning errors of the proposed method and the traditional method are compared on the basis of the above experiments. The calculation formula of positioning error is as follows:

$$d = \sqrt{(x_i - x_r)^2 + (y_i - y_r)^2}$$  \hspace{1cm} (18)

In equation (18), 1 is used to describe the real position coordinates of the target; 2 is used to describe the target position coordinates obtained by different methods. On the basis of the above experiments, the positioning errors of the method in this paper and the spring method
are calculated through equation (18), and the comparison results are described in table 3 and table 4.

Table 3. The comparison results of positioning errors between two positioning methods of target position 1

<table>
<thead>
<tr>
<th>angle of view</th>
<th>The method of this paper/m</th>
<th>conventional method/m</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.06</td>
<td>0.13</td>
</tr>
<tr>
<td>2</td>
<td>0.09</td>
<td>0.18</td>
</tr>
<tr>
<td>3</td>
<td>0.07</td>
<td>0.20</td>
</tr>
<tr>
<td>4</td>
<td>0.11</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Table 4. The results of positioning error comparison of two positioning methods of target position 2

<table>
<thead>
<tr>
<th>angle of view</th>
<th>The method of this paper/m</th>
<th>conventional method/m</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.03</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>0.06</td>
<td>0.21</td>
</tr>
<tr>
<td>3</td>
<td>0.05</td>
<td>0.17</td>
</tr>
<tr>
<td>4</td>
<td>0.07</td>
<td>0.25</td>
</tr>
</tbody>
</table>

It can be seen from table 3 and table 4 that the maximum positioning error of the method in this paper is 0.11, lower than the minimum positioning error of the traditional method, which indicates that the positioning accuracy of the method in this paper is higher than that of the traditional method and the positioning performance is excellent.

4 Conclusion

This paper analyzes the target positioning technology of campus video surveillance system based on WBAN and Internet of things, and locates the target through the method of WBAN and Internet of things. The local function is obtained from the video data in different angles. According to the generation of panorama and perspective to achieve information fusion, to achieve the goal of positioning. Experimental results show that the proposed method has high positioning accuracy. It is hoped that the research in this paper can provide theoretical basis for the target positioning method of campus video surveillance system based on WBAN and Internet of things.

Headings. Please follow the formatting instructions for headings given in Table 1.

Tables. All included tables must be referred to in the main text and the table title and caption are to be positioned above the table. The captions need to be written in Times New Roman, 9pt.

5 Fund project

Science and technology project of Jiangxi Provincial Department of education.
Project No.: GJJ171481. Project Name: Research on Key Technologies of campus video monitoring system based on Internet of things. Project type: Youth Project

References

Research on multimedia digital real-time signal transmission based on probability and statistics

LI Man\(^1\), YU Yu-min\(^2\), MA Feng\(^3\)

\{liman98855@163.com\(^1\), yuyumin68544@163.com\(^2\), mafeng3213@163.com\(^3\)\}

(1. School of Mathematics and Statistics, Nanyang Institute of Technology, Nanyang 473000 China; 2. School of Mathematics and Statistics, Nanyang Institute of Technology, Nanyang 473000 China; 3. School of Mathematics and Statistics, Nanyang Institute of Technology, Nanyang 473000 China)

Abstract. Aiming at the problem of poor transmission effect and low transmission integrity of multimedia digital real-time signal, this paper studies the optimization, and puts forward a method of multimedia digital real-time signal transmission based on probability and statistics method. Based on the principle of information feature collection, a real-time signal transmission model of multimedia digitization is constructed, and the transmission rate of multimedia digitization real-time signal is effectively controlled according to the collected information, so as to ensure the integrity of signal transmission. Finally, the experiment proves that the multimedia digital real-time signal transmission method based on probability statistics method has high integrity and can better improve the signal transmission effect.

Keywords: Probability statistics; multimedia; digitization; real time signal; signal transmission

1 Introduction

The accurate transmission of multimedia digital real-time signal has good benefits for long-distance real-time signal transmission, but at the same time, the long-distance real-time signal transmission link is interfered by many factors in the process of signal transmission, which is easy to cause the problem of transmission delay. Because the multimedia signal is very large, it needs a good compression method for digital multimedia to guarantee the signal quality. Only when the signal is compressed to a certain extent can real-time transmission be carried out, but there is no reliable guarantee in the current real-time signal transmission link. In the transmission process of weak network signal, a long time delay will cause more compression packets to be lost, resulting in the sluggish and unclear phenomenon of multimedia signal. Therefore, it is necessary to adjust the transmission delay in the long-distance real-time signal transmission link to meet the smoothness and reliability of multimedia digital real-time signal. The key problem of real-time signal transmission delay is pointed out. Furthermore, the influence of transmission delay of multimedia digital real-time signal is discussed. After a series of comparison, analysis, research and discussion, the problem
points of transmission delay are obtained. Finally, the problem of multimedia digital real-time signal delay is improved to solve the problem of transmission delay and improve the signal sensitivity of transmission delay.

2 Multimedia digital real-time signal transmission

2.1 Multimedia digital real-time signal acquisition

According to the location relationship, size and total number of spatial objects of the multimedia digital real-time signal's neighbor objects, the specific feature collection process is as follows:

It is assumed that in the process of multimedia digital real-time signal transmission, the order of signal space filling curve is m, and the spatial range of signal set $S$ can be divided into $2^M \times 2^M$ grid, each grid has four-dimensional Hilbert coding [1]. The spatial four-dimensional Hilbert coding linear filling curve aggregation feature is used to decompose the massive signal coding blocks, mark the corresponding storage sequence of each coding block $H_2$, and form the corresponding spatial signal partition matrix, as shown below:

$$ F = \begin{bmatrix} H_{2de0} & H_{1a0} & s_0 \\ ... & ... & ... \\ H_{2den} & H_{1em} & s_n \end{bmatrix} \quad (1) $$

In the above algorithm, $H_{2de0}, H_{1a0}, s_0$ represents the space element of real-time signal. If the source signal \(S = [s_1(t), s_2(t), ..., s_n(t)]^T\) of real-time signal is an unknown n-dimensional source signal vector, \(a\) is an unknown mixed matrix, \(S = [s_1(t), s_2(t), ..., s_n(t)]^T\) is a dimensional noise vector, and \(X = [x_1(t), x_2(t), ..., x_n(t)]^T\) is the output dimensional observation signal vector, then:

$$ X = AS + Fm \quad (2) $$

According to this element, the corresponding signal coding and the corresponding massive signal block storage label after matrix matching are obtained, thus completing the signal element division [2]. According to the single division of the corresponding signal feature set, the collected multimedia digital real-time signal assumes that the length is larger than the width, then the transverse direction of the massive signal is a positive direction set, and the longitudinal direction is a reverse direction set, so as to calculate the spatial element interval of
the massive signal. Through this interval, the coding block is decomposed to realize the acquisition and division of signal characteristics [3]. According to the results of real-time signal feature division of multimedia digitization, the massive signal is controlled through protocol transformation to realize real-time collection, analysis and combination of massive signal in the process of large-scale transmission. The model construction is shown in Figure 1.

Take communication server as the core, maintain the connection between signal converter and client, and provide support for signal transmission control [4]. All the non current signals inside the server are converted into the history signal library, and the real-time signals are read through calling and communication interface [5]. There are usually two kinds of signals in the history signal library, namely control command signal and monitoring command signal [6]. Once the multimedia starts or stops, the control command of signal transmission forms the signal frame through the communication device, and then it is transmitted to the charger which can automatically cut off the power through the protocol. The problem of signal transmission concurrency is solved by multi thread scheduling, which provides a good environment for the realization of control technology. In the process of mass multimedia real-time data transmission, the problem of path blocking of transmission nodes is very easy to occur. In the process of signal characteristic data packet transmission, large-scale collision is the main cause of data packet loss. At the same time, due to the limitation of data link communication time limit, multimedia real-time signal transmission is incomplete. In order to achieve the integrity of multimedia real-time signal transmission, transmission integrity control technology needs to be studied. The specific research content of control principle is as follows:

① Determine the coordinates of signal transmission initialization node, determine the grid and conflict area of each signal transmission node, and set the node counter;
② Count the data transmission speed of the signal transmission node;
③ Check whether all nodes are of the same type of signal transmission node. If so, record the transmission speed of each node in the signal transmission grid. If not, the transmission speed of the signal transmission node itself needs to be calculated;
④ Check whether the transmission cycle ends. If yes, the average speed of the new
composite node and grid of this type of node needs to be published. If not, the transmission speed of the node itself needs to be calculated;

⑤ According to the signal transmission routing protocol, the transmission priority of the signal transmission node is calculated;

⑥ Check whether all signal transmission nodes are of the same type. If so, record the transmission speed of each node in the grid. If not, adjust the current channel competition window;

⑦ Check whether the energy is exhausted. If so, the node exits. If not, check whether the cycle ends, if so, it ends. If not, the new cycle begins. Further standardize the real-time information transmission control level, as shown in the table below

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Control method</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t'_s &lt; t'$</td>
<td>No control measures</td>
</tr>
<tr>
<td>$t'_\text{min} &lt; t' &lt; t'_s$</td>
<td>Changing data transmission frequency</td>
</tr>
<tr>
<td></td>
<td>Adjust transmission rate</td>
</tr>
<tr>
<td></td>
<td>Take anti-interference measures</td>
</tr>
<tr>
<td>$t' &lt; t'_\text{min}$</td>
<td>Replace the data link network</td>
</tr>
</tbody>
</table>

According to the information in the table above, the multi-level division of multimedia digital real-time signal features is carried out to provide data support for the integrity acquisition control value.

2.2 Multimedia digital real time signal transmission rate control

Based on the above signal acquisition results, further analyze and control the wave frequency of non concurrent signal transmission, and study the integrity of mass information transmission and the precise effect of signal control technology of multimedia digital real-time signal in the same network storage environment [7]. In view of the congestion problem in the transmission process of massive multimedia digital real-time signals, it is necessary to analyze the congestion situation of each transmission node and select the optimal signal output path [8]. Draw and display the specific signal congestion situation, as shown in the figure below:
Based on the above characteristics of congestion wave frequency, the collected inflow signal wave frequency is detected and recorded to control and process the transmission speed of morning signal packet. It is equally allocated to its own nodes and other sub nodes to ensure the accuracy of the control of the integrity of parallel signal transmission [9]. Record the transmission time and interval of the signal packet, and calculate the signal service time and arrival time by using the mobile weighting method. When the arrival time is less than the service time, the signal transmission congestion is more serious; otherwise, it means that the signal congestion is not serious and the signal transmission speed is faster [10]. According to the above analysis results, cl-aptc protocol and probability statistical method are used to control the single node signal transmission rate, actual transmission rate and signal chain transmission effectiveness. The specific control steps are as follows:
Based on the above steps, the wave frequency generated in the process of signal transmission is further managed and controlled. When the storage space of a node is less than the transmission rate adjustment threshold, the signal node will not be congested; When the storage space of a node is larger than the transmission rate adjustment threshold and smaller than the maximum storage space of the node, the signal node may be congested. The signal transmission rate is adjusted based on the wave frequency of single node signal transmission [11]. Determine the size of the signal transmission wave frequency, and record the information transmission wave frequency rate of the most node as 0.

The expected value of the average transmission signal through \( n-1 \) -period grid is recorded as \( e \), the threshold value of signal transmission rate adjustment is \( T_1 \), the maximum storage space is \( N \), the total number of information transmission nodes in a grid is \( r \), when \( E < r^*T_1 \), the signal transmission will not be congested; when \( r^*T_1 < E < r^*N \), the signal transmission will be congested, the node transmission rate needs to be adjusted.

By setting the node transmission rate weight, the optimal input and output speeds of different nodes are obtained [12]. Based on the above-mentioned principle and the commonly used DSP algorithm, the two-way address of information transmission is accessed in one way. In the process of signal transmission, it is easy to generate some abnormal signals and interference values, which leads to differences in relatively fixed signal transmission [13]. Therefore, it is necessary to improve and optimize the signal transmission path in combination with the principle of neural network, and further optimize the common neural...
network structure, as shown in the figure below.

Fig. 4. Information transmission neural network structure

Based on the above structure, the frequency conversion parameters of the received multimedia digital real-time signal are further optimized. If $q$ is the baseband signal formed by the inherent frequency conversion of the equipment, it can be expressed in $iR_n \leq p \leq (i+1)R_n$ as follows:

$$s_i = X + \frac{r^* T + r^* N - 1}{2F(q_i e^i + z_i)} (3)$$

In the above algorithm, $R_n$ represents the signal transmission period; $z_i$ represents the mean value of information transmission; $q_i$ represents the baseband form of multimedia data. The multimedia signal received by VDE receiver of multimedia data has the problem of time delay and frequency offset. Therefore, it is necessary to estimate the time delay and frequency offset of the signal and correct them [14]. For the modulated baseband signal of multimedia wireless ad hoc network, the baseband form $q_i$ in $iR_n \leq p \leq (i+1)R_n$ can be
expressed as:

\[ q_i = \sqrt{2p'e^i} \quad (4) \]

Where \( p' \) is the monitoring power of multimedia digital real-time signal transmission. The current algorithm for the transmission symbol phase of multimedia signal can be recorded as follows:

\[ f_i = f_{i-1} + \Delta a \quad (5) \]

In the above algorithm, \( f_{i-1} \) represents the characteristic phase at the end of the previous signal transmission symbol, and \( \Delta a \) represents the phase change of the current information transmission symbol. The transmission rate of multimedia real-time signal is further calculated, and the specific algorithm is as follows:

\[ v_0 = \log_2 \left( \frac{f_i(K'-I)}{H_1 + D_0} \right) H - k \quad (6) \]

In the above algorithm, \( D_0 \) represents the total amount of signals; \( H_1 \) represents the storage size of signal blocks. The information set of statistical coding signal elements is recorded as \( K' \). Assuming that the total number of coding blocks of multimedia digital real-time signal is \( I \), if the storage size \( H \) of the signal block is greater than the maximum threshold percentage of the storage mass signal block, then the coding speed should be divided into sample set \( k \). Based on the above algorithm, it can effectively optimize the transmission rate of multimedia digital real-time signal and improve the effectiveness of multimedia digital real-time signal transmission.

2.3 The realization of multimedia digital real-time signal transmission

Based on the above algorithm, the multimedia digital real-time signal transmission method is further optimized. If the standard value of multimedia digital real-time signal transmission in a secure environment is \( e \), in order to avoid other factors interference, ensure the information security of users, and avoid the problem of transmission data leakage. Combined with probability and statistics algorithm, the transmission signal is authenticated and judged. In the process of signal transmission, the user's identity information and password, scanning information and related digital password are authenticated and retrieved [15]. After the realization of identity authentication, the real-time signal transmission
protocol is further optimized. Aiming at the non coordinated real-time signal transmission characteristics, fuzzy inference technology is used to provide technical support for automatic modulation recognition of signal. After acquiring the new real-time signal transmission characteristics, it is necessary to judge whether the previous priority thread of real-time signal transmission is blocked and whether the latter priority thread is occupied, and observe the likelihood of different real-time signal transmission signals. The specific calculation formula is as follows:

\[
S = f_1(a) \sum_{i=1}^{n} k \omega_i (v_0 - 1)^n
\]

In the above algorithm, \( n \) is the real-time signal transmission quantity; \( \omega_i \) is the weight of signal component \( i \); \( f_1(a) \) is the analysis model function. According to the above algorithm, the likelihood degree of real-time signal in different periods can be obtained. According to the likelihood degree, the transmission path of real-time signal transmission can be accurately simulated, and then the output path of signal can be selected and tracked.

![Fig. 5. Real time signal output path](image)

The signal is identified by the signal tracking results obtained above. The tracking signal path is divided into three paths, which are the transmission path of the source, the intermediate transmission path and the receiving path of the terminal. The three paths are transmitted to \( f_1(a) \), and the signal likelihood is judged in the analysis model function. The method of product reasoning is used to identify the signal. The recognition process is as follows:

1. Obtain multimedia real-time signal through A / D converter;
2. Analog signal digitization, signal processing, acquisition of complex signal and spectrum;
3. The amplitude and frequency of instantaneous transformation are obtained by using
complex signal, and the variance and accumulation of instantaneous amplitude and the maximum value of power spectral density are obtained respectively;

(4) Set the parameters of each threshold, and use the automatic modulation method to identify the signal characteristics, and display the identification results through the screen.

The periodic rotation system is used to update the real-time signal, construct the signal feature cluster, and divide it into two output stages, namely the initial stage and the stable stage. In order to fast cluster transmission and management of real-time signal acquisition nodes, the specific process is shown in the figure below.

---

**Fig. 6** Real time signal acquisition and cluster transmission management steps

In this process, the multimedia control layer protocol is used to limit the data link for data transmission. In view of the channel conflict during signal transmission, the relevant node transmission rate and data packets are sent. The real-time signal transmission time slot is allocated by using the multiplexing method of shared transmission medium. Multimedia real-time signal transmission module is mainly realized by path covering in the network. In the process of data transmission, we need to avoid the congestion caused by heterogeneous
network as much as possible, and treat the transmission information and multimedia output data for classification processing. Because the order of the packets arriving at the destination is very strict in multimedia transmission, TCP socket is used to realize the order transmission of multimedia. For signal transmission, the arm node should be used as the connection client and the server as the server. The user needs to improve the speed of signal acquisition through the interface. Therefore, the picture transmission process is designed as shown in the figure.

![Multimedia real-time signal transmission process](image)

**Fig. 7.** Multimedia real-time signal transmission process

Based on the above process, it can effectively realize the accurate transmission and management of multimedia digital real-time signal, and improve the accuracy and effectiveness of data transmission management.

### 3 Analysis of experimental results

In order to verify the effect of multimedia digital real-time signal transmission based on probability and statistics method, the traditional methods are compared and detected, and the detection results are recorded. In order to ensure the validity and confirmation of the experimental detection, the experimental environment and experimental parameters are set
uniformly.

3.1 Experimental environment

Using MATLAB software to analyze the integrity of mass data transmission under big data storage, and keep the location of network nodes unchanged. The experimental environment is set as follows: fix the nodes on the plane of $120m \times 120m$, the number of nodes is 80, and the running PC host is configured as Pentium (R) 4cpu2.40 GHz. Further set the experimental parameters, as shown in the table below:

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Parameter setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adapter frequency</td>
<td>12Mb/s</td>
</tr>
<tr>
<td>communication mode</td>
<td>TCP/IP Protocol</td>
</tr>
<tr>
<td>Number of keys</td>
<td>3-5</td>
</tr>
<tr>
<td>Security protocol level</td>
<td>A-F</td>
</tr>
<tr>
<td>Hard disk memory</td>
<td>16GB</td>
</tr>
<tr>
<td>Number of experiments</td>
<td>4</td>
</tr>
</tbody>
</table>

3.2 experimental result

In the above experimental environment, carry out comparative detection, and record the detection results of the two groups of methods. In the experimental process, the data transmission integrity and data transmission security are taken as the reference basis, the higher the two values, the better the signal transmission effect. The specific experimental detection results are as follows.

![Comparison test results](image)

Fig. 8. Comparison test results

3.3 empirical conclusion
Based on the above detection results, it can be seen that compared with the traditional signal transmission method, the multimedia digital real-time signal transmission method proposed in this paper has higher transmission integrity and data transmission security in the practical application process, which fully meets the research requirements.

4 Concluding remarks

In order to solve the problem of poor transmission effect of multimedia digital real-time signal, this paper optimizes and improves the transmission delay and the loss of compression packet in the long-distance real-time signal transmission link. To maximize the quality of signal transmission. Through the research on the influence of transmission delay in real-time signal transmission link, the influence value of long-distance real-time signal transmission can be obtained to improve the reliability, safety and efficiency of transmission link.

Reference


Multimedia self-learning behavior monitoring data mining system based on Web

DENG Chun-Sheng¹, CHENG Hai-Bo², LIN Jian-Ming³, YANG Ju-Rong⁴, ZHU Yan-Chen ⁵*

{hy20160401@163.com¹, jwcxjz163@163.com², xj20160608@163.com³, cl20160629@163.com⁴, dddd24616563@163.com⁵*}

(1.School of Computer, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, China; 2.School of Computer, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, China; 3.School of Computer, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, China; 4.College of Humanities, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, China; 5.School of Computer, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, China.)

Abstract: In view of the poor clustering effect of information data in the use of the original English multimedia independent learning behavior monitoring data mining system, the original data mining system is optimized according to the Web architecture, and a web-based English multimedia independent learning behavior monitoring data mining system was designed. Install the monitoring sensor, data transmission channel and decoder into the original system hardware framework to complete the system hardware design; According to the requirements of Web architecture design, this paper constructs the English multimedia independent learning behavior monitoring database, adopts the data feature point extraction algorithm, and uses the Web mining technology to achieve the system data mining performance, so as to realize the design of the web-based English multimedia independent learning behavior monitoring data mining system. Through the comparison of data clustering effect, it is verified that the designed system can effectively improve the ability of data clustering and improve the performance of system data mining.

Keywords: Data mining; Autonomous learning; K-means algorithm; Correlation mining;

1 Introduction

With the increasing popularity of Internet services, multimedia data resources on the network are unprecedentedly rich. How to extract useful information from massive multimedia data and deal with the information in a way that can be understood by users, so as to dig out regular potential patterns is a research hotspot with practical significance and application prospect(1).

Monitoring and mining of English Multimedia autonomous learning behavior is a combination of data mining technology and multimedia autonomous learning technology. It is
an interdisciplinary research field of knowledge discovery, data mining, artificial intelligence, machine learning, database technology and multimedia technology. Data mining of English multimedia autonomous learning is an intelligent multimedia technology, which elevates multimedia processing and management from information access to knowledge acquisition. Multimedia data is unstructured or semi-structured. Each media data has different characteristics and its own way of expressing information. Each media can not only represent information independently, but also jointly represent different characteristics of the same event, and jointly describe the existence, development and results of the event\(^2\). Therefore, there must be some characteristics, attributes and relations of the information subject in the multimedia data set, or some modes that people can't get intuitively. In previous studies, data mining systems used in this study often had poor data mining effects due to poor clustering performance. Therefore, in this study, Web technology was used to improve the data mining ability of the system.

Web mining the process of extracting patterns and hidden information of interest and potential value from Web documents, media, structures, and user interactions\(^3\). There are three types of Web mining: Web content mining refers to the summarization, classification, clustering and correlation of Web page text and various media contents; Web structure mining refers to the discovery of knowledge in Web chain structure, which can reflect the flow and distribution of information in Web space, as well as the nature and characteristics of Web elements. Web usage mining refers to the use of access path analysis, classification and clustering and other data mining methods to track the interaction and manipulation between users and the Web, including the mining of server access records and the analysis of access path. This technique is used to optimize the original data mining.

2 System hardware design

Aiming at the problem of poor clustering effect in the original English multimedia self-learning behavior monitoring data mining system, Web mining technology is used to optimize the original system. In order to realize the software module function in this design, the original system hardware is designed. The specific system hardware framework is shown below\(^4\).
The above framework is adopted as the basis and control scheme of hardware framework design to ensure the order and controllability of hardware design and improve the process of system design.

2.1 Design of monitoring sensor

In the process of data mining, data collection is the database of mining technology. In order to ensure the effective data source in this design, a data acquisition device based on sensors is set up. Sensor node is the core of data acquisition, mainly composed of sensor module, processor module, wireless communication module and power module, as shown in figure 2.
According to figure 2, in the sensor node, the sensor module is responsible for data collection. The module mainly senses various information through the sensor, and converts the analog parameters into digital parameters, which is convenient for the processor module to process the data\(^5\). The processor module is mainly used to receive and store data. Finally, the wireless communication module is used to realize the wireless data transmission; the power module provides stable voltage, provides energy for the system, and maintains the normal operation of the system.

According to the actual needs of the system, the CC3200 node master control chip is selected. It is the most widely used and powerful control chip. The internal integration of the two large cores provides user development with a high performance processor running at 80MHz, as well as a wide variety of external devices.

2.2 Design of data transmission channel

The design of data mining system is mainly implemented on digital hardware platform. In this design, FPGA (Field programmable gate array) is selected as the hardware implementation platform of English multimedia independent learning behavior monitoring communication function. FPGA has the advantages of low power consumption, abundant resources and high parallelism, which is very suitable for realizing the functions of the physical layer of the system in this paper. In this design, FPGA core digital processing chip is selected as the main processing chip (as shown in figure 3). The AD/DA chip is also connected through the external FMC (FPGA Mezzanine Card) extension interface to realize the transmitting and receiving at the rf level. In addition, crystal oscillator is required to provide a main reference clock for the internal use of the FPGA. In order to realize the data transmission to the upper computer, the external FPGA connects with the external DDR (Double Data Rate) memory through the ax bus, and then transfers the data to the arm processor.
In addition to the above Settings, FMC should also be added to the data transmission device to improve the function of the transmission device[6]. FMC card carrier connector pins are connected with FPGA chip pins that have just configured IO resources through PCB (printed circuit board) design; Connector pins on FMC subboard modules are also connected to IO interfaces by PCB design. Different IO (Input and output) interfaces can be designed on subboard PCB to realize various functions. In this way, the same carrier card can realize different extension functions through different subboard designs, making the chip more flexible to be applied in the scenarios designed by developers. In this project, FMC connector is used to realize the connection between FPGA chip and peripheral AD9361 chip, which can achieve high performance pins of Gb/s and expand FPGA chip to receive AD/DA data. Through the above steps, the construction of English multimedia independent learning behavior data feature extraction module is realized.

2.3 Decoder design

On the basis of the above hardware design, in order to realize remote control of the designed sensor, it is necessary to decode the control signal from the sensor end and the client end, and convert the control signal sent by the software into the actual action level signal[7]. The front end of the decoder is connected to the serial port of the video server through the RS-232/485 converter, and the serial port RS-232 control signal output by the video server is converted into RS-485 signal by the RS-232/485 converter, and then transmitted to the remote decoder. The decoder transcodes the received command signal, obtains the address and action mode of the corresponding cradle head and sensor, controls the action of the corresponding relay, and sends different level control signals to the cradle head and sensor control signal line at the output end of the decoder. The head and sensor are driven by the corresponding control signal level.
The decoder designed above is connected to the data transmission channel and front-end sensor in the paper, and installed into the original hardware framework to complete the hardware design process of the system in the paper. The optimized hardware framework is used as the software development environment to ensure the smooth operation of software modules.

3 System software design

In order to realize the corresponding effect of Web mining technology, the software framework structure in the original system is set as the hierarchy suitable for Web technology, as shown below.

In this design, the software results are set as three layers, namely: presentation layer, business logic layer and data access layer. Presentation layer (UI): this system uses ASP.NET
WEB application development, the client uses IE browser to achieve the user interface; Business logic layer (BLL) : responsible for the processing and data transfer of key businesses. The operation and logical judgment of the database are all processed in this layer. The evaluation process of the system is all processed in this layer. Data access layer (DAL) : responsible for database access, mainly providing data to the business logic layer and manipulating the database based on the parameters passed in. Using ADO.NET data adapters in ASP.NET and SQL server storage process is complete. The structure has the characteristics of convenient system operation, easy maintenance and short information processing cycle[9].

3.1 Build the behavior monitoring database

In view of the problems in the use of the original system, the monitoring data collected by the designed sensor is stored in the form of database, and the stored data is taken as the basis of data mining of the designed system. In order to control the designed database effectively, its internal data table structure is set as follows.

![Database Internal Data Table Connection Structure](image)

**Figure 6** database internal data table connection structure

The above connection results are used as the internal framework of the designed data, and the collected data are edited and stored in the form of data tables. In order to ensure the structural consistency of the information in the database, the following format is set for writing the contents of the data table[10].

<table>
<thead>
<tr>
<th>Serial number</th>
<th>field</th>
<th>form</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>data sources</td>
<td>int</td>
</tr>
<tr>
<td>2</td>
<td>Sensor number</td>
<td>varchar</td>
</tr>
<tr>
<td>3</td>
<td>Data type</td>
<td>varchar</td>
</tr>
<tr>
<td>4</td>
<td>Data information content</td>
<td>varchar</td>
</tr>
<tr>
<td>5</td>
<td>Data acquisition time</td>
<td>varchar</td>
</tr>
<tr>
<td>6</td>
<td>Characteristics of autonomous learning behavior</td>
<td>double</td>
</tr>
</tbody>
</table>
Use the above data table to organize and store the collected data, and use the above database structure to expand the connection, to ensure the effective data mining part.

3.2 Data feature extraction

Set the internal information of the database designed above in the form of data matrix, set there is a kind of learning behavior information in the database, these variables represent the objects in the database, and reflect the database content in the form of matrix, then:

\[
\begin{bmatrix}
  a_{i1} & \ldots & a_{ig} & \ldots & a_{ih} \\
  \ldots & \ldots & \ldots & \ldots & \ldots \\
  a_{i1} & \ldots & a_{ig} & \ldots & a_{ih} \\
  \ldots & \ldots & \ldots & \ldots & \ldots \\
  a_{n1} & \ldots & a_{ng} & \ldots & a_{nh}
\end{bmatrix}
\] (1)

Formula: \(g\) and \(h\) represent the constant parameter of the matrix, and the interval scale is used to obtain the phase differences of the features in the matrix. In order to improve the clustering accuracy of this clustering method, the scale variable is set as the minimum unit. The design of known variables has an absolute impact on the clustering result. Therefore, the standard metric value is set as a constant value.

Based on the above results, the distance between each object is set as \(d(m,n)\), and the characteristic distance between the database contents is obtained by using the minkowski distance function. In the process of using the distance function, the binary variables are represented by 0 and 1. According to the above formula, the degree of dissimilarity between evaluation objects \(m, n\) can be obtained. The specific formula is as follows.

\[
d(m,n) = \frac{e + f}{j + e + f + k} \quad (2)
\]
Formula: \( j \) represents the number of variables when the values of the object \( m, n \) are all 1; \( e \) represents the number of variables when the object value \( m \) is 1 and the object value \( n \) is 0; \( f \) represents the number of variables when the object value \( m \) is 0 and the object value \( n \) is 1; \( k \) represents the number of variables when the values of the object \( m, n \) are all 0.

The calculation formula of the internal information heterogeneity of the system database is as follows:

\[
d(m, n) = \frac{e + f}{j + e + k} \tag{3}
\]

Through the above formula, the feature extraction of data information is completed, and the extracted features are calculated by using the mismatch rate. The specific formula is shown below.

\[
d(m, n) = \frac{p - q}{p} \tag{4}
\]

Formula: \( q \) represents the number of variables with the same state value as \( m, n \); \( p \) represents the total number of variables. Through the above formula, the feature extraction of English multimedia independent learning behavior data is completed, which is used as the auxiliary data support of data mining.

### 3.3 System data mining based on Web mining technology

The extracted monitoring data feature points are used to classify the data. Combined with the content mining technology in Web mining technology, the data are stored in different data warehouses, and the feature weights are obtained. The data are mined with different feature weights according to the actual needs. Part of the code to implement data mining is shown below.
Using the clustering algorithm, all the monitoring data of English multimedia autonomous learning behavior are divided into multiple clusters. Objects in the same cluster have relatively high similarity values, while objects in different clusters have relatively small similarity values. According to this feature, we use the feature weight of the data samples to mine the data, and divide the data into the most similar classes, so as to realize the autonomous learning behavior monitoring of English Multimedia in the process of data mining.

Combining the software module and computing process set above with the system hardware framework designed in this paper, the design of the web-based multimedia English learning behavior monitoring data mining system is realized.

4 Experimental test

Combined with the above hardware design results and software module design, the design of the web-based multimedia English learning behavior monitoring data mining system is completed. In order to ensure the effectiveness of the design, the system testing process was constructed to compare the performance differences between the original system and the designed system.

4.1 System test platform design

In order to understand the performance difference between the designed system and the original system, carry out system test, and ensure that the designed data mining system and the original data mining system can be compared on the same platform, the experimental environment is composed of two parts, namely hardware environment configuration and

```java
class BehaviorMonitor {
    get data() { // receive data
        centroids = new double[]{init_data} // initialize cluster center
        print('Start iteration');
        target = []
        stop = false // initial parameter setting
        while(!stop) { // start traversing data
            stop = true
            points = new double[]{same data} // for i in range
            target = []
            for point in self.data: // calculate the nearest cluster center to the data
                target.append(point)
            for s in range(self.com) // calculate the distance
                if(i == 0) { // count the points contained in cluster
                    count(to end) // count all points contained in cluster
                }
            }
        }
    }
}
```
software environment configuration. During the experiment, the consistency of all equipment should be ensured to avoid the inaccuracy of the experimental conclusion.

<table>
<thead>
<tr>
<th>software environment</th>
<th>hardware environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ubuntu Linux 14.04 LTS</td>
<td>CPU: i7-4790</td>
</tr>
<tr>
<td>Server side</td>
<td>Memory16GB</td>
</tr>
<tr>
<td>RStudio0.98.1091</td>
<td>Hard disk8TB</td>
</tr>
<tr>
<td>Python 3.4</td>
<td></td>
</tr>
<tr>
<td>Windows 7</td>
<td>CPU: i7-4790</td>
</tr>
<tr>
<td>User side</td>
<td>Memory16GB</td>
</tr>
<tr>
<td>Chrome</td>
<td>Hard disk8TB</td>
</tr>
</tbody>
</table>

The above hardware and software were used to construct the experimental environment to ensure the validity of the system test results. The collected English multimedia autonomous learning behavior information database is set as the data sample of this test. The data distribution in the database is as follows.

Figure 8 sample distribution of system test data

The above database is taken as the data sample of this experiment. The original data mining system and the designed mining system are used to process the test sample database, and the data clustering ability of the two databases is compared.

4.2 Test result analysis

Through the above design, the system test process is completed. The specific system test results are shown below.
The experimental results show that the original data mining system is unable to realize the high-precision clustering convergence for the monitoring information of English multimedia independent learning behavior, and some information points are in discrete state. It can be seen from the above that the clustering effect of the original data mining system is poor.

The data mining system designed in this paper has good convergence of clustering results and can complete complete clustering mining within the data set without data dispersion. The data can prove that the original data mining system can not reflect students' learning situation.
comprehensively and timely, and the application effect is not ideal. The data mining system designed in this paper can reflect the learning status of all students and is more effective. To sum up, Web mining technology can effectively improve the classification clustering ability of the system, and the system performance of the data mining system designed in this paper is better than the original data mining system.

5 Conclusion

The web-based multimedia self-learning monitoring data mining system can quickly, efficiently and accurately understand the self-learning ability and behavior status of English multimedia in school. For students, this system can stimulate the motivation of independent learning, strengthen the awareness of independent learning, guide and establish the self-confidence of independent learning, develop correct learning habits, and train the ability of self-evaluation process. For teachers, the system can timely feedback the learning situation, promote teachers to continue to learn relevant theoretical knowledge, and actively explore, adjust and improve learning methods in practice, so as to achieve the educational purpose of constantly improving the ability of independent learning and improve the quality of teaching.

In order to further promote the development of this technology, other performance indicators of the system can be tested, so as to continuously improve and improve the system performance, better realize the mining of English multimedia independent learning monitoring data, continuously improve students' English scores, and promote the development of education in China.

6 Fund projects

Science and Technology Research Project of Jiangxi Provincial Department of Education (GJJ160815)

References


Research on data transmission method of new media marketing analysis based on Web4.0

WU Qi-xiang¹, HE Lin², ZHOU Xian-chun³
{ainiyiwantian255@tom.com¹, hupeng12123@sohu.com², wangyang19111@163.com³}
(1. College of Information Engineering, Fuyang normal university, fuyang 236041, China;
2. College of Information Engineering, Fuyang normal university, fuyang 236041, China;
3. School of Information and Intelligence Engineering, Sanya University, Hainan Analysis and Application Engineering Research Center of Software Gene Virus, Sanya 572022, China)

Abstract: The energy consumption of existing data transmission methods is too high, so a new media marketing analysis data transmission method based on Web4.0 is designed. By designing the structure diagram of transmission method, the reliable data transmission in maintaining system and at the same time as far as possible to reduce the node energy consumption and data transmission of Steiner tree model is set up, cut it off and distributed data do not need the link, reduce the amount of calculation, the design parameters of transmission channel, the research of inductive coupler circuit inductance and the resistance value, on the basis of reducing one hop communication distance, multipath allocation process design, complete the multipath optimization allocation mechanism. So far, the research on the data transmission method of new media marketing analysis based on Web4.0 has been completed. The experimental results show that the designed transmission method can save 176 KWH compared with the existing data transmission method.

Key words: New media marketing; Data transmission; Transmission channel parameters;

1 Introduction

With the development of information technology and Internet technology, web4.0 has gradually integrated into our lives. Under the premise of identity recognition, it integrates all the digital resources and information resources that can be integrated to provide the best information services for users. In this context, many products are also faced with the environment changed by new production technology and new marketing methods[1-2]. For consumers, the new media marketing under Web4.0 allows everyone to experience products from all over the world without leaving home, shortening the time, space, and home delivery experience, which is wrapping up more and more people. Becoming a new force in new consumption patterns has also allowed the online sales scale of various products to continue to break through.

In order to increase product sales and understand the status of online marketing,
marketing analysis data needs to be transmitted. In the existing transmission method, the energy consumption of the transmission node is too high, so a new media marketing analysis data based on Web 4.0 is designed. Research on transmission methods. By designing the structure block diagram of the transmission method, the Steiner tree model of data transmission is established to reduce and balance the energy consumption of nodes as much as possible while ensuring the reliable transmission of the system data. In the process of data distribution, unnecessary links are cut off, the calculation amount is reduced, the parameters of the transmission channel are designed, the circuit of the inductive coupler is studied, and the inductance value and impedance value are obtained. On the basis of reducing the single hop communication distance, the multi-path allocation process is designed and the multi-path allocation mechanism is optimized.

2 Research on Data Transmission Method of New Media Marketing Analysis Based on Web4.0

With the increasing application of Internet Web 4.0, the requirements for data transmission reliability have become higher and higher. Due to the limited energy resources of the nodes in the Internet, when the energy of the nodes is reduced, the communication capacity of the nodes is significantly reduced, which further reduces the reliability of the forwarded packets. Once the nodes do not work properly, the network topology will change, which will affect the data transmission efficiency and overall reliability of the sensor network. Therefore, it is necessary to reduce and balance node energy consumption as much as possible while improving reliability [3-4]. Because the energy-saving of nodes and the reliability of the system depend on the network structure to a great extent, only from the perspective of network structure can we design an efficient and reliable data transmission strategy, and establish a network structure that takes into account the energy-saving of nodes and the reliability of the system. The block diagram of the design transmission method structure is shown below:
The network structure is mainly composed of time structure and space structure. Its model is usually expressed by network parameters. Different network structure determines the properties of each network parameter, and the different values of each network parameter will make the network structure save energy on the nodes with different performance, and make the reliable transmission of system data different. Therefore, the key to the construction of the network structure is how to consider from the time and space structure of the network to reduce and balance the energy consumption of the node as much as possible while maintaining the reliable transmission of system data. The basis for reliable transmission applications.

2.1 Building a data transfer model

Modern data center networks provide multiple discontinuous Steiner trees for reliable group data transmission. Using these disjoint trees to perform data transmission in parallel can effectively improve the transmission rate of reliable group data transmission \([5-6]\). Therefore, multiple data disjoint Steiner tree models in the data center are used to complete the data distribution. The Steiner tree structure is shown below:
As can be seen from the above figure, the Steiner tree is a tree with the data source as the root node and all receiving nodes. In the figure above, the data transmission method based on the packet buffer is also shown. The solid line is the Steiner tree A, and the dashed line is the Steiner tree B. Assume that 00 is the source node, 12, 13, 21, and 33 are the data receiving nodes. Compared with the general creation of multiple spanning trees, the structural model has the advantage that the time complexity is much smaller than the general model, and the calculation can be completed in a short time. Secondly, the spanning tree height calculated by this model is higher than that of the general model. The model is small and similar to the network radius. After computing multiple edge disjoint spanning trees for the data center network, we calculated multiple Steiner trees by pruning, that is, cutting off the links that are not needed for data distribution. The process is shown in the following figure 3.
Fig3 Link pruning process

After computing multiple edge disjoint spanning trees for the data center network, we calculated multiple trees by pruning, that is, pruning the links that are not needed for data distribution. In the above figure, when 0001 is not a receiving node, the link from switch 000 to node 0001 can be reduced. In order to prune the spanning tree, we only need to calculate the set of paths from each node to the source node in the spanning tree. This can reduce the amount of calculation and complete the establishment of the data transmission model.

2.2 Designing the parameters of the transmission channel

The data transmission channel is formed by multiple inductive couplers through wireless high-speed data transmission connections. The electromagnetic inductive coupler consists of a ferrite core and a coil. This article uses nickel-zinc ferrite to combine the two ferrites. Its physical dimensions are shown below:

Fig4 Coupler size chart

In the coupler, the initial magnetic permeability is $\mu_i = 200$, the specific resistance is $\rho = 140 \Omega \text{m}$, and the Curie temperature is $T_c > 130 \degree \text{C}$. Because the demodulation chip used in the previous stage is unstable, the carrier frequency should be selected at the low and medium frequencies, and the frequency of the signal at 270kHz Skin depth is:

$$\Delta = \sqrt{\frac{2}{\mu \rho}} \quad (1)$$

The skin depth can be calculated, and it is determined that the skin diameter is doubled, and the number of coil turns of the coupler is 50 turns. The circuit of the inductive coupler is very similar to the high-frequency transformer. Based on the high-frequency transformer, the
The circuit of the inductive coupler is designed:

![Fig5 Circuit design of inductive coupler](image)

Because the inductive coupler is symmetrical in the primary and secondary, all parameters in the circuit model can be considered to be symmetrical. The amplifier $D_1$ in the circuit represents the DC loss of the inductive coupler coil. Because it has leakage inductance, the inductive coupler and the ideal transformer model differ in some characteristics \[7,8\]. Use an RLC automatic measuring instrument to measure one side of the inductive coupler at different frequencies $f$. It is equivalent to the inductance and the coil impedance in series.

The inductance and impedance values are shown in the following table:

<table>
<thead>
<tr>
<th>$f$ (kHz)</th>
<th>$L$ (mH)</th>
<th>$R$ (Ω)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>1.042</td>
<td>7.56</td>
</tr>
<tr>
<td>100</td>
<td>1.073</td>
<td>11.8</td>
</tr>
<tr>
<td>150</td>
<td>1.121</td>
<td>17.98</td>
</tr>
<tr>
<td>200</td>
<td>1.172</td>
<td>25.33</td>
</tr>
<tr>
<td>250</td>
<td>1.26</td>
<td>35.64</td>
</tr>
<tr>
<td>300</td>
<td>1.34</td>
<td>54.14</td>
</tr>
<tr>
<td>350</td>
<td>1.49</td>
<td>67.52</td>
</tr>
<tr>
<td>400</td>
<td>1.55</td>
<td>71.32</td>
</tr>
<tr>
<td>450</td>
<td>1.68</td>
<td>80.36</td>
</tr>
</tbody>
</table>

Through the RLC automatic measuring instrument at the frequency of 450kHz, the
self-inductance deviation of the coupler was measured within 1.04-1.6mH. Considering the frequency range of the data transmission channel of the experimental system, \(L\) was selected as 1.3mH, and \(R\) was selected as 50\(\Omega\). This completes the design of the channel parameters.

2.3 Optimize multi-path allocation mechanism

In the practical application of data transmission, the energy consumption between different links will be different due to a series of effects. Therefore, it is necessary to optimize the multi-path allocation mechanism and add transmission consumption weights for each link, so that the network can better integrate with the objective environment in which it is located, thereby better improving the reliability of data transmission. Due to the influence of the topography, such as mountains, buildings, obstacles, and natural environments such as storms and lightning, the performance of wireless communications may change frequently, and communication interruptions frequently occur. These factors determine the value of the transmission consumption weight. Ability to calculate the transmission consumption weight matrix:

\[
W = (w_{ij})_{n \times n} \quad (2)
\]

among them:

\[
w_{ij} = \begin{cases} w_{ij}^* & e_{ij} \in E \\ 0 & e_{ij} \notin E \end{cases} \quad (3)
\]

In the above formula, \(e_{ij}\) is the link between nodes \(V_i\) and \(V_j\), \(w_{ij}^*\) is the transmission consumption weight of the link between nodes \(V_i\) and \(V_j\), and \(0 \leq w_{ij}^* \leq 1\). \(E\) is the link set, \(n\) is the number of middle nodes, and \(2 < n < 4\). It is known that the energy consumption \(P\) of the data transmission and the communication distance \(d\) exist:

\[
P = kd_{ij}^n \quad (4)
\]

In the above formula, \(k\) is the energy parameter and \(d_{ij}\) is the distance between nodes \(V_i\) and \(V_j\). It can be seen that as the communication distance increases, the energy
consumption will increase sharply, and the single-hop communication distance should be minimized on the premise of satisfying the communication connectivity. Therefore, a multi-path allocation process can be designed:

![Multi-path allocation process diagram](image)

The base station extracts the node information traversed by the data packet and the corresponding signal strength parameters, and converts all the path request data to the base station into a graph. The algorithm designed in this paper is used to generate the path from any node extracted from the node set to the base station. In order to obtain the multi-path routing from a node to the base station to the greatest extent, the idea of breadth-first algorithm is used in the path generation process\(^{[9,10]}\). This completes the optimization of the multi-path allocation mechanism.

3 Experiment

In order to verify the effectiveness of the Web4.0-based new media marketing analysis data transmission method designed in this paper, it is necessary to perform a power consumption test to verify its energy saving effect.

3.1 Experiment preparation

In different branch lines of the same data transmission line, the transmission method designed in this paper and the existing transmission method are used to randomly select a
continuous period (24 hours) as the test stage, and the energy consumption monitoring equipment gem-view is used to monitor the energy consumption of the two systems respectively. Firstly, the energy consumption of the system designed in this paper is monitored, and then the energy consumption of the traditional system is monitored. The equipment has the functions of basic energy consumption measurement and billing and report analysis, and provides online energy consumption detection (up to second level), energy consumption diagnosis and analysis, etc. Its parameters are as follows:

<table>
<thead>
<tr>
<th>Features</th>
<th>project</th>
<th>parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>measuring</td>
<td>Current</td>
<td>Range 0-20A</td>
</tr>
<tr>
<td></td>
<td>power</td>
<td>Range 0-10kw</td>
</tr>
<tr>
<td>Calculation</td>
<td>Electric energy</td>
<td>Four-quadrant energy metering, multi-rate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>energy accumulation, maximum demand</td>
</tr>
<tr>
<td>monitor</td>
<td>Voltage crest factor, current K factor, voltage and current, Imbalance</td>
<td>-</td>
</tr>
<tr>
<td>Measure</td>
<td>Itemized metering circuit</td>
<td>Separately measured external power supply circuits, special area power supply circuits</td>
</tr>
</tbody>
</table>

In the above experimental environment, the real-time energy consumption results of the two data transmission methods are statistically analyzed.

3.2 Experimental results and analysis

Under the arranged experimental conditions, the continuous period (24h) was randomly selected for testing, and the energy consumption of the existing data transmission method and the design method of this paper was recorded. The experimental results are shown in the following figure:
In the selected continuous period, observe and record the energy consumption every 2h, clear it after recording, and start recording the energy consumption within the next 2h. During the period of 0:00-8:00, during the Internet gap period, the transmission link is basically in a dormant state. Between 8:00-20:00, the transmission link begins to continuously transmit marketing analysis data and is at work. In this mode, energy consumption is gradually increasing. According to the comparison chart of the experimental results, it can be clearly seen that the current data transmission method consumes more energy than the method in this paper, whether in the sleep state or the working mode. The energy consumption within 24h is calculated in detail. The total energy consumption of the existing transmission method is 869 kWh. The total energy consumption of the transmission method in this paper is 693 kWh. In 24 hours, the transmission method designed in this paper saves 176 kWh than the existing data transmission methods.

In order to further verify the effectiveness of this system, the traditional system and the new media marketing analysis data transmission accuracy of this system are compared and analyzed, and the comparison results are shown in Figure 8.

![Figure 8](image)

**Figure 8** Comparison of transmission accuracy between two systems

According to figure 8, the new media marketing analysis data transmission accuracy of this system is higher than that of the traditional system.

### 4 Concluding remarks

Web4.0 can integrate all digital and information resources that can be integrated to provide users with the best information services under the premise of identification. Data transmission is inseparable during this period. In the context of the whole society advocating energy saving and emission reduction, and building a green business environment, the existing data transmission methods consume too much energy and no longer meet the requirements of a green society. Therefore, a new media marketing analysis based on Web4.0 is designed...
transmission method. By designing the structural block diagram of the transmission method, the Steiner tree model for data transmission is established, the transmission channel parameters are designed, and the optimization of the multipath allocation mechanism is completed. The experimental results show that the transmission method designed in 24 hours saves 176 kWh than the existing data transmission methods. This method does not consider the problem of transmission time in the process of verifying the accuracy of data transmission in new media marketing analysis, which results in the increase of data volume and the slower transmission time. Therefore, in the next study, we will further study the data transmission in new media marketing analysis and improve the transmission efficiency.

References


Designing and Testing of a Semi-Physical QUIC Video Streaming Simulation Platform

Jun Zhao¹, Siqie Zhang², Ben Liu³, Hongjie Yu⁴, Mingwu Yao⁵, and Fuzheng Yang⁶

{zj410728@163.com, sqzhang1994@163.com, liuben@stu.xidian.edu.cn³}

State Key Laboratory on Integrated Services Networks, Xidian University, Xi’an, China

Abstract. QUIC (Quick UDP Internet Connections) is a streaming control protocol that was proposed by Google in recent years. In order to study the ability of QUIC to support streaming video in a wireless mobile environment, we combine a DASH (Dynamic Adaptive Streaming HTTP) server and QUIC source code to design and implement a virtual machine-based testing and simulation system. The system can not only use real video traffic for streaming video playback, but also use the simulation system such as NS3 to introduce wireless network features. Based on the system, we perform a simulation analysis on the proposed QUIC stream video improvement strategy for large-delay ACK discarding (LDAD). The results demonstrate that the simulation platform can help to understand and investigate the interaction between streaming video and QUIC in-depth.

Keywords: QUIC; DASH; LDAD; NS3; Simulation Platform

1 Introduction

Video exchanging and browsing via social networking website and mobile apps is becoming a real killer application in recent years [1]. Aiming at improving the QoE (Quality of Experience) of users, the next-generation protocols such as HTTP/2 [2], QUIC [3] and HTTP/3 [4] have been proposed. A conventional way to support such applications is a combination of HTTP/1.1 and TCP (HTTP over Transportation Control Protocol) [5]. However, they often incur huge connection establishment delay with complex multi-connection management overhead, and suffer severe Head-of-Line Blocking (HOLB) [6] issues. HTTP/2 adopts a series of new features to solve the application-layer HOLB [2][5] but triggers blockage in the transport layer[7]. The latency of TCP Connection Establishment is another problem. Some enhancements such as TCP fast open transport-layer [8] have been proposed long before but are rarely used in practice. Moreover, TCP’s close coupling with the kernel of popular Operating Systems (OSs) hampers its updating and optimization speed in a view of ordinary users.

In 2012, Google presented QUIC [3], a new type of multiplex and secure transmission protocol based on UDP. QUIC have serval new significant features including multiplexing and flow control that is equivalent to HTTP/2, encryption equivalent to TLS and connection semantics, reliability and congestion control equivalent to TCP. Fig. 1 presents the functionality of the QUIC in the OSI reference model. At the same time, the HTTP-over-QUIC experimental protocol has been renamed to HTTP/3 and is expected to be the third official HTTP release.

QUIC is purposely designed to overcome the shortcomings of TCP. QUIC’s stream multiplexing greatly alleviates the transport layer HOLB. There is usually 0-RTT (Round Trip Time) in the connection establishment of QUIC. Furthermore, a QUIC connection is identified...
by a 64-bit Connection ID instead of the quad of IP address and port number, which is significantly suitable for NAT rebinding and network switching. QUIC adopts a strict authentication and encryption mechanism. QUIC does flow control more efficient than TCP. QUIC’s congestion control algorithm is pluggable and implemented in the user space of the OS. QUIC replaces TCP’s Sequence Number with a strictly monotonically increasing Packet Number, avoiding the retransmission ambiguity of the TCP. Every QUIC packet contains an unencrypted public header and an encrypted payload. And the latter part contains one or more data or control frames, carrying application data and control messages respectively [9]. For the latest features, please refer to [10]. Finally, QUIC ACKs explicitly carry the delay incurred at the client receiver, explicitly inform the sender the latency.

Fig. 1. The architecture of QUIC.

QUIC has arisen widespread research interests. Megyesi et al. compare the QUIC with HTTP and SPDY[11], and proposed a method to know when it is appropriate to use QUIC. Somak gives a detailed analysis of QUIC access to Web pages [12]. Kakhki Arash Molavi et al. test QUIC in a large number of environments, showing that, though QUIC is generally better than TCP, its performance is significantly reduced under mobile and cellular networks [13]. Jan Rüth et al. show that QUIC usage is increasing as a protocol deployed in user space [14]. Szabo Geza et al. report the QoE of QUIC for media application experience [15]. Clark proposes a non-multiplexed relay transmission protocol QUUX based on QUIC [16].

To study the performance of QUIC more effectively, we build an emulation/simulation platform combing DASH, QUIC and NS3 (Network Simulator). DASH technology is a hybrid media distribution method, which uses HTTP protocol to download and distribute content, likes the HTTP progressive download method. NS3 provides a wireless network simulation scenario. QUIC is used to transmit video service streams, and DASH is used to request video service streams and play received video segments. The DASH, QUIC and NS3 are independently developed and sophisticatedly connected. To the best of our knowledge, there is no published literature depicts such a design so far. In the system, DASH plays the role of the application layer entity and QUIC takes the responsibility of the function of the transport layer. NS3 works as a network. We use the subclass rewriting and dynamic library technology to connect DASH
and QUIC, and use the datagram Socket technology and multithreading multitasking parallel mechanisms to attach QUIC and NS3.

Using the platform, we verify our improvement towards QUIC. Different to known methods, we implemented the improvement at AP (Access Point)/Base Station (BS) of the QUIC based on the feature of the wireless network. The performance is verified on the designed system, especially detailing the performance of the algorithm in the scenario of LTE and WLAN.

The rest of this paper is organized as follows. We detail the simulation platform of QUIC and do some test work on it in Section 2. In section 3 we introduce the mechanism of LDAD. Performance evaluation is presented in Section 4. Section 5 concludes this paper.

2 Simulation platform

As shown in Fig. 2, the semi-physical QUIC platform consists of three major modules: DASH, QUIC and NS3. The DASH system is a suite of video streaming tools. The server is used to generate the video stream and carry out the video code. The DASH client receives, decodes, and plays the streaming video. The client code include also the functions of QoE evaluation to track and evaluate the playback effect. The video stream data output by the DASH server are delivered by QUIC. After being multiplexed and sorted by the QUIC server, the video stream data are assembled into UDP frames, and then enter into the NS3 simulation network. The simulated network will incur delay, jitter, and loss to the UDP frames, to simulating the affections of a real network on the QUIC transportation. Afterwards the data enter into the QUIC client, where they are error corrected, resorted, confirmed, and/or required to be retransmitted. The resulting stream is passed to the DASH client where the video is decoded and played.
2.1 System building

The building blocks of the platform are DASH, QUIC, and NS3, arranged in the order from top to bottom. The DASH and QUIC are implemented in the same OS of one virtual machine, and the NS3 is running at another virtual machine. One of the key tasks of platform construction is to connect the above three blocks so that the three platforms operate in synchronization and cooperation.

![Diagram showing the connection between DASH, QUIC, and NS3](image)

**Fig. 3. DASH-QUIC connection scheme.**

*The Connection between DASH platform and QUIC platform:* In this scheme, subclass rewriting technology and dynamic library technology are used. The DASH-QUIC connection scheme is based on adaptive streaming media and changes the selected media block from TCP transmission to QUIC transmission. The DASH-QUIC as shown in Fig. 3. A QUIC client is created in the DASH client, and the DASH client interacts with the DASH server through the QUIC server-client interactions that providing a transportation link.

*The Connection between QUIC platform and NS3 platform:* This scheme is implemented based on the datagram socket communication technology. The specific connection scheme implementation diagram is shown in Fig. 4.

As NS3 has difficulties to deliver the real stream data packets due to its capacity, the data packets are treated in a modified way. When the QUIC client sends a packet, it first transmits the packet size information to the NS3 client through the socket process communication mechanism (In theory, QUIC needs to transmit packet size and packet number informations to NS3, but since the packet number in QUIC increases successively, and the packet information transmitted by QUIC to NS3 will not be lost, so we can set a self-increment variable as the mapping of the QUIC package number. Therefore, QUIC only needs to transmit packet size information to NS3); then the actual data packet is transmitted to the receiving buffer of the QUIC server via local transmission. The NS3 client first puts the packet size information into the cache, and the simulate gateway node will generate a dummy data packet with the same size and header and sends it through the simulated network.

When the dummy packet is received by the NS3 receiving node, the NS3 server’s HandleRead function sends the correctly received packet number and corresponding
transmission delay to the QUIC server via the socket. The lost packet in NS3 will incur a real packet loss in the QUIC. The QUIC server puts the received NS3 message into the NS3 message receiving buffer. Each time the QUIC server receives a local transmission packet from the QUIC client, if the packet is not lost, after a certain delay according to the delay information, the packet is forwarded upward. When the packet processing is completed, the NS3 information buffer will be traversed again until NS3 message cache is empty. The QUIC server enters a state waiting to receive the actual data packet transmitted locally by the QUIC client.

The packet sending process of the QUIC server is basically the same as that of the QUIC client, except that the corresponding processing interfaces are different.

If NS3 retransmission fails, the corresponding NS3 client or NS3 server cannot successfully receive packets from the simulated network. At this time, if NS3 only sends information about successfully received packets to QUIC, it is obvious that the actual packet receiving buffer of QUIC may be filled with the failed transmission in the simulated network which will make it impossible to receive the actual packets transmitted from the QUIC opposite end. Therefore the information of the underlying retransmission failure packet needs to be submitted upwards so that QUIC can delete the actual data packet corresponding to the network transmission failure.

The current solution is: during the underlying retransmission process of the NS3 server and client, if the number of retransmissions exceeds the maximum retransmission threshold set by the source code, socket communication is triggered to send corresponding packet information.
to the corresponding QUIC, and QUIC clears the corresponding receiving buffer according to the information.

2.2 System Verification

We do some verification tests after the system is built. In the first verification, the connectivity of the simulation platform is tested without packet loss. In this case, we have tested that QUIC, NS3, and DASH are connected as we expected (after a lot of debugging and rewriting). The smoothly displayed video showing our treatment of the simulated packet streams is correct.

The second verification is showing the effects of losses introduced by the NS3 network on the video playback qualities. As shown in Table 1, the playback of DASH was tested with the packet loss rate of 0.1, 0.2, 0.3, 0.4, and 0.5 respectively, with different QoE performances.

Table 1. Packet loss rate and video playback test.

<table>
<thead>
<tr>
<th>Packet loss rate</th>
<th>DASH playback</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>Can be played, no stutter, no errors</td>
</tr>
<tr>
<td>0.2</td>
<td>Can be played, some freezes, no errors</td>
</tr>
<tr>
<td>0.3</td>
<td>Can be played, obvious freeze, no errors</td>
</tr>
<tr>
<td>0.4</td>
<td>Can be played, obvious freeze, no errors</td>
</tr>
<tr>
<td>0.5</td>
<td>Can be played, obvious freeze, poor playback effect</td>
</tr>
</tbody>
</table>

3 Using the platform to test LDAD

3.1 Why LDAD

Chrome-based experiments show that packet loss can be reduced by the Pacing mechanism. Pacing mechanism reduces packet flow fluctuations, thereby reducing congestion-based losses (i.e. packet drops in routers due to overflow). The current QUIC is based on bandwidth estimation to dynamically adjust the pacing rate. The pacing rate in QUIC is calculated as:

\[
r = \begin{cases} 
2 \times \frac{C_w}{SRTT} & \text{slow start phase} \\
1.25 \times \frac{C_w}{SRTT} & \text{not slow start phase} 
\end{cases}
\]

(1)

Here \( C_w \) refers to the current congestion window, and SRTT is the Smooth Round Trip Time.

We want to improve the performance of QUIC in a wireless network where the user terminals are poorly covered by the BS (Base Station) or AP (Access Point). The essence of increasing the QUIC transmission rate is to increase the Pacing rate of QUIC. According to the source code of the QUIC protocol, the QUIC packet rate calculated is based on the current updated congestion window and the current updated SRTT. Each time the QUIC receives an
ACK, the SRTT will be updated by the current RTT. When the congestion or sudden interference occurs in the network, it will cause the RTT to become larger, thereby reducing the rate of Pacing.

In LDAD, parts of large delay ACKs that may reduce the transmission rate of QUIC are blocked on a certain rule to prevent it from continuing to submit. Conventionally, the sending rate will drop adaptively when the network suffers from the congestion or a sudden interference. The drop in rate can prevent the network from further congestion, which is positive in the congestion case. But for the sudden interference, which is almost instantaneous, so the drop is not necessary. And the sudden interference is frequent in the wireless network, especially when its wireless coverage is poor. On the one hand, our algorithm can improve congestion sensitivity, that is to say, the congestion can be detected more easily. On the other hand, by discarding ACKs with large delay, the unnecessary drop of the rate caused by burst interference can be avoided.

3.2 Overview of LDAD

The algorithm can be deployed at BS/AP, where it just needs a ACK blocking model as detailed in Fig. 5. Moreover the algorithm can be easily transplanted to the QUIC server.

![Algorithm flow](image)

**Fig. 5. Algorithm flow.**

As shown in Fig. 5, the packets will be judged by several modules: ACK-only Packet Determination Module, Downlink Packet and Uplink ACK Mapping Module, ACK Interception Module, Sender Congestion Judgment Timer Module, Sender Timeout Determination Timer Module and State Estimation Module. The purpose of this algorithm is to generate a mapping of downlink data packets and uplink ACKs, and then calculate the current instantaneous RTT and intercept the ACK which with abnormally large delay RTT, so as to prevent the QUIC
transmission rate from deteriorating when there is a sudden interference in the current wireless network. By the way, there is a different processing flow for uplink and downlink packets. For the downlink packet received by the current BS, it will be judged whether is a QUIC packet according to the port number. Then it enters the Downlink Packet and Uplink ACK Mapping Module if it is a QUIC packet, where it is joined with the current packet to the mapping table according to the judgment mechanism of the module. Then it enters the State Estimation Module to update the parameters, and finally goes into the timer module to perform necessary timer updates after updating the parameters.

For an uplink packet, the process is approximately the same. The BS determines the type of the received uplink packet. If it is a QUIC packet, the ACK-only Packet Determination Module is activated, and the ACK-only packet estimation algorithm in the module determines whether the current packet is an ACK-only packet. If it is a QUIC ACK-only packet, then it enters the ACK Interception Module, where the instantaneous RTT of the current ACK is calculated according to the mapping table. The mapping result will determine forwarding the ACK-only packet directly or not. The subsequent process is the same as the downlink process. That is, the packet enters into the State Estimation Module and timer module orderly. The function of Abnormal ACK-only Cache in the module is to store the ACK-only packet currently intercepted. The ACK-only packet reckoned as with large delay is purposely cached instead of being directly discarded, in case of that the QUIC server may not receive any ACK for a long time, which will be more harmful than receiving a delayed one. If that happens, the congestion window will be completely occupied and the data exchange process between the QUIC server and the QUIC will be totally stucked, resulting in a very poor user experience.

4 Experimental results

In order to verify the effectiveness of the LDAD, tests are performed using the designed QUIC semi-physical simulation platform with the LTE and WLAN scenario respectively. To be more specific, we implement the algorithm at eNB in the LTE scenario and at the AP in the WLAN scenario respectively, and then test the QUIC average throughput with various BLER/PER, with and without using the LDAD. Here the QUIC server throughput is defined as the total amount of data transmitted by the QUIC server per second. It is worth mentioning that the average throughput is obtained after abandoning the outliers due to the systematic error. The simulation structure refers to the LTE design document of the NS3, as shown in Fig. 6.

1) The performance in LTE: We first test the idea in the scenario of LTE, as is shown in Fig. 6. The LTE scenario includes: a remote host, an eNB, and an UE, whose nodes are all stationary. The remote host communicates with the UE through the eNB. The remote host in the simulation scenario is connected to the QUIC server of the QUIC platform, and the UE is connected to the QUIC client of the QUIC platform. We use the virtual box to operate two virtual machines running QUIC and NS3 respectively. The QUIC client requests 256 kB data from the QUIC server. The uplink and downlink transmission delays of the LTE fixed network are all set to 50 ms. Testing is performed in the case of 5%, 10%, 15%, 20%, 25%, and 30% BLER of the LTE wireless network. The change average throughput rate of the QUIC server with the various packet lost in physical layer of the LTE wireless network is shown in Fig. 7.

The blue line (solid or dotted) in the Fig. 7 are the changes of the average throughput of the QUIC server with the BLER when the algorithm is used or not respectively. From the figure, it can be seen that the throughput decreases with the increase of the BLER in both cases, the reason
is that as the probability of packet loss at the physical layer increases, the number of retransmissions of the underlying data packet increases, and the end-to-end instant delay jitter increases and the time delay fluctuates more frequently. As can be seen from the figure, compared with the original case where the algorithm is not used, almost all the algorithms used in the current test scenario can bring gains, especially when the BLER is 0.05, 0.1, and 0.15, and the gain is about 30%. Obviously, the gains brought by using the scheme in the LTE scenario are significant.

2) The performance in WLAN: As shown in Fig. 6, the WLAN consists of a remote host, a sta (station) and an AP, all stationary. Remote host interacts with sta through the AP. The remote host in the simulation scenario is connected to the QUIC server in the platform, and the sta connects to the QUIC client of the QUIC platform. To simulate the actual situation, the remote host is connected to the AP through a point-to-point channel in the current scenario. The uplink and downlink transmission delays of the point-to-point channel are set to 60 ms and the transmission rate is 5 Mbps. The WLAN network uses the 802.11b protocol, and the WLAN transmission bandwidth is a constant 1 Mbps. The QUIC client requests 256 kB data from the QUIC server. The packet loss rate at the bottom layer (physical layer) of the WLAN network is 0%, 5%, 10%, 15%, 20%, 25%, and 30%, respectively. Under the condition that the tested 802.11b protocol has a constant 1 Mbps transmission rate, the average throughput of the QUIC server changes with the various physical layer packet loss rate as follows:

As shown in Fig. 7, the red line (solid or dotted) indicates the variation of the QUIC server’s throughput with the physical layer packet loss rate when the algorithm is not used or not. Similar to the test in the LTE scenario, the QUIC server throughput under two cases both decreases with the increase of the packet loss rate at the physical layer. The transmission performance of the performance is basically the same when the packet loss rate at the physical layer is small. However, as the packet loss rate increases, the performance gain after the use of the LDAD becomes obvious, especially in the case of a packet loss rate of 0.15, the usage scheme brings about a 10% performance gain.

In summary, the proposed algorithm can bring about certain gains in the LTE scenario or WLAN scenario and the gain of the algorithm in WLAN is about 10%, and the gain is 20%~35% in LTE.
3) The performance with estimation error: There is a high probability of an uplink ACK-only packet loss in the actual poor wireless environment. At this time, the estimation error would occur in the algorithm. Therefore, the impact of the estimation error on the performance of the algorithm is tested. In the tests, the QUIC server throughput performance with some estimation error in the scenario of LTE is shown.

![Fig. 7. The algorithm performance in WLAN and LTE.](image)

The simulation settings are basically the same as before. In NS3, we manually set an estimation error of 20% when the node receives the upstream ACK and calculates the current RTT. Here, the estimation error is defined as 20% delay disturbance. For example, if the current calculated instantaneous RTT is 100 ms, there is a 20% probability that the instantaneous RTT submitted to the QUIC is 120 ms.

From the Fig. 7, the black solid line indicates the results of using algorithm with 20% estimation error in LTE. It can be seen that using the algorithm can bring a large gain no matter there is an estimation error or not. The RTT of the wireless segment can obtain real-time updated data at the eNB, and therefore can accurately estimate that there is a certain estimation bias in the fixed RTT currently set. Based on the performance analysis above, it can be seen that deviations in fixed RTT will only affect the threshold setting of the two timer modules. This will cause the duration of the algorithm to be several milliseconds longer than that of the ideal case. Compared with the second-level test time, it is apparent that the wired-link RTT exists. The deviation has little effect on the performance of the algorithm. The simulation test results are in good agreement with the performance analysis. Therefore, it can be concluded that under the current settings of the simulation scenario parameters, estimation error of the wired-link RTT has little effect on the performance of the algorithm.

5 Conclusion

We design and implement a QUIC simulation system that combines DASH, QUIC, and NS3 platforms and test its effectivity. DASH is used to send service requests and play successfully received video segments. QUIC plays the role of transmitting DASH-based video
service streams; and NS3 is used to provide emulated networks including WLAN and LTE. After the simulation platform passing some initial verifications, we use the platform to implement the proposed LDAD algorithm on the BS of the LTE and the AP of the WLAN, and test its performance. According to the result, the proposed algorithm can bring significant gains in poor wireless environment. Among them, the performance gain of the algorithm in the WLAN scenario is about 10%, and 20% to 35% in the LTE scenario. The results show our semi-physical QUIC simulation platform is powerful in investigating such problems. The platform can be improved to have more features to modify and observe the interactions between QUIC and user applications in the future.

Acknowledgment. The research on the protocols and QoS of networks is supported by the National Science Foundation of China (No.61671353), and Huawei Technology Co., Ltd.
References

Signal processing method of campus mutual aid platform based on mobile terminal

HUANG Yi1, GAO Jie2
{jahg245@163.com1, ojsfdiosd984@163.com2}
(1. NanChang Business College of jxau, Gongqingcheng 332020, China;
2. Modern Economics & Management College of JXUFE, Gongqingcheng 332020, China)

Abstract: Intelligent terminal equipment is closely connected with wireless LAN, which makes the study and life of students and teaching staff more convenient, but its signal processing has a certain degree of lag. Based on this, the signal processing method of campus mutual aid platform based on mobile terminal is proposed. The channel is encoded and decoded so that the center frequency and noise of the signal can be corrected. After the timing and code optimization of the signal processing, the centralized processing of the signal can be realized. Experiments show that the method designed in this paper can guarantee high processing results and better stability under different degrees of interference.

Key word: Mobile terminal; Campus mutual aid platform; Signal processing method; Campus website; Educational administration management system

1 Introduction

With the rapid development of information technology, information technology has been integrated into all aspects of education. Educational informationization has become the main development direction of educational reform[1]. With the rapid development of information technology, "Internet of things", "cloud computing concept", "mobile smart campus" and other new things are gradually known by people, and affect the higher education environment and concept mode. In this profound change, the university digital campus has entered a new stage, towards the development of intelligent campus. On this basis, colleges and universities continue to research and develop digital campus service platform, such as the establishment of campus network, campus mutual platform, research office system and information management system. These systems and platforms can not only meet the daily office and teaching needs, but also facilitate the interaction between teachers and students. But there are also some problems: for example, in the early stage of construction, due to the lack of overall planning and guidance of the campus service platform, the poor connection between the school portal website, the provision of information services and OA, the utilization rate of information resources is low, and it is difficult for users to obtain the information and functions they want through a main page, not to mention the important information services and analysis and judgment Service.
As one of the products of the development of information technology, intelligent terminal equipment is more popular and comprehensive. It is closely connected with wireless LAN, which makes the study and life of students and teaching staff more convenient. Intelligent terminal is gradually replacing desktops and laptops, and plays an important role in people's office, leisure and communication[2]. As the main user group of smart phones, college students have a certain knowledge base and are easy to accept new things. Because smart phones have become one of the main tools for communication between teachers and students, teachers and students can make full use of the fragmented time to acquire the knowledge they want at any time and anywhere, so they are not bound to sit in front of the classroom. At present, there are many colleges and universities in most parts of the country, each of which has its own campus portal. The functions of campus websites in Colleges and universities are relatively simple, mainly including the publicity of school running characteristics, professional introduction, teacher introduction, etc. The function of campus website in some colleges is more comprehensive, including data collection and educational administration management system. However, there is a lack of interaction between teachers and students, which can not meet the multi-faceted application needs of college students. Higher vocational college students need more in learning, entertainment, social and other aspects. Combined with the commonly used smart phones, a campus mutual aid platform suitable for the use of mobile terminals of higher vocational college students is designed, so that students can solve daily learning and life problems only by using mobile phones on campus. At the same time, the platform can also realize the interaction between students and teachers in their spare time. Teachers can manage students and release information. Students can also ask for leave from teachers in time, and can also query courses and scores independently, which makes teachers' teaching and students' knowledge learning more convenient and not limited by time and space.

2 Signal processing method of campus mutual aid platform based on mobile terminal

2.1 Channel coding and decoding

According to the traditional signal processing method, there are all kinds of interferences in the channel, which may lead to the distortion of the transmission information. In general, we use channel error correction coding to code baseband signal. In order to prevent signal distortion[3]. Since the American scholars first proposed the reliable coding method of transmitting information in noisy channels in the early 20th century, error correction codes have been gradually developed and improved.

Channel coding technology improves error correction performance to some extent by adding redundant information bits to the transmitted information sequence. Through channel
coding, when a certain instantaneous fading occurs in the channel, the transmitted data can still be recovered at the receiving end. At the beginning of baseband processing, channel coding transforms the original information sequence into a new code sequence containing more information bits by a specific algorithm. Then, the coded sequence is modulated so that it can be transmitted normally in the wireless channel.

In the digital communication system, the methods of information error control include feedback retransmission and forward error correction coding. In the feedback retransmission mode, the error code will be fed back, and then the transmitter will retransmit. Forward error correction code has no feedback mechanism, but encodes the error correcting information bits to the information sequence to be sent. The receiver receives the transmitted data, and the received redundant bits can find the error and achieve the purpose of error correction at the receiver. So this kind of forward error correction code reduces the efficiency of the channel by adding redundant bits, but it can make a lot of error codes be correctly recovered at the receiver. When the forward error correction code is not used, the transmitter that generates the error code must resend the information bits. If the efficiency loss caused by the increase of redundant bits is less than the loss caused by the retransmission of error codes, the forward error correction code actually improves the efficiency of the channel. One of the most important parameters of FEC code is the code rate, which describes the ratio of loaded input bits to output bits with a decimal. For example, for a code rate of 1/2, two information bits are output for each input of one information bit, and one bit of redundant bit information is added.

![Diagram](image)

Fig. 1 forward error correction coding communication

Like the previous 802.11a/g, in 802.11n, the forward error correction code continues to use convolutional coding. Convolutional code has high speed and simple hardware structure\(^4\). In this paper, the parameters of convolutional coding are (2,1,6). This parameter is based on an example of 802.11n. The selection of convolutional coding parameters in 802.11n is based on the experimental results considering the factors such as bit error rate and transmission rate. After coding, the transmitter needs to decode at the receiver. At present, viterbi decoding is most commonly used. Viterbi decoding is a decoding method based on “maximum similarity”. By reproducing the path of information bits on the grid during convolutional coding, the most likely path is found as the output of decoding. When the constraint length is not large, the
advantages of Viterbi decoding are particularly obvious. The hardware structure is relatively simple and the speed is relatively fast. Through the optimization of some key modules, a Viterbi decoder with high performance is designed to facilitate the real-time processing of signal data later.

2.2 Adjustment of center frequency

To filter signals with different frequencies, a filter with different center frequencies is needed, and the center frequency of the wavelet filter is related to the sampling frequency and the scale of the wavelet. At the same time, signal frequency and sampling frequency are closely related, that is, different sampling frequency is required for different frequency signals. If the sampling frequency is selected too high, there will be too many sampling points in a signal cycle, resulting in oversaturation. However, if the sampling frequency is too low, it may not satisfy the sampling theorem and can not accurately reflect the signal components, resulting in the decrease of measurement accuracy. On the premise of accurately reproducing the measured signal, we can adjust the center frequency of the filter through the following three paths to realize the filtering of different frequency signals.

First of all, changing the series of wavelet decomposition, that is, changing the scale to move the passband of the filter on the frequency axis, so that the signal components of different frequency bands pass through, which is very good for the signal with little frequency change.

Secondly, the data is extracted at intervals. To adapt to the change of signal frequency, the method of extracting data from each other is used. Sampling data every certain number of points, which is equivalent to reducing the sampling frequency. For the extracted data, wavelet transform of different series is still used to measure the value at lower frequency.

Finally, change the sampling frequency. In practical application, the data extraction method based on the number of sampling points is limited. The center frequency of wavelet filter is adjusted by directly changing the actual sampling frequency to meet the requirements of different frequency signals. As long as the sampling frequency is set correctly, the measurement of signal frequency is quite accurate. By changing the sampling frequency, very low frequency band (below 1Hz) and very high frequency band (above 2000Hz) can be measured. Through a large number of practices, it is found that the sampling frequency is set to 20 times of the signal frequency, and the accurate frequency measurement results can be obtained by calculating the period of the data after decomposing to level 4 wavelet without sampling points.

It should be noted that the effective adjustment of the center frequency is for the follow-up signal noise reduction processing, so that the campus mutual aid platform can play a
better role.

2.3 Signal noise correction

In the process of signal acquisition, transmission and storage, it is often interfered and influenced by various kinds of noise, which leads to signal degradation. Moreover, the quality of signal preprocessing algorithm is directly related to the effect of subsequent signal processing, such as signal segmentation, target recognition, edge extraction, etc. In order to obtain high-quality digital signal, it is necessary to reduce the noise of the signal, keep the integrity of the original information as much as possible (i.e. the main features), and remove the useless information in the signal. Therefore, noise reduction has always been the focus of signal processing and computer vision research.

The ultimate goal of signal video denoising is to improve the given signal and solve the problem of signal quality degradation caused by noise interference. The denoising technology can effectively improve the signal quality, increase the signal-to-noise ratio, and better reflect the information of the original signal. Among the existing de-noising algorithms, some of them have achieved good results in low-dimensional signal processing, but they are not suitable for high-dimensional signal processing; or they have good de-noising effect, but they lose part of the signal edge information. At the same time, some researches are devoted to studying the edge information of the detected signal and retaining the signal details [5]. How to find a better balance between resisting noise and keeping details has become the focus of this paper.

Wavelet shrinkage method is the most widely used method at present. There are two kinds of wavelet shrinkage method: the first one is threshold shrinkage. Because threshold shrinkage is mainly based on the following facts, that is, the larger wavelet coefficients are generally based on the actual signals, while the smaller ones are largely noise. Therefore, by setting a suitable threshold value, the coefficients smaller than the threshold value can be set to zero, while the wavelet coefficients larger than the threshold value can be retained. Then the estimated coefficients are obtained by the threshold function mapping, and finally the estimated coefficients are inversely transformed to achieve denoising and reconstruction [6]. The second type is proportional shrinkage, which is called proportional shrinkage by judging the degree of noise pollution of coefficients and introducing various measurement methods (such as probability and membership) to determine the proportion of shrinkage. The research of signal processing method based on PDE is also a hot research direction of signal denoising, and has made some achievements in theory and practical application. Its denoising process is to establish the initial condition that the noise signal is a certain nonlinear PDE, and then solve the PDE to get the solution at different times, that is, the filtering result. Perona and Malik
proposed a PDE based nonlinear diffusion filtering method (hereinafter referred to as P-M). The anisotropic denoising model determines the diffusion speed according to the gradient value of the signal, so that it can take into account the requirements of noise elimination and edge preservation. These methods, represented by P-M model, have been widely used in signal enhancement, signal segmentation and edge detection, and have achieved good results. P-M is a non-linear anisotropic method, which aims to overcome the shortcomings of fuzzy edge and edge position moving in linear filtering method. Basic idea: reduce the diffusion coefficient where the signal feature is strong, and enhance the diffusion coefficient where the signal feature is weak. Although the p-m equation has achieved some results in suppressing noise and retaining important features of signal, it is ill conditioned and unstable. CATT et al. Improved the equation by convoluting the Gaussian kernel with the signal and then using its gradient modulus to estimate the edge information of the signal. The optimized symmetric exponential filter is used to smooth the signal, and then the gradient modulus is used to estimate the edge information of the signal. The basic idea of these two estimation methods is to reduce the interference of noise, extract the edge feature information of signal more truly, so as to control the diffusion behavior of P-M equation better with the edge information. TV method is proposed by Rudin Osher and Fatemi. Based on the idea of variational method, it determines the energy function of signal, and achieves the purpose of smooth denoising by minimizing the energy function of signal. It is now a popular signal restoration method.

In this paper, the noise reduction algorithm of P-M equation and TV method is mainly considered. In the new algorithm, not only the important features of noise suppression and signal retention are studied, but also the minimization of signal energy function to achieve smooth noise reduction is studied. According to the characteristics of mobile intelligent terminal camera, such as uneven, small size and so on, the denoising algorithm is studied. For many low and medium pixel cameras, after denoising, the camera yield can be improved, which conforms to the goal and direction of high performance and low cost.

2.4 Signal processing timing and code optimization

The processing timing of each module is arranged in our method. The signal is processed in two channels (high frequency band channel and low frequency band channel). Each frame of the low frequency band channel is equivalent to the K frame of the high frequency band signal in the time span [7]. When the high-frequency band signal has enough K frames, one frame of low-frequency band signal processing can be carried out. If all the calculations are completed in the next high-frequency band signal frame, the calculation is too centralized. In order to enable the algorithm to handle higher speed control, it is necessary to arrange the processing time sequence of each module of the algorithm appropriately, so as to spread the
calculation as far as possible, so as not to occur the situation that DSP is too late to handle. The FFT of the low-frequency band signal can only be processed when all the points of the frame are complete, so it can not be distributed to the intra processing of the high-frequency band signal, so the corresponding transfer function estimation, IFFT and other processing related to the output of the next frame before interpolation need to be completed in a high-frequency band signal frame. However, the decimation filtering with a large amount of computation can be distributed in each high-frequency band signal frame. When a high-frequency band signal frame data is ready, the corresponding decimation filtered signal can be calculated. The interpolation filtering of the output signal can also be distributed to each high-frequency band signal frame to complete. Each time the output stack is needed, the interpolation filtering is carried out to calculate the corresponding The output of the low-frequency band and the output of the high-frequency band are superposed. Therefore, the frame with the largest amount of calculation occurs when the data of the low-frequency band signal frame is ready. In this high-frequency band signal frame, it is necessary to extract the corresponding signal, estimate the spectrum of the low-frequency band input and output signal, estimate the transfer function and solve the output signal with positive ft. 1 service frame low-frequency band signal output interpolation filtering. Every k high-frequency band signal frames, there will be a large amount of computation. Under this arrangement and after the code manual optimization mentioned immediately, the control algorithm can complete the control of 8-channel input and 1-channel output with a control frequency of up to 9.skhz. The original single resolution control algorithm can only realize the optimization of 9.skhz control algorithm in the system of 8-Input and 1-output, which in fact includes the module calculation timing problem mentioned above. In addition, for each module, the optimization process mainly includes two steps: first, the optimization of C program before assembly; second, the optimization of assembly code. This is a good idea of algorithm optimization. High level language is relatively easy to write, which can make some optimization work that computers can do, and then optimize the code that some computers can't achieve efficient optimization manually.

![Fig. 2 Calculation timing of each module](image_url)

In the whole implementation process of the method, the modules that account for a large
amount of calculation include: decimation filtering module, FFT, IFFT and interpolation filtering module. The cores of FFT and IFFT are provided by TI company and have been optimized manually. Therefore, only decimation filtering module and interpolation filtering module are needed to be optimized. Considering that the decimation filtering in the control algorithm is carried out by frame and does not require continuous output, the direct filtering method can be used. After filtering, one k-1 signal in each k-point will be discarded and only one point will be reserved. Therefore, in the filtering calculation, we only need to calculate the point that needs to be reserved. In this way, the 160 order decimation filter is actually equivalent to a 160 order ordinary filter, so the calculation is not too large. In the filtering of each frame data, there are two sources of original data: current frame data and part of historical data.

The filtering process is composed of two layers of loops, the outer loop is the number of output points, and the inner loop completes the product of the filter sequence and the corresponding data, and sums it. The number of instructions in the inner loop determines the amount of calculation. If we need to judge whether the original data of convolution operation comes from the current frame or the historical data in the inner loop, then we know that at least seven or eight instructions are needed to make such a judgment in assembly code, and the calculation amount will increase greatly. Therefore, before filtering, it is necessary to combine the current frame data and historical data into a group. The DSP’s on-chip RAM supports dual access, that is, each instruction cycle can be accessed twice, while the filtering operation accesses the data very frequently, so it is better to copy the original data and filter coefficients into the on-chip RAM before filtering. If the on-chip RAM is not large enough, then at least copy the filter coefficients into the on-chip, which is good for increasing the filtering speed. The next step is to use the assembler and optimizer provided by TI company for preliminary optimization to get the assembly code. On this basis, artificial optimization is carried out. The focus of artificial optimization is to reduce the number of instructions in the inner loop. The result of optimizer optimization is that the inner loop contains three instructions. In fact, C3x and c4x series DSP contain a multiplier and an adder, which can handle one multiplication and one addition at the same time. Therefore, the optimal result is that the inner loop should contain only one instruction, so the effect of artificial optimization is very obvious. The optimization process of interpolating filter is basically the same, and it also goes through two processes: C language optimization and assembly code optimization. Table 1 and table 2 show the changes of calculation amount in the optimization process of decimation filter and interpolation filter respectively. Note that the data given here is the calculation amount used in the whole algorithm, including FFT and IFFT, because it is difficult to directly measure the
calculation amount of a single module through experiments. In the experiment, the control mode is 8 inputs (1 control channel and 7 monitoring channels), 1 output, the sampling frequency is 12khz, the maximum control frequency is 4680hz, and the data per frame is 1024 points.

**Table 1** change of calculation amount of control algorithm in the process of decimation filter optimization

<table>
<thead>
<tr>
<th>Optimizing content</th>
<th>Calculations for the whole control algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Optimization</td>
<td>0.9</td>
</tr>
<tr>
<td>Merge historical data with current frame data</td>
<td>0.83</td>
</tr>
<tr>
<td>Filter coefficient copied to on-chip RAM</td>
<td>0.805</td>
</tr>
<tr>
<td>The original data is copied to the in-chip RAM.</td>
<td>0.793</td>
</tr>
<tr>
<td>Assembly code optimization</td>
<td>0.709</td>
</tr>
</tbody>
</table>

**Table 2** change of calculation amount of control algorithm in the process of interpolating filter optimization

<table>
<thead>
<tr>
<th>Optimizing content</th>
<th>Calculations for the whole control algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Optimization</td>
<td>0.709</td>
</tr>
<tr>
<td>Merge historical data with current frame data</td>
<td>0.67</td>
</tr>
<tr>
<td>Filter coefficient copied to on-chip RAM</td>
<td>0.662</td>
</tr>
<tr>
<td>The original data is copied to the in-chip RAM.</td>
<td>0.57</td>
</tr>
<tr>
<td>Assembly code optimization</td>
<td>0.532</td>
</tr>
</tbody>
</table>

Through the optimized algorithm, the highest frequency of 9.8khz can be controlled (other conditions remain unchanged, such as 8 input, 1 output, 1024 points per frame, etc.), at this time, the sampling frequency is 24khz, and the calculation amount reaches 0.9944. It can be seen that our filter design enables the system to achieve the best performance. Table 5.6 shows the calculation amount comparison between the original single resolution control algorithm and the new multi-resolution control algorithm under the control of different highest control frequency, 1 input control channel, 7 monitoring channels and 1024 points frame, so as
to facilitate the subsequent centralized processing of signals.

2.5 Centralized signal processing

After the completion of the above several links, the final centralized signal processing is mainly realized by the empirical mode method. Empirical mode decomposition method can decompose non-stationary and non-linear signals into a set of stable and linear sequences, i.e. eigenmode function \[^8\]. According to Huang's definition, the IMF of each stage should meet two conditions:

1. The extreme point and zero crossing point of data appear alternately, and the number is equal or at most different at any point;
2. At any point, the mean value of envelope defined by local maximum and local minimum must be zero. The filtering algorithm is as follows:

A. For input signal \(x(t)\), determine all extreme points of \(x(t)\).
B. The upper and lower envelope lines of \(x(t)\) are obtained by fitting the maximum and minimum points with cubic spline function.
C. Subtract the mean value of the upper and lower envelopes from the original data series.

Average curve:

\[
m(t) = \frac{e_{\min}(t) + e_{\max}(t)}{2}
\]  \(1\)

\(m(t)\) represents the average curve, \(e_{\min}(t)\) and \(e_{\max}(t)\) represent the minimum and maximum value of the input signal respectively.

Detail signal:

\[
s(t) = x(t) - m(t)
\]  \(2\)

\(s(t)\) stands for detail signal.

D. Generally, \(s(t)\) does not meet the conditions of IMF, so the above steps need to be repeated for iterative processing \[^9-10\] and the iterative stop criteria given are:

\[
SD = \sum_{t=0}^{T} |S_{n-1}(t) - S_n(t)|^2 / \sum_{t=0}^{T} S(t) \leq \sum_{t=0}^{T} S(t)
\]  \(3\)

SD is the screening threshold value, which is generally 0.2-0.3. If SD is less than this threshold value, the filtering iteration will end. After the iteration meeting the stop criteria, the \(s(t)\) is the effective IMF, and the remaining signals enter the next screening process. After several times of screening, the original data sequence is decomposed into a group of IMF components and a residual, and the IMF obtained is stable. The results obtained by Hilbert
transform can well analyze non-linear and non-stationary signals.

3 Experiment test and analysis

In order to verify the improvement of the signal processing method of the design platform, a simple and intuitive test experiment is designed by using the GUI tool of MATLAB. The function of this system is to decompose the input signal with traditional EMD and EEMD, to display the IMF components and instantaneous frequency of each modal function after signal decomposition, and to describe the Hilbert time spectrum.

3.1 Parameter setting

The variance of white noise and the number of noise groups (range 1-500) can be set freely on the parameters. When the variance is set to 0 and the number of noise groups is selected to 1, the system realizes the function of traditional EMD decomposition. The EEMD decomposition function adds the above set white noise to the signal and depicts the Hilbert time spectrum of the input signal. Display IMFs function can display IMF components and instantaneous frequency after signal decomposition in the form of pop-up Fig.

3.2 Test results and analysis

The specific test results are as follows:

Firstly, the multi-component ideal sample signal is decomposed, and the signal structure is as follows:

\[ x(t) = \sin(2\pi f_1 t) + \sin(2\pi f_2 t) + \sin(2\pi f_3 t) + \sin(2\pi f_4 t) \]  \hspace{1cm} (4)

The normalized frequency is: \( f_1 = 0.02, f_2 = 0.05, f_3 = 0.1, f_4 = 0.2 \).

EMD decomposition method should decompose the signal containing four frequency components into four IMF components containing single frequency information. The decomposition results are shown in Figure 3.
It can be seen that the traditional EMD decomposition method has a very good effect for the ideal signal without interference. The four frequency components are clearly displayed on the Hilbert spectrum. A group of actual signals with interruption interference are decomposed, and the results are shown in Fig. 4 and Fig. 5.

![Fig. 4 Time domain diagram of actual signal](image)

It can be seen from the spectrum that the low frequency components are mixed together and difficult to distinguish.

After analyzing the EEMD processing method, 100 groups of Gaussian white noise with
standard deviation of 0.2 are added, and the results are shown in Figure 4. Through the comparison of Hilbert spectrum, it can be seen that the processing results have been greatly improved.

4 Conclusions

Through the research and update of the signal processing method of campus mutual aid platform of mobile terminal, the interaction between university students and teachers in their spare time is realized, which provides convenience for teachers' classroom management. At the same time, students can also conveniently ask teachers for leave. Students can check their course scores here. Teachers and students can also understand the important information of campus through the platform. In this paper, the signal processing method of campus mutual aid platform based on mobile terminal is designed. By comparing with the traditional signal processing method, it is proved that the method designed in this paper can guarantee higher processing results and better stability under different degrees of interference.

5 Fund projects

Science and technology project of Jiangxi Provincial Department of education in 2019. Project name: Research and implementation of intelligent delivery terminal based on mobile Internet and AI (GJJ191579)

Reference


[8] Fan Xiaochun, Lu Yong, Tao Liang. Research on the Construction of Intelligent Campus in


Joint scheduling of audio teaching resources in NFV service chain deployment

XU Wei-dong¹, YU Bing-yan²

{XuWeidong3812@163.com¹,YuBingyan3812@163.com²}

(1. Jiangxi institute of applied science and technology, Nanchang 330100, China;
2. Jiangxi institute of applied science and technology, Nanchang 330100, China)

Abstract. With the increasing network demand of users, the original middleware equipment shows the defects of single function and large geographical limitation, so the network function virtualization technology (NFV technology) should be introduced. Based on this, an audio teaching resource joint scheduling method for NFV service chain deployment is proposed. The audio resources are first virtualized and combined on demand. Then, the algorithm is deployed according to the NFV service chain to realize the joint scheduling of audio resource data. By comparing with the traditional method, it is proved that the method designed in this paper can accurately distinguish audio resources to a certain extent, ensure a higher scheduling utilization rate of resources, and have better stability and effectiveness.

Keywords: NFV service chain; Audio teaching resources; Joint dispatching;

1 Introduction

With the increasing network demands of users, the original middleware devices show the defects of single function and large geographical limitation. Network Function Virtualization (NFV) technology aims to use the general equipment in the data center to realize network functions, to replace the traditional proprietary equipment, so as to realize the flexible deployment and removal of network functions, and improve the intelligence of network management. Therefore, network function virtualization technology (NFV technology) should be introduced [1]. A single network request is arranged in a reasonable order so that it has several virtual network functions. And under the support of standard x86 equipment, greatly improved the system performance. NFV technology migrates the traditional physical network into the virtual network to realize the large-scale scheduling and management of virtual equipment, which reflects the characteristics of good flexibility, security and reliability. NFV system services consist of a series of virtual network functions that can be flexibly deployed, authenticated, and enabled on demand through decoupling of network functions from
proprietary hardware devices to manage hardware resources extremely efficiently. The NFV service function chain mainly consists of the following workflow: first, the VNF chain composition (VNFS) phase. In an NFV environment, different types and quantities of NF are virtualized and deployed in the virtual network infrastructure to allocate computing resources and link bandwidth resources according to the needs of different users. Related to the number of virtual network functions, the order of association, the distribution of service chains in the physical network and other factors, to provide users with high-performance network services. The second SFC mapping phase. This is the mapping of virtual resources to candidate physical resources, including the mapping of virtual nodes and virtual links, reflecting the typical virtual network embedding problem or virtual data center embedding problem. When each VNF is mapped to a physical node, it is necessary to ensure the fit of the attributes, so as to better realize point-to-point chain network services such as load balancing, data encryption and decryption [2]. At the same time, in the mapping process involving choreographer, service, virtual layer and NFVI, the corresponding network service can be provided for users when the virtual resource is fully mapped to the physical network. Taking an SFC mapping stage as an example, the choreographer reasonably determines the mapping strategy through the operation of the SFC mapping algorithm, and carries different VNFS to the physical nodes in the virtual node mapping stage. In the virtual link mapping phase, the virtual links of VNFs are mapped to the physical network through the mapping algorithm. NFVI is used to calculate, store, and link support throughout the mapping process [3]. Third VNFS scheduling phase. That is, virtual network function scheduling. In this stage, the focus is on efficient scheduling of different SFCS. On the premise of maintaining system service performance, SFC can be deployed and scheduled reasonably based on different user requests to save resource consumption and reduce system running time. In at the same time, in an age of artificial intelligence technology constantly infiltration, build a set of nonlinear signal processing system, can using the correlation algorithm and neural network in the system, realize the optimization of NFV service chain deployment and scheduling and better solve NFV service chain nodes and bandwidth resources waste problem in deployment, enhance the user's web experience.

2 Joint scheduling of audio teaching resources in NFV service chain deployment

2.1 Virtualization of audio resources

The virtualization of audio resources based on overlay network is based on the detection results of audio resources [4]. Access nodes distributed in different geographical areas send detection information to each other and obtain performance data of each rasterized access

797
node through interactive detection results. VNF based on each node to obtain access to the network connection state (related to audio resources distribution and coverage), network interface queue occupancy (with audio resources distribution of processing power and business related), link transmission rate (associated with the transmission technology and the transmission bandwidth of audio resources), packet loss rate (related to the interference environment of audio resources) and other periodic detection data to build the network performance database. Abstract the service transmission and processing service capability of different link transmission services or network units to realize the service abstraction of network transmission capability. VNF upload abstraction of the underlying network transmission ability to VCD, VCD performance in relation to any location from the VNF normalized processing, sample values under different based on the Internet in Beijing to support the big data technology and application of hybrid teaching research business types of different transmission requirements and comprehensive performance measurement mechanism, as shown in figure 1.

![Diagram](image.png)

**Fig. 1.** Audio resource virtualization process based on performance database

### 2.2 Audio resources are combined on demand

First, according to the detection results of service transmission requirements and virtual link performance, select the appropriate cross-network cooperative forwarding entity, construct packet forwarding and track and monitor the forwarding capability of cooperative entity. Second, when the service capability of the collaboration entity changes, the collaboration entity is reselected to complete the flexible reorganization of the business transfer resources and smooth switching of the business transmission path. After receiving the intra-domain routing query business, DMC obtains the source node identity, destination node identity and business type, and then queries the corresponding routing table according to
If a corresponding routing record exists in the local routing table, the channel reservation instruction is generated and sent to the VNF. If no routing record exists, send the inter-domain collaboration request and wait for the collaboration reply to be received before looking for the inter-domain routing table. When DMC performs intra-domain/inter-domain routing calculation, it calculates the link cost corresponding to the current source node and destination node identity based on the latest link state information between nodes maintained by VCD. Query the connection table maintained by DMC. If there is already a routing record for this source node and destination node, the link cost is the same as the current calculated result, there is no need to recalculate the route. Otherwise, update the link cost for this side, then use Dijkstra algorithm to calculate the route and regenerate the route table. According to the current QOS requirements for service transmission, DMC can choose multiple cooperative forwarding paths to provide parallel transmission services. After receiving the channel instruction returned by DMC, the VNF directly providing access to the source node reserves the resource and replies for confirmation. The combination of on-demand resources across domains is different from within domains in that the DMC distributes the business QOS reply to each VNF participating in the collaborative transport on the selected path based on the generated inter-domain routing. If the corresponding route record (unavailable resource) cannot be found, the VNF route establishment is notified of failure.

2.3 Design of NFV service chain deployment algorithm

The complexity of mathematical models in large-scale networks is relatively high and it is difficult to be fully applied \(^6\). Therefore, this section will design an efficient algorithm to solve the NFV service chain deployment problem. By expanding the concept of the longest common subsequence and its dynamic programming method, an efficient algorithm based on the longest common subsequence was designed to deploy the NFV service chain. Given two sequences, sequence \( X \) and sequence \( Y \). If sequence \( X \) is a subsequence of both \( X \) and \( Y \), then sequence \( X \) is said to be a common subsequence of sequence \( X \) and sequence \( Y \). Of all common subsequences of sequences \( X \) and \( Y \), the sequence \( X \) with the maximum length is called the longest common subsequence of sequences \( X \) and \( Y \).

The longest common subsequence is solved by dynamic programming. Given two sequences, the \( X_m = \langle x_1, x_2, \ldots, x_m \rangle \) sequence and the sequence \( Y_n = \langle y_1, y_2, \ldots, y_n \rangle \),
assume that the sequence $Z_k = <z_1, z_2, \ldots, z_m>$ is any of the longest common subsequences of sequence $X$ and sequence $Y$. So if $x_m = y_n$, then $z_k = x_m = y_n$, then the sequence $Z_{k-1}$ is the longest common subsequence of the sequence $X_{m-1}$ and the sequence $Y_{n-1}$.

If $x_m \neq y_n$, and $z_k \neq z_m$, then the sequence $Z_k$ is the longest common subsequences of the sequence $X_{m-1}$ and the sequence $Y$. If $x_m \neq y_n$, and $z_k \neq y_n$, then the sequence $Z_k$ is the longest common subsequence of the sequence $X_m$ and the sequence $Y_{n-1}$.

Through the above analysis, if we use $lcs[i, j]$ to represent the length of the longest common subsequence of sequence $<x_1, x_2, \ldots, x_i>$ and sequence $<y_1, y_2, \ldots, y_j>$, we can obtain the recursive expression to solve the length of the longest common subsequence of sequence $<x_1, x_2, \ldots, x_i>$ and sequence $<y_1, y_2, \ldots, y_j>$ as formula (1).

$$lcs[i, j] = \begin{cases} 0, & i = 0 \text{ or } j = 0 \\ lcs[i-1, j-1] + 1, & i, j > 0 \text{ and } x_i = y_j \\ \max\{lcs[i, j-1], lcs[i-1, j], i, j > 0 \text{ and } x_i \neq y_j\} \\ \end{cases}$$

Therefore, the longest common subsequence length of sequence $X_m = <x_1, x_2, \ldots, x_m>$ and sequence $Y_n = <y_1, y_2, \ldots, y_n>$ only needs to be $lcs[m, n]$ [7]. According to the recursive expression formula (1), the longest common subsequence length of the two sequences can be obtained. In the process of solving the longest common subsequence length of the two sequences, a table of constructing the longest common subsequence is maintained to construct the longest common subsequence of the two sequences. For the dynamic programming method to solve the longest common subsequence length, the time complexity is $\theta(m, n)$, and the time complexity of constructing the longest common subsequence is
For the VNF sequence $\varphi$ in a VNF service chain, the expression of $\varphi$ is $\varphi = <VNF(a_1),...,VNF(a_n)>$. If there is another VNF series in which $\varphi = <VNF(c_1),...,VNF(c_k)>$ is the VNF sub-sequence of $\varphi$, if and only if there is a strictly increasing sequence $<i_1,...,i_k>$ such that $VNF(a_{i_j}) = VNF(c_j)$. The $VNF(a_{i_j}) = VNF(c_j)$ here indicates that their VNF types are the same. For the VNF sequences $\varphi$ and $\varphi'$ of any two NFV service chains. If there exists a VNF series $\hat{\varphi}$ is a VNF common subsequence of the two sequences $\varphi$ and $\varphi'$, if and only if the VNF sequence $\hat{\varphi}$ is a VNF subsequence of both sequence $\varphi$ and sequence $\varphi'$.

In the calculation of the longest common subsequence length of VNF, for the VNF sequence $\varphi$ and $\varphi'$ of any two VNF service chains. In their VNF common subsequence, the VNF common subsequence with the maximum length is called the longest common subsequence of VNF($\text{VNF}\_\text{LCS}$). We write down the longest common subsequence of VNF of the VNF sequence $\varphi$ and $\varphi'$ as a structure $\text{LCS}$, which contains the longest common subsequence of VNF itself $\text{LCS}(\varphi, \varphi').se$ and the length $\text{LCS}(\varphi, \varphi').len$ of the longest common subsequence of VNF.

As mentioned above, the longest common subsequence can be solved in polynomial time using dynamic programming algorithm. If we assume that $\varphi$ is the VNF sequence of NFV service chain request $R_i(s_i, d_i, T_i, B_i)$, and $\varphi'$ is the VNF sequence already deployed on the path $s_i \rightarrow d_i$, then $\text{LCS}(\varphi, \varphi').len$ is the VNF matching degree of sequence $\varphi$ and $\varphi'$. 

$O(m+n)$.
For example, if you have the VNF sequences $\varphi = <VNF_1, VNF_6, VNF_8>$ and $\varphi' = <VNF_1, VNF_5, VNF_8, VNF_9>$, then their $<i_1, ..., i_k> = <VNF_1, VNF_8>$, and $LCS(\varphi, \varphi') \text{len} = 2$. That is, the VNF match between sequence $\varphi$ and $\varphi'$ is 2. We will apply this VNF matching degree in the later algorithm design. For the deployment of NFV service chain in Inter-DC EON network, an algorithm based on longest common subsequence design (LCS-Based Algorithm, LBA) is proposed. Above, the details of the LBA algorithm are given, which considers the VNF sequence deployment of NFV service chain and the spectrum resource allocation in the deployment process.

2.4 Implementation of audio resource data joint scheduling

On the basis of the above resource scheduling arrangement and related algorithm design, the audio sharing resources are reasonably scheduled to ensure the audio data sharing of open courseware teaching resources. In order to overcome the disadvantages of the traditional method, this paper adopts the conjugate hierarchical scheduling method. Firstly, the input resource information data is FFT at N points, which is expressed by equations (2) and (3):

$$x_i(k) = FFT\left[ x_i(k), x_i(k+1), ..., x_i(k+N-1) \right]^T \quad (2)$$

$$x_2(k) = FFT\left[ x_2(k), x_2(k+1), ..., x_2(k+N-1) \right]^T \quad (3)$$

The above input resource information data are used to carry out a balanced design for the audio resource scheduling process, and formula (4) is used to obtain the scheduling data set of the finite conjugate hierarchy.

$$X = \{x_1, x_2, ..., x_n\} \subset R^l \quad (4)$$

Using the finite conjugate hierarchical scheduling data set to schedule n audio resources. Schedule and use equations (5) to (7) for layered design of audio resources:

$$X = \{X[1], X[2], ..., X[N]\} \quad (5)$$

$$X[1] = (id_i, n_i) \quad (6)$$
The characteristic decision rule base is formed by mining and processing the audio resource data at different levels. Formula (8) is used to calculate the power spectrum characteristics of audio resources in the cluster.

\[
M = \sigma_1 \sum_{i=1}^{m,n} (H_i - S_i) + M_2 \sigma_2 \sum_{i=1}^{m,n} (S_i - V_i)
\]

\[
\sigma_3 \sum_{i=1}^{m,n} (V_i - H_i)
\]

(8)

Enter the initial value \( r = 1 \), \( k_j = k_j \), \( b_j = b_j \), \( (j = 1, 2, ..., m) \), \( X = \emptyset \), 錦

\[
K' = \bigcup_j \{ k_j \}, \quad B' = \bigcup_j \{ b_j \}
\]

Y = \emptyset , 錦

Once new data features enter the system, the feature library is constantly enriched. So if \( A_x = \sum_{i=1}^{n} a_i S_{ir} > 0 \) is equal to \( x^r = \left[ \frac{A_x}{b_j^r} \right] \). Carry out the classification attribute decomposition of R class items of audio resources in L time slices, when \( j \neq t \), take the following formula.

\[
A_r = A_r - x^r * b_j^r, \quad k_j^r = \begin{cases} 0, & \text{if } \tilde{t} \leq t \\ k_j^r, & \text{if } \tilde{t} > t \\ \end{cases}, \quad t = t + 1
\]

(9)

When \( y_{ij} = 0 \) is equal to, let's take \( Y_i = Y \cup \{ y_{ij}' \} \).}

Through the above mentioned conjugation layer scheduling, the audio Shared resource data can complete the joint scheduling.

3 Experimental results and analysis

In this paper, a method of joint scheduling of audio teaching resources in NFV service chain deployment is designed. In order to verify the performance of this method, a comparative experiment is designed with the common method of joint scheduling of audio teaching resources. Through comparison, the experimental hypothesis is verified.

3.1 The experiment content

In this section, the experiment will judge the advantages and disadvantages of the two methods by comparing the resource utilization of the two audio teaching resources. The
experimental method is a comparative experiment, in which the experimental verification process is completed by controlling other variables, except the experimental variation, to remain unchanged.

The audio resources for data integration and resource partitioning, each resource scheduling of sampling time is 12 ms, audio resources are divided into several regions, each region with GridSim storage capacity expansion, the number of each audio resource scheduling task file access to 2500 gb, task delay for 10223 ms, according to the simulation environment and parameters for audio Shared resource scheduling simulation data set.

3.2 Analysis of experimental results

First, the time-frequency feature partition of audio resources is constructed. The resource Scheduling and load balancing fusion algorithm with deep learning based on cloud computing algorithm is used as the traditional comparison method[11]. The results of the two methods are compared. The more unified the scheduling resources, the better the scheduling effect. The specific experimental results are shown in Figure 2.

![Fig. 2. Time-frequency characteristic partition of audio resources.](image)

It can be seen from the analysis figure 2 that the algorithm in this paper is used for audio resource scheduling, which has a good grid partitioning ability.

In order to verify the performance of this method, the utilization of resource management mobile nodes is taken as the test index. The comparison results of data utilization ratio of audio sharing resources by different methods are obtained, as shown in Figure 3.
It can be seen from figure 3 that the algorithm proposed in this paper is used for audio source scheduling, with high utilization rate and coverage of data management nodes, high space occupancy and superior performance to traditional methods.

4 Conclusion

This paper designed the NFV service chain deployment under the combined audio teaching resource scheduling method, combined with traditional audio teaching resource scheduling method of contrast experiment, proves that this design method can to a certain extent, accurately distinguish between audio resources, ensures high utilization rate of scheduling resources, has better stability and effectiveness.

5 Fund projects

Research on the artistic characteristics and cultural protection of gongs in Yichun (Project approval No.: yg2017121)

References

[4] Liu Yitzen, Lu Yu, Wang Shan, Chen Xingkai, Qiao Wenzin. An optimal deployment mechanism of


A New OOK Wake-Up Receiver with Sampling Technique

Zhiyuan Tao, Tao Ma, Jicheng Zhang
{zhiyuantao@stu.xidian.edu.cn, tma@xidian.edu.cn, jichzhang@stu.xidian.edu.cn}
Xidian University, Xian, 710071

Abstract. In this paper, we proposed OOK wake-up receiver (WuR) with the sampling technique. In this method, the power consumption of RF envelope detector, which is the most power hungry component in the receiver, is greatly reduced by the signal-level duty cycling (or sampling), and the original OOK modulated IF signal is recovered by a designed active IF bandpass filter, which consume only micro power. Using this novel method, a new wake-up receiver with the high sensitivity and the low power consumption is designed and implemented. According to our test, the lowest input RF power that the OOK modulated signal can be recovered is -70dBm, which is almost same as the sensitivity of envelope detector LT5538. Meanwhile, the power consumption of our WuR is about 310μW.

Keywords: Wake-up receiver, sampling, RF envelope detector.

1 Introduction

Wireless sensor networks (WSNs) have been recognized as an enabling technology for a large variety of applications, including smart homes and cities, agriculture, transportation, health and fitness, entertainment, and structural health monitoring [1]. Reducing the communication power consumption of WSN nodes is important, as the radio transceiver is one of the components with the highest power consumption. Optimizing the power consumption of the wireless transceiver can free up power budget to add much more functionality. In addition, since the battery size is the decisive factor for the entire size of battery, operated systems, low-power circuits could enable smaller batteries and lead to miniaturization required by many applications such as wearable WSNs, medical body area networks and implantable devices.

To establish communication, two radios (the receiver and the transmitter) need to be synchronized as a message can be received only if the radio is in its listening state, and idle listening consumes significant power. Hence, a significant design effort is required to alleviate this power waste. To reduce communication power consumption, several techniques have been proposed [2]-[7] for lowering or eliminating the power wasted due to idle listening of the transceiver. Duty cycling is a common technique to reduce the idle mode energy consumption which consists of switching from listening mode to sleep mode [2]. However, while duty cycling helps save power, it can severely limit the network reactivity as the radios are OFF (or in the sleep state) and they cannot receive messages. Different from general duty cycling receiver, bit-level duty cycling receiver samples only a portion of each transmitted bit to reduce the power consumption of RF front end [8]. The paper [9-10] designed a logarithmic power detector based WuR with the bit sampling technique, obtaining sensitivity of -77dBm. In this
paper, we proposed a signal-level sampling wake-up receiver, whose power consumption can be dynamically adjusted by MCU. Furthermore, we proposed an active IF filter to recovery the original continuous IF signal from the samples of the output of RF envelope detector. The attenuation caused by sampling can be alleviated by an IF amplifier, which consumes only micro watt power due to its low frequency. Therefore, the overall power of WuR can also keep in several micro watts by the duty cycling (or sampling) the high-power component RF envelope detector. According to our test, the lowest input RF power that the proposed WuR can successfully recovery the OOK modulated signal is about -70dBm, which is almost same as the sensitivity of envelope detector LT5538. Meanwhile, we could calculate the power consumption of WuR is about 310μW.

The rest of paper is organized as follows: Section 2 gives the description of the proposed WuR with sampling, mathematically analyze attenuation of the output of the IF bandpass filter. Section 3 presented a simulation design of IF filter, and conducted the hardware experimental test to evaluate the performance of the proposed WuR. Finally, the conclusion remarks were given in Section 4.

2 The proposed OOK WuR

2.1 The description of the proposed OOK WuR

We proposed a new WuR with sampling technique, which is shown in Fig.1. The OOK modulated RF signal is chosen as the input of our WuR. In our WuR, the envelope detector is enabled only short duration every period T. Then the output of our envelope detector is the samples of original demodulated OOK signal, which is shown in Fig.2b. After the bandpass filter with a proper setting and the comparator, the original OOK signal can be recovered.

![Fig. 1. The architecture of WuR.](image)

By signal-level sampling the envelope detector, which is the most power hungry component, the overall power of WuR is greatly reduced. After an bandpass filter and an IF amplifier, the original OOK signal is recovered from the samples with limited power consumption. Therefore, compared with the tuned RF receiver, the proposed wake-up receiver is able to demodulate OOK signal with much lower power consumption.

2.1 The Theoretical Analysis

Let $S(t)$ denote the periodic square wave signal with period T, $S_p(t)$ denote the periodic pulse signal with period $T_p$ and pulse width $\tau$. The output of envelope detector can be expressed as follows.
Let \( F_s(w) \) be the Fourier transform of \( S_s(t) \), \( F_w \) denote the Fourier transform of \( S(t) \).

\[
F_s(w) = \frac{1}{2\pi} S(w) \otimes F[S_p(t)] 
\]

Where \( F(S_p(t)) \) are

\[
F[S_p(t)] = 2\pi \sum_{n=0}^{N} \frac{\pi n \tau}{T_p} S_a \left( \frac{\pi n \tau}{T_p} \right) \delta \left( w - n \frac{2\pi}{T_p} \right) 
\]

Combining (3) and (4), we have:

\[
F_s(w) = \frac{\tau}{T} (S(w)) + \sum_{n=1}^{N} S_a \left( \frac{\pi n \tau}{T_p} \right) S \left( w - n \frac{2\pi}{T_p} \right) 
\]

When \( \frac{1}{T_p} \geq \frac{2}{T} \), by using a perfect bandpass filter, we can recover \( F_f(w) \).

\[
F_f(w) \approx \frac{\tau}{T} S(w) 
\]

Thus, by using a proper bandpass filter, we could recover original signal. As we can see from equation (6), the strength of the recovered signal is related with duty cycle of the signal \( S_p(t) \), which is the ratio between \( \tau \) and \( T \). By decreasing this ratio, the average power of envelope detector is reduced, however at the same the amplitude of the recovered signal is also decreased. Nevertheless we could use low power IF amplifier to enlarge the recovery signal since its frequency is not high.
3 The Simulation and Experimental Results

We use TINA-TI software simulation tools provided by TI company [13] to simulate the active filter. We design a bandpass filter with cut-off frequency of 3KHz to filter out the alias frequency of the sampled signal. The designed active bandpass filter is shown in fig.3.

![Fig. 3. The active IF bandpass filter.](image)

It is worth noting that the output impedance of envelope detector in our WuR changes over time. The output impedance of LT5538 is 150Ω when it is enable and is 30KΩ when it is off [11]. These changes can significantly impact the characteristic of designed filter. In order to eliminate this influence we add a signal follower in front of IF active filter to stabilize the input impedance of designed bandpass filter. Its frequency characteristic of our designed filter is shown in Fig.4. We can see that it has -3db gain at point A where the frequency is 7HZ and point B where the frequency is 3KHZ. The purpose of filtering out low frequency signals below 7HZ is to avoid excessive DC voltage caused from the envelope detector. On the other hand, filtering out low frequency signals avoids DC voltage changes due to communication distance changes. The purpose of filtering out frequency signals above 3KHz is to recover IF signal.

Furthermore, we conducted the hardware experiment to evaluate the proposed idea. We adopted LT5538 as the RF envelope detector, which has the sensitivity of -71.5dBm at 880 MHZ. It consumes 29 mA current at active stage, while consumes 1μA at disabled stage. LMV551 is chosen to compose bandpass filters. It consumes 49μA at 3V. MSP430FR5969 is chosen as microcontroller in our WuR. It consumes 210μA at 3V when it is active, and only 0.4μA at low power mode 3 where only auxiliary clock is on. One of its IO pin is connected to the enable port of LT5538 to control its working stage. The output of filter is connected to another pin in our microcontroller. The measurement setup shows in Fig.5.

![Fig. 5. The measurement setup.](image)
In this simulation, the RF signal generator creates the 1kbps OOK modulated signal. We samples the envelope detector LT5538 by a rectangular pulse signal with a frequency of 4 kHz and a pulse width of 1μs, which is shown in Fig.6. It means that power consumption of RF envelope detector is 1/100 of the original power. This sampled signal is selected as the input of our designed filter. Fig.7 shows output of the filter. Finally, the recovered signal is shown in Fig.8. The designed filter successfully recovered the original signal from its samples as analyzed earlier in this article.

As you can see from our test, OOK signal is successfully recovered. Now we analyze the power of the proposed WuR in our implementation. In this paper we only consider the power of RF front end, since it is most power hungry part of the receiver. Let $P_a$ and $P_c$ be the power of envelope detector in active stage and inactive stage, $P_f$ be the power of filter, $P_a$ be the power of comparator. We assume that $n = \frac{T}{\tau_p}$. Then the power of our WuR can be expressed as:

$$P_{all} = \frac{\pi}{T} \times P_{ea} + \left(1 - \frac{\pi}{T}\right)P_{ei} + P_f + P_a \tag{7}$$

If we set $\tau = 1\mu s$, $n = 4$ and we select the hardwares in Table 1, we could easily calculate the power consumption of WuR is about 310μW.
According to our test, the lowest input RF power that the proposed WuR can successfully recovery the OOK modulated signal is about -70dBm, which is almost same as the sensitivity of envelope detector LT5538.

<table>
<thead>
<tr>
<th>Hardware</th>
<th>Power(active)</th>
<th>Power(idle)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active filter</td>
<td>111μW</td>
<td>111μW</td>
</tr>
<tr>
<td>Envelope detector</td>
<td>90mW</td>
<td>3μW</td>
</tr>
<tr>
<td>Comparator</td>
<td>152nW</td>
<td>152nW</td>
</tr>
</tbody>
</table>

**Fig. 6.** The sampled OOK signals. (-60dbm input)

**4 Conclusion**

In this paper we proposed wake-up receiver with the sampling technique. In this method, the RF frontend component power consumption is greatly reduced by signal-level sampling. On the other hand, the original IF OOK modulated signal is recovered by a designed active bandpass filter with limited power consumption. By this method, the overall power consumption of the receiver is greatly reduced without lowering the sensitivity. Using this new method, we implemented the OOK wake-up receiver with the off-the-shelf components. According to our experimental results, the implemented wake-up receiver can consume only less than 310μW and achieve a -70 dBm sensitivity.
Fig. 7. The output of the filter amplification. (-60dbm input)

Fig. 8. The recovered signal. (-60dbm input)

References


Research on node importance evaluation of railway logistics distribution network based on Rough Set

LIU Xiu-miao
{LiuxiuMiao22@163.com}
HEBEI VOCATIONAL COLLEGE OF RAIL TRANSPORTATION, Shijiazhuang 050000, China

Abstract: The traditional evaluation method of the node importance of railway logistics distribution network has not eliminated redundant data, resulting in errors in the evaluation results. Therefore, the evaluation method of the node importance of railway logistics distribution network based on rough set is proposed. This method uses the rough set theory to eliminate redundant node data, obtain the collection and distribution uncertainty index, establish an importance evaluation index, obtain the node importance evaluation weight, and achieve the higher accuracy of the network node importance evaluation. The experimental results show that compared with the traditional evaluation method, the evaluation results of network node importance based on rough set are closer to the expected results. It can be seen that the performance of this method is superior.

Keywords: rough set; collection and distribution network; node importance

1 Introduction

With the continuous development of national economy and the continuous optimization of industrial structure adjustment, China's railway transportation industry has grown rapidly. In order to improve the transportation efficiency and save the transportation cost, the railway logistics collection and distribution requires the importance evaluation of the network nodes, so as to set up the railway transportation scheme.

According to this requirement, the traditional evaluation method directly evaluates the node importance of railway logistics with complex road conditions and overlapping lines, and the railway transportation department formulates the collection and distribution scheme according to the evaluation results. However, as the length and complexity of the route become longer, there are some drawbacks in the scheme under the design of the evaluation results. The elimination of redundant data is not complete, resulting in errors in the evaluation results.

Therefore, in view of the shortcomings of the traditional methods, this paper puts forward the evaluation method of node importance of railway logistics collection and distribution
network based on rough set [1]. As a reliable technology to deal with uncertainty, rough set theory is of great significance to remove redundant nodes in the network. This method not only solves the existing problems of the traditional evaluation method, but also improves the accuracy of the evaluation results to a new level, ensures the feasibility of the implementation of the railway logistics collection and distribution scheme, and provides strong technical support for the development of the national railway transportation industry.

2 Evaluation of node importance of railway logistics distribution network based on Rough Set

2.1 obtaining uncertainty index of collection and distribution based on Rough Set

On the premise of using rough set theory to keep the classification ability unchanged, the redundant attributes of railway logistics network nodes are eliminated, and the hidden potential laws in the data are explored. Given an information system dimension \( IS = (W, K) \), where \( W \) is the domain and \( K \) is the set of conditional attributes on the domain \( W \). For any given conditional attribute \( k \in K \), there exists a function \( k : W \rightarrow T_k \), \( T_k \) expressed as a value field with attribute \( k \). Each element in \( W \) is called an individual, an object, or a row.

For any given attribute subset \( I \subseteq K \) and any given individual \( x \subseteq K \), the following information functions are satisfied:

\[
\text{Inf}_I(x) = \{(k, k(x)) : k \in I\} \quad (1)
\]

Suppose that a set of conditional attributes on the domain \( W \) is \( K \), if there are \( I \subseteq K \) and \( I \neq \emptyset \), then \( I \cap K \) is an equivalent relation on domain \( W \), which is called the indiscernible relation on \( I \), and it is recorded as \( \text{IND}(I) \). According to the above, we can know the concept \( X \subseteq W \) of network node and the attribute subset \( I \subseteq K \). Then we can define the upper and lower approximate domain of the concept \( X \):

\[
\begin{align*}
\overline{I}(x) &= I(IS, X) = \{x | (x \in W) \land ([x] \cap X \neq \emptyset)\} \quad (2) \\
\underline{I}(x) &= I(IS, X) = \{x | (x \in W) \land ([x] \cap X)\}
\end{align*}
\]

It can be concluded that the set \( \text{IND}_I(X) = \overline{I}(X) - \underline{I}(X) \) is a \( I \) boundary region of \( X \); the set \( \text{POS}_I(X) = \underline{I}(X) \) is a \( I \) positive region of \( X \); the set \( \text{NEG}_I(X) = W - \overline{I}(X) \) is a \( I \) negative region of \( X \) [3]. Figure 1 below shows the
upper and lower approximation domain, boundary domain and negative region control diagram of set $X$ under the control of rough set.

![Diagram of upper and lower approximation domain, boundary domain and negative region](image)

**Fig. 1.** Schematic diagram of upper and lower approximation domain, boundary domain and negative region

Using the parameters in the figure above, the attributes of railway logistics collection and distribution network nodes are reduced. Through the upper and lower approximation domain, boundary domain and negative region control of the set, the final region is delineated in the display area of the graph, reducing the possibility of nodes’ disorderly connection, so as to reduce the attributes of network nodes. Attribute reduction needs to keep the classification ability of database unchanged and delete unnecessary redundant attributes. Suppose that given a database $M = (W, S)$ and an equivalent relation family $G \in S$ in the database, for any $H \in G$, if there are:

$$\text{IND}(G) = \text{IND}(G - \{H\})$$  \hspace{1cm} (3)

When the above formula holds, it is unnecessary to call logistics data $H$ as $G$, otherwise it is necessary to call logistics data $H$ as $G$. If it is necessary for each $H \in G, H$ to be $G$, $G$ is called independent, otherwise $G$ is called dependent or not independent, so as to realize attribute reduction of logistics network nodes [3].

Rough set is a method to deal with the uncertainty problem, so the uncertainty index is measured for the uncertainty of the nodes in the distribution network of iron flow logistics. It is known that in the information system $IS = (W, K)$, there is a condition $X \subseteq W$, then under the condition of attribute subset $I \subseteq K$, the membership calculation expression
of \( x \in W \) relative to concept \( X \subseteq W \) is:

\[
\phi(I, X, x) = \frac{|[x] \cap X|}{|[x]|} \quad (4)
\]

Where \(||\) represents the potential of the set. If there is a conceptual condition \( X \subseteq W \) in the logistics collection and distribution decision system \( DS = (W, K, a) \), which is called decision attribute \( a \) depends on the condition attribute set \( K \) with degree \( h(0 \leq h \leq 1) \), then:

\[
h = \lambda(K, a) = \frac{|POS_K(a)|}{|W|} \quad (5)
\]

Where \( \lambda \) is the degree of dependence. According to the formula and formula (4), the uncertainty index is obtained:

\[
\beta = (I, X)\left(1 - \frac{\phi}{\ln h} \times \kappa \cdot \frac{I(X)}{I(X)}\right) \quad (6)
\]

In the formula, \( \kappa \) represents the influence factor of the uncertainty degree index; \( \beta \) is the uncertainty index of the logistics network node [4].

2.2 Establishment of importance evaluation index

According to the structural characteristics of railway logistics collection and distribution network, the evaluation indexes of the node importance are established based on the central characteristics and structural hole characteristics of network nodes. According to the idea of constructing the contribution matrix of node importance, a new method of node importance ranking is proposed. This method uses the importance index value of the structure hole and the \( k \) core of the adjacent nodes to get the importance contribution relationship between the nodes. At the same time, the \( k \) core importance of the nodes is used to describe the global position information of the nodes. Based on the analysis of the local importance of the network, combined with the global importance, the importance evaluation index of the nodes is more comprehensive and the evaluation result is more accurate True. From a sociological point of view, structural holes exist in the gap between data without redundant connections, as shown in Figure 2 [5].
After the rough set processing in the previous step, there is no redundant connection between \( u \) and \( v \) in the figure. The logistics data with structural holes between them can bring cumulative rather than overlapping network route planning. As can be seen from Figure 2, node \( k \) has more network revenue than its neighbor nodes \( u \), \( v \) and \( w \), that is, in the network, node \( k \) is more important than other nodes. However, if nodes \( v \) and \( w \) are connected, the network revenue of node \( k \) will be reduced. From the point of view of complex network, the more structural holes a node has, the more conducive to information dissemination.

Let \( Q = (A,B) \) be a logistics collection and distribution undirected network without self loop, which has \( n \) nodes and \( m \) edges, of which \( A = \{a_1, a_2, \ldots, a_n\} \) is the set of all nodes in the network and \( B = \{b_1, b_2, \ldots, b_m\} \) is the set of edges between nodes. The adjacency matrix is recorded as \( C_{nxn} = (d_y)_{nxn} \), where:

\[
d_y = \begin{cases} 1, & \text{if connected to } j \\ 0, & \text{if not connected to } j \end{cases}
\] (7)

According to the above formula (7), the degree of node \( i \) can be recorded as:

\[
f(i) = \sum_{j \in Q} d_y \] (8)

According to the degree of node, the adjacency degree of node \( i \) can be expressed as:
In the formula, $\Upsilon(i)$ represents the $i$ set of adjacent nodes, and $\gamma$ represents the adjacency influence parameter. It is known that the input energy of node evaluation index $i$ to evaluation node $j$ can be shown in Figure 3 [6].

\[ h(i) = \sum_{j \in \Upsilon(i)} k(\gamma) \quad (9) \]

According to the above figure, the network constraint coefficient can measure the constraints of adjacent nodes when the nodes in the network form the structural hole, and it is an indicator to measure the structural hole. The smaller the constraint coefficient is, the greater the degree of structural hole is, and the more important the nodes are. Then the constraint coefficient can be expressed as:

\[ RC_i = \sum_{j \in \Upsilon(i)} \left( g_{ij} + \sum_q g_{iq} g_{qj} \right)^2 \quad (10) \]

Where: $q$ represents the number of common neighbors of node $i$ and node $j$. Considering the influence of the node degree and the topology of the number of neighbors on the node, the input energy of node $i$ to node $j$ can be recorded as $g_{ij}$:

\[ g_{ij} = \frac{RC_i Q(j)}{\sum_{s \in \Upsilon(i)} Q(s)} \quad (11) \]
Where: $Q(j)$ represents the separation function of node $j$; $Q(s)$ represents the separation function of the desired node similar to the attribute of node $j$. According to the formula, the evaluation indexes of importance are as follows:

$$p_i^c = g_{ij} \left( p_i^c + \alpha p_i^m \right) \quad (12)$$

In the formula: $p_i^c$ represents the evaluation index with the evaluation degree of $e$; $p_i^z$ represents the evaluation standard index with the constraint amount of $z$; $p_i^m$ represents the evaluation reference value with the update times of $m$; $\alpha$ represents the controllable parameters of the node to achieve the accurate establishment of the evaluation index of the node importance [7].

2.3 Determine the evaluation weight and evaluate the importance of network nodes

According to the regional characteristics of the railway logistics network, taking the average degree of nodes, the centrality of mediators and the rate of message forwarding of the whole network as the attributes of the evaluation key nodes, and combining the evaluation importance index, the evaluation weight of the network nodes is calculated, so as to realize the evaluation of the importance of the network nodes.

According to the subjective information of logistics transportation, AHP is used to weight the importance evaluation indexes of each attribute. On the basis of determining the decision-making objectives and evaluation attributes, a suitable hierarchical structure model is built for the next calculation; Establish the judgment matrix, assume $f = \{f_1, f_2, f_3, \ldots, f_n\}$ is the set of decision attributes, $D$ is the goal, $s_i$ is the value of $f_i$ for the quantitative expression of the relative importance of $f_j$ [8]. By comparing the relative importance of the attributes between the two, the judgment matrix is established as follows:

$$S = \begin{bmatrix}
1 & s_{12} & s_{13} & \cdots & s_{1m} \\
s_{21} & 1 & s_{23} & \cdots & s_{2m} \\
s_{31} & s_{12} & 1 & \cdots & s_{3m} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
s_{m1} & s_{m2} & s_{m3} & \cdots & 1
\end{bmatrix} \quad (13)$$
In the above formula: $s$ represents the node attribute in the logistics collection and distribution network. According to the above matrix, calculate the attribute weight of the evaluation node. The judgment matrix $S$ is calculated, and the accumulation of elements in each row is $R_j$:

$$R_j = \prod_{i=1}^{m} s_{ij} \quad (14)$$

Where, $j = 1, 2, \ldots, m$. The $m$ power root of $R_j$ is calculated to obtain $w_j$:

$$w_j = m^{\sqrt{R_j}} \quad (15)$$

According to the above formula, the weight $\overline{w}_j$ corresponding to attribute $f_j$ is obtained by normalizing $w = (w_1, w_2, w_3, \ldots, w_m)$ vectors:

$$\overline{w}_j = \frac{w_j}{\sum_{c=1}^{m} w_c^2} \quad (16)$$

Where: $\overline{w}_j$ represents the weight of the evaluation of the importance of the network nodes under the control of the boundary constraint $c$ [9]. Use the consistency index test method to test the consistency of the weight. The following formula (17) is the consistency index test result:

$$\begin{align*}
CI &= \frac{\zeta_{\text{max}} - m}{m - 1} \\
CR &= \frac{CI}{RI}
\end{align*} \quad (17)$$

In the above formula: $\zeta_{\text{max}}$ represents the maximum consistency of the weight index; $CR$ represents the consistency index test coefficient; $RI$ represents the average consistency index of the weight, and the value of this index can be determined according to table 1.

*Table 1.* $RI$ list of values

---

822
Because the judgment matrix is generated based on subjective judgment, and there may be inconsistencies in the judgment matrix, $CR$ is introduced to check whether the judgment matrix is consistent. If $CR = 0.1$ is calculated, it is considered that the matrix is completely consistent; if $CR < 0.1$, it is considered that the matrix basically meets the requirements; if $CR \geq 0.1$, it is considered that the matrix does not meet the consistency requirements, and the above calculation must be repeated until the consistency requirements are met [10]. At this point, the weight can be used to evaluate the importance of railway logistics network nodes based on rough set.

3 Experimental test

The evaluation method proposed in this paper is compared with the traditional evaluation method, and the evaluation effect of the importance degree of railway logistics collection and distribution network nodes is compared through test experiments. Set up the experimental test platform, select a section of area as the basic environment for the experimental test, using MSE to measure the difference between test results and expectations. And Figure 4 below is the schematic diagram of the network topology of the railway logistics collection and distribution area selected for the experimental test.

<table>
<thead>
<tr>
<th>Judgment matrix dimension</th>
<th>RI</th>
<th>Judgment matrix dimension</th>
<th>RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.00</td>
<td>6</td>
<td>1.24</td>
</tr>
<tr>
<td>2</td>
<td>0.00</td>
<td>7</td>
<td>1.32</td>
</tr>
<tr>
<td>3</td>
<td>0.58</td>
<td>8</td>
<td>1.41</td>
</tr>
<tr>
<td>4</td>
<td>0.96</td>
<td>9</td>
<td>1.45</td>
</tr>
<tr>
<td>5</td>
<td>1.12</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
According to the railway collection and distribution line in the figure above, the road area network control nodes are set to connect the basic information of the road with each node, as shown in Table 2 below.

**Table 2. Basic information of sections in the area**

<table>
<thead>
<tr>
<th>Link number</th>
<th>Target</th>
<th>Road grade</th>
<th>Link length</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>5</td>
<td>Highway</td>
<td>700</td>
</tr>
<tr>
<td>S2</td>
<td>5</td>
<td>Main road</td>
<td>700</td>
</tr>
<tr>
<td>S3</td>
<td>4</td>
<td>Secondary road</td>
<td>650</td>
</tr>
<tr>
<td>S4</td>
<td>5</td>
<td>Secondary road</td>
<td>600</td>
</tr>
<tr>
<td>S5</td>
<td>4</td>
<td>Secondary road</td>
<td>645</td>
</tr>
<tr>
<td>S6</td>
<td>4</td>
<td>Main road</td>
<td>710</td>
</tr>
<tr>
<td>S7</td>
<td>2</td>
<td>Branch road</td>
<td>220</td>
</tr>
</tbody>
</table>
It is known that there are 22 nodes in the above regional network. In this experiment, nodes 1 to 10 are taken as the importance evaluation objects. The data in Table 3 below is the basic road data of these nodes.

Table 3. Road data information at nodes

<table>
<thead>
<tr>
<th>Node number</th>
<th>Traffic</th>
<th>Capacity</th>
<th>Saturation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5301</td>
<td>8943</td>
<td>0.62</td>
</tr>
<tr>
<td>2</td>
<td>4049</td>
<td>6950</td>
<td>0.67</td>
</tr>
<tr>
<td>3</td>
<td>5486</td>
<td>9528</td>
<td>0.58</td>
</tr>
<tr>
<td>4</td>
<td>6792</td>
<td>9516</td>
<td>0.73</td>
</tr>
<tr>
<td>5</td>
<td>12485</td>
<td>16827</td>
<td>0.79</td>
</tr>
<tr>
<td>6</td>
<td>11397</td>
<td>12795</td>
<td>0.82</td>
</tr>
<tr>
<td>7</td>
<td>2495</td>
<td>4910</td>
<td>0.5</td>
</tr>
<tr>
<td>8</td>
<td>1046</td>
<td>1795</td>
<td>0.65</td>
</tr>
<tr>
<td>9</td>
<td>779</td>
<td>1204</td>
<td>0.68</td>
</tr>
<tr>
<td>10</td>
<td>11834</td>
<td>6133</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Based on the above data as the experimental basic data, it is assumed that the expected value of the evaluation result is shown in Figure 5 below.
Taking the data in Figure 5 as the experimental reference value, two evaluation methods are used to evaluate the importance of network nodes, among which the test results of the proposed evaluation method are group A, the test results of the traditional evaluation method are group B, and Figure 6 is the comparison results of the five tests of this experiment.

According to the above figure, compared with the expected test results, the test results of the proposed evaluation method are very similar to the expected curve; while the evaluation results of the traditional evaluation method are similar to the general curve twice and the disappointed curve thrice, which shows that the evaluation effect of the network node importance evaluation method based on rough set is better.

4 Conclusion

Based on the evaluation method of rough set, the redundant data in the network node is eliminated by rough set theory, and the evaluation weight is obtained according to the established importance evaluation index, so as to realize the more regional and comprehensive importance evaluation of railway logistics network node. The test results of the evaluation method are very similar to the expected curve. However, this evaluation method is proposed for the collection and distribution of railway logistics, which may not be suitable for the evaluation of the importance of other transport network nodes, so it can not be used blindly.

Reference

[1] Xu Fei, Ren Yongtai, Yi Ke-han.: Study of Evaluative Method of Graduate Training Quality Based on


Multi objective optimization model of railway transportation logistics network based on improved ant colony algorithm

LIU Xiu-miao
{LiuxiuMiao22@163.com}
HEBEI VOCATIONAL COLLEGE OF RAIL TRANSPORTATION, Shijiazhuang 050000, China

Abstract. In view of the problem that the transportation cost of the original multi objective optimization model of railway transportation logistics network is too high, this paper introduces the improved ant colony algorithm and designs the multi objective optimization model of railway transportation logistics network based on the improved ant colony algorithm. Set the logistics warehouse as the logistics transportation node and calculate the node optimization cost; use the improved ant colony algorithm to obtain the optimal transportation path in the railway transportation logistics network; set the optimal path as the form of particles, and use the particle swarm algorithm to complete the multi objective optimal solution process of the railway transportation logistics network, and build the multi objective optimization model. So far, the multi objective optimization model design of railway transportation logistics network based on improved ant colony algorithm is completed. Compared with the original optimization model, the application effect of this optimization model is better, which can effectively reduce the cost of logistics network transportation.

Keywords: improved ant colony algorithm; multi objective optimization; railway transportation; logistics network;

1 Introduction

Railway transportation is a basic industry related to the national economy and people's livelihood. Its security, security, coverage and advanced technology play a vital role in our daily life and the country's basic economy. [1] In recent years, the pressure and challenges faced by the railway transportation industry are more and more serious. Therefore, the railway logistics company must strengthen its service capacity, reduce costs and increase efficiency, so as to meet the requirements of social and economic development. With the development of logistics industry, railway logistics plays a more and more important role in the development of logistics industry. Reducing logistics cost and improving logistics service is a great
opportunity for the development of railway logistics company. The service ability of logistics is based on the logistics network of railway transportation. Therefore, the optimization of the node layout of the logistics network of railway logistics is the basis and key to improve the service level of railway logistics. It can effectively reduce the logistics cost of logistics companies, improve the efficiency of logistics operation, and achieve the purpose of cost reduction and efficiency increase of logistics companies [2].

Based on the above background, aiming at the existing problems of railway logistics network node layout, this paper designs a multi objective optimization model of railway logistics network. Firstly, based on the existing hierarchical structure of railway logistics network nodes, the improved structure of "virtual warehouse + regional warehouse + turnover warehouse" is proposed, which provides the premise and foundation for further research; then, through the theoretical research of warehouse location and regional aggregation, the regional aggregation model and cost model suitable for the optimization of railway logistics network node layout are proposed, aiming at the optimization of railway logistics network nodes At the same time, it determines the optimization of the regional warehouse, and analyzes the storage cost, transportation cost and service penalty cost to realize the optimization of the total cost of the railway logistics network, and determines the optimization of the turnover warehouse; then, it uses the modified ant colony algorithm to program the model; finally, it uses the experimental demonstration analysis method By analyzing the optimization results, the feasibility and effectiveness of the optimization method of railway logistics network node layout proposed in this paper are verified, which provides a new idea and method for optimizing the node layout of railway logistics network.

2 Multi objective optimization model design of railway transportation logistics network based on improved ant colony algorithm

Through the study of the original railway transportation logistics network, it can be seen that the original railway transportation network multi objective model can not select the optimal logistics route, resulting in the problem of high logistics cost. In this model design, the original logistics network model is optimized by using the improved ant colony algorithm to reduce the cost of railway transportation logistics. In order to ensure the effectiveness of the design process, the construction process model is used to control the optimization model design process. The specific process is as follows.
Use the above process to complete the design of the optimization model. In this design, in addition to improving the ant colony algorithm, particle swarm optimization is also added to ensure the accuracy and effectiveness of the optimization model design.

2.1 Optimization of railway transportation logistics network nodes

According to the related problems of the railway transportation logistics network proposed above, firstly, the hierarchical structure of the railway transportation logistics network node is adjusted, and the virtual warehouse in the logistics network is used as the information center of the railway logistics material warehouse, which is responsible for the overall management and coordination of the materials in the logistics network, so as to break the administrative barriers and improve the regional self-management of the material warehouse in the logistics network In order to achieve the top-down management goal of logistics network materials, we should give full play to the advantages of logistics standardization, intensification and scale
[3-4], and further establish an integrated management system of "virtual warehouse + regional warehouse + turnover warehouse". The hierarchical structure of the optimized logistics network node is shown in the figure below.

![Hierarchical structure of railway logistics network nodes](image)

Fig. 2. Hierarchical structure of railway logistics network nodes

When optimizing the node layout of the logistics network, on the one hand, the service level of the optimized logistics network should be as high as possible, on the other hand, the lowest cost of the logistics network should be taken into account, so it is necessary to analyze the cost of the logistics network. In the previous section, the location of the regional warehouse of the logistics network is determined. Then, in the optimization of turnover warehouse, the location of regional warehouse is taken as the known condition, and the rest warehouses except the logistics network material warehouse which is selected as the regional warehouse are taken as the alternative warehouse of turnover warehouse to optimize the storage, transportation and service penalty cost [5], so as to determine the location of turnover warehouse. The cost parameter settings required by the network node are shown in the table below.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Parameter Value</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>3</td>
<td>Inventory turnover</td>
</tr>
<tr>
<td>B</td>
<td>4</td>
<td>Regional Library Collection</td>
</tr>
<tr>
<td>C</td>
<td>15</td>
<td>Alternative node set of turnover</td>
</tr>
<tr>
<td>Symbol</td>
<td>Value</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>-------</td>
<td>-------------</td>
</tr>
<tr>
<td>$E$ (Ten thousand yuan/m²)</td>
<td>0.3</td>
<td>Material stock per square meter</td>
</tr>
<tr>
<td>$M_1$ (yuan/m²)</td>
<td>1600</td>
<td>Reconstruction cost of regional warehouse</td>
</tr>
<tr>
<td>$N_1$ (yuan/m²)</td>
<td>1100</td>
<td>New cost of Regional Library</td>
</tr>
<tr>
<td>$M_2$ (yuan/m²)</td>
<td>1200</td>
<td>Transformation rate of turnover warehouse</td>
</tr>
<tr>
<td>$N_2$ (yuan/m²)</td>
<td>850</td>
<td>New rate of turnover warehouse</td>
</tr>
<tr>
<td>$P_1$ (yuan/m²)</td>
<td>260</td>
<td>Average labor cost per square meter of regional warehouse</td>
</tr>
<tr>
<td>$Q_1$ (yuan/m²)</td>
<td>16.5</td>
<td>Average storage and maintenance cost per square meter of regional warehouse</td>
</tr>
<tr>
<td>$P_2$ (yuan/m²)</td>
<td>230</td>
<td>Labor cost per square meter of turnover warehouse</td>
</tr>
<tr>
<td>$Q_2$ (yuan/m²)</td>
<td>8.5</td>
<td>Average storage and maintenance cost per square meter of turnover warehouse</td>
</tr>
<tr>
<td>$F_1$ (m)</td>
<td>60</td>
<td>Service radius of regional warehouse</td>
</tr>
<tr>
<td>Parameter</td>
<td>Value</td>
<td>Description</td>
</tr>
<tr>
<td>------------</td>
<td>-------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>F2(m)</td>
<td>25</td>
<td>Service radius of turnover warehouse</td>
</tr>
<tr>
<td>T1(yuan/m)</td>
<td>13</td>
<td>Service radius of turnover warehouse</td>
</tr>
<tr>
<td>T2(yuan/m)</td>
<td>9</td>
<td>Service penalty rate of turnover warehouse</td>
</tr>
</tbody>
</table>

Through the above parameter setting, the cost of the optimized logistics network node is calculated, and the cost is scheduled as a constant value to ensure the optimized logistics network node, which can effectively reduce the logistics cost. The optimized nodes are taken as the design basis of the optimization model.

2.2 Using improved ant colony algorithm to solve the optimal logistics path

Using the above set logistics network nodes to complete the initialization work, combined with the characteristics of the original railway transport logistics, the optimal solution flow of the logistics path is formulated as follows.
According to the above process, ant colony algorithm is used to solve the optimal route [6-7]. Because ant colony algorithm is an evolutionary algorithm that simulates the behavior of ant colony in nature, it sets the logistics transport vehicle as an ant in the ant colony algorithm, and sets the feasible route from the material distribution center to the goods demand.

The action route of logistics vehicles is \( a_j \), and the initial coordinate of the route is set as \( [(m_i, n_{i,ord}), (m_j, n_{j,ord})] \). In the route construction, there is a set of material demand points \( \mathcal{W} \), in which there are all the logistics points in the corresponding region, and these logistics points can meet the needs of railway transportation, logistics cost budget and transportation
time. Select the appropriate material demand point in the set and insert it in front of \( M_n \), and set it to connect with \( M_n \). By implementing this step repeatedly, a new railway logistics transportation route can be obtained. In order to ensure that when the logistics transport vehicle goes to the next material demand point \( f \) at a demand point \( c \), variable \( X \) is generated in the influence conditions of logistics transport. When \( X \leq X_0 \), the logistics vehicle will go to the next material demand point. The formula can be expressed as follows:

\[
\delta = \arg\max_{i \in T} \beta_j \Xi_j (1)
\]

In the above formula, \( \beta_j \), \( \Xi_j \) are the influencing factors in the route advance, and \( \delta \) is the information functions. When \( X > X_0 \), the logistics vehicle will go to the next material demand point. The formula can be expressed as follows:

\[
z_{m}^i = \begin{cases} 
\sum_{i=1}^{\beta_j} \Xi_j, & t \in T \\
0, & t \in T
\end{cases} (2)
\]

Through the above formula, the order of material demand points can be set, and the obtained route is the current optimal solution. Combined with the above process cycle calculation, the correctness of the optimal solution can be verified.

2.3 Reference multi objective fitness function to build multi objective optimization model of railway transportation logistics network

Set the optimized route as the form of particles to complete the construction of the optimization model. After the path particles are encoded, their fitness can not get the optimal result in the program according to the objective function. It is necessary to adjust the fitness function [8-9] to make it optimal. The multi objective optimization problem is expressed as follows:

\[
y(r^*) = [y_1(r^*), y_2(r^*), \cdots, y_n(r^*)] (3)
\]
In the formula, $y(r')$ represents multidimensional objective function and $r' \in p^i$, $i$ is decision variable. According to the above content analysis, assuming that there are $a$ vectors $O_1, O_2, \cdots, O_a$ in the calculation process, randomly select a particle to check its corresponding coding, carry out cyclic iteration, and identify multiple factors mentioned above. If the corresponding value of the factor represented by $r'_a$ is 1, then add the factor to $O_1$, if the corresponding value is equal to 2, then add the element to $O_2$, and so on in order Row derivation placement, arranging the components of $N$ vectors in the above order.

Considering that the fitness value calculated by each factor is different, the fitness function with smaller value is in a disadvantageous position. Therefore, the balance coefficient $\zeta$ is introduced for processing, and the balance coefficient is determined by the analysis of the value magnitude of the function in the actual calculation example. Then the calculation formula of multi objective fitness function for railway transportation is as follows:

$$fitness(r) = \zeta \times y_n(r') \quad (4)$$

Through the above calculation, the fitness function corresponding to the particle is obtained, but the particle has different advantages and disadvantages. For the poor particles, help them get rid of the bad position as soon as possible, and speed up the development to other areas; for the high-quality particles, strengthen the development of their corresponding areas, and excavate better solutions. The fitness function is adjusted by the mean fitness of particle population. The mean value of population fitness is calculated as follows:

$$fitness(x) = \frac{\sum_{i=1}^{n} y(r')}{q} \quad (5)$$

In the formula, $q$ represents the population size, $\alpha$ is a constant and represents the number of key factors. If the result is $fitness(r'_a) > fitness(r'_a)$, it means that the particle is in a better position, close to the optimal solution, so reduce the search step size [10] of this part of particles, avoid the step size is too large, and miss the optimal solution, adjust the step size by reducing the speed limit of the increase of inertia weight. The formula is as follows:
If $\text{fitness}(r_i) < \text{fitness}(r_j)$, it shows that the particle is in a worse position and far from the optimal solution. By increasing the search step size of the particle, the search is avoided to be too slow. The method of adjusting step size is to increase inertia weight and reduce speed limitation. The formula is as follows:

$$
\begin{cases}
\text{fitness}(r_i) = (3_{\min} - 3) \cdot \frac{\text{fitness}(r_i)}{\text{fitness}(r_j)} \cdot S_{\max} \\
S_{\max} = 1 + \frac{1}{S_0}
\end{cases}
$$

In equations 6 and 7, $3$ represents the inertia weight, $3_{\min}$ represents the minimum inertia weight, $3_{\max}$ represents the maximum inertia weight, $S_0$ represents the initial velocity of particles, and $S_{\max}$ represents the maximum velocity of particles. Through the above process, the optimal fitness function is calculated, and the optimal solution of railway logistics is obtained by using this function. The above formula is used as the multi objective optimization model of railway transportation logistics network.

3 Experimental demonstration and analysis

Through the above part, the construction process of the multi objective optimization model of railway transportation logistics network based on the improved ant colony algorithm is completed. In order to ensure that the design model in this paper has the ability to optimize the railway transportation network, the use effect of the design model in this paper is obtained by means of comparison.

3.1 Design of experimental platform

In order to ensure the consistency of the experimental process and the authenticity of the experimental results, the experimental platform is set to transmit and control the experimental process.

Hardware environment configuration:
Gddr7 SSD, 1556mhz graphics card, embedded CPU, dual power socket, LED LCD.

Software environment configuration:

The experiment uses SQL 2016 software to complete data storage, C++ programming language, and 32-bit Windows 8.0 operating system. The specific experimental environment is as follows.

![Experimental environment](image)

**Fig. 4. Experimental environment**

Using the above hardware and software to build the experimental environment, the experimental object is set as the network transportation cost of the optimized model and the original optimized model.

In this experiment, we use the form of railway transportation logistics to complete the experimental analysis process. It is set that there are 20 railway logistics points in a certain area, each of which has a construction cost of 100000 yuan and a railway transportation cost of 500 yuan per kilometer. The original model and the optimization model are used to develop the logistics network design, and the optimal route is obtained. The cost calculation is carried out for the design results of the two models, and the cost differences are compared.

**3.2 Analysis of experimental results**

Through the above design, the experimental process is completed, and the specific experimental results are as follows.
Through the above experimental results, we can see that the logistics network designed by the optimization model is better than that designed by the original model. The logistics network node location designed by the optimization model in this paper is more representative and more scientific. The original model is more random in the location of logistics network nodes, which causes the problem of increasing logistics transportation distance. It is known...
that for each kilometer increase, the logistics transportation cost increases by 500 yuan, and the transportation cost of the original model design result increases by 500 kilometers compared with the design optimization model design result in this paper, then the transportation cost of the original model design result increases by 250000 yuan compared with the design model in this paper. Using the original model results in a significant increase in transportation costs. In conclusion, the optimization effect of the design optimization model in this paper is better. Using the design model in this paper can effectively reduce the transportation costs of the railway transportation network.

4 Conclusion

The multi objective optimization problem generally exists in the railway transportation logistics network. In this paper, the multi objective optimization problem in this field is studied. Based on the theory of supply chain management, logistics planning and design, the research system of railway logistics network optimization is summarized. The basic problem model of railway transportation logistics network multi objective optimization is constructed; the logistics path algorithm of improved ant colony algorithm is designed, and the multi objective optimization problem model is solved by combining particle swarm algorithm. The effectiveness and feasibility of the algorithm are verified by case simulation. The results show that the model and algorithm are in line with the reality of the development of logistics enterprises and play an important role in reducing the cost of railway logistics transportation.

Reference


Hierarchical planning control model of smart city based on urban micro community

WEN Wen*, LI Yan-dong, CHENG Shu

{wenwen8779@163.com, sdfds8574@163.com, saefceas874@163.com}

1. Yaha School of Built Environment, Haikou University of Economics, Haikou, 571127, China;
2. Haikou University of Economics, Haikou, 571127, China;
3. Haikou University of Economics, Haikou, 571127, China

Abstract: In view of the low control intensity of the traditional intelligent urban spatial hierarchical planning control model, which leads to the problem that the control effect is not obvious, a kind of intelligent urban spatial hierarchical planning control model based on urban micro-community is proposed. First of all, using the "urban microcommunity theory" to calculate the spatial hierarchical planning index, construct a hierarchical matrix, determine the intensity parameters of the control model, define the spatial stratification index as the search to meet a series of equality, inequality constraints, establish a mathematical control model, and complete the design of the intelligent urban spatial hierarchical planning control model based on the urban micro community. The experimental results show that compared with the traditional control model, the intelligent urban spatial hierarchical planning control model based on urban microcommunity has the highest control intensity grade and the best control effect.

Keywords: urban micro community; intelligent city; spatial stratification; planning control model;

1 Introduction

The concept of "smart city", which emerged 10 years ago, is a concept of how to apply communication and information technology to improve urban function, enhance urban efficiency, enhance competitiveness, and provide us with a new concept of urban development to solve the problems of urban equity, poverty, social collapse and environmental pollution [1]. Smart cities are often described as digitally simulating the shape and operation of real cities, using the combination of various infrastructure and electronic devices, and connecting through the network to provide real-time urban dynamics, and to achieve equal and safe quality of life through data integration and analysis and feedback. In the future information and communication technology and urban research, more attention will inevitably be paid to how to serve these facilities in cities, citizens and various social organizations, and urban planning as the overall planning of urban operation will be more and more closely linked to the emergence of intelligent cities [2].

At present, our country does not attach importance to the integration of urban design
related concepts in the process of regulation compilation, but to the mandatory content in the control index, and despise the guiding content of urban design to the control index. At present, urban design is the best optimization method of urban space and urban form, the main purpose of which is to transform two-dimensional plane into the control of urban three-dimensional space and urban environment \[3\]. In order to solve the contradiction between the control and execution of the control regulations, we should combine the relevant concepts of urban design, start from the essence of the control rules, optimize the control system of land use intensity, and make the control regulations have a systematic control over the development of urban land \[4\].

At present, how to integrate urban design into the statutory regulation system, especially how to achieve remarkable results in shaping the characteristics of urban space environment, and effectively guide the establishment of land use intensity index more scientific and reasonable, has become a hot topic in academic research \[5\]. The control of land use intensity is the core content of control regulation, social, economic and policy all belong to the influencing factors of land use intensity, but these factors focus on the feasibility of implementation and the purpose of implementation effect, attach importance to the management and control of land use intensity index to the "quantity" of space, lack the shaping of urban space environment, and lack the guidance of "quality" of space. Urban design is to shape the physical and spatial environment and characteristics of the city for the purpose of making up for the current regulations in urban aesthetics, urban environment. It cannot only create a pleasant and livable urban overall space environment, but also ensure and achieve the macro-level planning objectives such as urban master planning and zoning planning.

2 Control Model of Intelligent Urban Spatial Stratified Planning Based on Urban Microcommunity
2.1 Using "Urban Microcommunity Theory" to Calculate Spatial Stratified Planning Index

Whether it is human life or production activities, it needs the support of resources and environment, but also by the constraints of resources and environment. In combination with the actual situation, the resources and environment support mainly considers the land resources, the water resources support index, the restraint aspect mainly considers the geological environment restraint index. In calculation, the per capita available land area is expressed by the ratio of usable land area to population, that is, the per capita usable land area is two usable land area / rural population, which can be used by land area = slope less than 25° area = slope less than 25° one extremely important and important ecological space area \[6\].
When calculating the planning index, it mainly measures the quantity and quality of the ecological space of the evaluation unit. The ecological protection index uses the results of ecological importance evaluation, in quantity, mainly considers that the extremely important ecological space area accounts for the single element proportion of the evaluation and the important ecological space area accounts for the proportion of the evaluation unit quality separate.

There are often different units between indicators, and there are great differences in values, which are not comparable and cannot be directly calculated. The standardization of indicators is to quantify non-comparative indicators into equivalent values in the same way, and to carry out operational analysis. In the index calculation, the stratified potential index represents the "wisdom" degree of urban space by the average of the three indicators of population density, urban economy and traffic superiority. The formula is as follows:

\[
P_1 = \sqrt[3]{\frac{1}{3} \sum_{i} a_i^2} \tag{1}
\]

Where \( P_1 \) denotes the degree of wisdom and \( a_i \) denotes the population density. The ratio of the weighted average of available land resources \( b_1 \) and water resources \( b_2 \) to the geological environment index \( b_3 \) is used to express the comprehensive support and constraints.

\[
P_2 = \sqrt[2]{\frac{\frac{1}{2} (b_1^2 + b_2^2)}{b_3}} \tag{2}
\]

Where, \( P_2 \) denotes the comprehensive support constraint coefficient. The ecological conservation index is used to calculate the quantity and quality of urban space.

\[
P_3 = 3\alpha + 2\beta \tag{3}
\]

Among them, \( \alpha \) and \( \beta \) denote the proportion of extremely important ecological space and the proportion of important ecological space. Finally, the method for calculating the hierarchical indicators is shown in the following table:
Table 1 Method of calculating stratified index

<table>
<thead>
<tr>
<th>Criterion level</th>
<th>Computing method</th>
<th>Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development potential index</td>
<td>$P_1 = \sqrt{\frac{1}{3} \sum_i a_i^2}$</td>
<td>Population density urban Economy Traffic advantage</td>
</tr>
<tr>
<td>Resource environment support and constraint index</td>
<td>$P_2 = \sqrt{\frac{1}{2} (b_1^2 + b_2^2)}$</td>
<td>Available land resources b_1 Available water resources b_3 Geological environment constraints</td>
</tr>
<tr>
<td>Ecological protection index</td>
<td>$P_3 = 3\alpha + 2\beta$</td>
<td>Proportion of important ecological environment Proportion of important urban space</td>
</tr>
</tbody>
</table>

Using the above calculations, the results of the indicators are shown in table 2 below:

Table 2 Calculation results of indicators

<table>
<thead>
<tr>
<th>Code</th>
<th>Potential index</th>
<th>Constraint index</th>
<th>Protection index</th>
<th>Evaluation value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.862</td>
<td>0.359</td>
<td>1.586</td>
<td>-0.856</td>
</tr>
<tr>
<td>2</td>
<td>0.775</td>
<td>0.425</td>
<td>0.965</td>
<td>-0.862</td>
</tr>
<tr>
<td>3</td>
<td>0.751</td>
<td>0.368</td>
<td>0.149</td>
<td>-0.582</td>
</tr>
<tr>
<td>4</td>
<td>0.865</td>
<td>0.458</td>
<td>0.865</td>
<td>-0.956</td>
</tr>
<tr>
<td>5</td>
<td>0.695</td>
<td>0.562</td>
<td>0.741</td>
<td>-1.535</td>
</tr>
<tr>
<td>6</td>
<td>0.965</td>
<td>0.258</td>
<td>0.625</td>
<td>-0.753</td>
</tr>
<tr>
<td>7</td>
<td>0.856</td>
<td>0.684</td>
<td>0.965</td>
<td>-0.635</td>
</tr>
<tr>
<td>8</td>
<td>0.742</td>
<td>0.631</td>
<td>0.598</td>
<td>-0.452</td>
</tr>
<tr>
<td>9</td>
<td>0.842</td>
<td>0.965</td>
<td>0.475</td>
<td>-0.692</td>
</tr>
<tr>
<td>10</td>
<td>1.268</td>
<td>0.589</td>
<td>0.465</td>
<td>-0.652</td>
</tr>
<tr>
<td>11</td>
<td>1.297</td>
<td>0.287</td>
<td>0.685</td>
<td>-0.485</td>
</tr>
</tbody>
</table>

The calculated results from the various indices calculated from the above table show that
the larger the potential index is between 0.2 and 2.8, the greater the value indicates that the better the programmable potential is, and the greater the value of the constraint index is between 0.132 and 1.000, the stronger the supporting effect of resource and environment conditions on production and life, whereas the smaller the value indicates that the resource and environment support is weaker or even overloaded [7]. The more important the ecological conservation index values are distributed at 0.393-1. On the other hand, the smaller the value, the lower the proportion of ecological protection space is, the weaker the importance of ecological protection is, but the proportion of important ecological space is not reduced. The value of the ecological protection index is distributed between 0.39 and 1.484, and the larger the value, the higher the proportion of the ecological protection space required by the evaluation unit.

After calculating the spatial hierarchical planning index, the intensity parameters of the control model are determined according to the numerical value, and the hierarchical planning control model is established to realize the intelligent city stratification.

2.2 Determination of control model strength parameters

Before determining the parameters of the control intensity model, calculate the urban spatial index, according to the results of table 2 above. The results are plotted in ascending order, and five inflection points are divided into six intervals. The values of the five inflection points are -0.1014,-0.832,-0.546,-0.224 and 0.368, respectively. For different layered units, using 1 and 2 as the living function space, 3 and 4 as the production function space, and 5 and 6 as the ecological function space, the following line diagram is drawn:

As can be seen from figure 1 above, the intelligent city space contains two kinds of planning control parameters, which are combined (3) the objective function of the optimal planning problem at this time. To construct a hierarchical matrix, assuming that there are \( n \) control elements, the influence on the command urban spatial hierarchical planning \( Z \), determine the proportion of each control factor in the planning, the expression of the control
element can be expressed as:

\[ y = (y_1, y_2, \ldots, y_n) \quad (4) \]

The two factors \( y_i \) and \( y_j \) in the above formula are denoted by \( a_{ij} \), so the final \( n \) control indexes are compared to form a matrix:

\[ A = (a_{ij})_{n \times n} \quad (5) \]

The interlocking formula (4)(5) ultimately constitutes a control positive and inverse matrix, as follows:

\[
A = \begin{bmatrix}
    a_{1_1} & a_{1_2} & \cdots & a_{1_n} \\
    a_{2_1} & a_{2_2} & \cdots & a_{2_n} \\
    \vdots & \vdots & \ddots & \vdots \\
    a_{n_1} & a_{n_2} & \cdots & a_{nn}
\end{bmatrix} \quad (6)
\]

In determining the value of \( a_{ij} \) in the preceding formula, the reference number 1 to 9 is used as a scale. The meaning of the scale is shown in the following table:

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Scale</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>The two factors are of the same importance</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>The former is slightly more important than the latter</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>The former is more important than the latter</td>
</tr>
<tr>
<td>4</td>
<td>7</td>
<td>The former is more important than the latter</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>The former is more important than the latter</td>
</tr>
<tr>
<td>6</td>
<td>2/4/6</td>
<td>Mean value of the above adjacent judgments</td>
</tr>
</tbody>
</table>

The final control model parameters are obtained by selecting the scale corresponding to the ordinal number 5 from the data in the scale in the table above. The formula is as follows:

\[
P_i = \sum_{j=0}^{n} f_i(x)g(x) \quad (7)
\]

Where, \( f_i \) is the \( i \)-th urban maturity. using the final calculated parameters of the control model, the mathematical computational model of spatial hierarchical planning control of mathematical wisdom city is established [8].
2.3 Establish mathematical control model

When the mathematical control model is established, the spatial stratification index is defined as a set of decision variable values of the optimal objective function when the search satisfies a series of equality and inequality constraints \( [9] \). The multi-objective optimization is not to find the optimal solution of a given sub-objective, but to find the optimal solution set satisfying multiple objectives. To establish a minimum objective planning program with a mathematical model of:

\[
\begin{align*}
\min F(x) &= [f_1(x), f_2(x), \ldots, f_k(x)]^T \\
\text{s.t. } &g(x) \geq 0 \\
h(x) &= 0
\end{align*}
\]

Of which, \( \min F(x) \) Represents the minimum optimization target for hierarchical planning control indicators, \( f_i(x) \) For the sub-objectives included in the planning issue, \( i \) Subtarget number, \( i = 1, 2, \ldots, k \), \( k \) Number of sub-targets. The solution vector of multi-objective programming is: \( x = (x_1, x_2, \ldots, x_n)^T, x \in R \), \( R \) Represents the real domain, \( g(x) \) Represents inequality constraints, \( h(x) \) denotes the equality constraint condition. If \( x_0 \) is the optimal solution of the programming control model, the solution for any solution \( x \neq x_0 \), \( f_i(x) \geq f_i(x_0) \), that is, the optimal solution \( x_0 \) ensures that any subtarget in the control model function is the smallest in space.

If \( x_0 \) is a non-inferior solution to a planning problem, there should be no one solution \( x \neq x_0 \), make \( F(x) \leq F(x_0) \), that is, when taking \( x_0 \), at least one or more sub-objectives are minimized, and no other solution makes all sub-objectives in \( f(x) \) superior to that non-inferior solution. In general, multi-objective programming can be translated into single-objective programming for individual sub-objectives:
\[
\begin{align*}
\begin{cases}
\min f_i(x) \\
s, t, g(x) \geq 0 \\
h(x) = 0
\end{cases}
\end{align*}
\] (9)

Where \( f_i(x) \) is the sub-goal. If \( x^i = (x_1, x_2, \ldots, x_n)^T \) is the optimal solution of the sub-objective, then the optimal solution \( x_0 \) of the multi-objective programming is equal to the optimal solution \( x^i \) of each sub-objective.

\[
x_0 = x^1 = \cdots = x^i \quad (10)
\]

In general, each sub-goal restricts each other, and finds a \( x_0 \) that satisfies the requirement of each sub-target taking the minimum value at the same time, selects the optimal solution from it, and establishes an intelligent urban spatial hierarchical planning control model. The final control flow is shown in the following figure:
Fig. 2 Model control flow

By the control flow above, and finally complete the design of the intelligent urban spatial hierarchical planning control model based on the urban microcommunity [10].

3Simulation experiment

3.1 Experiment preparation

The model prepared for the experiment is carried out in a small example of three urban layers, four planning routes, and the layer-distributed network vertices of OD in 36 residential areas. The plan is shown in the following figure:

Fig. 3 Planning of experimental preparation

The experimental preparation for the processing of computer parameters for quantization of urban space is shown in the following table:

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Name</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Operating System</td>
<td>Window7</td>
</tr>
<tr>
<td>2</td>
<td>System type</td>
<td>64 place</td>
</tr>
<tr>
<td>3</td>
<td>Processor</td>
<td>Intel (R) Core (TM) i5-4590</td>
</tr>
<tr>
<td>4</td>
<td>CPU</td>
<td>3.30GHz</td>
</tr>
<tr>
<td>5</td>
<td>Memory</td>
<td>4GB</td>
</tr>
<tr>
<td>6</td>
<td>Hard disk capacity</td>
<td>500G</td>
</tr>
<tr>
<td>7</td>
<td>Graphics card</td>
<td>Integrated graphics</td>
</tr>
</tbody>
</table>

Based on the above preparation experiment preparation, two traditional planning control models and the intelligent urban spatial hierarchical planning control model based on the urban micro community are used respectively to carry out experiments, and the quantitative results of the three control models are compared.

3.2 Experimental result
Three control models are required to control the urban space of the same area, and the final results of the three control models are shown in the following figure:

(a) Experimental results of traditional control model 1

(b) Experimental results of traditional control model 2
(c) Experimental results of intelligent urban spatial hierarchical planning control model based on urban microcommunity

Fig. 4 Experimental results of three kinds of control models

As shown in the experimental results shown in the figure above, the traditional control model shows that the intensity level of control model is 1, the intensity level is low, the intensity level of traditional control model is 2, the intensity level is higher, and the control level of intelligent urban spatial hierarchical planning control model based on urban microcommunity is 3, the control intensity is higher, and the final control effect is the most obvious, which is suitable for practical use.

4Conclusion

Intelligent urban spatial stratification planning is one of the components of urban planning, which runs through the whole process of urban planning. Starting from the vertical relationship of urban design itself, urban design is divided into overall urban design, zoning urban design and detailed urban design. Starting from the subordinate relationship between urban design and urban planning system, urban design is divided into "separately compiled urban design" and "urban design that runs through all stages of urban planning ". The expression of urban design should be integrated with the requirements of urban planning, corresponding to the planning content of each stage of urban planning.

5Fund projects

supported by the philosophy and Social Sciences Planning of Hainan Province, Project No.: HNSK(YB) 18-52

University Project No.: hjky18-13
Reference
Research on Real-time Mining of New Energy Vehicle Fault Diagnosis Data under Cloud Computing

WEI Jia-tong\textsuperscript{1}, XU Yan\textsuperscript{2}, KONG Chun-hua \textsuperscript{3}
\{ouyangmeixue100@163.com\textsuperscript{1}, xuyan464803398@163.com\textsuperscript{2}, 13596157766@163.com\textsuperscript{3}\}

(1. Department of Automotive Engineering, Jilin Communications Polytechnic, Changchun 130012, China; 2. Department of Automotive Engineering, Jilin Communications Polytechnic, Changchun 130012, China; 3. Department of Automotive Engineering, Jilin Communications Polytechnic, Changchun 130012, China)

Abstract. In order to effectively improve the semantic retrieval ability and information analysis ability of network database, and improve the accuracy of data mining of new energy vehicle fault diagnosis, a real-time data mining method of new energy vehicle fault diagnosis under cloud computing is proposed. The modern signal mining technology is used to analyze the fault diagnosis data of new energy vehicles, and a data signal analysis model is established. On this basis, the fault diagnosis data of each section of new energy vehicles are segmented matched and filtered, and the characteristic input of the fault diagnosis data of new energy vehicles is obtained, combining the optimal classification surface of the fault diagnosis data characteristics of new energy vehicles and the fault of new energy vehicles. The feature vector track of the fault diagnosis data can realize the extraction of the feature value of the fault diagnosis data of new energy vehicles, and complete the real-time mining of the fault diagnosis data of new energy vehicles under the cloud computing. Experimental results show that the proposed method has high accuracy and real-time performance.

Keywords: cloud computing; new energy vehicles; fault diagnosis; data mining;

1 Introduction

The development of information technology has led to the explosive growth of information production scale and transmission speed, which has brought human social life and scientific research into the era of cloud computing. Research shows that the amount of data generated by fault diagnosis of new energy vehicles in recent years is more than the total data created by the whole traffic history, and this trend is still accelerating. In the era of cloud computing, thanks to the continuous progress of information technology, every process of human life can be recorded in the form of data and displayed in different ways. Therefore, these massive data have the characteristics of heterogeneous types, low value density, huge data model and fast propagation speed [1]. Traditional data analysis and processing technologies have been unable to meet the requirements of cloud computing analysis. How to design data mining and analysis technologies that meet the characteristics of cloud computing,
explore the hidden laws and knowledge in cloud computing, and enhance the timeliness of
decision-making services are the key contents of current research in the field of data mining
and analysis.

At present, one of the main contradictions of data analysis and data processing is that the
increase of data scale and transmission speed cannot be synchronized with the increase of
computing power, that is to say, the requirements of analyzing and processing large-scale
dynamic fault diagnosis data of new energy vehicles exceed the current computing power,
which seriously affects the effectiveness of decision-making services. The traditional methods
and models of fault diagnosis data analysis of static new energy vehicles have been unable to
meet the processing requirements of large-scale dynamic new energy vehicle fault diagnosis
data [2]. In the application of sensor network monitoring, sensors in network nodes transmit
large-scale fault diagnosis data of new energy vehicles to the central processor in real time.
The monitoring system needs to make timely and accurate statistical analysis on the feedback
data of sensors, and make corresponding judgments and decisions. In social network analysis,
social network media such as Tweet, Facebook and so on are available at all times Massive
social information release and reception. Among them, public opinion monitoring is an
important research content. The real-time analysis of public opinion of social hot spots, key
events and issues can accurately grasp the context of the event development, predict the
direction of the event development, and timely make reasonable disposal. In the management
and analysis of fault diagnosis data of new energy vehicles, the real-time analysis of massive
fault diagnosis data of new energy vehicles can quickly and accurately mine out various
modes of fault diagnosis of new energy vehicles, and timely discover new energy vehicle
faults [3]. In the new energy vehicle market, various activities produce large-scale new energy
vehicle fault diagnosis data in real time, and these data have the characteristics of wide
coverage, rich content, strong real-time, and most of them are unstructured text data, such as
stock index data recording the fluctuation of new energy vehicle stock price, and related
financial news information. The new energy vehicle market is driven by information, and the
change of fault diagnosis information of new energy vehicles will be reflected in the
fluctuation of new energy vehicle market in varying degrees. Therefore, the relevant analysis
of large-scale new energy vehicle fault diagnosis data can effectively discover the operation
law of new energy vehicle market, and the relevant new energy vehicle fault diagnosis
information and new energy The connection between the automobile market.

The fault diagnosis data of new energy vehicles generated in the above application fields
have the characteristics of massive, high-speed and dynamic, and the timeliness of the data is
strong. It needs the relevant analysis system to make a quick response, accurately mine the
hidden laws and knowledge, and make a decision in time [4]. Therefore, according to the characteristics of these application areas, researchers put forward a new data model, the new energy vehicle fault diagnosis data flow model, which can meet the requirements of real-time and efficient data mining tasks.

The fault diagnosis data stream of new energy vehicles is composed of a series of data which arrive in sequence according to the time axis. It can also be regarded as a digital signal string which is encoded and processed in the process of information transmission. It has the dual attributes of time and space. Its spatial attribute is the same as the static new energy vehicle fault diagnosis data set, which can be considered as the physical meaning expressed by the attribute value of the data generated by the system. Its time attribute is mainly reflected in the arrival order of data. Each stream of data has a matching time index, that is, the order of reaching the analysis system, which is the main difference with the static new energy vehicle fault diagnosis data set. Excellent new energy vehicle fault diagnosis data analysis algorithm should be insensitive to the order of the input new energy vehicle fault diagnosis data when analyzing the static new energy vehicle fault diagnosis data set. When processing the new energy vehicle fault diagnosis data flow pattern, the relevant algorithm must consider the time sequence relationship between the data and mine the new energy vehicle fault diagnosis The evolution law of broken data [5]. The data flow model of new energy vehicle fault diagnosis is similar to the time series model, which has time and space attributes, but the time series data is a kind of static data, while the data flow model of new energy vehicle fault diagnosis represents the dynamic data, which has the following characteristics: the arrival sequence of new energy vehicle fault diagnosis data is independent and uncontrolled; the fault diagnosis data of new energy vehicle When it arrives in real time, the analysis system must respond quickly; the scale of fault diagnosis data of new energy vehicles tends to be infinite, and it can not be saved in memory, in principle, it can only read the data once.

The complex pattern mining on the data stream of new energy vehicle fault diagnosis has practical and theoretical significance. First of all, due to the rapid development of Internet and communication technology, a large number of new energy vehicle fault diagnosis data have been automatically generated in the aspect of new energy vehicle fault diagnosis, and data analysis technology has entered the era of cloud computing. Cloud computing has the characteristics of huge data scale, various data types, low data value density and fast data processing speed. Therefore, data flow mining, which can quickly respond to large-scale fault diagnosis data of new energy vehicles, is an important research content of cloud computing analysis. Secondly, the research fields such as intrusion mining, anomaly analysis, trend monitoring, exploratory analysis are very sensitive to the response speed of new energy
vehicle fault diagnosis data, which needs real-time complex analysis of the data. Finally, due to the characteristics of dynamic new energy vehicle fault diagnosis data, it is impossible to obtain the overall distribution information of data in the process of data mining. Dynamic data mining is a process from local to global, from details to the whole. Therefore, through the analysis of data flow, we can explore the internal evolution mechanism of the system, as well as the evolution rules of various knowledge or patterns hidden in the data flow law [6].

Therefore, this paper analyzes the data flow of new energy vehicle fault diagnosis, and proposes a real-time mining method of new energy vehicle fault diagnosis data under cloud computing. Combined with modern data processing technology to extract the characteristic value of fault diagnosis data of new energy vehicles, the analytical model of fault diagnosis data signal of new energy vehicles is decomposed through segmented pre whitening processing, and the fault diagnosis data of new energy vehicles is matched and mined, so as to extract the characteristic value of fault diagnosis data of new energy vehicles and complete the fault diagnosis of new energy vehicles under cloud computing Data mining in real time. The experimental results show that the proposed method improves the mining accuracy of fault diagnosis data of new energy vehicles, and achieves the purpose of improving the real-time data mining.

2 Real time mining of fault diagnosis data of new energy vehicles under cloud computing

2.1 Data flow relevance

This paper analyzes the data flow density of new energy vehicle fault diagnosis. After the data flow density of new energy vehicle fault diagnosis is determined, the region is divided according to the data flow density of new energy vehicle fault diagnosis. The correlation between the regional network and the data flow of new energy vehicle fault diagnosis is calculated by using the incremental subspace data mining theory, and the node and time relationship of the data flow are analyzed. The number of contacts can accurately determine the nodes of fault diagnosis data flow of new energy vehicles, and complete the real-time mining of fault diagnosis data of new energy vehicles under cloud computing. The specific process is as follows:

Assuming that the number of network types in the cloud computing environment is the number of regions, an undirected ergodic graph $G = (V, E)$ is used to describe the network regions in the whole cloud computing environment, among them, $V$ represents the fault diagnosis node set of new energy vehicles, and $E$ represents the fault diagnosis link set of new energy vehicles, In the cloud computing environment, the network consists of $m$
regions, and  \( V = (v_1, v_2, \cdots, v_n) \) is the number of nodes in the new energy vehicle fault diagnosis network region. Then, the data flow density of the regional network new energy vehicle fault diagnosis can be calculated by formula (1):

\[
\phi(k) = \left( \sum_{i=1}^{n} \frac{M(k)}{N(k)} \right) \left( \frac{M(k)}{nN(k)} \right)^2
\]

Where, \( \phi(k) \) represents the data flow density of fault diagnosis of new energy vehicles in the \( k \) th region, \( M(k) \) represents the number of fault diagnosis data flow of new energy vehicles in the \( k \) th region, and \( N(k) \) represents the number of nodes in the \( k \) th region, \( n \) represents the number of network loops in the cloud computing environment, \( i \) represents the fault diagnosis data flow of new energy vehicles, \( M(p) \) represents the number of fault diagnosis data flow of new energy vehicles in the \( p \) th region, and \( N(p) \) represents the number of nodes in the \( p \) th region.

Assuming that \( \sigma(i) \) represents the size of the new energy vehicle fault diagnosis data stream \( i \) that is known to be mined, and \( D(i) \) represents the feature set of the new energy vehicle fault diagnosis data stream \( i \), the correlation calculation formula of the new energy vehicle fault diagnosis data stream \( i \) and area \( k \) can be obtained as follows:

\[
L(k,i) = \exp \left[ \frac{\sum_{j \neq i} (\phi(u) - \phi(k))}{1 + \alpha} \right] \times D(i)
\]

In the formula, \( \alpha \) represents the correlation degree factor, which is related to the size of the average new energy vehicle fault diagnosis data stream of the regional network, \( \phi(u) \) represents the new energy vehicle fault diagnosis data stream density in the \( u \) region, \( j \)
represents a new energy vehicle fault diagnosis data stream, and \( m \) represents the number of new energy vehicle fault diagnosis data streams in the \( k \) region.

After using the above process to obtain the region of the data flow of fault diagnosis of energy vehicles, the data mining method based on multi incremental space is used to mine the loop of the data flow of fault diagnosis of new energy vehicles, and the correlation coefficient between the node and time of the data flow of fault diagnosis of energy vehicles is obtained, so as to accurately determine the node of the data flow of fault diagnosis of new energy vehicles. Real time mining of fault diagnosis data of new energy vehicles under cloud computing [7].

2.2 Extract eigenvalues of fault diagnosis data of new energy vehicles

In the process of real-time mining of fault diagnosis data of new energy vehicles, firstly, a new energy vehicle fault diagnosis data signal model is established, and the modern signal mining technology is used to analyze the discrete data of new energy vehicle fault diagnosis data signal. Then, a new energy vehicle fault diagnosis data signal analysis model is established under cloud computing, and the new energy vehicle fault diagnosis data in the network under cloud computing are analyzed. The high frequency signal simulation is carried out, and the fault diagnosis data of each new energy vehicle is mined by segmented matching filter. The specific process is as follows:

Firstly, a new energy vehicle fault diagnosis data signal model is constructed, which integrates modern signal mining technology to make the new energy vehicle fault diagnosis data signal discrete data analysis, and establishes the new energy vehicle fault diagnosis data signal analysis model in the cloud computing environment:

\[
\begin{align*}
H_0: x_{k+1}(t) &= r_{k+1}(t) \\
H_1: x_{k+1}(t) &= s(t - \tau) + r_{k+1}(t) \\
0 &\leq t \leq T_B
\end{align*}
\]

(3)

Where, \( x_{k+1}(t) \) represents the fault diagnosis data signal of new energy vehicles, \( x(t) \) represents the real part of the analytical model of the fault diagnosis data signal of new energy vehicles, \( x_{k+1}(t) \) represents the LFM signal when the pulse width of the fault diagnosis data of new energy vehicles is \( T_p \), the maximum pulse width is \( T_B \), \( s(\cdot) \) represents the high-order cumulant, \( t \) represents the sampling time interval of the fault diagnosis data signal of new energy vehicles, and \( \tau \) represents the time change parameter.
Suppose that the fault diagnosis data of network new energy vehicles in the cloud computing environment is time-varying model $AR$, $p$ represents the corresponding order of the model, $a_i(t)$ is the model parameter, and formula (4) is used to represent the interference noise $n(t)$ in the process of fault diagnosis data mining of network new energy vehicles

$$n(t) = -\sum_{i=1}^{p} a_i(t) \cdot n(t) + \sigma(t) \cdot w(t) \quad (4)$$

In the formula, $\sigma(t)$ represents the Gaussian asymmetric function. From the above formula, $w(t)$ is the Gaussian white noise with mean value of 0 and variance of 1. In the cloud computing environment, the network reverberation data is obtained by formula (5):

$$r(t) = an(t) \quad (5)$$

By pre whitening the fault diagnosis data of network new energy vehicles, the analytical model of fault diagnosis data signal of new energy vehicles is decomposed into multiple narrow-band signals, and the matching mining of fault diagnosis data of new energy vehicles is carried out [8], and the interference frequency characteristic formula of fault diagnosis data of new energy vehicles is given by formula (6):

$$f(t) = \frac{1}{2\pi} \times \frac{d\theta(t)}{dt}, \beta \quad (6)$$

In the formula, $\theta(t)$ represents the phase information of fault diagnosis data signal of new energy vehicles, $dt$ represents the interference characteristic amplitude of fault diagnosis data of new energy vehicles, and $\beta$ represents the segment pre whitening matching parameter. In the process of estimating the spectrum of $k$ segment of fault diagnosis data of new energy vehicles, the reverberation spectrum in $k$ segment is estimated by $AR$ model, and the current fault diagnosis data segment of new energy vehicles is smoothed That is:
Where, $w_k(t)$ represents the weight of new energy vehicle fault diagnosis data under cloud computing, which represents a Gaussian white noise with variance of $\sigma_w^2$, $a_{k,i}$ represents the lattice recursive model parameters of the $k$-th segment of new energy vehicle fault diagnosis data, and $p_k$ represents the corresponding order of Lattice recursive model.

Based on the effective recognition results of new energy vehicle fault diagnosis data, the least square method is used to mine the features of new energy vehicle fault diagnosis data, and the feature input of new energy vehicle fault diagnosis data is obtained. Combined with the optimal classification surface and feature vector track of new energy vehicle fault diagnosis data, the feature value extraction of new energy vehicle fault diagnosis data is realized. Based on this, real-time mining of fault diagnosis data of new energy vehicles under cloud computing is completed [9]. The specific process is as follows:

Based on the effective identification results of the fault diagnosis data of new energy vehicles given above, the least square method is used to mine the fault diagnosis data features of new energy vehicles, and the input amount of the fault diagnosis data of new energy vehicles is obtained as follows:

$$E = \sum_{p=1}^{k} E_p = \frac{1}{l} \sum_{p=1}^{k} \sum_{k=1}^{l} \left[ r_p(k) - y_p(k) \right]^2$$

In the formula, $l$ represents the magnitude of fault diagnosis data of new energy vehicles, $E_p$ represents the termination frequency of fault diagnosis data of new energy vehicles, and $r_p(k)$ represents the initial data. According to equation (8), the feature value of fault diagnosis data of new energy vehicles is extracted and the phase space Fourier transform is carried out in the cloud computing environment according to the input value of fault diagnosis data of new energy vehicles. The optimal classification surface of fault diagnosis data of new energy vehicles needs to be calculated. The optimal solution of the
The problem can be represented by $\chi^*$, and the fault diagnosis of new energy vehicles can be obtained by equation (9). The optimal classification plane of fault features:

$$H(x, y) = \sum_{i=1}^{\infty} J_m(t) \ast [f_m + mf, + \eta] \exp \left[ j(m \varphi + \theta + \psi) \right]$$ (9)

In the formula, $J_m(t)$ represents the fault diagnosis data of new energy vehicles at time $t$, $f_m$ represents the decreasing trend function of fault diagnosis data of new energy vehicles, $\eta$ represents the initial classification time of fault diagnosis data of new energy vehicles, $f_s$ represents the carrier frequency of fault diagnosis data of new energy vehicles, $\varphi$ represents the inclination angle of fault diagnosis data input of new energy vehicles, $\theta$ represents the line of sight angle of data, $\psi$ represents the data output. The initial angle of input time, $j$ represents the interval of fault diagnosis data of new energy vehicles. If the characteristic component $z$ of fault diagnosis data of new energy vehicles is set in the radius $T$ region, the optimal classification surface of fault diagnosis data of new energy vehicles under cloud computing should meet the following constraints:

$$j \leq \min \left[ \left( \frac{T^2}{\Delta^2} \right), n \right] + 1$$ (10)

In the formula, $\Delta$ represents the change component of the fault diagnosis feature of new energy vehicles, normalizes the above results, transforms the quadratic programming problem in the support vector machine algorithm into a linear equation for solution from another perspective, the time series of the fault diagnosis data of new energy vehicles is $\{x(t_0 - i\Delta t), i = 0, 1, \cdots N - 1 \}$, and the vector track after the quantification of the fault diagnosis data of new energy vehicles is:

$$X = x_t \left[ x(t_0), x(t_0 + \Delta t), \cdots, x(t_0 + (K - 1)\Delta t) \right]$$ (11)

In the formula, $\Delta t$ represents the sampling time interval of fault diagnosis data of new energy vehicles, $x_n$ represents the feature directivity of fault diagnosis data of new energy vehicles, $x(t_0)$ represents the vector track after quantization of fault diagnosis data of new energy vehicles.
energy vehicles at time \( t_0 \), \( X(t_0 + \Delta t) \) represents the vector track after quantization of fault diagnosis data of new energy vehicles at time \( t_0 + \Delta t \), and \( K \) represents the coefficient of non-linear normalized basic parameter.

Through the above processing process, combined with the optimal classification surface of new energy vehicle fault diagnosis data features and the vector track after the new energy vehicle fault diagnosis data quantification under cloud computing, the new energy vehicle fault diagnosis data feature value extraction is realized, which is shown in formula (12)

\[
z(r) = X \cdot E \cdot H_\psi (\eta, \psi) \quad (12)
\]

In the formula, \( X \) represents the vector trajectory correlation parameter, and regards the above results as the basic fault diagnosis data of new energy vehicles, and combines the deep learning theory to complete the real-time mining of fault diagnosis data of new energy vehicles.

### 2.3 Real time mining algorithm design under cloud computing

According to the Apriori feature, the selection set is pruned to reduce the support count of the extra selection set. In the phase of parallel operation, according to the load balancing characteristics of Hadoop distributed architecture, the unreasonable data division of fault diagnosis of new energy vehicles can be avoided to the greatest extent.

For the new energy vehicle fault diagnosis data file \( D \), assume its scale is \( G \), transaction number is \( N \), the new energy vehicle fault diagnosis data block size of the new energy vehicle fault diagnosis database is \( M \), Hadoop divides the new energy vehicle fault diagnosis data file \( D \) into \( G / M \) block new energy vehicle fault diagnosis data. Here, \( N_{\text{max}} \) is used to describe the maximum number of parallel map tasks in Hadoop distributed file system. When \( G / M > N_{\text{max}} \), the 1-term set with support lower than \( N_{\text{min}} \) is removed, and the fault diagnosis data of new energy vehicles are described in vertical form. The vertical partition method is used for output, and the transaction identifiers of each range are stored in each file, and then the frequent 2-item set is calculated.

Based on the above analysis, a real-time data mining algorithm for fault diagnosis of new energy vehicles based on Hadoop distributed architecture is proposed.

Input: new energy vehicle fault diagnosis data file \( D \), minimum support threshold \( \text{minsup} \);

Output: new energy vehicle fault diagnosis data set.
Step 1: Organize the fault diagnosis data set $G$ of new energy vehicles into $G / M$ small data blocks. All small data block nodes need to be calculated. The calculation method is based on the fault characteristics of new energy vehicles. The output results of all map tasks are temporary files containing the fault diagnosis data of new energy vehicles;

Step 2: Combine the local output results in the map phase to reduce the data exchange volume of new energy vehicle fault diagnosis between nodes. The format of each line is $<$ item, Tid-Set $>$, TidSet is the transaction set containing the project;

Step 3: Merge the temporary files of all calculation nodes into vertical 1-item sets through the reduce function, and calculate the item support degree. If the support degree exceeds the product of $N$ and $\minsup$, it is considered as frequent 1-item set. Split the transaction sets of corresponding items and send them to different types of files. In order to ensure the consistent size of all vertical block data and HDFS data blocks, this section divides the transaction set of all 1-item sets into $G / M$ data blocks, in which the first $G / M$ files cover the transaction volume of $N / G / M$, and the last file covers the transaction volume of $N - N \cdot G / M - 1 / G / M$, so the transaction $i$ is in the $j$ file, where:

$$j = \left\lfloor \frac{i}{N / G / M} \right\rfloor, \quad i \leq N \cdot \frac{G / M - 1}{G / M}$$

(13)

Step 4: $m \leftarrow 2$;

Step 5: Divide the transaction set of the project into several data blocks and transfer them to the map process, so that all processes can calculate the support of candidate $m$-item set in parallel. Through the reduce process, the output results of all map processes are collected together to obtain the global $m$-item set. According to the idea of block, the global frequent $m$-item set is sent to each file;

Step 6: $m \leftarrow m + 1$;

Step 7: Repeat step 5 to step 7 until the fault diagnosis data output of new energy vehicles without more frequent item sets, and the output result is the fault diagnosis data of new energy vehicles.

In the process of parallel mining, the new energy vehicle fault diagnosis data real-time mining algorithm based on Hadoop distributed architecture uses mobile program rather than
parallel mining program. At the same time, through step (2), the mining results of all nodes are gathered together, so as to reduce the traffic and communication cost. The proposed method divides the fault diagnosis data set of new energy vehicles into non-overlapping data blocks, so that the intersection operation can be executed in parallel, greatly improving the real-time performance of fault diagnosis data mining of new energy vehicles [10].

2.4 Real time mining of fault diagnosis data of new energy vehicles

This section conducts correlation analysis on the fault diagnosis data of new energy vehicles, and mines the fault diagnosis data of new energy vehicles. The flow chart is shown in Fig. 1.

![Flow chart of real-time mining of fault diagnosis data of new energy vehicles](image_url)

Fig. 1 flow chart of real-time mining of fault diagnosis data of new energy vehicles

The real-time mining process of fault diagnosis data of new energy vehicles is as follows:

1) Preprocessing of fault diagnosis data of new energy vehicles. Before the correlation analysis, the fault diagnosis data of new energy vehicles are processed, and the fault diagnosis data of new energy vehicles are imported into the database for correlation analysis;

2) Correlation analysis. Association analysis is an effective data mining technology for fault diagnosis of new energy vehicles, which is mainly realized through the following two processes:

   ① Form frequent item set. In this section, the frequent itemsets of new energy vehicle fault diagnosis data are mined by Apriori method, which is an effective method to mine frequent itemsets of Boolean association rules and is widely used.

   ② Form strong association rules. Strong association rules are association rules that meet the minimum confidence. After obtaining frequent item sets, strong association rules are formed through the following process:

   1) For each frequent term set \( q \) with cluster label \( r \), form subset \( v = (q - r) \) of \( q \).
2) If \( \frac{S_N(q)}{S_N(r)} \geq C_{\text{min}} \), strong association rule \( q \rightarrow r \rightarrow r \) will be formed.

Here \( S_N(q) \) and \( S_N(r) \) are used to describe the support numbers of \( q \) and \( r \) in turn; \( C_{\text{min}} \) is used to describe the minimum confidence level.

Each association rule described a certain attribute of the fault diagnosis data of new energy vehicles. All rules can be regarded as references and applied to the real-time mining of fault diagnosis data of new energy vehicles. In the process of mining, support and confidence are calculated according to the rules, and the similarity of association rules is compared, which can realize the real-time mining of fault diagnosis data of new energy vehicles.

Compare rule \( Z_1(A \rightarrow B.S.C_1) \) with rule \( Z_2(A \rightarrow B.S.C_2) \) by the following formula:

\[
w(Z_1, Z_2) = |S_1 - S_2| + |C_1 - C_2| \quad (14)
\]

The larger \( w(Z_1, Z_2) \) value is, the smaller the similarity between strong association rule \( Z_1 \) and \( Z_2 \) is. A threshold \( \delta \) should be set in advance. If \( w(Z_1, Z_2) \geq \delta \), the corresponding data is the fault diagnosis data of new energy vehicles.

3 Comparative analysis of experiments

In order to verify the performance of the proposed real-time mining method for fault diagnosis data of new energy vehicles under cloud computing, experiments are carried out. Matlab simulation software is used to design the data mining algorithm of new energy vehicle fault diagnosis. The experimental test data set is from a large-scale cloud storage database. The number of samples is 2500, and the sampling period is \( T = 0.08s \). In the process of data real-time mining, the intensity of interference between network data codes is SNR = 0-24db, and the base frequency of scalar time series is 200Hz. According to the above experimental environment and parameter settings, the simulation analysis of real-time mining method of feature data is carried out, and the time-domain waveform of original new energy vehicle fault diagnosis data flow is obtained as shown in Fig. 2.
Fig. 2 Time domain waveform of information flow of original new energy vehicle fault diagnosis data

Based on the discrete data feature extraction results of the new energy vehicle fault diagnosis data signal in Fig. 2, the real-time mining performance of different types of new energy vehicle fault diagnosis data sets is measured. New energy vehicle fault diagnosis data set includes three types: data set A (including 602142 pieces of data), data set B (including 495203 pieces of data), data set C (including 584762 pieces of data). Under the condition of false alarm rate $P_f = 0.05$, in the range of $-50 \sim 20$dB, 500 Monte Carlo tests are used to conduct real-time mining simulation of new energy vehicle fault diagnosis data, and the network under cloud computing is obtained. The comparison results of the accurate mining probability (%) of the fault diagnosis data of China new energy vehicle are shown in Fig. 3.
According to the analysis of Figure 3, the mining probability of fault diagnosis data mining method of new energy vehicles based on segmented pre whitening matching mining under cloud computing is high under different types of data sets, which shows that the method in this paper can reasonably configure parameters $\alpha$, $\beta$, effectively improves the probability of mining, improves the accuracy of the new energy vehicle fault diagnosis data identification significantly, shows the better performance of the new energy vehicle fault diagnosis data mining, improves the semantic retrieval ability and information analysis ability of the network database.

Table 1 shows the real-time mining method of fault diagnosis data of new energy vehicles based on segment pre whitening matching mining under cloud computing, and the mining accuracy results under 500 Monte Carlo experiments on three test data sets.

<table>
<thead>
<tr>
<th>Number of experiments</th>
<th>New energy vehicle fault diagnosis data set A</th>
<th>New energy vehicle fault diagnosis data set B</th>
<th>New energy vehicle fault diagnosis data set C</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>92.15</td>
<td>97.27</td>
<td>95.21</td>
</tr>
<tr>
<td>200</td>
<td>93.12</td>
<td>96.89</td>
<td>94.89</td>
</tr>
<tr>
<td>300</td>
<td>91.02</td>
<td>97.65</td>
<td>93.52</td>
</tr>
<tr>
<td>400</td>
<td>90.28</td>
<td>95.41</td>
<td>95.21</td>
</tr>
<tr>
<td>500</td>
<td>93.58</td>
<td>96.87</td>
<td>94.21</td>
</tr>
</tbody>
</table>

From the experimental results in Table 1, the accuracy of the data mining method based on segmented pre whitening matching mining in all test sets is higher than 90%. In order to further verify the effectiveness of this method, the mining accuracy of the traditional method and this method is compared and analyzed, and the comparison results are shown in Figure 4.
According to figure 4, the mining accuracy of the method in this paper is mainly due to the traditional method. In the process of segmented matching filter mining for each section of fault diagnosis data of new energy vehicles, the method realizes the fault diagnosis data of new energy vehicles by combining the optimal classification surface of fault diagnosis data characteristics of new energy vehicles and the feature vector track of fault diagnosis data of new energy vehicles. The extraction of eigenvalues improves the performance and calculation efficiency of the subsequent new energy vehicle fault diagnosis data classification, and reduces the packet loss rate of non-new energy vehicle fault diagnosis data.

4 Concluding remarks

In view of the shortcomings of current data mining methods in dealing with fault diagnosis data of new energy vehicles, a real-time data mining method based on segmented pre-whitening matching mining for fault diagnosis data of new energy vehicles in cloud computing environment is proposed, and the real-time performance and accuracy of the proposed method in dealing with fault diagnosis data of new energy vehicles are proved by simulation. In the process of researching the real-time mining method of new energy vehicle fault diagnosis data under cloud computing, the problem of mining time is not considered. In the next research, mining time is taken as the experimental index to improve the effectiveness of real-time mining of new energy vehicle fault diagnosis data.

5 Fund projects

Research and Innovation Fund of Science and Technology Development Center of Ministry of Education—“Beichuang Assistant” Fund Project: 《Research on Application of New Energy Vehicle Training Platform Intelligent Network Technology》（number: 2018A05024）

Reference

University of Technology. Vol. 40, pp. 60-64 (2018)


Numerical simulation and analysis of heat dissipation characteristics of battery box of electric vehicle based on coupling model

WEI Jia-tong1,KONG Chun-hua2,XU Yan3
{ouyangmeixue100@163.com1,13596157766@163.com2,xuyan464803398@163.com3}

1.Department of Automotive Engineering, Jilin Communications Polytechnic, Changchun 130012, China;
2. Department of Automotive Engineering, Jilin Communications Polytechnic, Changchun 130012, China;
3. Department of Automotive Engineering, Jilin Communications Polytechnic, Changchun 130012, China

Abstract: In order to reduce the heat loss of battery in the process of electric vehicle moving, a numerical simulation analysis method based on coupling model is proposed. According to the actual consumption demand of the heat dissipation structure of the battery box, the center difference value level of the heat dissipation coefficient is calculated, and then the heat dissipation coefficient of the battery box of the electric vehicle is processed according to the separation principle of the boundary heat dissipation authority. On this basis, a standardized simulation grid is established by combining the characteristic coupling numerical value, and the comprehensive simulation processing of the thermal characteristic numerical value is realized by calculating the angle deflection of the thermal domain of the battery, and the successful application of the numerical simulation analysis method of the thermal characteristic of the battery box of the electric vehicle based on the coupling model is completed. The results show that, compared with the conventional numerical processing, the DSI and PSI coefficients of the new numerical simulation method show a significant downward trend, and the heat loss of the battery in the process of electric vehicle moving is effectively controlled.

Keywords: coupling model; electric vehicle; heat dissipation characteristics; numerical simulation; consumption demand; center difference value; simulation grid; heat dissipation domain deflection;

1 Introduction

In electronics, coupling refers to the transfer of energy from one circuit part to another. For example, through electrical conductivity coupling, energy is transmitted from a voltage source to the load. The AC (Alternating Current) part and DC (Direct current) part of the circuit can be coupled by using the properties of the capacitor that allows the AC part to pass through and block the DC part. The transformer can also act as the coupling medium, and the appropriate impedance matching can be achieved by configuring the appropriate impedance at both ends. In practical engineering, temperature field, gravitational field and humidity field are...
all physical fields, and many of the problems we need to solve are the superposition of these physical fields, because these physical fields interact with each other. For example, the temperature has an effect on the stress distribution. The problem of superposition of multiple physical fields is called multi field coupling problem, which is also a kind of coupling. Simply put, in software engineering, the coupling degree between objects is the dependency between objects [1]. The main problem guiding the use and maintenance of objects is the multiple dependencies between objects. The higher the coupling between objects, the higher the maintenance cost. Therefore, the design of objects should minimize the coupling between classes and components.

The heat dissipation technology is derived from the traditional heat dissipation technology, which is based on conduction and convection. In the field of security, the heat dissipation technology is applied to the heat dissipation of security monitoring equipment to improve the working performance and stability of the equipment. The heat dissipated in this way depends on the temperature difference between the body temperature and the contact object, the contact area, and the thermal conductivity of the object in contact with the body to dissipate heat. Combined with the convective heat dissipation technology, the heat from the body is exchanged through the gas flow. The amount of heat dissipated by convection depends on the temperature difference between the fuselage and the surrounding environment and the effective cooling area of the fuselage, which is greatly affected by the wind speed [2]. The greater the wind speed, the more heat dissipation; on the contrary, the smaller the wind speed, the less heat dissipation. In general, the contact surface between the electronic components and the heat sink should be coated with a layer of thermal conductive silicone grease, so that the heat from the components can be more effectively transmitted to the heat sink, and then distributed to the surrounding air through the heat sink. In the conventional simulation scheme, the part close to the front cover of the battery box and the car is changed to copper block, which has the characteristics of fast heat absorption and strong heat conduction ability, so as to quickly a large amount of heat energy generated by the operation of the module is brought to the copper block with nickel plating on the surface, and the copper block and the front cover aluminum block are closely combined to dissipate heat, so that a large amount of heat energy can quickly spread to the aluminum alloy body, and then be taken away by the flow of gas. Combined with the heat dissipation technology of the battery, an open space is provided for the heat loss of the battery module, so as to ensure that the equipment has enough space to realize air convection and effectively The battery box of electric vehicle is ensured to work in the daily environment. The heat dissipation of battery box is a hot research issue in the field of heat dissipation technology. Under the condition of high-speed driving of electric vehicle, a
new numerical simulation analysis method of heat dissipation characteristics of battery box is designed based on the theory of domain angle deflection. The practical application value of this method is highlighted by the way of comparative experiment.

2 Heat dissipation coefficient treatment of battery box of electric vehicle

The treatment of the heat dissipation coefficient of the battery box of the electric vehicle consists of three parts: the analysis of the heat dissipation structure of the battery box, the calculation of the center difference of the heat dissipation coefficient and the separation of the boundary heat dissipation authority. The specific operation method is as follows.

2.1 Thermal structure analysis of battery box

The working cycle of the electric vehicle engine is carried out under high temperature, and the temperature of the mixture entering the cylinder can reach over 2000 °C. At this time, the piston, cylinder block, cylinder head, valve and other parts of the engine are in contact with the high-temperature combustible mixture and are heated intensively. If the engine is not cooled effectively, its mechanical strength will become worse. At the same time, the inflation coefficient of the cooling structure will decrease, resulting in the imbalance of air-fuel ratio and abnormal combustion of the engine. However, the high temperature in the battery box will cause the mixture to burn early (burn in advance), resulting in the knock phenomenon that seriously damages the electric vehicle [3]. Too high temperature will also cause the lubricating oil to burn and deteriorate. At high temperature, it will reduce the gap in the heat dissipation structure, damage the protection of the oil film, reduce the lubricating capacity, and even cause adhesion wear and seizure (cylinder pulling) fault.

Fig. 1 Heat dissipation structure of battery box

2.2 Calculation of center difference value of heat dissipation coefficient

The central difference value is the necessary performance condition of the heat
dissipation structure of the battery box, because the heat dissipation coefficient of the battery box of the electric vehicle not only retains the Taylor series information in the first-order upwind, but also retains the second term of the Taylor series, so the precision of the central difference value is higher than that of the first-order upwind format. The first-order upwind scheme considers that the value on the upstream cell control point is equal to the value of the local cell boundary point. The second-order upwind scheme considers that the value of the boundary point of the local element is equal to the sum of the value and increment of the control point of the upstream element[4-5]. The second-order upwind scheme is commonly used to solve vehicle flow field. In order to obtain the second-order accuracy of the flow field variables on the boundary point, the control point of the local grid element can also be taken as the base point. Similarly, the flow field variables are expanded by taylor series and the first two terms are retained. Generally, the value of the second-order upwind scheme is different from that of the boundary point. The value of the flow field variable calculated by the central difference scheme at the boundary point is the arithmetic mean of the two. Let $q_b$ represent the two lower limit accumulation conditions of the heat dissipation coefficient of the battery box, $q_t$ represent the two upper limit accumulation conditions of the heat dissipation coefficient of the battery box, and $\lambda$ represent the control increment of the battery unit. The central difference value of the heat dissipation coefficient can be expressed as:

$$\delta = \frac{1}{\Delta x} \sqrt{\frac{1}{2} \left( \frac{j}{f} \right)^2} \| \beta \| \Delta t \quad (1)$$

Among them, $\gamma$ represents the boundary value of the local cell, $j$ represents the first-order upwind scheme, $f$ represents the second-order upwind scheme, and $\beta$ represents the characterization index of flow field variables in Taylor series.

2.3 Separation and solution of boundary heat dissipation authority

As the name implies, the separation and solution of boundary heat dissipation authority is to deal with different heat dissipation structures of battery box of electric vehicle. Because of the nonlinear characteristics of the control equation, the convergence solution can be obtained only after several iterations. Figure 2 is the solution flow chart. The process is as follows:

1. Update the heat dissipation characteristic value of the battery box of the electric vehicle. In the first calculation, the initial value of the variable is assigned by initialization. In the later calculation process, each iteration gets a new solution.
(2) A new characteristic field strength is obtained by solving the momentum equation from the current mass and pressure flux values of the cell box.

(3) Because the numerical solution of the characteristic field strength can't meet the continuous equation completely, it is necessary to solve the Poisson equation derived from the continuous equation again, and solve the equation to modify the heat flux, heat pressure and driving speed field, so as to meet the continuous equation.

(4) Through the solutions obtained above, the equations of components, energy and turbulence are solved.

(5) If interphase interference is considered in multiphase flow calculation, the source term solution in the continuous phase equation needs to be solved again.

(6) Check whether the convergence condition is met. If so, the calculation is terminated. On the contrary, continue to iterate.

![Flowchart](image)

**Fig. 2** Flow chart of boundary heat dissipation authority separation

### 3 Numerical simulation method of heat dissipation characteristics based on coupled model

Under the support of the treatment condition of the heat dissipation coefficient of the electric vehicle battery box, according to the process of building the numerical simulation grid
of the characteristics coupling, calculating the angle deflection of the magnetic domain of the heat dissipation of the battery, and comprehensively simulating the numerical value of the heat dissipation characteristics, the design of the numerical simulation analysis method of the heat dissipation characteristics of the electric vehicle battery box based on the coupling model is completed.

3.1 Grid construction of characteristic coupled numerical simulation

In the characteristic numerical simulation grid, the numerical simulation of the heat dissipation characteristics of the electric vehicle battery needs to discretize the spatial domain of the calculation range to solve the discretized equations, and the discretization process is based on the grid. In the process of calculation, all kinds of physical information of flow field are stored in the grid nodes. Therefore, the grid has a direct impact on the accuracy of the calculation results. In reality, when the car is running in the air, the surrounding air tends to be infinite, and the simulation can not include all the space, and in the distance from the car, the air disturbance by the car is very small [6]. In the characteristic numerical simulation grid, the heat sink is the same as the battery electron, and can only transfer from one direction to another. This trend will never be reversed throughout the analysis. In the application of the characteristic devices of the electric vehicle battery radiator, the external detection coil is generally used to realize the sensing of magnetic signal, and the analysis of the influence of the device structure, effective magnetic field distribution, material permeability and other factors on the sensing results is simple. In the process of numerical analysis of self-sensing characteristics, there is no external detection coil, so it is necessary to extract the magnetic sensing signal from the excitation signal of the excitation coil. How to reduce the interference of the excitation signal to the magnetic sensing signal is the first choice; and the sensed battery signal is the change of the magnetic field in three parts, including giant magnetostrictive material, nylon framework and excitation coil, in addition to optimized implementation. In the optimized structure, besides the uniformity of the magnetic field inside the device, the influence of the uneven distribution of the magnetic field inside the giant magnetostrictive material and the different values of the permeability of the material on the magnetic sensing results should also be considered [7]. Therefore, it is necessary to establish a precise mathematical sensing relationship including actuator structure and magnetic field optimization, effective magnetic field calculation in the actuator and the influence of permeability of giant magnetostrictive materials.
3.2 Calculation of domain angle deflection for cell heat dissipation

In the process of domain deflection of battery cooling materials, the effective path of domain deflection is relatively complex, which mainly depends on the very large anisotropic properties of the materials. In addition to the magnetocrystalline anisotropic energy of the electric vehicle itself, the application of stress field will also produce a stress anisotropic property in the materials, which will make the deflection process of the domain under the load More complex [8]. Under the action of stress coupling model and heat dissipation load, the change of free energy in each part of the material will make the domain angle deflect. In the study of the domain deflection model of giant magnetostrictive materials, table 1 lists a group of typical constraints between the heat dissipation coefficient and the domain angle deflection.

<table>
<thead>
<tr>
<th>Heat dissipation coefficient of battery</th>
<th>Domain angle deflection</th>
<th>Restrictive relationship</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_1$</td>
<td></td>
<td>Inhibition</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td></td>
<td>Promote</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$d$</td>
<td>Inhibition</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td></td>
<td>Promote</td>
</tr>
<tr>
<td>$p$</td>
<td>$w$</td>
<td>Promote</td>
</tr>
<tr>
<td>In the above table, the heat dissipation coefficient of the battery from top to bottom successively represents the lower limit characteristic value, upper limit characteristic value, minimum coupling error, maximum coupling error, heat dissipation coefficient of the battery</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
box and the traveling parameters of the electric vehicle. With the support of the above physical quantities, the joint formula (1) can express the calculation result of the angle deflection of the magnetic domain of the battery heat dissipation as follows:

\[
\delta = \frac{\int_{t_1}^{t_2} (p - tt')^2 dt}{\sqrt{\int_{A} W \cdot \|l + l'\| dl}}
\]  

(2)

Among them, \( t \) represents the thermal load condition of the battery box of the electric vehicle, \( t' \) represents the numerical supplementary authority index of \( t \), \( l \) represents the primary deflection coefficient of the thermal domain angle, and \( l' \) represents the numerical supplementary condition of \( l \).

### 3.3 Comprehensive simulation of heat dissipation characteristics

From the point of view of the structural behavior of deflection from the perspective of the thermal domain of the battery, the excitation magnetic field provided to the giant magnetostrictive material is controlled by adjusting the magnitude of the excitation current in the battery box, so as to drive the electric vehicle to generate the mechanical stress, in which the mechanical structure will bear the application of the preload stress and the output of the mechanical stress, thus affecting the actual application effect of the coupling model. The actuator of battery box just uses the magnetostrictive mechanism of giant magnetostrictive material itself to realize the conversion of magnetic energy to mechanical energy, and outputs the mechanical stress and displacement through the car top bar [9]. Considering the working characteristics of giant magnetostrictive materials, the actuator of the battery box will sometimes be equipped with a heat dissipation magnetic field, which is mainly provided by characteristic bias coil or coupling device to eliminate the "dual frequency characteristics" of the electric vehicle itself; the automobile top bar is mainly used to provide a prepressing stress required for the work of materials with heat dissipation characteristics, so as to make the materials work in a better strain area, so as to improve the energy conversion efficiency of the battery box actuator; the cooling mechanism is mainly used for the temperature rise caused by the internal loss of the battery box actuator, so as to reduce the impact of the change of heat dissipation temperature on the performance of the actuator, including forced water cooling and phase change Temperature control and semiconductor refrigeration. The specific processing flow is shown in Figure 4.
The self-sensing application of the battery box actuator is an organic collection of execution and perception. The driving and characteristic sensing functions of the electric vehicle are displayed in the material in real time. The self-sensing application is completed through the coupling model analysis and characteristic numerical simulation of the material. Therefore, the coupling of heat dissipation materials and the identification and extraction of electromagnetic parameters will be the key to realize the application of material self-sensing [10]. Among them, the coupling model of heat dissipation material and the application model of battery box actuator are the premise of realizing self-sensing application. In terms of material coupling model, the internal coupling relationship between input heat dissipation parameters and output characteristic values is established, and the change of state parameters under different loads is the basis of defining the coupling relationship model, which is applicable to the constitutive relationship establishment of self-sensing application. It is the foundation of actuator application model research; In the aspect of the application model of battery box actuator, the mathematical establishment of heat dissipation characteristics and the perception of coupling parameters will be the key to the application of actuator perception. The optimal design of heat dissipation characteristics and the mathematical calculation of accurate indicators directly affect the perception accuracy of the model.

Let $f$ represent the coupling coefficient of the heat dissipation characteristic material,
and $\tilde{k}$ represent the disposal conditions of the application model of the battery box actuator. The combined formula (2) can express the comprehensive simulation results of the heat dissipation characteristic values as follows:

$$c = \sqrt{\frac{\sum_{n=1}^{k} (f + f')^2 \hat{g}}{3k \beta / \lambda |\hat{x}|}}$$

(3)

Among them, $n$ represents the lower deflection coefficient of the characteristic heat dissipation value of the electric vehicle battery heat sink, $b$ represents the upper deflection coefficient of the characteristic heat dissipation value of the electric vehicle battery heat sink, $f'$ represents the opposite characteristic number of the coupling coefficient, $\hat{g}$ represents the comprehensive simulation authority, $\beta$ represents the simulation cycle of implementing the heat dissipation characteristic value, $\lambda$ represents the heat dissipation index vector of the electric vehicle battery box, $\hat{x}$ represents the heat dissipation characteristic. The average simulation condition of the numerical value.

4 Simulation analysis and verification

In order to verify the effectiveness of the numerical simulation analysis method based on the coupling model, the following comparative experiments are designed. Two electric vehicle battery heat sinks with the same model and output effect are selected as the experimental objects. The heat loss behavior of the heat sink in the experimental group is monitored by the characteristic numerical simulation analysis method, and the heat loss behavior of the control group is monitored by the conventional numerical processing method. The specific changes of the relevant index parameters are recorded in the established experimental time.

4.1 Establishment of simulation inspection environment

By means of manual intervention, the numerical monitoring methods applied in the experimental group and the control group were changed. With the aid of transmission wire, the recorded index parameters were fed back to the core simulation host.
4.2 DSI coefficient

DSI coefficient has a positive effect on the heat loss of electric vehicle battery, that is, the higher the value level of the former, the greater the total consumption value of the latter, and vice versa. The following table reflects the specific changes of DSI coefficient in the experimental group and the control group within the experimental time of 50min.
According to table 2, during the first 35 minutes of the experiment, the DSI coefficient of the experimental group kept a declining trend. After reaching the minimum value of 0.31, it appeared a stable state for 10 minutes, and the global maximum value was only 0.48.

According to table 3, DSI coefficient of the control group remained stable in the first 20 minutes of the experiment. From the 25th minute, there was a step-by-step rising trend. The global maximum value reached 0.74, far higher than the maximum value of 0.48 in the experiment group.

### 4.3 PSI coefficient

The PSI coefficient also has a positive effect on the heat loss of electric vehicle battery,
that is, the higher the value level of the former, the greater the total consumption value of the latter, and vice versa. The figure below shows the specific changes of PSI coefficient in the experimental group and the control group within the experimental time of 50min.

![Comparison of PSI coefficient](image)

**Fig. 7 Comparison of PSI coefficient**

It can be seen from Figure 7 that with the increase of experimental time, the PSI coefficient in the experimental group basically keeps a relatively stable trend of change, with the global maximum value only reaching 0.61; the PSI coefficient in the control group keeps a state of change of decreasing first and then rising, with the global maximum value reaching 0.98, far higher than the value level of the experimental group.

**5 Concluding remarks**

Due to the high psi coefficient of the traditional numerical simulation analysis method of the heat dissipation characteristics of the battery box of electric vehicle, the heat dissipation loss of the battery in the process of the electric vehicle moving is high. Therefore, the numerical simulation analysis method of the heat dissipation characteristics of the battery box of electric vehicle based on the coupling model is proposed. Under the support of the coupling model, the numerical simulation analysis method of the thermal characteristics of the battery box of the electric vehicle deals with the comprehensive simulation behavior of the thermal characteristics according to the calculation results of the central difference value, and then obtains the accurate value of the magnetic domain angle deflection. From the practical point of view, DSI coefficient and PSI coefficient show a significant downward trend, which plays a
very strong control role in reducing the heat loss of battery in the process of electric vehicle moving. In order to further enhance the heat dissipation characteristics of the electric vehicle battery box, the next research will focus on the heat dissipation characteristics of the electric vehicle battery box under different arrangements.

6 Fund projects
Research and Innovation Fund of Science and Technology Development Center of Ministry of Education—"Beichuang Assistant" Fund Project : 《Research on Application of New Energy Vehicle Training Platform Intelligent Network Technology》（number: 2018A05024）

Reference
A Constellation Diagram Design Method with Minimum Average Energy: An Idea from Structural Chemistry

Yan Li
{yan98li@yeah.net}
College of Communications Engineering, Army Engineering University, Nanjing 210007, China,

Abstract. Based on the theory of crystal structure in chemistry, introducing a method to optimize constellation diagram by means of maximizing space utilization ratio (SUR) when the Euclidean distance between signal points is constant. After studying the average system energy (ASE) and SER of the method and N-QAM system at different points, it is found that the ASE of the method is only 80% to 90% of that of the latter, and the SER of system based on it is lower when their ASE are the same. These results show the superiority of the arrangement method in reducing system energy or SER.

Keywords: Space Utilization Ratio (SUR), Average System Energy (ASE), SER, Hexagonal Rules

1 Introduction

The intuitive representation of signal Constellation Diagram is the modulation state of signal points.[1] The anti-interference ability and energy of communication system are closely related to the arrangement of signal points in its constellation. We often judge the performance of Constellation Diagram by SER and ASE. Generally speaking, the large Euclidean distance between adjacent signal points means that they have less interaction on each other under the influence of noise. So while keeping this distance constant, we can adjust the position of each signal point so that they can be as close as possible to the origin of coordinates, thus having a smaller Average System Energy.

Space utilization is a concept in structural chemistry. It can judge the efficiency of the arrangement of various elementary particles through theoretical analysis. Some arrangements take up more space, while others make good use of space. From this inspiration, we use this concept to explore how to plan the arrangement of signal points in the Constellation Diagrams so that they have the highest energy efficiency, thus reducing the average energy of the system while maintaining the distance between signal points.

QAM has been widely used because of its simple modulation, but the Spatial Utilization Ratio of the signal point arrangement still has a larger space to improve.

The design schemes independently discovered in this paper are partly similar to those in reference [1], but both have the purpose of reducing system energy or improving anti-jamming capability. Reference [1] introduced a efficient 3-D Constellation Diagram model and examplified how to divide a 3-D Constellation Diagram to several 2-D Constellation
Diagrams. But did not provide a practical algorithm for calculating 2-D Constellation Diagram. This paper will introduce the principle and method of designing low-energy Constellation Diagram from a new design point of view and concept which is called Spacial Utilization Ratio and give the algorithm of Constellation Diagram design for given points. During the research, we found that it is a good model. Unlike the general literature, it guides our design from the geometric level. At the same time, it has guiding significance for the further interpretation and optimization of the 3-D Constellation Diagram in reference 1. I also hope that more readers can put forward their own ideas, communicate with each other and make progress together. The contributions of this paper are mainly reflected in the following aspects:

- Introduction of SUR to optimize the constellation structure from the geometric level.
- A general algorithm for constellation design based on the principle of minimum energy is presented.
- Exemplify the Hexagonal Arrangement has the highest SUR so that has the lowest ASE.

In this paper, we first introduce the concept of space utilization and explore how to improve it by making the arrangement of signal points more compact without reducing the distance between adjacent points, which analyses the performance of common arrangement modes at the geometric level. Then the hexagonal rule is proposed according to the arrangement method with the maximum space utilization rate, which is the fundamental basis for the constellation design in this paper. Finally, we perform performance analysis. By comparing the method in this paper with the average energy of QAM system under the condition of the same Euclidean distance between adjacent signal points, the superiority of the method in reducing the average energy of QAM system is more quantitatively and intuitively reflected.

2 Related Work

People have done a lot of work in the design of Constellation Diagrams in the past. At present, the most common Constellation Diagrams are rectangular QAM Constellation Diagram and cross QAM Constellation Diagram.[2,7] QAM modulation is a very popular modulation method in digital communication. Its efficiency and SER performance are good. At present, QAM modulation is widely used. Its constellation is square. However, the energy of the signal points at each corner of the system is relatively large, which will affect the average energy and maximum energy of the system.[3] So sometimes the points in the corner of the improved QAM constellation are moved closer to the origin of the coordinates to reduce ASE.

In the improvement scheme of Constellation Diagram, the expansion method of Constellation Diagram has also attracted the attention of some researchers. The constraints of this method are to expand constellation points outward without reducing the Euclidean distance between constellation points. It achieves a peak-to-average power ratio about 1dB lower than traditional OFDM.[6]

We can see that almost all of above methods are based on QAM or QPSK. The model proposed in this paper makes the design of constellation diagram have a new space for improvement under the viewpoint of these authors. I believe that future authors can make new progress on the basis of the concept of this paper.
Some people also want to expand the Constellation Diagram to 3-D or even higher dimensions to design.[4,5,8] This is a very creative idea. In our analysis, we can think that this method folds a planar Constellation Diagram so that the signal points are closer to the origin of the coordinates, thus greatly reducing the ASE. For example, in a 3-D Constellation Diagram, the signal points are arranged according to the hexagonal densest accumulation of crystals in chemistry to achieve the maximum space utilization. Of course, the higher the dimension, the higher the degree of folding, and the lower the ASE. However, the geometric properties of high-dimensional constellations intersect with those of planar constellations. Because high dimensional Constellation Diagram can be divided into several lower-dimensional and eventually decomposed into twodimensional Constellation Diagram for analysis.

3 SPACE UTILIZATION RATIO

Firstly, we might as well set the minimum distance between adjacent signal points in a constellation as. Then, we can think of the "field" of each signal as a circular region with its center and radius. Then our task is to put as many disjoint circles as possible in a limited area. Next, we analyze two representative methods of ranking.

![Fig. 1. Two Arrangement Methods.](image)

(a) (b)

Shown above, in the Square Arrangement in Fig. 1(a), each signal point has four points adjacent to it, which belongs to the square arrangement (N-point QAM is exactly this arrangement). In the Hexagonal Arrangement in Fig. 1(b), each signal point has six nearest adjacent points. The stacking pattern in these two graphs is very similar to that in chemistry. Chemical knowledge tells us that hexagonal arrangement has a higher space utilization rate. Now we will prove it by calculating the space utilization rate of the two arrangements. We can see from the smallest area unit that:
\[ \eta_1 = \frac{4 \cdot \frac{1}{2} \pi A^2}{4(2A)^2} \]  

(1)

In the Hexagonal Arrangement:

\[ \eta_2 = \frac{3 \cdot \frac{1}{2} \pi A^2}{A \cdot \sqrt{3} A} \]  

(2)

According to formula (1) and (2):

\[ \frac{\eta_1}{\eta_2} = \frac{\sqrt{3}}{2} \approx 0.8760 \]  

(3)

At this time, we have come to the conclusion that hexagonal arrangement does have higher space utilization. In addition, this arrangement is the most space-efficient of all the arrangements on the plane. It may be called Planar Closest, hereinafter referred to as PLC. Because a circle’s outer regular polygon satisfies the hexagon with the largest number of edges that can be mosaic in the plane, and the area of the outer regular polygon will gradually decrease and approach the area of the circle with the increase of the number of edges.

Similarly, the Spatial Utilization Ratio of permutations in three or higher dimensions can still be achieved using the above methods. And each dimension will have one of the most space-efficient solutions. For example, under 3-D conditions, the Hexagonal Closest Packed, which is shown in Fig.2, has the highest Spatial Utilization Ratio rate (about 74%).

Fig. 2. Hexagonal Closest Packet.

4 Hexagonal Rules

Starting from the hexagonal arrangement, we can arrange all the signal points in the intersection points as shown in the Fig 3. When we select a point on the plane of the
graph as the center point (not necessarily a black point), the signal points occupy the nearest point to the center first, thus ensuring the lowest energy.

Generally speaking, for the choice of the center point, we first briefly analyze: when using the most intensive arrangement, the closer the center of gravity is to the center point, the lower the energy is, which is very similar to the change of inertia of rigid body when choosing different fulcrums. For the convenience of discussion, we set the coordinates of one of the black points to be zero. Four methods of selecting the center point are recommended here. As shown in the Table 1.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Figure</th>
<th>The coordinates of central point</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>· · ·</td>
<td>(0,0)</td>
</tr>
<tr>
<td>II</td>
<td>· · ·</td>
<td>(A,0)</td>
</tr>
</tbody>
</table>
For a given numSER of signal points, we first calculate the average energy $E_1 (N)$, $E_2 (N)$, $E_3 (N)$, $E_4 (N)$ of the four permutation methods respectively. Then select the lowest energy arrangement scheme, whose energy is $\min \{E_1 (N), E_2 (N), E_3 (N), E_4 (N)\}$. Generally speaking, the algorithms for calculating energy and confirming schemes are as follows:

1) Input the number of signal points.
2) Enter enough coordinates of possible signal points on the coordinate plane.
3) For the above four models, the coordinates of the possible signal points are arranged in ascending order according to the distance from the center point, and the first $N$ coordinate points are selected as the occupied coordinate points.
4) The Average System Energy of the above four methods is calculated by dividing the energy of the first $N$ points by the number $N$.
5) The serial number of the lowest energy scheme and its corresponding energy value can be output, and the coordinates of $N$ signal points can also be output according to the need.

For the above algorithm, the amount of manual calculation will be relatively large. It is suggested that the program be implemented according to the above algorithm.

For example, here we give the Constellation Diagram at 32 and 64 points in Fig. 4.
At this point, you may find that the contour of the whole Constellation Diagram is somewhat circular. Indeed, the more points, the closer the contour of the PLC Constellation Diagram is to the circle. For example, when the number of points is 1024, the contour of the constellation in Fig. 5 is very close to the circle, which is the inevitable result of making the signal point as close as possible to the origin of the coordinate.

When the number of points $N$ is large, the relative error of any two of the four schemes will be very small (the number will be smaller as the number of points increases), and there are:
\[
\lim_{N \to \infty} \frac{E_{N-PLC}}{E_{N-QAM}} = 1
\]  

In addition, we sometimes need to follow the above general method of Constellation Diagram planning based on the actual situation and not necessarily follow the above. For example, we can plan the Constellation Diagram in a square area like a square QAM. In this way, we ensure that each row has the same number of signal points, just as the QAM Constellation Diagram is staggered between rows and rows, and then refer to the equilateral triangular grid for design. In this way, while improving the Space Utilization Rate, it can also take the QAM’s Gray code characteristics into account. Here we show the Constellation Diagram of Square QAM and Square PLC over 64 points in Fig 6. If you are careful enough, you will find that the red points and blue points in the square PLC Constellation Diagram can be obtained by stretching the vertical axes of the two QAM Constellation Diagrams. Therefore, when analyzing or implementing the system, you can refer to QAM and you can get it with a slight change.

![Fig. 6. Constellation Diagram of Square QAM and Square PLC over 64 points](image)

Besides, the Performance Analysis of Constellation Diagrams below are based on generally arranged model. The square constellations or other variants are described only, and no further analysis is provided below.

5 Performance Analysis

For the following energy analysis and SER analysis, we set the distance between adjacent signal points is 2.
5.1 Energy Analysis

We still keep the distance between the adjacent signal points equal to 2. Let’s first observe the average energy changes of N-QAM and N-PLC from 2 to 2048, as shown in Fig. 7.

We find that the average energy of the two schemes increases approximately linearly as the number of points increases. And the average energy of PLC is lower than that of N-QAM, which preliminarily reflects the effect of hexagonal arrangement method to reduce the Average System Energy. In addition, the energy estimation function at N point is given here (where the distance between signal points is set as 2A)

\[ E_{QAM}(N) \approx 0.6367NA^2 \]  
\[ E_{PLC}(N) \approx 0.5513NA^2 \]

We can see from (5) and (6) that:

\[ \frac{E_{PLC}}{E_{QAM}} = \frac{\sqrt{3}}{2} \approx 0.8660 \]

More specifically, the variation of the ASE ratio of N-PLC and N-QAM systems in the range of 2 to 2048 is given in Fig.8.
When there are few signal points, the graph fluctuates sharply because the characteristic of arrangements is not obvious. But the arrangement model becomes more obvious with the increase of the number of signal points, the ratio of the ASE of the two schemes tends to be stable after 130 points, which is about 0.8660, that is, the ratio of their SUR.

**Table 2.** Performance comparison when the number of signal points is 2 integer power.

<table>
<thead>
<tr>
<th>Bit Number</th>
<th>N-PLC</th>
<th>N-QAM</th>
<th>ASE Ratio(N-PLC/N-QAM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.000</td>
<td>2.000</td>
<td>0.500 0</td>
</tr>
<tr>
<td>2</td>
<td>2.000</td>
<td>2.000</td>
<td>1.000 0</td>
</tr>
<tr>
<td>3</td>
<td>4.500</td>
<td>6.000</td>
<td>0.750 0</td>
</tr>
<tr>
<td>4</td>
<td>9.000</td>
<td>10.000</td>
<td>0.900 0</td>
</tr>
<tr>
<td>5</td>
<td>17.625</td>
<td>20.000</td>
<td>0.881 3</td>
</tr>
<tr>
<td>6</td>
<td>35.250</td>
<td>41.000</td>
<td>0.859 8</td>
</tr>
<tr>
<td>7</td>
<td>70.562</td>
<td>81.750</td>
<td>0.863 1</td>
</tr>
</tbody>
</table>
In practical applications, we often take the positive integer power of 2 as the number of points to design the communication system. In this paper, the ASE change table in the case of N power of 2 is given in Table 2.

Though Energy comparison, we find that the design of Constellation Diagram of communication system by using hexagonal arrangement method can ensure that the distance between signal points remains unchanged, so that the system has lower ASE. In other words, if the ASE is the same, the hexagonal arrangement can make the allowance between signal points more specific, thus improving the anti-noise performance.

5.2 SER analysis under AWGN

Let the independent random variables X and Y have the same variance $\sigma^2$. If X and Y obey the two-dimensional normal distribution, their probability density function is:

$$f(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2 + y^2}{2\sigma^2}}$$  \hspace{1cm} (8)

And more, if the Energy of AWGN is $E = \sigma^2_x + \sigma^2_y = 2\sigma^2$, the probability density function can also be described as:

$$\varphi(x, y) = \frac{1}{\pi E} e^{-\frac{x^2 + y^2}{E}}$$ \hspace{1cm} (9)

Set the distance between adjacent signal points is 2A. If each received signal point is decided to the nearest standard signal point, we can think that the decision range of each signal point is the circumscribed regular polygon of the circle with the radius of A and the signal point as the center. For example, the Decision Area of Square Arrangement and Hexagonal Arrangement, surrounded by red curves, is shown in Fig. 9.
Let the decision area is \( C \), then the SER is the probability of \((X,Y)\) appears out of \( C \) and it can be calculated by

\[
SER = 1 - \int_{\mathbb{R}^2} \frac{1}{\pi E} e^{-\frac{x^2 + y^2}{E}} \, dx \, dy
\]  

(10)

For the following discussion, let the distance between adjacent signal points of Square QAM is 2. Slightly enlarge the distance between adjacent signal points in PLC, which has lower ASE, to make the Energy of them the same. According to equation (5) and (6).

\[
0.5513 \cdot A_{PLC}^2 = 0.6367 \cdot 1^2 \Rightarrow A_{PLC} \approx 1.0747
\]  

(11)

In above case compare their SER according to equation (10), the SER variation is shown in Fig.12.

![Fig. 9. Decision Area of Square Arrangement and Hexagonal Arrangement](image1)

![Fig. 10. SER Variation of PLC and QAM under AWGN](image2)
We can see that as the noise power becomes small, the SER becomes lower, and the SER of PLC is lower than SER of QAM. Besides, the optimization effect of PLC is more obvious in low noise power case.

6 Summary and Outlook

In this paper, the Constellation Diagram of communication system is designed to improve the SUR. It improves the power utilization efficiency and reduces the ASE or SER. And with the increase of the number of points, the energy optimization effect tends to be stable with the SER ratio approach to 0.8660. Besides, the SER of PLC is lower when ASE are the same. The automatic Constellation Diagram design algorithm in this paper provides a solution for readers to sort out ideas and solve the problem of large amount of calculation when the number of points is high, and fills the blank of the existing literature in this respect. Besides, if we slightly enlarge the distance of signal points in PLC to make its energy equals to QAM, the PLC model has lower SER.

The conceptual approach and method in this paper provides a new idea for the research and design of Constellation Diagrams. At the same time, they are of great significance to the research and design of high-dimensional Constellation Diagrams. Although the technology in this paper fully embodies the advantages of reducing the average energy of the system, there are only four schemes for choosing the center point, so that we can get a good but not the best center point. Next, we intend to study the precise selection of the optimal center point algorithm and the design of high-dimensional Constellation Diagram, so as to make the theory more perfect.

References

Location of feature points in 3D reconstruction of multi vision color image based on principal component analysis

FENG Huan-ping1, ZHANG Li-wei2
{aini2125@tom.com1, lihaixia09@sina.com2}
(1. Heibei Institute Of Communication College, shijiazhuang 051430, China; 2. Heibei Institute Of Communication College, shijiazhuang 051430, China)

Abstract: Traditional image feature point location methods, due to the existence of calculation errors, lead to the accuracy of the location of feature points decreased, so based on principal component analysis, a new multi vision color image 3D reconstruction feature point location method is proposed. In this method, the gray level of color image is transformed by principal component analysis, and the color features of the image are obtained according to the gray level difference of local color areas; According to the changing characteristics of environment light coefficient and diffuse reflection light coefficient, the control parameters affecting the three-dimensional light shadow effect of multi vision color image are set; The least square method is used to eliminate the position error of feature points, and according to the confidence degree of association rules between nodes, more accurate feature point positioning results are obtained. Experimental results show that the accuracy of the proposed method is 11.27% higher than that of the traditional method. Therefore, the proposed method is more suitable for feature point location.

Keywords: Keywords principal component analysis; multi vision color image 3D reconstruction; feature point location;

1 Introduction

At this stage, with the continuous innovation of network technology, the sharpness of image is getting higher and higher, and the effect of light and shadow is getting better and better. More and more people use image to store information, and use three-dimensional technology to pursue image visual effect and experience. However, due to the large number of colors in the color image and the small difference between colors, 3D reconstruction is more difficult. At the same time, because the figures and animals in the picture will show different movements, the plants have extremely irregular outlines, and the solid objects such as
buildings and vehicles will present complex structures, which will also increase the difficulty of 3D reconstruction. In addition, the scene in the figure above will show different light and shadow effects in different environments due to the influence of light, so it is necessary to accurately locate these feature points to achieve the accurate reconstruction of the three-dimensional image\(^1\).

In reference \(^1\), a feature-based 3D reconstruction method based on visual slam anti depth filtering is proposed. A more real-time scene structure is gradually established by using video sequence. The key frame tracking method based on motion model is used to provide accurate relative attitude relationship. Map points are no longer directly calculated by two frame triangulation, but by using the depth inverse filter based on probability distribution By accumulating and updating the multi frame information, a back-end hybrid optimization framework composed of feature and direct method and a point selection strategy based on adjustment constraints are obtained. This strategy can accurately and effectively solve the camera attitude and structure. However, due to the calculation error of this method, the final positioning result is not accurate enough, and the reconstructed three-dimensional image does not have the stereoscopic feeling of multi vision.

Therefore, aiming at the problems of traditional methods, a location method based on principal component analysis is proposed. In this method, principal component analysis is used to obtain more accurate color features of the image and realize the accurate positioning of feature points.

2 Feature point location method for 3D reconstruction of multi vision color image based on principal component analysis

2.1 Obtaining color features of image by principal component analysis

To locate the feature points in 3D reconstruction of multi vision color image, the first step is to obtain the color features of image by principal component analysis. The first step of this method is gray-scale transformation of multi vision color 3D image. In principal component analysis, gray-scale transformation is a very basic image processing method in spatial domain. According to the transformation relationship of gray-scale transformation, the gray value of each pixel in the image is transformed, so as to change the dynamic range of gray-scale of the image, and finally get the method of processing the image. It can enlarge the dynamic range of the image, expand the contrast of the image, make the image clearer and more obvious\(^2\). The grayscale transformation formula is as follows:

\[
 f(x, y) = D[f(x, y)] \tag{1}
\]

Formula: \(D\) represents the gray transformation function, representing the transformation
relationship between the input gray value and the output gray value; \( x, y \) represents the horizontal and vertical coordinates of the local gray area; \( f(x, y) \) represents the result of gray-scale transformation. According to the different forms of transformation function, gray-scale transformation can be divided into linear transformation, piecewise linear transformation, nonlinear transformation, and other gray-scale transformation. Because of the segmented linear feature of 3D image, segmented linear transformation is used to preprocess the gray value. The image inversion transform function is shown in Figure 1.

**Figure 1** image inversion transform function

Image inversion transform function image inversion is simply to make black white, make white black, turn the gray value of the original image over, so that the gray value of the output image decreases with the increase of the gray level of the input image. This processing is particularly effective for enhancing white or gray details embedded in a dark background, especially when the amount of color difference in the image is very small. According to the image inversion transformation relationship shown in Figure 1, combined with the tangent form of the linear equation, when there is \( k = 1, b = d - 1 \), whose expression is as follows:

\[
g(x, y) = kf(x, y + b) = f(x, y + d - 1) \quad (2)
\]

Formula: \( g(x, y) \) represents image reversal function; \( k \) is the slope; \( b \) represents a constant change. On the basis of image inversion, the linear gray-scale transformation is
carried out, which transforms each pixel in the image into another pixel value by using the linear relationship, that is, processing each pixel, transforming the image pixel range to the specified range, improving the multi-visual effect of the three-dimensional color image[3]. In the actual operation, it is assumed that the gray range of the image before and after transformation has been given, as shown in Figure 2 below.

![Figure 2 Schematic diagram of linear gray scale transformation](image)

It is known that the range of gray scale \( f(x, y) \) of the original image is \([u, v]\), The range of gray scale \( g(x, y) \) of the transformed image is \([s, w]\). Therefore, the following linear transformation is used to realize gray-scale transformation:

\[
g(x, y) = \frac{w-s}{v-u} \left[ f(x, y) - u \right] + s \quad (3)
\]

According to the above formula, set the single color range under a certain brightness value in the input image to \([u, v]\). Through proportional linear gray-scale transformation, the gray-scale of each pixel in the image is stretched linearly, which effectively improves the visual effect of the image[4]. On the basis of the above transformation, the gray scale of the image is divided into two or more intervals, and each interval is transformed linearly to achieve the purpose of segmentation. The following figure 3 is the result of segmented transformation.
With the piecewise linear transformation method, the gray level needed in the image can be stretched and the gray level not used in the image can be compressed at the same time. The mathematical expression is as follows:

\[
g(x, y) = \begin{cases} 
\frac{s}{u} f(x, y), & 0 \leq f(x, y) < u \\
\frac{w-s}{v-u} \left[ f(x, y) - u \right] + s, & 0 \leq f(x, y) < v \\
\frac{f-w}{g-v} \left[ f(x, y) - v \right] + w, & v \leq f(x, y) < g 
\end{cases}
\]  

(4)

According to the above content, the color difference enhancement of 3D color image is completed\cite{5}. Establish RGB color cube, as shown in Figure 4 below, and obtain image color features by principal component analysis.
According to the gray value, the three-dimensional image without obvious color difference is re-segmented. The color difference characteristic value of the image can be described by the following algorithm:

\[
\begin{align*}
R &= Y + 1.041V \\
G &= Y - 0.347U - 0.721V \\
B &= Y + 1.782U
\end{align*}
\]  

In formula, \(Y, U, V\) represents the color space value after segmented gray-scale conversion, and obtains the image color characteristics through the above steps\(^6\).

2.2 Set up the 3D lighting effect of multi vision color image

Based on the acquired color features of the image, the three-dimensional light and shadow effect of multi vision color image is set up. The illumination model is applied to the 3D image reconstruction algorithm, so it is necessary to set the illumination attributes of the 3D image, including four parameters: ambient light coefficient, diffuse light coefficient, specular light coefficient and specular index.

According to the above program setting parameters, analyze the influence of the changes of ambient light coefficient and diffuse light coefficient on the image effect. Figure 5 is a schematic diagram of the light and shadow effect of the image under different ambient light coefficients.
According to the above diagram of light and shadow effect change, the brightness of the reconstructed image increases gradually with the increase of the ambient light coefficient. However, when the brightness is too large, the inner details of the image cannot be distinguished. When the ambient light coefficient is set to 0.4, the reconstructed color image has the best 3D visual effect. In the same way, obtain the light shadow effect of the change of diffuse light coefficient on the image. After analysis, when the diffuse light coefficient is 0.9, the image reconstruction effect is the best\(^7\). According to the above analysis results, set the sampling parameters and reconstruction period. The specific figures are shown in Table 1.

**Table 1** matching table of sampling parameters and reconstruction period

<table>
<thead>
<tr>
<th>Match group</th>
<th>Sampling parameters</th>
<th>Rebuild cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>2048</td>
<td>0.02</td>
</tr>
<tr>
<td>A2</td>
<td>1024</td>
<td>1.1</td>
</tr>
<tr>
<td>A3</td>
<td>512</td>
<td>2.3</td>
</tr>
</tbody>
</table>
According to Table 1, the sampling period increases with the decrease of sampling parameters. This is because the smaller the sampling parameter, the more pixels emitted by 2d pixel points in the plane, and the slower the reconstruction. When the sampling parameter is less than 64, the speed of the increase of sampling time is more obvious. According to the above analysis results, the control parameters of 3D light and shadow effect of multi-vision color images were set by using the illumination model. Since the surface of the object in the image will reflect light with different intensity, the intensity formula of diffuse reflected light at a certain point in the image is as follows:

\[ p \mu = p' \]  \hspace{1cm} (6)

Formula: \( p' \) represents the intensity of ambient light; \( \mu \) represents nodes at different locations; \( i_1 \) represents the diffuse reflection coefficient of ambient light at the node; \( p \) represents the light intensity after the diffuse reflector interacts with the ambient light. As far as an ideal diffuse reflector is concerned, there are reflection rays of equal intensity in all directions, but the intensity of the light on the surface of the object also depends on the intensity and direction of the incident ray. This phenomenon is described quantitatively by Lambert's law. Suppose that the intensity of diffuse light is directly proportional to the cosine of the incident angle when a certain direction of light is irradiated on the Lambert mirror, then the control parameters of image shadow effect are obtained:
\[
\sigma_a = p \ln(e + \sqrt{\frac{\sin \beta^2 - 1}{e \cos \beta}})
\]  

(7)

Formula: \( \sigma_a \) represents the control parameter with control intensity of \( a \); \( e \) represents the base value of the exponential function; \( \beta \) represents the incidence angle of the light source, whose value ranges from 0° to 90°; \( \varepsilon \) represents the light intensity index generated by the interaction reflection between the diffuse reflector and the incident light in a certain direction. Through the above process, the control parameters of 3D light and shadow effect of multi-vision color images are set[8].

2.3 Eliminating error to realize the feature point location of 3D reconstruction

According to the set light and shadow effect control parameters, the least square method is used to set the position coordinates of the image feature points to achieve the accurate positioning of the reconstructed 3D image. Assuming that the center of the feature point is \( c \) and there are \( n \) anchor nodes from the center. Within the determined defect target range, each coordinate can be represented by \( a_1, a_2, \ldots, a_n \), so there are

\[
a_u = \{(x, y) | (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}\ 	ext{for each anchor node} \]

Set an unknown node as \( a_k \) and its coordinate as \((x_k, y_k)\). Assume that the node is the real coordinate of the feature point, then the relationship between the unknown node and the anchor node is:

\[
\begin{align*}
(x_k - x_1)^2 + (y_k - y_1)^2 &= d_1^2 \\
(x_k - x_2)^2 + (y_k - y_2)^2 &= d_2^2 \\
&\vdots \\
(x_k - x_n)^2 + (y_k - y_n)^2 &= d_n^2
\end{align*}
\]  

(8)

Formula: \( d_1, d_2, \ldots, d_n \) represents the estimated distance between the unknown node and the anchor node. The coordinate boundary obtained by this formula is adjusted according to the control parameters, and the adjusted equation is:

\[
\lambda = s_1 \times \left(2 \times \frac{\sigma_a d}{s_0} + q\right)
\]  

(9)
Formula: \( \lambda \) represents the adjusted value, which is usually between \((0,1)\); \( s_1 \) represents the traversal times of image code phase; \( s_0 \) represents the total spatial step size during image positioning; \( d \) represents the frequency step of image feature data; \( q \) is the non-zero constant of change\(^9\). Figure 6 shows the boundary conditions of the location region of feature points.

\[ \text{Min} \| \lambda x - \omega \|^2 \] is obtained. The specific formula is as follows:

\[ P = \lambda \begin{bmatrix} x' \\ y' \end{bmatrix} \quad (10) \]

Formula: \( x' \) and \( y' \) represent the abscissa and ordinate of the feature points of the reconstructed image; \( P \) represents the feature points. Thus, the location of the feature points in 3D reconstruction of multi-visual color images is realized. The correlation confidence between feature points is calculated as follows:
Formula: \( CL(P) = \frac{\zeta_{\text{support}}(P_1 \cup P_2)}{\zeta_{\text{support}}(P_2)} \) (11)

points; \( P_1 \) and \( P_2 \) represent any two adjacent random feature points in all feature points \( P_n \). When the above calculation results are greater than 0.96, it proves that the positioning is accurate. When it is less than the reference value of the standard, it indicates that there is error in the positioning result, and the deviant data need to be removed and repositioned. Through the above content, based on principal component analysis, the feature point location of 3D reconstruction of multi-vision color image is realized[10].

3 Experiment and analysis

In order to prove the reliability and feasibility of the proposed positioning method, a comparative experiment was proposed for this study. Based on the experimental test results, the accuracy of feature point positioning results was analyzed during 3D reconstruction of multi-vision images. At the same time, the traditional image feature point localization method is introduced to compare the differences between the two methods.

This experiment adopted the evaluation system as the test software, and loaded the software into the experimental test computer, whose operating system is Windows 10, the memory capacity is 8G, the hard disk capacity is 500G, and the browser version is IE11.0, which meets the test requirements of this experiment. A relatively complex photo was randomly selected as the experimental test object, as shown in figure 7 below. Two feature positioning methods were used to reconstruct the image in three dimensions.
Two positioning methods are used to mark the contours and features of objects such as pigeons, ground, trees, grass and buildings. Figure 8 below shows the marking results of pigeons with different flight attitudes.

(a) the pigeon contour mark with attitude A
In this experiment, the test results of the proposed feature point positioning method were taken as experiment group A, while the test results of the traditional feature point positioning method were taken as experiment group B. Test run the experimental test software for 30min, and there are no data anomalies or system hardware problems, so the experiment can start. Figure 9 shows the comparison results of this experiment.

According to the above test results, the positioning accuracy of the proposed positioning
method in group A is above the standard value, and the average accuracy is 95.56%. However, in the positioning results of feature points in group B of the traditional positioning method, 6 groups of data were outside the standard value, and the average positioning accuracy was only 84.29%. Therefore, the feature point positioning method based on principal component analysis can obtain higher accuracy of feature points.

4 Conclusion

In order to solve the problem of low accuracy of traditional positioning methods, a method of locating feature points in 3D reconstruction of multi-vision color images is proposed. Through principal element analysis of image color difference features and image light and shadow effect, accurate positioning of feature points during the reconstruction of 3D image is realized to enhance the 3D visual effect of the image. However, the proposed positioning method, the analysis process is more, the calculation steps are relatively complex, so need to pay special attention. In the future research, some calculation steps can be simplified to improve the positioning efficiency of the method.

References


[8] Zhang Hongyun, Zhao Quanhua, Li Yu.: Geometric feature extraction of scattered targets from remote sensing images based on irregular mark point process. Control and Decision. 34(09):1840-1846 (2019)


Edge tracking method of damaged mural images based on deep learning

ZHANG Li-wei1, FENG Huan-ping2
{lihaixia09@sina.com, aini2125@tom.com}
(1. Hebei Institute Of Communication College, shijiazhuang 051430, China;
2. Hebei Institute Of Communication College, shijiazhuang 051430, China)

Abstract: The traditional edge tracking method of damaged mural images has the phenomenon of high noise variance in the process of tracking, which affects the actual performance of tracking. Therefore, an edge tracking method based on deep learning is proposed. Through image preprocessing, the image gray mean value was unified and image quality was enhanced. Roberts edge detection operator was used to detect image edge features. RCF model in deep learning was used to fuse edge features and output them to find the starting point of fusion features. The test results show that the noise variance of the edge tracking method for damaged mural images designed based on deep learning is between 0.010 and 0.015, which is lower than the noise variance generated in the traditional tracking method, indicating that this method is better than the traditional tracking method.

Keywords: Deep learning; Image edge; Edge tracking;

1 Introduction

As the mural has a long history, the wall is not conducive to the long-term preservation of the mural, but also has the influence of natural disasters and human factors, which has caused various damages, such as fading, discoloration, falling off, etc., so it is important and necessary to protect the mural. When using different methods to repair murals, the edge tracking of damaged murals is a very important step[1-2].

From the perspective of the image boundary tracking implementation method, the traditional boundary tracking methods are mainly divided into two categories: one is based on run length, and the other is based on chain code. The run-length tracking method is used to divide multiple target areas into run-length representations in a row, and analyze the relationship between adjacent runs to obtain the area outline. The chain code-based tracking method is relatively intuitive. It only needs to track on the edges of the image, which is more efficient. However, for multi-region and multi-connected images, the above two methods are prone to excessive noise variance. In reference [3], a target tracking scheduling method based on wireless sensor networks is proposed. Considering the characteristics of wireless sensor...
networks and the requirements of extended clarity of monitoring system functions, the original LCFs algorithm is improved, and a simple and effective target detection and tracking scheme is designed and implemented. A scheduling algorithm (dolb) for the definition optimal load balancing is proposed. However, the noise variance of this method is high in the process of image tracking.

In view of the problems of the above methods, this paper proposes an edge tracking method of damaged mural image based on deep learning. By using the hidden nodes in deep learning and feature learning, the purpose of feature fusion is achieved, and then the purpose of edge tracking of damaged image is solved.

2 Edge tracking method of broken mural image

2.1 Raw image preprocessing

Low-quality images will inevitably be produced due to the effects of damaged mural images due to insufficient light sources during the shooting process. In order to improve the visual effect of such images and the quality of the images, a series of enhancement processes are performed on such images.

Convert the original image into a grayscale histogram and set an initial label for each gray level of the grayscale image histogram in order to find a local maximum, to set the initial label, establish constraints:

\[ y(\alpha) = \begin{cases} 
-1 & x(\alpha) < x(\alpha - 1) \\
1 & x(\alpha) \geq x(\alpha - 1) 
\end{cases} \]

Where \( y(\alpha) \) is the label on the \( \alpha \) gray value, and \( y(\alpha) \) is the number of pixels on the \( \alpha \) gray value.

The above process is that if the total number of pixels on the \( \alpha \)-th gray value is greater than or equal to the total number of pixels on the \( \alpha - 1 \) gray value, the \( a \)-th label is set to 1, otherwise it is set to -1. Set all possible values of \( \alpha \) satisfying \( y(\alpha - 9) = 1, \ldots, y(\alpha + 1) = 1 \) and \( y(\alpha + 1) = -1, \ldots, y(\alpha + 9) = -1 \) conditions as the local maximum we are looking for[4]. As shown in Figure 1.
It can be seen from the figure that there are 4 extreme points that meet the conditions in the histogram, but the difference between the third extreme point and the minimum points on both sides is very small. In order to better enhance the image, these relative points need to be removed. Inconspicuous extreme points, so we set a condition, that is, the extreme point that meets our set conditions is the final optimal segmentation point. The conditions are as follows:

\[
\begin{align*}
\frac{(x(\alpha) - u(\alpha - 1))}{T} & \geq \eta \\
\frac{(x(\alpha) - u(\alpha + 1))}{T} & \geq \eta
\end{align*}
\]

In the above formula, \( x(\alpha) \) indicates the number of pixels of the maximum point, and \( u(\alpha + 1) \) indicates the number of pixels of the interval minimum point. \( T \) represents the size of the gray interval, and the parameter \( \eta \) is the set threshold. The intensity of the false extreme points can be removed by setting the value of \( \eta \).

According to the obtained optimal segmentation value, the gray interval of the original image is divided, and a certain enhanced gray interval is reassigned for each gray interval. The interval mapping function is as follows:

\[
\begin{align*}
Q_i &= h_i - l_i \\
t_i &= \frac{Q_i \cdot \log_{10} \left( \frac{S_i}{Q_i} \right)}{T} \\
T &= \sum_{i=0}^{n} Q_i \cdot \log_{10} \left( \frac{S_i}{Q_i} \right)
\end{align*}
\]
In the above formula, \( Q_i \) represents the range of the gray interval of the \( i \)-th sub-histogram, and \( h_i \) represents the maximum gray value of the gray interval. \( l_i \) represents the minimum grayscale value of the \( i \)-th interval, \( S_i \) is the total number of pixels in the grayscale interval, and \( t_i \) is the size of the redistributed grayscale interval of the \( i \)th sub-histogram[6]. Assume that the gray interval of the sub-histogram of the first output image is \([0, t_i]\), then the range of the \( i \)th gray interval can be obtained by the following function:

\[
\begin{align*}
\text{start}_i &= \sum_{a=0}^{i-1} t_a + 1 \\
\text{end}_i &= \sum_{a=0}^{i} t_a
\end{align*}
\] (4)

\( \text{start}_i \) represents the starting gray value of the \( i \)th gray interval, that is, the minimum gray value; \( \text{end}_i \) represents the maximum gray value of the \( i \)th interval. If \( i = 1 \) is the first interval, then \( \text{start}_1 = 0 \); if \( i \) is the last interval, \( \text{end}_i \) is 255.

In order to further ensure the stability of the gray value of the gray image after the enhancement process, the histogram of each interval is cut to prevent the pixels in the gray image from being too concentrated in a small gray interval Causes excessive enhancement in the gray interval that takes up too much in the mapping process, and processes the histogram of each interval. According to the segmented histogram and segmentation value obtained in the previous step, the shear closure value of each interval is calculated respectively. The shear threshold of each interval can be calculated according to the following formula:

\[
C_i = \frac{1}{\alpha_i - \alpha_{i-1}} \times \sum_{s=\alpha_{i-1}}^{\alpha_i} x(s)
\] (5)

In the formula, \( C_i \) is the calculated threshold for each interval, where \( \alpha_i \) represents the maximum gray value of the \( i \)th interval, \( \alpha_{i-1} \) represents the minimum gray value of the
gray interval, and $x(s)$ represents the total number of pixels in the gray interval. The process of clipping the input gray image histogram is to perform the following processing on the number of pixels at each gray level according to the clipping threshold of each interval. If the number of pixels is less than the clipping threshold, the gray level value unchanged, otherwise change the value of this gray level to the clipping threshold. According to the histogram processed by the segmentation interval and the peak clipping operation, an independent histogram grayscale unification operation is performed on the sub-images of each interval separately to maintain the stability of the gray average value before and after the gray image processing. Assume that the gray value of the input gray image $P(u,v)$ is $\omega_i$. For each pixel of the output image, multiply its gray value by an offset coefficient, that is, do the following processing:

$$P'(x,y) = \omega_i P(u,v) \quad (6)$$

In the formula, $P'(x,y)$ represents the final enhanced image. After the above formula is processed, the gray average value of the output image can be further adjusted to be close to the average gray value of the input image to achieve the purpose of enhancing image quality.

### 2.2 Detect image edge features

In the process of image edge feature detection, the processed image is regarded as a two-dimensional matrix, and the points in the matrix are regarded as sample points of the continuous change function of the gray intensity of the image. The edge of the image is the set of points where the gray value changes significantly. Therefore, the edge of the image can be determined by obtaining the first derivative of the gray intensity function. The magnitude of the gradient is expressed as the intensity of the edge, and the direction of the gradient is perpendicular to the direction of the edge. In general, the gradient magnitude of a continuous image function $f(u,v)$ is expressed as:

$$\text{grad}(f(u,v)) = \left[ \frac{\partial f(u,v)}{\partial x}, \frac{\partial f(u,v)}{\partial y} \right] \quad (7)$$

Since the pixels in the image are discrete, the partial differential in the formula can be approximated by difference. The simple gradient approximation is:
To quote Roberts edge detection operator, calculate the difference of pixels in the diagonal direction in the field of 2 * 2:

\[
\begin{align*}
    f_u &= f[u,v+1] - f[u,v] \\
    f_v &= f[u,v] - f[u+1,v]
\end{align*}
\]  

(8)

The gradient of Roberts edge detection operator in the image \( f(u,v) \) is:

\[
g(f(u,v)) = \max \{ f(u,v) - f(c,s) \}
\]  

(10)

In the formula, \( f(c,s) \) represents the four domain points of the pixel \( (u,v) \). Its convolution template in the \( u \) and \( v \) directions is shown below.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

**Figure 2** Convolution template of edge detection operator

According to these two convolution templates, the Roberts gradient amplitude \( r(u,v) \) can be calculated. An appropriate threshold is set to classify the image pixels. When the amplitude is greater than \( T \), it is an edge feature point. Roberts edge detection operator uses the gray difference between two adjacent pixels on the diagonal to approximate the edge features of the image. There are some breakpoints in the detected edge features. Deep learning technology is used to fuse the edge features of the image.

2.3 Design of image edge feature fusion based on deep learning

The RCF model in deep learning is used to output fusion features. The internal structure of the first stage of the model is shown in the figure below.
Figure 3 internal structure of RCF model

The figure shows the feature fusion mode of the RCF model in phase 1. In the figure, Conv1 1 and Conv1 2 represent the first two convolutional layers of the first stage. Conv1x1-1 and conv1x1-21 respectively represent convolutional layers with convolution kernel size of 1x1 and output channel of 1 and 21[10]. Deconv represents a deconvolution layer, also called a transposed convolution layer, whose role is to upsample features to the original picture size.

Using this model, using a multi-level feature fusion strategy, feature extraction is performed as a level feature for each stage, and it is finally used for the detailed feature compensation of the final prediction output. The fusion of hierarchical features and preliminary classification features, the fusion results are shown in the following figure.

Figure 4 Multi-level feature fusion

The result of the first stage fusion shown in the figure above, the stacked feature fusion is
used in the second stage, so that the fused features can contain more effective detailed information, and the features between the stages are further fused. The stacked structure is shown in Figure 4.

![Figure 4 Stacked Structure](image)

**Figure 5** Phased feature fusion of the stacked structure

In Figure 5, S1 represents the first stage of feature extraction in the model, and S1-edge represents the hierarchical features extracted at this stage. The specific stacking method is: the next stage feature is first stitched all the above stage features, and then a Conv1x1-1 convolution layer is used for feature fusion. Because it is necessary to ensure that all feature sizes are consistent when stitching feature channels, the features of the upper layer must be down-sampled to the same size as the stitched layer before stitching. In order to save the important information of the subject in the original feature during the downsampling, the downsampling method uses the maximum pooling operation to obtain the final fusion feature.

### 2.4 Implement image edge tracking

The edge tracking of the damaged mural image first extracts the starting point in the fused feature. This starting point must be the end point of the extracted edges. The set edge tracking strategy is tracking from small scale to large scale, so the starting point is searched for the fused features at the smallest scale.

Let \( e_i \) be an edge point in the fusion feature at the minimum scale. If there is only one edge point in the neighborhood of \( e_i \), then \( e_i \) is considered as a starting point. Suppose \( e_i \) is a point in the fused feature. Let \( E(e_i) \) be the parent node corresponding to \( e_i \) in the
gradient image. Define a detection operator \( \gamma(e_i) \), if \( e_i \) is a local maximum point, \( \gamma(e_i) = 1 \); otherwise, \( \gamma(e_i) = 0 \).

Define the tracking indicator operator \( r(e_i, e_j) \):
\[
r(e_i, e_j) = \gamma(e_i) \land \gamma(E(e_i)) \land \gamma(e_j) \land \gamma(E(e_j))
\]
(11)

Among them, \( \land \) represents a logical AND operator. If \( r(e_i, e_j) = 1 \), it indicates that the edge point \( e_j \) after the tracking algorithm is selected to be connected to \( e_i \). At this time, there is a path between the edge points \( e_i \) and \( e_j \). Otherwise, point \( e_j \) is treated as a non-edge point and is set to 0. In \( r(e_i, e_j) = 1 \), the path between edge points \( e_i \) and \( e_j \) is shown in the figure below.

![Image edge point tracking](image)

**Figure 6** Image edge point tracking

If the path of two non-adjacent edge points is finally obtained from the above process, the purpose of image edge tracking is achieved. At this time, the edge tracking method of the damaged mural image based on deep learning is designed.

3 Simulation test of edge tracking method of broken mural image

3.1 test environment and data set
The test hardware environment is Intel Xeon CPU E5-2620 2.10GHz, 256DDR4 memory, 1TB solid-state hard disk; the software environment is MATLAB software, Caffe deep learning framework.

In order to verify the actual performance of the designed edge tracking method based on deep learning for broken mural images and the traditional tracking method. A unified SBD data set is adopted, which is an open source data set. This dataset contains more than 20,000 image data, which belong to 20 categories. The original image data in the SBD dataset is from the VOC2011PASCAL dataset.

3.2 Test data training hyperparameter settings

Because the image data in the SBD data set is too large, before the parameter setting, the data set required for testing is reasonably divided, and 8498 of them are used as the training set and 2857 are used as the test set. At the same time, the data was enhanced and the data in the SBD data set was scaled down and scaled to achieve the effect of increasing data samples. At the same time, the above data is divided into ten groups in an increasing relationship for future testing.

The mean reduction operation is performed on the data in the data input layer to ensure that the network speeds up the convergence of each layer's weight when backpropagating. At the same time, in order to enhance the generalization ability of the data, the input is also subjected to random mirror flip processing. After random flip, each input is not a fixed value, so that the tracking method based on deep learning has a different input situation. The better the learning, the more robust the data obtained after training.

The data was randomly cropped with a fixed size of 472x472. This size can be adjusted according to the consumption of resources, but it must be a multiple of 8, because the input is continuously down-sampled during the test. Each stage will down-sample to half of the previous stage, so a multiple of 8 is good for data processing.

Set the maximum number of training iterations on the SBD dataset to 22000/08000 times. The basic learning rate is set to 2.5e-8, the weight update impulse is 0.9, and the weight attenuation is 5e-4. After the setup is completed, test the variation of noise variance in the process of tracking the edge of the damaged image with different tracking methods.

3.3 Noise Variance Test Results

Different edge mural tracking methods were used to test the variance of noise generated during the tracking process in the same environment. Using third-party software to statistically test the results, the results are shown below:
Because the amount of data in the ten test groups has an increasing relationship, the data amount gradually increases during the test by default. Observing the results in the figure, it shows that using the chain code-based tracking method, as the amount of data increases, the noise variance gradually increases, reaching the peak in the second set of tests, and has been in a higher position since then, without a downward trend; the test results based on the run-length tracking method show that with the increase of the amount of data, the noise variance fluctuates greatly, there is no obvious law, the overall fluctuates between 0.01 ~ 0.04, and most of them are in a higher position; the tracking method based on deep learning shows that as the amount of data increases, the noise variance does not change significantly, always between 0.010 and 0.015, which is in a lower position.

In summary, the noise variance of the edge tracking method of the damaged mural image based on deep learning is lower than the other two methods, which indicates that the designed tracking method is more suitable for edge tracking of the damaged mural image.

4 concluding remarks

The generation of edge tracking method for broken mural images is of great significance for the repair and protection of broken murals. With the development of technology, traditional tracking methods can not meet the actual needs of today's edge tracking of broken murals. Therefore, deep learning technology is used to design a method for tracking broken mural images based on deep learning, and fuse images through network models in deep learning Edge features to reduce the variance of noise generated during the tracking process. The designed comparative test proves that the method in this paper can effectively solve the problems existing in the traditional method, and has certain practical significance for the
subsequent development of image edge tracking methods.

References


A Novel 3D Reconstruction Algorithm of CT Images Based on Improved Marching Cubes Algorithm

Dongxue Su, Lin Ma*

hitsudongxue@163.com, malin@hit.edu.cn

School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China

Abstract. In 3D reconstruction for CT images, the Marching Cubes (MC) algorithm is a popular used surface rendering algorithm. However, current MC algorithm has to calculate a large amount of data and triangular patches, which leads to a lower speed of the algorithm. This paper proposes an improved MC algorithm by improving the representation of voxels. By increasing the volume of voxels, the number of voxels that need to be traversed in 3D reconstruction is greatly reduced. Before the isosurface is drawn, the triangle mesh is subtracted. At the same time, in the process of calculating the equivalent points, this paper proposes to introduce smoothness and distortion coefficients to control the smoothness and distortion of the 3D reconstruction model. Simulation results show that the proposed algorithm is more efficient than the MC algorithm, and improves the real-time interaction and the controllability of smoothness and distortion.

Keywords: Marching Cubes algorithm; CT image; 3D reconstruction

1 Introduction

In recent years, with the development of modern medical imaging technology, computer tomography (CT), magnetic resonance imaging (MRI), ultrasound (US) and other medical imaging technologies have been widely used. In the current medical diagnosis, the doctor creates a 3D model in his mind by observing the CT sequence pictures. And make corresponding treatment plans with their own experience. However, this approach obviously increases the difficulty for doctors to diagnose. The accuracy of the diagnosis made in this way depends largely on the clinical experience of the doctor. It is difficult to reach an accurate judgment. In order to reduce the negative impact of subjective judgment and insufficient clinical experience on diagnosis, medical image segmentation and 3D reconstruction are performed with the help of computers.

Generally, 3D reconstruction methods are mainly divided into two categories: 3D surface rendering and 3D volume rendering. Volume rendering operations are huge and cannot meet the needs of clinical interaction. Therefore, surface rendering is still the mainstream of clinical 3D reconstruction. Surface rendering is the reconstruction of intermediate geometric units from the original 3D medical image data. The geometric unit is used to describe the 3D structure of the object. In 3D reconstruction of medical images, the main algorithm used for surface rendering is the Marching Cubes (MC) algorithm proposed by Lorensen [1,2]. Its essence is to extract an isosurface from a 3D data field. So it is also called isosurface extraction algorithm. At present, with the development of medical equipment, medical image
data sets are getting larger and larger. Massive data brings huge challenges to traditional 3D reconstruction techniques.

The MC algorithm needs to extract a large number of triangular patches included in the surface geometric model constructed by the isosurface, which results in a slower algorithm. This limits the application of the MC algorithm, so the academic community has proposed a lot of improved algorithms to solve this type problem. Montani et al. In [3] directly used the midpoint of the voxel edge instead of the intersection point of the isosurface and the cube voxel. This algorithm effectively reduced the calculation amount of the MC algorithm when calculating the vertices of the triangular patches. But the reconstruction results are inaccurate. The grid simplification algorithm is a representative improved algorithm to reduce the operation time [4]. But this algorithm improves the visual effect while improving. [5] is a commonly used algorithm that improves the traversal efficiency of voxels and effectively shortens the processing time of the MC algorithm. The Decimation algorithm reduced the number of triangles generated [6]. Xie Z et al. proposed an adaptive MC algorithm in [7], which greatly reduced the number of triangular patches, and the display effect was not much different from the standard algorithm. Liu et al. uses the tetrahedral partitioning method to reduce the number of triangular patches, which improves the speed of the rendering algorithm [8].

Aiming at the above problems, this paper proposes an improved MC algorithm. We first improve the representation of voxels. This reduces the number of voxels that the algorithm needs to traverse during the 3D reconstruction process and the 3D reconstruction time. Finally, the efficiency of the algorithm is improved. At the same time, based on the MC algorithm, smoothing and distortion coefficients are introduced. The smoothness and distortion of the 3D reconstruction model are controlled by setting the smoothing and distortion coefficients. Finally, before isosurface drawing, a mesh simplification algorithm based on quadratic measurement errors is introduced to reduce the output triangular mesh. The improved algorithm proposed in this paper responds faster to zoom, pan, and rotate operations in subsequent interactive applications. So it has better real-time interactivity.

The chapters of this article are arranged as follows. This article will first briefly introduce the basic principles of the MC algorithm in Section 2. In Section 3, an improved algorithm based on MC algorithm is introduced. The experimental comparison results of the traditional MC algorithm and the proposed algorithm are given in Section 4. We analyze the experimental results and give the advantages of the algorithm proposed in this paper. Finally, the full text is summarized in Section 5.

2 MC Algorithm

2.1 MC Algorithm Overview

MC algorithm is a kind of 3D reconstruction surface rendering algorithm. In 3D reconstruction of medical images, a CT image is a tomographic slice, and two adjacent CT images in the sequence are taken to form the upper and lower slices. The 2D CT tomographic image is described by 3D coordinates to obtain a 3D image of the human body. A 3D image or volume data can be described by a 3D array with corresponding values. Each of these elements becomes a volume element, referred to as a voxel. Each fault section takes four
adjacent points, and a total of 8 points form a voxel. The basic idea of MC algorithm is to divide the 3D space into multiple voxels. Then iterate through each voxel to find a voxel that intersects the isosurface. Intersection points of voxels and isosurfaces are calculated by linear interpolation. Intersection points form triangle patches of different configurations within the voxels. The triangle patches are connected to complete the 3D reconstruction.

A voxel has 8 vertices, and each vertex has 2 states, in-plane and out-plane, so there are \(2^8 = 256\) cases of the relationship between voxels and iso-surfaces. By enumerating these 256 cases, we can get a triangle patch configuration index table to record the information that the isosurface intersects with the sides of the voxel and the configuration of the triangle patch in the voxel. Utilizing the symmetry of the cube, the relationship between each vertex and surface on the voxel is reversed at the same time, without affecting the configuration of the triangular patch in the voxel. In other words, the 256 cases are complementary to each other, so you only need to consider the cases where 0-4 kinds of vertices are on the surface. In addition, according to the symmetry of the rotation in the three directions of the cube, the voxel is arbitrarily rotated by a multiple of 90 degrees in the three directions. The configuration of the triangular patch in the voxel is unchanged. Utilizing the characteristics of the above cube, 256 cases are finally reduced to 15 types of triangular patch configuration index table, as shown in Figure 1.

Each cube in Figure 1 represents a voxel. The numbers below the voxels represent the index values of the 15 triangle patch configurations. The shaded plane inside the cube represents the isosurface. The 15 cases in Figure 1 are arranged according to the number of voxel vertices in the isosurface from as few as possible. In each case, several triangle patches inside the voxel divide the voxel vertices into two categories, one inside the isosurface and the other outside the isosurface.

The key of MC algorithm is to find the isosurface in the original data. To find the isosurface with a threshold value of \(c\), you must find the points where the gray value function
value is \( c \). These points are called equivalence points \((x, y, z)\). The isosurface can be expressed as a collection of isosurfaces, as shown in equation (1):

\[
\{(x, y, z) \mid F(x, y, z) = c\}
\]

where \( c \) is the threshold of the isosurface, \( c \) is a constant, \( F \) is a gray value function, and the range of \( F \) is \([0, 255]\).

The equivalence points constitute several triangular patches. Then the triangular patches are fitted to the isosurface output. In order to display the isosurface image, the normal vector of each triangular patch forming the isosurface must be calculated. However, directly calculating the normal vector of each triangular patch is not only computationally intensive but also complicated. The MC algorithm uses gradient vectors instead of normal vectors. The gradient vector of each point on the isosurface represents the normal vector of the isosurface at that point. The normal vector \( g(x, y, z) \) for a point \((x, y, z)\) in a 3D data field can be expressed as:

\[
g(x, y, z) = \nabla F(x, y, z)
\]

where \( \nabla F(x, y, z) \) represents the gradient of the function \( F \) at the point \((x, y, z)\).

The gradient of the vertices of the triangular patch is obtained by linear interpolation of the gradient of the voxels of the voxels. In actual calculation, the central difference method is used to calculate the gradient of each voxel vertex. For a certain voxel vertex \((x_i, y_j, z_k)\) in a 3D data field, its gradient \( g(x_i, y_j, z_k) \) is calculated as:

\[
\nabla F(x_i, y_j, z_k) = \left( g_x, g_y, g_z \right)
\]

where \( g_x, g_y, g_z \) represent the values of the normal vector in the \( x \), \( y \), and \( z \) directions.

\[
g_x = \frac{F(x_{i+1}, y_j, z_k) - F(x_{i-1}, y_j, z_k)}{2\Delta x}
\]

\[
g_y = \frac{F(x_i, y_{j+1}, z_k) - F(x_i, y_{j-1}, z_k)}{2\Delta y}
\]

\[
g_z = \frac{F(x_i, y_j, z_{k+1}) - F(x_i, y_j, z_{k-1})}{2\Delta z}
\]

where \( \Delta x \), \( \Delta y \), \( \Delta z \) represent the side lengths of voxels in the \( x \), \( y \), and \( z \) directions, \((x_{i+1}, y_j, z_k)\) and \((x_{i-1}, y_j, z_k)\) are voxel vertices adjacent to point \((x_i, y_j, z_k)\) in the \( x \) direction, \((x_i, y_{j+1}, z_k)\) and \((x_i, y_{j-1}, z_k)\) are voxel vertices adjacent to point \((x_i, y_j, z_k)\) in the \( y \) direction, \((x_i, y_j, z_{k+1})\) and \((x_i, y_j, z_{k-1})\) are voxel vertices adjacent to point \((x_i, y_j, z_k)\) in the \( z \) direction.

### 2.2 System Block Diagram of MC Algorithm

The advantages of the MC algorithm are obvious, but it also has the following disadvantages. It can be known from the above analysis that in the MC algorithm, all voxels in the original 3D data need to be traversed. However, the number of voxels is often huge, which leads to a large amount of calculation and low efficiency of the MC algorithm. In addition, the
calculation of the intersection between isosurface and voxel in MC algorithm uses a single interpolation method. 3D reconstruction cannot be performed according to user needs. The output of the MC algorithm is a triangular mesh. The data during 3D reconstruction is often very large. The number of 3D triangular meshes created in this way is huge, which greatly reduces the real-time interaction of the 3D reconstruction model in actual use.

Therefore, this paper proposes an improved CT image 3D reconstruction algorithm based on Marching Cubes algorithm. First, the CT image sequence is input as the original 3D data required for reconstruction. And select voxels to divide the original 3D data space. Then take out a certain voxel. The configuration of the triangular patch in the voxel is determined according to the triangular patch configuration index table. Calculate the coordinates of the equivalence points and the normal vector of the triangular patches. After traversing all the voxels in the 3D data space, all the obtained triangle patches are output to obtain isosurfaces. Finally, the triangle mesh of the isosurface is simplified to obtain the final 3D reconstruction model. This paper improves the traditional MC algorithm in three aspects: the voxel representation of the algorithm, the interpolation method and the simplification of the triangular mesh. The system block diagram of the improved algorithm is shown in Figure 2.

In Figure 2, a CT image sequence is used as an input of a 3D reconstruction process. The 3D reconstruction model of the CT image is displayed as the final output result. Among them, the MC algorithm based voxel improvement algorithm will replace the original voxel selection process. The number of voxels to be traversed during the 3D reconstruction process is reduced to improve the operation efficiency of the 3D reconstruction algorithm. The smoothing and distortion coefficients are introduced in the calculation of the coordinates of the equivalent points by the MC algorithm to control the smoothness and distortion of the 3D reconstruction model. Finally, before the output of the 3D reconstruction model, a step of simplifying the triangular mesh is added to reduce the number of triangular meshes in the 3D reconstruction model. So that the 3D reconstruction model finally used in practical applications will have better real-time interaction.
3 Improvement of MC Algorithm

3.1 Improved Voxel Algorithm Based on MC Algorithm

The two adjacent CT images in the CT sequence are taken out to form the upper and lower slices. Four points were taken from each of the upper and lower slices to form a voxel. A voxel is the basic unit of volume data. It can be regarded as a small cuboid with a certain size or abstracted as an abstract point in a space with certain attributes, as shown in Figure 3.

In Figure 3, the 3D image or volume data can be represented by \( I(x, y, z, c) \), where \( x, y, \) and \( z \) respectively represent coordinates in a 3D space, and \( c \) represents a color component in a color image. When a certain dimension in a 3D image takes a fixed value, a 2D image can be obtained, which is commonly referred to as a tomographic image.

The traditional MC algorithm takes two adjacent images in the CT sequence to form the upper and lower slices. Each slice takes 4 adjacent points, and a total of 8 points form a voxel. Then the voxels need to be traversed, and the intersection coordinates need to be calculated by interpolation. Due to the huge number of voxels and the complicated calculation process of the interpolation method, the calculation amount of the MC algorithm is very large, and there will be a delay during 3D reconstruction. In order to improve the execution efficiency of the MC algorithm, this paper improves the voxel representation based on the MC algorithm. The improved voxel representation is shown in Figure 4.

In the improved MC algorithm, the length of one side of the voxel is increased. As shown in Figure 4 (a) and Figure 4 (b), the side length of the voxel in the \( x \) direction and the \( y \) direction is increased, resulting in a reduction in the number of voxels and a decrease in the complexity of the calculation process. This improves the execution efficiency of the MC algorithm and reduces the delay during 3D reconstruction.
direction is improved to 2 pixels. The improved voxel volume is doubled to the original voxel volume. Therefore, the number of voxels to be traversed during the 3D reconstruction process is reduced to half, and the time required for reconstruction is greatly reduced. In terms of the influence of 3D reconstruction accuracy, the improved voxel representation can still control the error range within 2 pixels, which has little effect on the reconstruction accuracy. In order to further reduce the number of voxels that need to be traversed, this paper changes the voxel side length to 2 pixel lengths in the \( x \) and \( y \) directions at the same time, as shown in Figure 5. The improved voxel volume is four times the original. The number of voxels that need to be traversed is reduced to the original \( 1/4 \), which further improves the operation efficiency of the proposed algorithm.

![Voxel representation with 2 pixels in the \( x \) and \( y \) directions.](image)

After improving the representation of voxels, the original 3D data will be divided into individual voxels. Compare the relationship between the gray value of the voxel vertex and the isosurface threshold. Determine the state of each voxel vertex, in-plane or out-plane. After determining the state of the vertices, find the configuration corresponding to the voxel in the triangle patch configuration index table and record it. Next we need to calculate the coordinate values of the equivalence points. In calculating the coordinates of the equivalent points, in order to control the smoothness and distortion of the 3D reconstruction model, this paper introduces smoothness and distortion coefficients based on the MC algorithm.

### 3.2 Smoothing and distortion control based on MC algorithm

When calculating the coordinates of the iso-points, the MC algorithm first calculates the case where a voxel intersects the iso-surface, and obtains several triangle patches. Then iterate through all the voxels to get all the intersecting triangle patches. Finally, these triangular patches are fitted to the isosurface output. When computing the intersection of voxels and isosurfaces, the 3D spatial data of CT images are usually discrete. The data field along the voxel edge changes continuously and linearly. If the two vertices of an edge of a voxel are greater than or less than the value of the isosurface, there is only one intersection on the edge. If both vertices on an edge are larger or smaller than the value of the isosurface, the edge has no intersection with the isosurface. MC algorithm finds the intersection of isosurface and voxel by processing voxels in the data field one by one. Then we find the isosurfaces contained in these voxels.

The traditional MC algorithm uses linear interpolation to calculate the coordinates of the equivalent points. When using the linear interpolation method, it is necessary to assume that the function value of the discrete data field in 3D space changes linearly along the voxel edge.
boundary. Due to the existence of the above assumptions, the linear interpolation method makes the display effect of the 3D reconstruction model smoother.

Suppose the two endpoints \( P_i = (x_i, y_i, z_i) \) and \( P_j = (x_j, y_j, z_j) \) of the edge on which the equivalence points lie. \( F \) represents the gray value function, then the gray values of endpoint \( P_i \) and endpoint \( P_j \) are \( F_i = F(x_i, y_i, z_i) \) and \( F_j = F(x_j, y_j, z_j) \). If the selected isosurface threshold is \( c \), according to the basic principle of the MC algorithm, it can be known that the coordinates of any equivalence point \( K = (x, y, z) \) can be obtained by equation (5) through linear interpolation.

\[
K = \left( x_i, y_i, z_i \right)^T + \frac{c - F_i}{F_j - F_i} \left( \left( x_j, y_j, z_j \right)^T - \left( x_i, y_i, z_i \right)^T \right)
\]

(5)

The median interpolation method is a simpler interpolation method. When calculating the coordinates of the equivalence points, it is only necessary to take the median of the coordinates corresponding to the edge where they are located. The median interpolation method can reduce distortion at the cost of smoothness. The median interpolation method is shown in equation (6):

\[
K = \left( x_i, y_i, z_i \right)^T + 0.5 \left( \left( x_j, y_j, z_j \right)^T - \left( x_i, y_i, z_i \right)^T \right)
\]

(6)

In order to control the smoothness and distortion of the 3D reconstruction model, parameter \( \alpha \) is introduced in this paper to represent the smoothness and distortion coefficient. The formula for calculating the coordinates of the equivalence point by introducing parameter \( \alpha \) is shown in equation (7).

\[
K = \left( x_i, y_i, z_i \right)^T + \alpha \left( \frac{c - F_i}{F_j - F_i} - 0.5 \right) \left( \left( x_j, y_j, z_j \right)^T - \left( x_i, y_i, z_i \right)^T \right) + 0.5 \left( \left( x_j, y_j, z_j \right)^T - \left( x_i, y_i, z_i \right)^T \right)
\]

(7)

Control the smoothness and distortion of the 3D reconstruction model by changing \( \alpha \) within the range of [0,1]. When \( \alpha = 0 \), the interpolation method at this time becomes the median interpolation method. The 3D reconstruction model will have the lowest distortion. When \( \alpha = 1 \), the interpolation method at this time becomes a linear interpolation method. The 3D reconstruction model will have the highest smoothness.

After the coordinates of the equivalent points are obtained, the triangle patches inside each voxel are determined according to the previously recorded triangle patch configuration. We get equivalent patches in the voxels. After outputting these isosurfaces together, the required isosurfaces are drawn.

The output of the MC algorithm is a triangular mesh. Generally speaking, these grids are huge and cannot meet the speed requirements of drawing and processing in interaction. In order to enhance the real-time interactivity of MC algorithm, this paper introduces a grid simplification algorithm based on quadratic measurement error to reduce the output triangular
grid. The purpose of introducing measurement error is to quantify the difference between the input model and the output model. The simplification of the model is guided according to the measurement error, so that the simplification error is within the error range allowed by the user. The quadratic measurement error is recognized as a better measurement error. On the premise of maintaining the reliability of the original mesh, the number of triangular patches in the triangular mesh is reduced. This reduces the response time of the 3D reconstruction model in operations such as panning, zooming, and rotating.

Adding the simplified steps of the triangular mesh in the 3D reconstruction process can effectively reduce the number of meshes to be drawn without affecting the overall rendering effect, which is of great significance to both real-time rendering and real-time interaction.

4 Experiment and Analysis

4.1 Efficiency Analysis of MC Algorithm Based on Improved Voxel Algorithm

In order to illustrate the effectiveness of the above algorithm, this paper implements a 3D reconstruction algorithm and process of CT images based on the improved Marching Cubes algorithm. CT images are generally in DICOM format. In order to facilitate the reading of matlab, the DICOM format is converted to the .png format before 3D reconstruction. In order to facilitate comparison, two sets of abdominal enhanced CT images are selected as experimental data, and only three-dimensional reconstruction of liver tumors is performed. The pixel size of the two sets of CT images is 512×512. We tested our method on the competitive dataset of MICCAI LiTS Challenge[9]. The first set of reconstruction data consisted of 29 CT cut edges, and the second set consisted of 250 CT slices.

First, the voxel representations for 3D reconstruction were selected, and 3 voxel representations were used to perform 3D reconstruction of the liver tumor surface. The first voxel representation is the voxel representation of the traditional MC algorithm, as shown in Figure 1. The second voxel is represented as an improved voxel representation, as shown in Figure 5. The third voxel is represented as an improved voxel representation, and the voxel sides in the x and y directions are both 3 pixels.

Fig. 6. Reconstruction results of the first set of data.
Figure 6 shows the results of processing the first set of reconstruction data, which can clearly reconstruct the surface of a single liver tumor. The reconstructed tumor surface consists of several triangular patches. In order to facilitate observation, the surface of the triangular patch is shown in red in this figure. The borders of the triangles are drawn in yellow. As can be seen from the figure, compared with the traditional MC algorithm, the proposed algorithm has a larger area and a smaller number of triangles.

In Figure 7, a three-dimensional reconstruction is performed using the second set of reconstruction data. The entire 3D reconstruction model is yellow. The second set of data has a larger amount of data than the first set of data, making the displayed triangular patches denser. The red surface area is compressed when displayed, and the yellow border is more prominent. From the overall effect of reconstruction, the MC algorithm based on voxel changes has little effect on the accuracy of 3D reconstruction.

**Table 1. Efficiency comparison of reconstruction algorithms**

<table>
<thead>
<tr>
<th>Voxel representation</th>
<th>The first set of data</th>
<th>The second set of data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reconstruction time/s</td>
<td>Improved efficiency</td>
</tr>
<tr>
<td>Voxel representation of MC algorithm</td>
<td>0.809010</td>
<td></td>
</tr>
<tr>
<td>Voxel representation with 2 pixels in the x and y directions</td>
<td>0.314780</td>
<td>61.0907%</td>
</tr>
<tr>
<td>Voxel representation with 3 pixels in the x and y directions</td>
<td>0.169684</td>
<td>79.0257%</td>
</tr>
</tbody>
</table>

By comparing the reconstruction time, it can be seen from Table 1 that no matter for the reconstruction of the first or second set of data, the operation efficiency of the MC algorithm based on voxel changes is much greater than the traditional MC algorithm. This is because when the side length of the voxel becomes longer, the number of voxels traversed during the 3D reconstruction process will be reduced. Therefore, the time of 3D reconstruction is greatly reduced, and the operation efficiency of the 3D reconstruction algorithm is greatly improved.
When the amount of data that needs to be reconstructed is larger, the difference in visual perception between the results of the proposed algorithm and the traditional MC algorithm is smaller, and the efficiency of the algorithm is significantly improved.

4.2 Effect Analysis of MC Algorithm with Smoothing and Distortion Coefficients

Three-dimensional reconstruction is performed on the same set of data under different smoothing and distortion coefficients. The reconstruction effect is compared. As shown in Figure 8, the smoothness and distortion can be controlled by introducing smoothness and distortion coefficients.

![Reconstruction results with different smoothing and distortion coefficients.](image)

(a) smoothing distortion coefficient =0  (b) smoothing distortion coefficient =0.25
(c) smoothing distortion coefficient =0.75  (d) smoothing distortion coefficient =1

**Fig. 8.** Reconstruction results with different smoothing and distortion coefficients.

In Figure 8(a), under the condition that the smoothing distortion coefficient is equal to 0, the interpolation method at this time becomes the median interpolation method. It can be seen from the figure that the triangles that make up the isosurface are of uniform size, consistent shape, and neat orientation. The 3D reconstruction model is more rigid in visual effects and the smoothness is reduced. In Figure 8(b) and Figure 8(c), the smoothing and distortion coefficients are set to 0.25 and 0.75. In Figure 8(d), under the condition that the smoothing distortion coefficient is equal to 1, the interpolation method at this time becomes a linear
interpolation method. It can be seen from the figure that whether on a flat surface or an uneven surface, the triangular patches that make up the isosurface will change in size, shape and direction according to the gradient of the gray value. The smoothness of the 3D reconstruction model is improved, but this is at the cost of the distortion of the 3D reconstruction model.

4.3 Effect Analysis of MC Algorithm with Smoothing and Distortion Coefficients

The experiment uses the first set of reconstruction data for 3D reconstruction. Before the isosurface is drawn, a simplified step of adding a triangle mesh is added. This paper chooses the rotation operation to test the real-time interactivity of the 3D reconstruction model. The 3D reconstruction model is continuously rotated, each rotation angle is 20°, the azimuth angle is changed from 0° to 360°, and a circle around the z axis. The experimental results are shown in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>First time/s</th>
<th>Second time/s</th>
<th>Third time/s</th>
<th>Fourth time/s</th>
<th>Fifth time/s</th>
<th>Average time/s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not simplified</td>
<td>4.000108</td>
<td>3.876800</td>
<td>3.749688</td>
<td>4.242814</td>
<td>3.701339</td>
<td>3.91415</td>
</tr>
<tr>
<td>Simplification rate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50%</td>
<td>3.708591</td>
<td>3.266038</td>
<td>3.421573</td>
<td>3.635367</td>
<td>3.217326</td>
<td>3.44978</td>
</tr>
<tr>
<td>90%</td>
<td>2.866155</td>
<td>2.916901</td>
<td>3.131374</td>
<td>2.718502</td>
<td>2.921755</td>
<td>2.91094</td>
</tr>
</tbody>
</table>

In order to make the experimental results more accurate, this article has measured the operating time several times. The average value is calculated for the analysis of the simplified effect of the triangular mesh. In this paper, the continuous rotation time is recorded in three cases: unsimplified, 50% simplified and 90% simplified. As can be seen from Table 2, the higher the simplification rate of the triangular mesh, the shorter the time taken for the continuous rotation operation. When the simplification rate is 50%, the average time of continuous rotation operation is reduced by 11.86% compared with the case without simplification. When the simplification rate is 90%, the average time of continuous rotation operation is even reduced by 25.63%. It can be verified that the triangular mesh simplification can well enhance the real-time interactivity of the 3D reconstruction model.

5 Conclusion

Aiming at the problems of low efficiency of 3D reconstruction algorithm of CT image and insufficient practicality of 3D reconstruction model, this paper proposes an improved algorithm based on MC. First, by improving the representation of voxels, the volume of voxels can be increased. This reduces the number of voxels that need to be traversed during the 3D reconstruction process, and further improves the operation efficiency of the 3D reconstruction algorithm. At the same time, smoothing and distortion coefficients are introduced in the process of calculating the equivalent points to control the smoothness and distortion of the 3D reconstruction model. Before drawing the isosurface, the triangulated mesh based on the quadratic measurement error is used to reduce the output triangular mesh. The real-time
interaction of the 3D reconstructed model is enhanced, so that the model has a faster response speed in the process of translation, enlargement and rotation.

Acknowledgments. This paper is supported by national nature science foundation of China (61971162, 41861134010).

References

Lightweight Deep Learning Model for Invoice Image Classification

Run Wang, Chonghu Cheng, Yang Meng, Jie Yang and Guan Gui
{1218012107@njupt.edu.cn, chengch@njupt.edu.cn, guiguan@njupt.edu.cn}

College of Telecommunications & Information Engineering
Nanjing University of Posts and Telecommunications
Nanjing 210003, China

Abstract. Deep learning based image classification usually needs a large amount of data and computing resources to achieve better classification results. Hence, image classification for small data sets has attracted more and more attention in recent years. we proposed a lightweight deep learning model in this paper, which can be used for image classification on small datasets. The experimental results show that this model can achieve high accuracy classification with small data sets.

Keywords: Image classification, deep learning, convolutional neural network, computer vision.

1 Introduction

In recent years, with the rapid development of deep learning [1], the application of deep learning has made breakthroughs in many aspects. For example, in the annual Image large-scale Visual Recognition Challenge (ILSVRC), the constantly updated deep convolution network model is used to classify 1000 categories of about 100,000 object images. When AlexNet won the 2012 ImageNet competition by an overwhelming margin [2], AlexNet only had 5 layers of convolutional layer. It developed to 19 layers of VGG-Net [3], 22 layers of GoogLeNet [4] [5], and ResNet [6], which created a new network structure. Its classification accuracy has been greatly improved every year.

The achievements of deep learning in recent years are mainly attributed to the significant improvement of data available for model training and the significant increase of computing resources (the continuous improvement of GPU performance) [7]. Along with the increasing depth of convolution model layer, that means need a large amount of data and computing resources, and deep learning in the practical application often need real-time and occupy less memory and computing resources of the lightweight model, such as hyperspectral images classification [8] [9], face recognition [10], water level observing [11]. Also a great deal of data for training model can be difficult to obtain, in order to solve this problem, how deep learning application on small data sets has attracted widespread attention.

As the depth of the model increases, the learning ability of the model does not increase. On the contrary, the deep model will produce higher training error rate than the shallow model. But it is not caused by overfitting. Because the neural network needs to propagate gradient continuously in the process of back propagation [12]. But when the network model becomes complex, the gradient will gradually disappear in the process of propagation, and optimization
will become more difficult, which leads to the so-called "degradation" problem. He proposed ResNet is to solve this problem.

This paper proposes a method of invoice image classification based on ResNet, which solves the problem of "degradation" caused by the increase of model depth. ResNet got the first place in the classification task of ILSVRC2015. However, when the data set is small, the model is still too large. Based on resnet-18, this paper modified the model and train in a data set with only 400 images for each type of invoice, and finally achieved a better classification effect in the test set.

2 DESCRIPTION OF THE MODEL STRUCTURE

In this section, the original structure of ResNet and the improved ResNet structure are described and analyzed in detail. The original model structure of ResNet is shown in Fig.1.

![Fig. 1. Structure of ResNet.](image1)

As illustrated in Fig.1, ResNet-18 contains 18 layers, which including 17 convolution layers and a full connection layer. The first convolution layer is followed by the Max pooling layer, and there is an average layer before the final convolution layer. The red number in the rectangle represents the size of kernel, the blue number represents the number of kernel, and the purple number represents the convolution stride. ResNet is different from the general deep learning model in that it does not stack the convolution layer directly.

![Fig. 2. A building block.](image2)
The curve above Figure 1 is the same as the curve in Figure 2 is called shortcut connection. It represents that the output $H(x)$ is equal to the input $x$ plus the $F(x)$. When $F(x)$ is inconsistent with the number of channels of $x$, it is represented by a dotted line, and a convolution operation needs to be added to make $F(x)$ consistent with the number of channels of $x$. When $F(x)$ is consistent with the number of channels of $x$, no operation is needed to connect directly with a solid line. Resnet-18 network performs well in many classification tasks, which is different from the simple stacking of convolution layer in traditional network. Shortcut connection is a revolutionary invention. Two layers of convolution make a building block. Such a structure can increase the depth of the network without causing degradation problems. With its unique model structure, the network ResNet has achieved convincing accuracy in many classification tasks. Nevertheless, the number of parameters contained in the 18-layer network still needs a lot of data for training. In the classified invoice task, there are only 400 invoice images of each category in the training data, which is far less than the number of images in ImageNet. It is difficult for the model to converge. If the model is not converge, the classification task cannot be well completed. Based on the resnet-18 network, some improvements are made in this paper, so that this model has better classification effect than the original model and other models. Based on the existing advantages of ResNet, this paper modified the model parameters and structure to further improve the model performance. The structure is described in Fig. 3.

**Fig. 3. Structure of proposed model.**
The activation function used here is \(+\) \(\sigma = \max(0, x)\). This function is constant in most cases, which is helpful to solve the vanishing gradient problem of the deep network. Moreover, it has the biological principle, which is usually better than other activation functions in practice. The operator of \(\sigma\) can be described as follows [13].

\[
\begin{align*}
\mu_{\sigma} &\leftarrow \frac{1}{m} \sum_{i=1}^{m} x_i & (1) \\
\sigma_{\sigma}^2 &\leftarrow \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_{\sigma})^2 & (2) \\
\hat{z}_i &\leftarrow \frac{x_i - \mu_{\sigma}}{\sigma_{\sigma} + \epsilon} & (3) \\
y_i &\leftarrow \gamma \hat{z}_i + \beta \equiv BN_{\gamma, \beta}(x_i) & (4)
\end{align*}
\]

Among these four equations, Eq. (1) calculate the mean of the samples in a mini-batch. Eq. (2) calculate the variance of the samples in a mini-batch. Eq. (3) calculates the normalized results, Eq. (4) trains the parameter \(\gamma\) and \(\beta\), and throw a linear transformation get a new \(y_i\). During forward propagation, new distribution values are obtained from learnable \(\gamma\) and \(\beta\) parameters. In the case of back propagation, by taking a chain derivative can find \(\gamma\) and \(\beta\) and their associated weights.

The BN layer is followed by the activation layer. The activation function used here is Rectified Linear Unit (ReLU). The vanishing gradient is particularly obvious when the number of network layers is large, and it is one of the main obstacles to deepen the network structure. The gradient of the ReLU function is constant in most cases, which is helpful to solve the convergence problem of the deep network. Moreover, it has the biological principle, which is usually better than other activation functions in practice. The operator of ReLU function can be described as follows.

\[ f(x) = \max(0, x) \]

where \(x\) is the input vector and \(f(x)\) is the output of ReLU as the activation function of neurons. With the introduction of activation function, the output of deep neural network is no longer a linear combination of input. The activation function can make the output of some neurons equal to 0, resulting in the sparsity of the network, reducing the interdependence of parameters and alleviating the overfitting problem.

3 EXPERIMENTAL RESULTS

3.1 Data set

In this invoice classification task, both the train set and the test set are photos taken by the mobile phone, with a total of 1200 images to training model. Instead of directly putting the data into the model for training, we first carry out the grayscale processing, then convert it into an
array, and finally carry out the normalization processing. This is a supervised learning task, so each kind of picture needs to be labeled as "one hot". Within each epoch, 1,200 images are divided into two parts, 840 of which are used for training and 360 for verification. To test the fitting degree of the model, 42 images different from the training set and the verification set are used for test model. If the various invoices can be accurately classified then the fitting is better.

The data set includes three types of invoices, namely value-added tax, train, and taxi invoices. Data imbalance will lead to deviation in the process of training model. Therefore, it is necessary to ensure that each type of images in the train set accounts for 1/3 of the total number when conditions permit, this means that there are 400 invoices images for each type of invoice. Similarly, the 42 images in the test set also account for 1/3 of each type. Considering that the model is used to classify the images taken by the mobile phone, the image will have various angles, so the train set need to perform data augmentation. Data augmentation is achieved by rotating, flipping, scaling, color jittering and adding noise to the train image. The significance of data augmentation is to increase the robustness of the model, so that the model has better performance for image classification tasks. Three of the raw Image after grayscale processing are provided in Fig.5.

![Fig. 5. Samples of the training images after grayscale processing](image)

### 3.2 COMPARISON OF MODELS IN CLASSIFICATION ACCURACY

In order to fairly compare the classification accuracy of the proposed model and other models, all experiments were performed on the same equipment and in the same data set, and the same test set is used for model testing.

Compared with the original model, the improved model is already lightweight, so the experiment only compares the classification accuracy of the model. In this experiment, in addition to comparing the model proposed in this paper with the original model, it will also be compared with other models that have won the ImageNET competition, they are Alex Net, ZFNet and VGG-16. Table 1 shows the classification performance of each model in the test set and the amount of parameters for each model. The source of these data is obtained in the same experimental environment through the same steps.

<table>
<thead>
<tr>
<th></th>
<th>AlexNet</th>
<th>ZFNet</th>
<th>VGG-16</th>
<th>ResNet-18</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>85.71</td>
<td>90.84</td>
<td>33.33</td>
<td>88.09</td>
<td>97.61</td>
</tr>
<tr>
<td>Number of parameters</td>
<td>24715071</td>
<td>24718467</td>
<td>3597795</td>
<td>11174787</td>
<td>701283</td>
</tr>
</tbody>
</table>
Results illustrate that the ResNet has natural advantages, compared with Alex Net, ZFNet, VGG-16. ResNet has more layers, but fewer parameters. The model proposed in this paper is based on ResNet. This model has the least parameters, but its classification accuracy is the highest. Only one of the forty-two test images was misclassified. In summary, in the invoice image classification task with high classification accuracy requirement, the accuracy of classification is the first consideration. The model proposed in this paper can accurately classify the images in the test set. And an image that is not accurately classified in the test set, even a person can’t determine which type of invoice belongs to. Therefore, this model is far superior to other models.

4 CONCLUSION

This paper proposes an invoice image classification model based on deep learning model ResNet, which can accurately classify value-add tax invoice images, train invoice images and taxi invoice images. In this paper. The kernel size and number of hidden neurons are modified based on the original model, and BN is added to 'shortcut'. The experimental results show that these modifications make the model perform better in the invoice classification task and perform better than other models, so the model is the best choice for invoice classification tasks. There are many types of invoices in the market. To better meet the actual needs, future research will increase the category of invoices and further optimize the model structure.

References

Research on image sensitive information recognition
based on machine learning

LI Kai-yong¹, MAO Duo-lu²
{WEIY1220@tom.com¹, zhushuang183@163.com²}
¹,²Qinghai Nationalities University Physics and Electronic Information Engineering, Xining, 810007
China;

Abstract: Aiming at the problem of low accuracy in traditional image sensitive information recognition methods, a new image sensitive information recognition method based on machine learning was proposed. The pre-processing operations of de-noising and detail sharpening are carried out for the recognition image, and the pre-processing image features are extracted from the three perspectives of color, shape and texture. The image sensitive information is retrieved by combining the extracted image features. Based on the principle of support vector machine (SVM) in machine learning, an image classifier is designed to realize the classification and recognition of image sensitive information. Compared with the two traditional image recognition methods, it is proved that this method has higher precision, shorter recognition time and is more suitable for the recognition of image sensitive information.

Key words: machine learning; Image sensitive information; Identification method; Support vector machine;

1 Introduction

With the rapid development of Internet technology and mobile communication technology, people's access to information becomes more convenient and diversified. There are a lot of sensitive information mixed in these information, including pornography, violence, contraband, cult and reactionary sensitive content information, which brings a huge challenge to the information security review of the network information regulatory department. Image recognition is widely used in real life, and image recognition technology is gradually improved with the development of science and technology. Images are closely related to human life [1-2]. As the main source of information about the external world, people rely more and more on pictures for communication, which brings convenience to life and highlights the importance of visual technology. People's dependence on network communication makes the amount of image information on the network increase continuously. With the continuous development of network and other related technologies, a large number of images containing sensitive information are widely spread on the network, which has a
negative impact on social development. Therefore, the research on image sensitive information recognition methods is more and more important [3-4]. The traditional image sensitive information recognition method uses emotion to distinguish, mainly for the text information contained in the image to identify, has certain limitations. The image sensitive information recognition method based on image feature extraction needs to compare the image features several times in the process of recognition.

Machine learning is an interdisciplinary subject involving probability theory, statistics, approximation theory, convex analysis, algorithm complexity and other theories. Machine learning is to simulate the human brain, so that the machine can learn autonomously just like human, so as to acquire new knowledge or skills, and reorganize the existing knowledge structure to continuously improve its performance [5]. Human mainly obtains the surrounding things through vision, presents the images through the eyes, and finally forms relevant concepts in the brain through certain abstraction and reasoning [6]. Machine vision in machine learning is to use machines to simulate the human brain, obtain pictures through equipment, and then after certain processing and feature extraction, finally identify relevant information. The core of machine learning is the training of classifier [7]. Through continuous training and adjustment, a good classifier model is finally obtained to realize accurate recognition of things. The good learnability of machine learning means that this method can be widely applied in the field of image recognition [8]. Based on the above analysis, this paper proposes an image sensitive information recognition method based on machine learning.

2 Research on image sensitive information recognition method based on machine learning

2.1 Image preprocessing

For the image to be recognized, image denoising and image detail sharpening are required. Since certain noise points will be included in the image, the existence of noise points will affect the accuracy of image feature extraction and subsequent recognition, so the noise needs to be filtered out before identifying sensitive information of the image [9-10].

In this paper, the image is denoised by means of mean filter and median filter. Let \( r(x, y) \) be the noisy image to be processed, and the image processed by mean filtering is \( q(x, y) \), which can be expressed by the following formula.

\[
q(x, y) = \frac{1}{K} \sum r(x, y) \quad (x, y) \in \Omega 
\]  

(1)

In formula (1), \( \Omega \) is the set of coordinates of all pixel points in the window.
neighborhood calculated by the mean filter template, and \( K \) is the total number of all pixel points in the window neighborhood. Mean filtering is suitable for denoising images with larger Windows, that is, images with larger sizes [11]. However, mean filtering can erase certain image details. Therefore, median filtering technology is used to filter images.

Set the center point of the image as \((x_0, y_0)\), arrange all pixel points in the neighborhood centered on point \((x_0, y_0)\) in descending order according to the gray value of the pixel, and take the middle value after sorting as the current pixel point, that is, the center point of the neighborhood. The 3×3 template was used to replace the original gray value of the central pixel with the median value, and the median value was filtered to obtain the image \( f(x, y) \) after two filters, and the denoising of the image was completed [12].

After denoising the image, the boundary contour of the image will be blurred, so the image needs to be sharpened. For denoising image \( f(x, y) \), the gradient at \((x, y)\) of a certain pixel in the image is represented as vector \( G \).

\[
G = \begin{bmatrix} f(x, y) \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \end{bmatrix}
\]

(2)

The direction of vector \( G \) gradient is the maximum change rate of \( f(x, y) \), and the difference processing method is used to calculate the gradient of image \( f(x, y) \). After calculating the gradient, the following formula is used to further correct the gradient.

\[
g(x, y) = \begin{cases} G[f(x, y)], & G(f) \geq T \\ f(x, y), & else \end{cases}
\]

(3)

In formula (4), \( T \) is the gradient threshold, and \( T \) is a non-negative value. An appropriate threshold is selected to highlight the edge contour of the object to be examined in the image [13]. After image preprocessing, image features are extracted.

### 2.2 Image feature extraction

Image feature is the key of image recognition. Image feature can be divided into global feature and local feature. In this paper, image features are extracted from color, shape and texture. Color histogram is used to extract the color features of the image. Color histogram is
to calculate the frequency of each gray scale in the color [14-15]. The gray scale of color
histogram represents the color brightness of the image. The horizontal coordinate represents
the gray scale of the image \( L \), and the vertical coordinate represents the probability of gray
scale appearing in the image \( p(l) \). The calculation formula of color histogram is as follows:

\[
p(l) = \frac{n_l}{N}, \quad l = 0, 1, 2 \cdots, L - 1
\]  

In formula (4), \( n_l \) is the number of pixels of the \( l \)-th grade gray scale, and \( N \) is the
sum of pixels in the image. After the color features of the image are extracted, shape features
in the image are further developed. In this process, the edges of things in the image need to be
detected first [16].

The Sobel operator is adopted to image edge detection, Sobel operator is a discrete
convolution template, according to the brightness of the image channel matrix for convolution
operation, detecting the edge of the object in the image to find the boundary of the pixel values
mutations, Sobel operator is through the horizontal, vertical and two direction of image edge
detection, so you need to \( \alpha, \beta \) convolution template in two directions, as shown below:

\[
G_x = \begin{bmatrix}
-1 & 0 & 1 \\
-2 & 0 & 2 \\
-1 & 0 & 1
\end{bmatrix}
\]  

\[
G_y = \begin{bmatrix}
1 & 2 & 1 \\
0 & 0 & 0 \\
-1 & -2 & -1
\end{bmatrix}
\]  

According to the above template, the operation process of convolution operation on the
image is as follows:

\[
\begin{align*}
I_x &= G_x \times A \\
I_y &= G_y \times A
\end{align*}
\]  

In formula (7), \( I_x \) and \( I_y \) are the gray values of edge detection in the horizontal and
vertical directions of the image respectively, and \( A \) are the image brightness channel matrix
[17-18]. The calculation formula of each gray value of the final edge detection matrix is as
follows:
After detecting the edges in the image, shape features in the image are extracted according to the edge feature map. Rotate from 0 ° to 360 ° with the image center as the origin, count the number of image pixel points whose gray value is not 0 on each Angle, draw the histogram with the Angle as the abscissa and the number as the ordinate, and normalize the histogram to make the distribution of the resulting histogram more uniform [19]. After the shape feature of the image is obtained from the histogram, the image texture feature is extracted.

Texture is an important visual feature of the image, because the change of gray level of the image shows certain rules. Texture shows the spatial information of image internal structure and pixel distribution [20-21]. In this paper, gray co-occurrence matrix algorithm is used to extract image texture features.

For a complete image, it can be seen that there are \( n \) points in the horizontal direction, represented by \( 1, 2, 3, \ldots, n \) and \( N \) respectively. There are \( m \) points in the vertical direction, represented by \( 1, 2, 3, \ldots, m \) and \( M \) respectively. There are \( l \) levels of pixels, \( 1, 2, 3, \ldots, l \) respectively. \( I' \) is used to represent the level. Therefore, the image can be regarded as a gray matrix of \( M \times N \rightarrow I' \).

After the image is converted into gray image, the gray image is quantized according to the size of the image. If the gray level co-occurrence matrix of the quantized gray level image is \( P \), take the specific direction \( W \), and the value of direction \( W \) is 0 degree, 45 degree, 90 degree and 135 degree [22]. The texture parameters are then calculated.

Image texture energy represents the thickness of texture. The larger the energy value is, the stronger the texture vein is. The smaller the energy value is, the more detailed the texture vein is. The calculation formula of image texture energy is as follows:

\[
F_1 = \sum_{i,j} p(i, j)^2
\]  

(9)

In formula (9), \( F_1 \) is the image texture energy, \( p(i, j) \) is the probability that the gray level of the image appears in a unit distance in a specific direction from level \( j \) to level \( j \) [23]. At the same time, the moment of inertia can also represent the thickness of texture. The formula of moment of inertia \( F_2 \) is as follows:

\[
G = \sqrt{G_x^2 + G_y^2}
\]

(8)
It can be seen from the above formula that the smaller the value of $F_2$, the stronger the texture vein of the image; the larger the value of $F_2$, the more detailed the texture vein of the image. Entropy is used to express the complexity of image texture, and its calculation formula is as follows:

$$F_3 = \sum_{i,j} p(i,j) \log_2 p(i,j)$$ (11)

It can be seen from the above formula that the smaller the entropy value of $F_3$ is, the simpler the texture is, the larger the value is, and the more complex the texture is. In order to recognize the sensitive information in the image, the similarity degree of the row and column directions of the texture in the image is calculated according to the following formula.

$$F_4 = \frac{\sum_{i,j} i \times j \times \log_2 p(i,j) - \mu_x \times \mu_y}{\sigma_x \times \sigma_y}$$ (12)

In the above formula, $\mu_x$ and $\mu_y$ are mean values, and $\sigma_x$ and $\sigma_y$ are variance. To sum up, the texture features of the image can be obtained by calculating the mean and variance of $F_1$, $F_2$, $F_3$ and $F_4$. After extracting the features of the image, the sensitive information contained in the image features is retrieved.

2.3 Image sensitive information retrieval

Image sensitive information can be divided into illegal bad information, involving personal privacy, images containing exposure behavior and text embedded in the image. Therefore, according to the image features extracted above, the image to be recognized is roughly divided into two categories, including text and not including text, and sensitive information retrieval is carried out respectively.

For the image sensitive information retrieval without text, the fusion convolution neural network and image color feature algorithm are used first, and the specific process of the algorithm is shown in the Figure 1.
Based on the color features extracted in the above section, in order to further characterize the sensitive information in the image, this paper uses the first-order moment, second-order moment and third-order moment of color to further characterize the color features. After the image is transformed from RGB space to HSV space, the first-order moment, second-order moment and third-order moment can be used as elements in a row of three column vectors to represent the color moment vector of a certain channel of the image.

The first-order moment value is the mean value of all elements of a channel matrix, the second-order central moment is the pixel variance of the image channel, and the third-order moment is the cube root of the ratio of the difference cubic matrix value of the channel to the total number of image pixels. Three channel vectors in HSV space are sequentially spliced, and the vector of image color moment is finally described as a vector with 1 row and 9 columns. Thus, the expression of normalized color feature vector of HSV spatial model is obtained as follows:

$$ R_i = \frac{V'_i}{\sqrt{\sum_{j=1}^{n} V''_i}} $$  \hspace{1cm} (13) 

Where, $i$ represents the number of elements in the color feature vector, $R_i$ represents the
The retrieval of image sensitive information containing text is mainly carried out by the texture width and color feature of a certain part of the image. In general, the text area in the image has relatively stable texture width and relatively uniform color features. According to the image features extracted above, the image containing text is screened. Sensitive information character detection based on context content is carried out for the screened images.

In order to save the operation cost of this method, the convolutional neural network is used to detect the sensitive characters in the image. The sample set of sensitive characters was established, which was used to train CNN and determine the parameters of the convolutional neural network. Assuming the training set is \( \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\} \), the neural network parameters are trained by the gradient descent method. The loss function of the whole sample set is defined as \( J(W, b) \), and the local minimum value is obtained by using the gradient iteration method, i.e.

\[
\begin{align*}
W^{(i+1)}_j &= W^{(i)}_j - \alpha \frac{\partial}{\partial W^{(i)}_j} J(W, b) \\
b^{(i+1)}_j &= b^{(i)}_j - \alpha \frac{\partial}{\partial b^{(i)}_j} J(W, b)
\end{align*}
\]  

(14)

Where, \( \alpha \) is the learning rate of the convolutional neural network. After determining the parameters of the convolutional neural network, the character detector is used for sliding detection, and the region containing text is extracted from the image. The similarity between the detected characters in this region and the elements in the sample set of sensitive information is calculated. If the similarity is greater than 0.5, the character sensitive information in the image is determined to be retrieved. If the similarity is less than 0.5, the sensitive information of character class is not detected. After the sensitive information in the image is detected, it is further processed in the classifier to complete the image sensitive information recognition.

2.4 Image sensitive information recognition based on machine learning
In this paper, the SVM principle in machine learning is adopted to design an image classifier and realize the recognition of image sensitive information. The kernel function and penalty coefficient are the core parts of SVM. The kernel function is to map the linear indivisible vectors of low dimension to the linear separable vector space of high dimension. Penalty factors can be adjusted accordingly when the samples are misclassified to reduce the case of misclassification, so the generalization ability of the classifier should be guaranteed.

SVM first performs classification calculation in low-dimensional space, then maps low-dimensional space to high-dimensional feature space, constructs optimal hyperplane in high-dimensional feature space, and separates data. The conversion process from 2d dataset to 3d space is shown in Figure 2.

Fig. 2 Spatial dimension transformation diagram

The final expression of the SVM can be converted into the following form.

\[ f(x) = \sum_{i=1}^{n} \alpha_i y_i x_i \cdot x + b \]  \hspace{1cm} (15)

In the above equation, \( x_i \) is the vector in the dataset, \( \alpha_i \) is the Lagrange multiplier corresponding to dataset \( x_i \), and \( y_i \) is the decision character of \( x_i \). In the process of data set classification, the following classification function is constructed.
Function $\phi(x)$ converts a dataset from two dimensions to three dimensions. The sensitive information in the image contains multiple categories, so it is necessary to design multiple classifiers for classification. According to the idea of binary tree, the multi-classifier as shown in the figure below is constructed. After $m-1$ classification divisions, the required $m$ categories can be identified.

\[
 f(x) = \sum_{i=1}^{m} w_i \phi_i(x) + b
\]  

(16)

Fig. 3 Binary tree multi-classifier structure

The image sensitive information feature set was used to train the multi-classifier, and the classification parameters of SVM were obtained after the training. Input the image to be recognized, and after the above processing, calculate the similarity between the image and the feature vector of the sample set in the SVM. If not, the classification is stopped to realize the recognition of the sensitive information of the image. So far, the research on the image sensitive information recognition method based on machine learning has been completed.

3 Experiment

In order to verify the effectiveness of the image sensitive information recognition method based on machine learning studied in this paper, comparative experiments will be conducted to evaluate the practical application effect of the research method in this paper by analyzing the experimental results.

3.1 Experimental contents
In the form of comparative experiment, the comparison group of the experiment is the image recognition method based on principal component analysis and the image recognition method based on template matching, and the test group is the image sensitive information recognition method based on machine learning studied in this paper. The comparison index of the experiment is the precision and recognition time of three different recognition methods, which are used to evaluate the advantages and disadvantages of the three recognition methods. In order to ensure the scientific validity of the experimental results, unique experimental variables were controlled during the experiment to complete the experimental verification.

3.2 Experimental preparation

The experiment was completed on three computers with the same configuration. The experiment was divided into three groups according to the data type for experimental verification. The specific parameters of experimental data are shown in the following table.

<table>
<thead>
<tr>
<th>Group</th>
<th>The data type</th>
<th>Total data</th>
<th>Quantity of sensitive information</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>The image contains clear text</td>
<td>15680</td>
<td>5372</td>
</tr>
<tr>
<td>A2</td>
<td>Images do not contain text</td>
<td>22134</td>
<td>5341</td>
</tr>
<tr>
<td>A3</td>
<td>Image information blur</td>
<td>10854</td>
<td>4375</td>
</tr>
</tbody>
</table>

Three identification methods were used to identify image sensitive information from different experimental data, and the accuracy of the three methods was recorded. The experimental results are as follows.

3.3 Experimental results

The experimental results are shown in the following table. By analyzing the data in the table, relevant conclusions of this experiment are drawn.

<table>
<thead>
<tr>
<th>Experimental data number</th>
<th>Method based on principal component analysis</th>
<th>Template matching based approach</th>
<th>This method</th>
<th>paper's method</th>
</tr>
</thead>
<tbody>
<tr>
<td>A11</td>
<td>54.70</td>
<td>68.10</td>
<td>90.30</td>
<td></td>
</tr>
<tr>
<td>A12</td>
<td>62.10</td>
<td>62.30</td>
<td>82.40</td>
<td></td>
</tr>
<tr>
<td>A13</td>
<td>67.60</td>
<td>66.50</td>
<td>83.80</td>
<td></td>
</tr>
<tr>
<td>A21</td>
<td>54.90</td>
<td>57.50</td>
<td>84.20</td>
<td></td>
</tr>
<tr>
<td>A22</td>
<td>62.70</td>
<td>65.20</td>
<td>86.50</td>
<td></td>
</tr>
<tr>
<td>A23</td>
<td>63.50</td>
<td>56.30</td>
<td>85.60</td>
<td></td>
</tr>
<tr>
<td>A31</td>
<td>61.60</td>
<td>62.40</td>
<td>82.40</td>
<td></td>
</tr>
</tbody>
</table>

Table 2 Accuracy rate(%) of sensitive information by different identification methods
It can be seen from the above table that the accuracy of the research method in this paper is higher when different experimental data are identified as image sensitive information. In the comparison group, the accuracy of the image recognition method based on principal component analysis is lower than that based on template matching. During the calculation, the average accuracy of the three identification methods was 61.5% for the identification method based on principal component analysis, 63.7% for the identification method based on template matching, and 85.83% for the identification method in this paper.

On the basis of the above experiments, compare the recognition time of different methods, and the results are shown in Figure 4.

![Comparison of recognition time of different methods](image)

**Fig. 4** Comparison of recognition time of different methods

Analysis of the above figure shows that the recognition time of the recognition method based on principal component analysis from 4.0s to 4.5s, and the recognition time of the recognition method based on template matching changes from 1.9s to 2.7s. The recognition time of the method in this paper is always less than 1.0s, with short recognition time and higher efficiency.
In conclusion, this method can realize the fast and accurate recognition of image sensitive information.

4 Conclusions

In order to solve the problem of low accuracy in traditional image sensitive information recognition methods, a new image sensitive information recognition method based on machine learning is proposed. The pre-processing of denoising and detail sharpening is carried out for the recognition image, and the image sensitive information is retrieved by combining the extracted image features. Based on the principle of SVM in machine learning, an image classifier is designed to realize the classification and recognition of image sensitive information. The simulation results show that this method has higher accuracy, better performance, better practical application and reliability, which provides a new idea for the further development of image sensitive information recognition technology. In the future, with the strengthening of the national government's supervision on the network information security, it is necessary to study the issue that many illegal institutions and individuals use image to release sensitive information, so as to further curb the image transmission with sensitive information, so as to purify the network environment, reduce the transmission of sensitive information on the physical and mental health of young people and social harmony adverse effects of stability.

5 Fund projects

science and technology plan of Qinghai province(Key Research&Development and conversion plans 2019-GX-170)

project name:Development of resource library of duixiu art Image digital protection in Huangzhong County, Qinghai Province

References


High speed linear array CCD image data acquisition system based on machine learning

LI Kai-yong
{WEIYI220@tom.com}
(Qinghai Nationalities University Physics and Electronic Information Engineering, xining, 810007 China)

Abstract: A high speed linear CCD (Charge Coupled Device) image data acquisition system based on machine learning was designed to solve the problem of poor signal stability of traditional high speed linear CCD image data acquisition system. Set up the overall system architecture. On this basis, complete the system hardware design through CCD sensor, A/D converter, CCD drive controller, RAM (Random Access Memory) read-write controller and PC; Through the CCD driver module, CCD signal processing module, image information reading module, image cache module, image information conversion mode and image information display module, the system software design was completed. Thus, the design of high-speed linear array CCD image data acquisition system based on machine learning was completed. Compared with the traditional high-speed linear CCD image data acquisition system, the experimental results show that the proposed high-speed linear CCD image data acquisition system based on machine learning has higher signal stability.

Key words: machine learning; CCD; Image data acquisition system;

1 Introduction

With the progress of science and technology, digital image acquisition and processing technology has been more and more applied in various fields. Traffic monitoring, dynamic target tracking and capture, robot navigation and other applications have also promoted the rapid development of real-time image processing technology. There are many methods to scan and identify the shape of high-speed moving objects. These methods firstly use CCD, CMOS (Complementary Metal Oxide Semiconductor) and other array photodetectors to image the objects, and then combine different data processing methods to process the collected images and store the data, so as to obtain relevant information of the measured objects [1-3]. With the advent of the information age, image sensors are increasingly used in scientific research and production. Among many image sensors and thermoelectric image sensors, CCD image sensor is widely used in high-precision measurement, detection, image analysis, spectrum detection and other fields for its excellent performance. CCD sensor has the characteristics of high precision, high resolution, stable performance, low power consumption, long life and self-scanning function, etc., and has been widely used in automatic and precise measurement [4-6].
As a new artificial intelligence technology, machine learning algorithm can analyze and store massive data and make intelligent decisions on complex problems for reference of technicians. Machine learning algorithms can simulate the process of human thinking, learning and creation. Deep learning method is the latest method in the field of machine learning. Based on artificial neural network, this new feature learning method uses deep (multi-layer) neural network structure to build models. In combination with effective parameter adjustment and optimization methods, it has achieved very good data processing results [7]. This kind of machine learning model and algorithm with multi-layer network structure is called deep learning method.

Image recognition is an early application of deep learning and a major breakthrough has been made [8]. In deep learning approach, there is a model for a multidimensional data such as image pattern recognition has excellent performance, around the connection of multilayer neural networks, this structure makes it easier to training and has better generalization performance, this is the convolutional neural network, its main characteristic is the network structure contains a large amount of convolution and pooling layer, through multiple convolution and pooling layer overlay in turn, can carry on the step by step to the complex data of feature extraction, combined with abstraction, thus learn more conducive to classification of high-level character description.

Based on the above analysis, a high speed linear array CCD image data acquisition system based on machine learning is designed.

2 High speed linear array CCD image data acquisition system based on machine learning

2.1 Overall system architecture

The overall architecture of the high-speed linear array CCD image data acquisition system based on machine learning is shown in Figure 1.
According to the overall architecture of the system, the hardware part and software part of the system are designed.

### 2.2 System hardware design

The hardware design of high-speed linear array CCD image data acquisition system based on machine learning is shown in Figure 2.
electrical signal that is convenient for processing [9-10]. Xilinx's FPGA (Field Programmable Gate Array) chip generates CCD driver timing pulses to generate, store, transfer and output the signal charge, and the light intensity distribution detected by the CCD will be reflected in the video signal output by the CCD. The output video signal of the linear tcd252d contains not only the effective charge signal, but also various noise clutter invalid signals. Three signal processing circuits are designed, including signal amplitude amplification, noise suppression and negative polarity reversal. Linear CCD outputs three parallel and independent analog video signals. A high-speed data acquisition system composed of three-channel high-speed analog channel selection, analog-digital conversion unit, dual-port RAM data cache unit, logic control unit and Xilinx integrated logic analyzer Chipscope Pro is designed to realize the acquisition, transmission and display of CCD output signals.

In the whole CCD data acquisition system, there is little need for FPGA I/O, and only the logical unit and block RAM storage capacity is required to meet the collected video data. Therefore, on the premise of meeting system requirements, low-cost chips are selected as far as possible. The FPGA used in this system is the low-cost XC3SSO0E programmable logic device of Xilinx company's spartan-3 E series. On-chip clock manager (DCM) provides advanced clock control functions for FPGA applications. Level standard: 4 I/O Banks, nuclear power voltage: 1.2v, port voltage: 3.3v, 2.5v, 1.2v; Support for 18 single-ended and differential I/O standards; Driving current up to 16mA; Support DDR storage interface, transmission rate can reach 622Mbit/s: support Xilinx Platform, BPI Flash(multi-boot), SPI Flash, JTAG configuration.

In order to make the light intensity of the object detected by CCD image better on the sensitive surface, it is necessary to add an optical lens or imaging objective lens in front of CCD to greatly improve the imaging quality. The imaging objective has a combination of fixed focal length, zoom, manual diaphragm and automatic diaphragm. In order to detect the object information, the user can calculate the focal length of the lens according to the distance and size of the detected object, so the appropriate CCD imaging objective lens should be selected according to the actual application environment. Image sensor is the core device of the whole image data acquisition system, which directly determines the performance of the system. TCD2252D is a kind of high sensitivity, low dark current three-wire, two-phase, double-channel color linear CCD camera device, with color filter inside, signal sharing, green and blue output, is a DIP22 packaging form of double-line directly inserted device. Each pin parameter setting is shown in Table 1.

<table>
<thead>
<tr>
<th>Pin</th>
<th>Symbol</th>
<th>Function description</th>
<th>Pin</th>
<th>Symbol</th>
<th>Function description</th>
</tr>
</thead>
</table>

Table 1  Pin parameter setting
<table>
<thead>
<tr>
<th>nu</th>
<th>mb number</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OS2</td>
</tr>
<tr>
<td>2</td>
<td>OS3</td>
</tr>
<tr>
<td>3</td>
<td>SS</td>
</tr>
<tr>
<td>4</td>
<td>NC</td>
</tr>
<tr>
<td>5</td>
<td>RS</td>
</tr>
<tr>
<td>6</td>
<td>φ₂B</td>
</tr>
<tr>
<td>7</td>
<td>SS</td>
</tr>
<tr>
<td>8</td>
<td>φ₂A₂</td>
</tr>
<tr>
<td>9</td>
<td>SH3</td>
</tr>
<tr>
<td>10</td>
<td>φ₁A₂</td>
</tr>
<tr>
<td>11</td>
<td>SH2</td>
</tr>
</tbody>
</table>

Inside the pin is a photosensitive diode containing 3 rows of 2700 effective pixels. The minimum pixel size is 8 m, the total length of the photosensitive array is 21.6mm, and the spacing between adjacent photosensitive columns is 64m. The device works under the condition of SV drive pulse and 12V power supply. The typical clock pulse frequency is 0.5mhz and the maximum drive frequency can reach 2MHz. At room temperature, typical sensitivity values of the three RGB arrays are 7.0v /lx·s, 9.1v/lx·s, and 3.2v/lx·s, respectively. The saturation exposure is 0.35lx·s. The typical saturation output voltage is 3.2V, which is composed of photodiode array, phase-shifting grid sh, two-phase CCD shift register, drive pulses (U 1, u 2) and output mechanism. In the camera area of the device, it has 2,700 effective photosensitive elements, carrying effective information, and 62 masked photodiodes are used to obtain dark current and other information for effective signal processing. D0-d12, which is not shown in the figure, is an imaginary unit, that is, no physical units exist, but there are 13 groups of corresponding units in the shift register. On both sides of the camera area are CCD shift registers, which are not sensitive to light, but only accept the transferred charge packet from the camera area and are separated from the photosensitive array. The two are
separated or communicated by the transfer pulse SH added to the transfer gate. At this point, complete the system hardware design.

2.3 System software design

On the basis of hardware facilities, the system software is designed. The software design diagram of high-speed linear array CCD image data acquisition system based on machine learning is shown in Figure 3.

![System software block diagram](image)

In the software system, the CCD driver module: the CCD module mainly realizes the image information collection and obtains the image simulation information. The CCD control driver performs this function by controlling and operating the CCD controller to drive the CCD chip. Since the main control logic of the CCD chip is implemented by the CCD controller, the CCD driver only needs to complete the simple initialization of the CCD controller. The initialization of CCD controller is divided into two aspects: sampling clock setting and moving pulse SH interrupt setting. CCD device converts charge signal into discrete voltage signal output, which includes a lot of noise from different places, including the noise caused by power supply and the noise generated in the acquisition process. The purpose of noise signal processing is to reduce noise and interference, improve signal-to-noise ratio, get high quality pixel signal and improve the detection accuracy. The output noise of CCD is analyzed and studied, and the analog signal of CCD output is denoised, and then the signal is output to ADS805 chip for analog-to-digital conversion. Uses a piece of drive to provide such a load current, using a transistor constant current source and op-amp composition, through the control the size of the output current in the collector of a transistor, according to the size of the R24 resistance to change the size of the transistor current to control the CCD output side of the collector is connected OS current, provide output signal OS need to load current.
CCD signal processing module: the output noise of CCD mainly consists of two parts, one is the noise generated by the photoelectric conversion inside the chip, and the other is the interference noise generated by the output process. Here are some common noises:

Firstly, granular noise. The process of the signal charge generated by the CCD device receiving the light is a random process. The signal charge generated by CCD device fluctuates up and down in a certain value in unit time, resulting in particle noise. However, in the actual use process, the scattered noise is inevitable, and can not be filtered out through subsequent improvement, so the scattered noise plays a decisive role in the size of the limit noise of CCD device.

Secondly, dark current noise. CCD devices can not receive photons or there is no other way to charge injection case, there will still be charge generation, thus forming a dark current, its generation process is also random. Dark current noise is mainly affected by temperature. For every 5-6 °C rise in temperature, the thermally excited noise will double. The existence of dark current noise reduces the dynamic range and sensitivity of CCD devices.

Thirdly, transfer noise. The transfer noise refers to the noise interference caused by the residual charge in the potential well transfer of the charge in the CCD potential well on the transmission of the next charge packet. Now the charge transfer efficiency of the CCD device is very high, basically reaching 99.99%, so the influence of the transfer noise on the signal noise can be ignored in general.

Fourthly, output noise. Output noise refers to the noise generated by the reset timing of the output circuit in the process of CCD photoelectric conversion. In the process of signal output, due to the reset pulse RST arrived ahead of schedule, the floating capacitor charging and discharging, waiting for the output charge transfer to the output side, in the case of CCD working frequency, capacitor charging and discharging speed to keep up with, cause incomplete discharge, which affects the output signal, the noise and called the reset noise, reset noise in the signal phase with reset signal is synchronized.

In view of the above noise, it is treated separately. CCD noise processing mainly deals with transfer noise and output noise. In order to suppress noise and interference, the following measures are taken:

In order to reduce the noise of DC power supply, two 10 frequency f and 0.1 frequency f filter capacitors are added to the input and output of the power supply, respectively, in order to reduce the noise generated during the transmission of DC power supply. In circuit design, the analog and digital ground is separated and connected by a single magnetic bead to reduce the interaction between digital noise and analog signal.
The simplest way to deal with the output noise is to use a low-pass filter for filtering. However, due to the limitation of its effect, the correlation double-sampling method is usually used to filter the output noise. Correlated double sampling is to like yuan signal sampling and reset noise, reset noise sampling is ahead of time, after the reset pulse signal pixels, sampling is in the charge packet arrival time, because the two sampling reset noise has always been there, in the two sampling results transformation, make a difference reset noise will be greatly weakened. When the CCD is reset, the clamping signal is set to a low level. When the CCD begins sampling, the output signal is kept at a low level for a short time before the CCD signal is output. The sampling clock of the box sampling method is related to the sampling interval of the CCD device and the phase of the reset signal, which makes the method more suitable for the processing of CCD output noise.

Image information reading module: in this module, the convolutional neural network degree in machine learning is used to extract image features. The multi-scale convolutional neural network is composed of three CNN model architectures in parallel. Each CNN model architecture consists of three stages. In the first two stages, a convolution kernel group is respectively included to generate a dense feature map, and a point-to-point nonlinear mapping is also included. In addition, after each feature map, a spatial pooling operation is connected for lower sampling. The final stage consists of only one set of convolution nuclei to produce the final set of feature graphs. The convolution kernel used in the model is also a weight matrix, which will be obtained by training. Each convolution kernel is applied to the input feature graph by using 2-dimensional convolution operation, and local features can be detected at all positions of the input. The convolution kernel groups vary as much as the input, but the rest of the network remains unchanged. The input of the multi-scale convolutional neural network is different images in the multi-scale pyramid, and the multi-scale hierarchical feature is expressed by connecting the output of all CNN networks f. A linear classifier is added to the model to learn the features and guide the training process of feature extraction, so that the multi-scale convolutional neural network can correctly predict the classification of all pixel positions in the image through the multi-scale hierarchical feature expression extracted after training. The sampling information reading program is mainly responsible for reading the image digital information output by CCD image processing chip. The image sampling information output by CSP chip AD9822 is written into the data cache module FIFO in high and low bytes. When the data stored in the cache reaches half of the cache capacity, the data cache module will notify the Nios II processor to read the data in the cache module by means of interrupt triggering through the output half-full control signal halffull. Therefore, image information reading is realized by FIFO interrupt processing.
Image cache module: Initialize the FIFO interrupt. The interrupt initialization of FIFO of cache module completes the two parts of interrupt initiation and interrupt registration. The interrupt open call function alt_irq_enable(AD9822_ALFFULL_IRQ), AD9822_HALFFULL_IRQ is the interrupt number allocated for the cached half-full signal.

In addition, the interrupt initialization of FIFO also calls the PIO register operation functions IOWR_ALTERA_AVALON_PIO_EDGE_CAP(AD9822_HALFFULL_BASE, 0x0) and IOWR_ALTERA_AVALON_PIO_IRQ_MASK(AD9822_HALFFULL_BASE,0xff) to clear the interrupt flag in the edge capture register and open the interrupt enable bit in the interrupt mask register. Interrupt registration call function Alt irq register(AD9822_HALFFULL_IRQ, 0, handle_fifo_interrupt), where handle_fifo_interrupt is an interrupt service program that reads the cached image information in FIFO.

Then, through FIFO interrupt service program to realize the main program of image information reading. Figure 4 shows a flowchart of interrupt service program processing.

**Fig. 4** Flow chart of interrupt service program processing

The FIFO interrupt handler first closes FIFO and clears the break flag to avoid nested calls to the FIFO interrupt service program. The image data in the FIFO is then read, four bytes at a time. In the process of reading, update the statistics of the read data, and determine whether the current data is a dummy unit, if so, discard it, otherwise write the globally defined image information array in order. When all valid data in FIFO are read out (28*4=1024 bytes),
the operation of image information reading is finished. Finally, start the FIFO half-full interrupt and exit the FIFO interrupt service program.

Image information conversion module: the main function of the image information processing program is to convert the read image data into LCD display format. When the CCD completes a round of image information collection, and the processor also transfers the effective image data to the image information array, the program will call the image information processing program to convert the data in the image information array. The conversion of image information data mainly considers the number of LCD horizontal or vertical display units and CCD effective image data. LCD adopts horizontal display to output the collected image data, so the number of display units per row is 320. Now it is necessary to output 10800 valid data collected by CCD to LCD display. In order to ensure uniform display of acquisition information, 320 pixel information is selected from 10800 valid data at an interval of 10800/32 points for display. After the data in the image information array is processed, it is stored in the global display information array.

Image information display module: image information display includes two parts: image statistics display and image data display. The image statistics display mainly displays the number of frames and the number of image data that have been collected. The image data display mainly transfers the data in the display information array to the LCD for display. Void LCD_Display_OneLine(unsigned char* lineLCDDate, int line), where lineLCDDate represents the starting address of the display row data, and line represents the line number displayed. The flowchart of its line display function is shown in Figure 5.
Fig. 5 Shows the flow chart of the function

Figure 5 describes the flow of the line display function LCD_isplay_OneLine(). Before writing data, the row display function first adjusts the current row/field display position, which is set to the line start address HSA set at initialization, and the field address to the line number line. After setting the display location, the row display function writes the lineLCDDate data in turn to the GDDRAM inside the LCD until it fills a line with hea-hsa +1. After displaying one line of sampled data, the image information display function will update the current line number, and wait for the next line to display the new display data after processing.

So far, according to the overall framework of the system, through hardware design and software design, the design of high-speed linear array CCD image data acquisition system based on machine learning is completed.

3 Experiment

The proposed high speed linear array CCD image data acquisition system based on machine learning is compared with the traditional high speed linear array CCD image data acquisition system to verify whether the proposed high speed linear array CCD image data acquisition system based on machine learning has more stable performance.

3.1 Experimental process

In order to ensure the reliability of the system drive signal, in addition to observing the simulation waveform, we should also check the actual output waveform in the oscilloscope.
Where, the actual driving signals of RS, CP and SP have the same frequency, and there is a certain phase difference between the three pulse signals, which conforms to the requirements of TCD2252D for these three driving signals. In general, the driving signals generated by FPGA all meet the requirements of TCD2252D for driving signals.

After debugging the system successfully, the stability of the signal of the proposed high-speed linear CCD image data acquisition system based on machine learning is compared with that of the traditional high-speed linear CCD image data acquisition system.

3.2 Analysis of experimental results

The signal stability comparison results of the proposed high-speed linear array CCD image data acquisition system based on machine learning and the traditional high-speed linear array CCD image data acquisition system are shown in Figure 6.

![Comparison results of signal stability](image)

**Fig. 6** Comparison results of signal stability

As can be seen from Figure 6, with the change of driving time, the fluctuation range of the signal amplitude of the traditional high-speed linear array CCD image data acquisition system is unstable, indicating that its signal is unstable. The wave amplitude of the high speed linear array CCD image data acquisition system based on machine learning is stable, indicating that the acquisition signal is stable. Through analysis, it is found that the proposed high-speed linear CCD image data acquisition system based on machine learning, based on machine learning, improves the system performance stability, compared with the traditional high-speed linear CCD image data acquisition system, its signal has higher stability.
4 Conclusions

A high speed linear array CCD image data acquisition system based on machine learning is designed to solve the problem of unstable acquisition signal in traditional high speed linear array CCD image data acquisition system. Compared with the traditional high-speed linear CCD image data acquisition system, the experimental results show that the signal stability of the proposed high-speed linear CCD image data acquisition system based on machine learning is higher, which is expected to provide certain reference value for the research of high-speed linear CCD image data acquisition system. However, there is still a long time for signal acquisition in this system. In the next research, we will focus on improving the acquisition system and improving the system efficiency.

5 Fund projects

science and technology plan of Qinghai province(Key Research&Development and conversion plans 2019-GX-170)

project name:Development of resource library of duixiu art Image digital protection in Huangzhong County, Qinghai Province

References


Automatic detection method of boundary data of
gemstone bearing image based on Fourier transform

LI Cai-hong¹, ZHU Xiao-Jun²

{lgf147852@tom.com¹, wz20160401@163.com²}

¹.Changchun Institute of Technology, Changchun 130021, China;
².Nantong Normal College, Nantong 226010, China

Abstract. At present, the automatic detection method of boundary data of gem bearing image has the defect of poor detection effect. Therefore, an automatic detection method of boundary data of gem bearing image based on Fourier transform is proposed. Based on infrared imaging technology, the image of gem bearing is obtained, and the image is segmented by Fourier transform algorithm. On this basis, the boundary information of image area is judged by information measurement. According to the result of boundary information measurement of gem bearing image, the boundary data of gem bearing image is extracted by log operator, and the boundary data of image is detected by ant colony algorithm. The automatic detection of boundary data of gem bearing image is realized. The experimental results show that compared with the traditional automatic detection method of gem bearing image boundary data, the proposed method greatly improves the detection effect, which fully shows that the proposed automatic detection method of gem bearing image boundary data has better performance.

Keywords: Fourier transform; Gemstone bearing; Image; Boundary; Testing;

1 Introduction

Gem bearing is a kind of sliding bearing made of gem and other hard materials. A natural or synthetic gem (usually sapphire) can be used as the bearing of the rotating shaft of the meter head or the bearing of other precision instruments by carefully grinding out a conical cavity [1].

Gemstone bearings are mainly used in instruments. The bearing of the instrument bears little load, but it needs high rotation accuracy, good sensitivity and long service life. Gemstone has the characteristics of small friction coefficient, high hardness, corrosion resistance, small thermal expansion coefficient, high compressive strength, and can meet the use requirements of instrument bearing. The materials for making gem shaft include corundum, agate, glass ceramics, etc. corundum is the main body of alumina. They are natural and artificial. Natural corundum has many impurities and uneven texture, so it is widely used in man-made corundum.

Gemstone bearing has certain technical requirements, for example, the conicity of the
outer circle of the groove gemstone bearing shall not exceed half of the allowable deviation of its outer diameter; the parallelism of the two end faces of the groove gemstone bearing shall not exceed half of the allowable deviation of its height; the perpendicularity of the outer circle of the groove gemstone bearing to the bottom surface shall not be greater than 0.005; the coaxiality of the groove gemstone bearing and the outer circle shall not be greater than 0.02mm, the spherical bearing is not more than 0.03mm; the groove bus of the tapered gem bearing is within 1 / 3 of the groove depth from the bottom; the groove bus of the spherical corundum bearing is within 2 / 3 of the groove depth from the bottom, etc.

Gem bearing needs high precision, especially the boundary precision of gem bearing. Therefore, higher requirements are put forward for gem bearing detection. Nowadays, infrared imaging technology is mainly used to obtain image boundary data of gemstone bearing, on this basis, to detect whether the gemstone bearing meets the standard. In reference [2], a method of automatic segmentation of SAR image target is proposed. Many algorithms can solve the problem of target segmentation in SAR image, one of them is the grabbing and cutting algorithm. The grab cut algorithm based on graph theory realizes the optimal image segmentation, and transforms the image segmentation problem into the problem of the maximum flow in the computational flow network. However, there are some defects in the automatic detection of boundary data of gem bearing image. In order to solve these problems, an automatic detection method of boundary data of gem bearing image based on Fourier transform is proposed.

Fourier transform is to use a digital computer to calculate in an efficient and fast way, referred to as FFT. Fourier transform was proposed by J.W. KULI and T.W. basis in 1965. By using this algorithm, the number of times of multiplication needed for computing DFT can be greatly reduced. Especially, the more sampling points n are transformed, the more significant the saving of FFT algorithm is. Through the application of Fourier transform algorithm, it can greatly improve the detection effect of gemstone bearing image boundary data automatic detection method, and design simulation contrast experiment to verify the performance of the proposed method.

2 An automatic detection method for boundary data of gemstone bearing image

2.1 Image segmentation of gem bearing

Based on the infrared imaging technology to obtain the gem bearing image, using the Fourier transform algorithm to segment the image to prepare for the detection of the gem bearing image boundary data, the specific process is as follows [3].

The image segmentation by Fourier transform algorithm is mainly divided into two
stages: interpolation stage and segmentation stage [4].

Among them, the main operation of the difference stage is to rotate the ray and circle the square. Rotating ray is to obtain ray sampling with even angle in polar coordinate system. It is necessary to rotate ray to get grid point. Its expression is as follows:

\[ BV = \begin{cases} 
  \xi_x = \frac{\pi l}{N} \text{ for } -N \leq l < N \\
  \xi_x = \frac{\pi l}{N} \tan \left( \frac{\pi m}{2N} \right) \text{ for } -\frac{N}{2} \leq l < \frac{N}{2} 
\end{cases} \]  

(1)

Among them, \( BV \) is interpolation grid point; \( \xi_x \) is Cartesian grid vertical frequency; \( l \) is gem bearing image length; \( N \) is gem bearing image matrix order; \( \xi_x \) is Cartesian grid horizontal frequency; \( m \) is algorithm parameter.

The rotating ray diagram is shown in Figure 1.

Circle the square to obtain the required concentric circle, which is completed by dividing each ray by a constant along \( \xi_x \) and \( \xi_y \), based on its angle. Therefore, set an algorithm parameter \( m \),

\[ R[m] = \sqrt{1 + \tan^2 \left( \frac{\pi m}{2N} \right)} \],

and the generated grid points are expressed as follows:
To circle a positive shape is equivalent to an operation along a ray. A group of equidistant points are replaced by a group of new points along the same line in different positions (marked as small orthonormal). This time, the target points are equidistant, but there is a gap between them. The first interpolation stage is applied to trigonometric polynomials to represent the expected accuracy [5].

Circle the square diagram as shown in Figure 2.

![Figure 2](image)

**Fig. 2.** Circle the square diagram

According to the grid points obtained from the above-mentioned gem bearing image interpolation, the segmentation threshold of gem bearing image is calculated, and the calculation formula is as follows:

$$T = \frac{BV \ast BH}{\alpha^2} \quad (3)$$

Among them, $\alpha$ is the segmentation threshold of gem bearing image; $B$ is the calculation parameter of segmentation threshold.

According to the above formula, the class variance between the target area and the background area of the gem bearing image is obtained, and the expression is as follows:

$$\delta^2(T) = w_0 (\mu_0 - \mu)^2 + w_1 (\mu_1 - \mu)^2 \quad (4)$$
Among them, \( w_0 \) is the proportion of the target area occupying the gem bearing image; \( \mu_0 \) is the average gray level of the target area; \( w_1 \) is the proportion of the background area occupying the gem bearing image; \( \mu_1 \) is the average gray level of the background area.

When \( \delta^2(T) \) reaches the maximum value, the corresponding threshold \( T \) is the best threshold of gem bearing image segmentation. Based on this, the gem bearing image segmentation is realized.

### 2.2 Image boundary information measure

Based on the above segmentation results of gem bearing image, the boundary information of image region can be determined by information measure [6].

In essence, the image noise points and boundary points have obvious characteristics differences: boundary points have structural gray variation and directional gray distribution in their neighborhood; and the gray level of noise points also has variation, but does not have some characteristics of boundary points[7].

The information measure of an image is a measure of the complexity of an image in a small region \( \sigma \). Generally, \( \sigma \) is a circular region with 3-8 pixels. If the image information measure value is large, the image has boundary in area \( \sigma \); otherwise, if the image information measure value is small, the image has no boundary in area \( \sigma \).

Let the gray level of the image correspond to \( f(x,y) \), the coordinate of the current pixel point is \( (x_0, y_0) \), with point \( (x_0, y_0) \) as the center, and \( \sigma \) is the neighborhood radius, then it meets the following requirements: \( \rho \), line \( \sigma = \{ (x, y) | |x-x_0| \leq \rho, |y-y_0| \leq \rho \} \) is a straight line passing through the pixel point \( L \), and the angle is \( (0^\circ, 180^\circ) \), and area \( \sigma \) is divided into two parts, \( \sigma_1 \) and \( \sigma_2 \), as shown in Figure 3.
The image boundary information measure is mainly divided into three aspects: structural boundary information measure, directional boundary information measure and neighborhood consistent boundary information measure.

In gem bearing image, the structural feature of boundary point is a significant difference between boundary point and non boundary point. The structural boundary information measure of point \( (x_0, y_0) \) is defined as:

\[
C(x_0, y_0) = \max_{x,y} \sum g(x,y) \left( \frac{\partial f(x,y)}{\partial x} \right)^2 + \rho \left( \frac{\partial f(x,y)}{\partial y} \right)^2
\]

Where \( g(x,y) \) is the gradient amplitude of point \( (x,y) \), the calculation formula is:

\[
g(x,y) = \sqrt{\left( \frac{\partial f(x,y)}{\partial x} \right)^2 + \left( \frac{\partial f(x,y)}{\partial y} \right)^2}
\]

The directivity of the boundary point is one of the remarkable characteristics of the boundary point. The expression of the directivity information measure of point \( (x_0, y_0) \) is:

\[
O(x_0, y_0) = \max \left| f_{x_0} - f_{x_j} \right| - \min \left| f_{x_0} - f_{x_j} \right|
\]

The gray distribution of boundary points and non boundary points of an image is different in their respective neighborhoods. The boundary of the image divides the neighborhood of the boundary point into two regions with different gray levels. The gray distribution of non boundary points in their neighborhoods is single. According to this feature, a neighborhood consistent boundary information measure is constructed [8].

Then the neighborhood consistency edge information measure of point \( (x_0, y_0) \) is:
\[ R(x_0, y_0) = \max \left\{ \left| f_{\sigma_1} - f_{\sigma_2} \right| \left[ \rho (\rho + 1) \right] \right\} \quad (8) \]

Through the above analysis, the result of boundary information measure of gem bearing image is \( Q = \{ C(x_0, y_0), O(x_0, y_0), R(x_0, y_0) \} \).

2.3 Image boundary data detection

Based on the above-mentioned measurement results of boundary information of gem bearing image, LOG operator is used to extract boundary data of gem bearing image.

The schematic diagram of LOG operator is shown in Figure 4.

\[ \text{Fig. 4. Schematic diagram of Log Operator} \]

In order to get the output value of LOG operator, Laplacian transform of Gaussian filter function and convolution operation with image are needed. Among them, the two convolution templates that log operators often use are shown in Figure 5.

(a)
2.4 Image Boundary Data Detection

According to the above extraction results, the image boundary data is detected based on ant colony algorithm, and the specific detection process is as follows.

Using ant colony algorithm to detect image boundary data, the first problem to be solved is how to transform the problem of image boundary data detection into a mathematical model which can be detected by ant colony algorithm. In the process of image boundary data detection, the gray gradient of image pixels is mostly selected as the heuristic information of ants. The basic idea of transformation is to use ant colony algorithm to transform image boundary data detection problem into combinatorial optimization problem. Images are considered maps with many pixels. Each pixel is the node that ants choose. From a certain node, ants can move in the neighborhood of pixels. According to the pheromone intensity and heuristic guide function of neighborhood pixels, ants select the pheromone concentration and heuristic guide function, take the point with the maximum calculated transfer probability as the node to be selected in the next crawling, release the pheromone on the previous node, and update the pheromone matrix with the pheromone formula, so the pheromone concentration on the boundary data is significantly higher than other nodes Point, so that most ants can quickly find the boundary data of the image [9].

The steps of detecting image boundary data based on ant colony algorithm are as follows:

Step 1: set the initial value of basic information.

In the detection image of $M \times N$, $m$ ants are randomly distributed. The size of $m$ is the square root of the number of pixels in the image, i.e. $m = \sqrt{M \times N}$. Set the total number of iterations $Z$ and $q_{\text{max}}$ of ant cycle, and the initial value of pheromone $\tau^0 = 0.0001$. In order to ensure the efficiency of the algorithm, if $\tau_j \geq T$, then $V(i, j)$
is the image boundary data, otherwise it is not.

Step 2: ant B selects the rules of the next node.

In the algorithm, each node moved by ant \( k \) is considered as an iterative process. Therefore, the total number of times of ant \( k \)'s movement \( q \) is the total number of iterations of ant \( k \)'s movement. In the iterative process of ant \( k \), the transfer probability function from the current node \((i, j)\) to the next node \((n, m)\) is:

\[
P_{(i,j),(n,m)}^{k} = \begin{cases} 
\left[ \frac{\tau_{ij}^{(n-1)}}{\sum_{(v,j)} \left[ \frac{\tau_{ij}^{(n-1)}}{\eta_{ij}} \right]^\beta} \eta_{ij} \right]^{\gamma}, & (i, j) \in allow_{ij}^k, \ (9) \\
0, & \text{Other}
\end{cases}
\]

Among them, \( \tau_{ij}^{(n-1)} \) is the pheromone value of the ant at the end of the \( n-1 \) iteration at node \((i, j)\); \( \beta \) is the pheromone influence factor, whose size has a certain impact on the probability of ant path selection; \( \eta_{ij} \) is the heuristic guidance function of the ant at node \((i, j)\), which is determined by the size of the gray value of the 8 neighborhood of the node; \( \gamma \) is the influence factor of the heuristic function, which is determined by the size of the gray value of the node. The size of ants has an effect on the probability of selecting neighborhood points with high gradient value.

Step 3: update rule of pheromone.

The pheromone value of each node in the gem bearing image needs to be updated locally and globally.

When ant \( k \) completes a move, it needs to update the pheromone locally, that is, after ant \( k \) moves to the next node \((i, j)\), it needs to update the pheromone value of node \((i, j)\). The calculation formula is as formula (10); otherwise, it will not be updated.
\[
\begin{align*}
\tau_{ij}^n &= (1 - \rho) \cdot \tau_{ij}^{(n-1)} + \rho \cdot \Delta \tau_{ij}^k \\
\Delta \tau_{ij}^k &= V_e(I_{(i,j)}) / C 
\end{align*}
\]  

(10)

Among them, \( \tau_{ij}^n \) is the concentration of pheromone on node \((i, j)\) after ants perform \( n \) searches; \( \rho \) is the volatility coefficient of pheromone, with a value of 0.076; \( \Delta \tau_{ij}^k \) is the increment of node \((i, j)\) pheromone of the \( k \) ant after the current cycle.

When all ants complete a cycle, the pheromone values of all pixels in the image shall be calculated according to formula (11).

\[
\tau_{ij}^n = (1 - \psi) \cdot \tau_{ij}^{(n-1)} + \psi \cdot \Delta \tau_{ij}^k 
\]  

(11)

Among them, \( \psi \) is the decay coefficient of the whole pheromone matrix, which is 0.05.

The advantage of using global update: ant colony algorithm is an algorithm that uses probability method to calculate. In order to make ants better find the edge of the image, avoid falling into the local optimum, concentrate too much on the edge with too much pheromone content, and lose some scattered and small edge information, thus expanding the scope of ant search [10].

Through the above process, the expression of boundary data detection of gem bearing image is obtained as follows:

\[
X_i = \frac{\sum T * Q}{P_{(i,j)(n,m)} \times \tau_{ij}^n} 
\]  

(12)

In conclusion, the automatic detection of boundary data of gem bearing image is realized, which provides a new technical support for gem bearing detection.

### 3 Test results

The above-mentioned process realizes the design of automatic detection method of boundary data of gem bearing image based on Fourier transform, but whether it can solve the problems existing in the existing methods is still uncertain, so the simulation contrast experiment is designed. The specific experimental process is as follows.

#### 3.1 Construction of Experimental Environment

In order to ensure the smooth progress of the experiment, the experimental environment is set up, as shown in Figure 6.
3.2 Selection of experimental indicators

The position accuracy, width and continuity of image boundary can be used to judge the performance of image boundary data detection. In this paper, a standard measure $F$ is introduced to evaluate the performance of boundary data detection, which is defined as:

\[
F = \frac{1}{\max(I_D, I_L)} \sum_{i=1}^{I_I} \frac{1}{1 + \zeta I_I} \quad (13)
\]

Among them, $I_D$ is the number of ideal boundary pixels; $I_L$ is the number of actual detected boundary pixels; $\zeta$ is the penalty factor between the value $[0,1]$; $I_I$ is the distance between the detected boundary point and the nearest ideal boundary point.

In the experiment, the test image of the measurement standard is used, as shown in Figure 7. The boundary information of the test image is known. Different methods are used to detect the test image. Compared with the boundary characteristics of the standard test image, the boundary detection performance index $F$ of each algorithm can be obtained. The value of $F$ can reflect the performance of boundary detection. The closer $F$ value is to 1, the better the effect of boundary data detection is.
3.3 Analysis of experimental results

The comparison of measurement $F$ is shown in Table 1.

<table>
<thead>
<tr>
<th>Number of experiments</th>
<th>Literature [2] method</th>
<th>Propose method</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-2.13</td>
<td>0.89</td>
</tr>
<tr>
<td>20</td>
<td>2.13</td>
<td>0.78</td>
</tr>
<tr>
<td>30</td>
<td>0.00</td>
<td>0.9</td>
</tr>
<tr>
<td>40</td>
<td>0.12</td>
<td>0.99</td>
</tr>
<tr>
<td>50</td>
<td>2.46</td>
<td>1.00</td>
</tr>
<tr>
<td>60</td>
<td>2.56</td>
<td>1.01</td>
</tr>
<tr>
<td>70</td>
<td>2.33</td>
<td>1.01</td>
</tr>
<tr>
<td>80</td>
<td>3.00</td>
<td>1.00</td>
</tr>
<tr>
<td>90</td>
<td>3.02</td>
<td>1.03</td>
</tr>
<tr>
<td>100</td>
<td>3.06</td>
<td>1.06</td>
</tr>
</tbody>
</table>

As shown in Table 1, the measurement $F$ of the proposed method is closer to 1, indicating that the detection effect of the proposed method is better.

The experimental results show that: compared with the traditional automatic detection method of gem bearing image boundary data, the proposed automatic detection method of gem bearing image boundary data greatly improves the detection effect, which fully shows that the proposed automatic detection method of gem bearing image boundary data has better performance.

4 Concluding remarks

Because the traditional automatic detection method of boundary data of gem bearing
image is not effective, this paper proposes an automatic detection method of boundary data of gem bearing image. This method greatly improves the detection effect and provides a new technical support for the detection of gem bearing. However, there is still room for improvement in the detection effect of the proposed method, which needs to be further optimized.

References

Signal Modulation Recognition Method based on Time-frequency Image

Yuqian Li¹, Cheng Chang², Chengzhuo Shi³, Sen Wang¹,*
{liyuqian@hrbeu.edu.cn, loeibx@163.com², scz_1990@163.com³, wangsen@hrbeu.edu.cn¹,*}

Harbin Engineering University, No.145 Nantong Street, Harbin 150001, China¹,*
China Academy of Launch Vehicle Technology, 1 South Dabongmen Road, Beijing 100076, China²
Nanjing Electronic Equipment Institute, No. 35 houbiaoying Road, Nanjing 210007, China³

Abstract. Signal modulation classification is an important technology for signal processing, in order to get higher recognition accuracy at low signal-to-noise ratios, this paper proposed a modulation recognition algorithm which combines CNN and time-frequency analysis methods. The method’s steps are as following, first, make the SPWVD transforms to the digital signals to obtain corresponding time-frequency images. Then, use image processing method to enhance the obtained images, such as image gray processing, perform the grayscale equalization and binarization the images. After that, the pictures are inputted into the CNN network for classification and recognition. The experimental results show that the recognition rate can achieve 90.44% at 0dB.

Keywords: Modulation recognition, time-frequency, CNN, image processing.

1 Introduction

Wireless communication has a fixed spectrum allocation strategy. As the requirements of wireless spectrum bandwidth continue to increase, researchers need to make full use of existing wireless spectrum resources to avoid signal problems faced by wireless networks[1]. As we all know, modulation identification is a method of optimizing spectrum allocation in cognitive radio monitoring, which plays a vital role in both military and civilian use[2], and its purpose is to provide wireless communication without almost any prior knowledge. The signals are classified and identified[3].In the military field, modulation identification is mainly used in electronic warfare to recover intercepted enemy signals and obtain corresponding intelligence. In the civil field, modulation identification is mainly used for spectrum monitoring and interference identification[4]. Spectrum resources are the most precious resources in modern wireless communications. The shortage of spectrum resources also causes some institutions or individuals to illegally occupy allocated frequency bands, which will seriously interfere with the normal communications of legitimate users and even normal users, besides its can cause security issues.

There are two main methods for signal modulation classification, one based on likelihood and characteristics. However, most likelihood methods require parameter estimation, which complicates the calculation, so methods based on easy-to-implement features have been widely used. So far, various feature-based methods[5-7] have been proposed. These features are based on spectrum analysis[8], time-frequency distribution[9], wavelet transform[10], constellation maps[11, 12], and so on.
The authors in [3] used a combination of semi-supervised learning[13] method and GAN network to modulate and identify signals. Experiments show that this method can achieve better classification performance with lower signal-to-noise ratio. In [14] used a generative adversarial network to obtain higher accuracy of classification for communication signals. The experimental results show that in the ACGAN-based data set, the classification accuracy of wireless signals is improved by 0.1~6%. In order to be able to accurately extract the individual characteristics of the radiation source in a complex wireless communication environment, a method for feature extraction of individual characteristics of communication signals based on fractal complexity is described in [15]. This method can reduce the signal noise compare the communication signals accurately.

2 Principle analysis

2.1 Short Time Fourier Transform

When the stationary signal is analyzed, the signal is recorded or observed at any time, and the Fourier transform is performed. The result obtained is same and it is a constant which is no relationship with time. When analyzing non-stationary signals, a time-frequency joint analysis method is required. The simplest and most direct one is the short-time Fourier transform (Short Time Fourier Transform, STFT). The STFT of the time signal can be defined as:

$$\text{STFT}(t, f) = \int_{-\infty}^{\infty} x(\tau)g(\tau-t)e^{-j2\pi f \tau}d\tau$$

$$=\{x(\tau)g_{t,f}(\tau)\}$$

where $\langle \bullet \rangle$ is the inner product, $t, \tau$ is frequency factor. $g_{t,f}(\tau) = g(\tau-t)e^{j2\pi f \tau}$, and the window functions satisfied $\|g(t)\| = 1, \|g_{t,f}(\tau)\| = 1$. The basic thinking of STFT is to build a window function $g(\tau)$ which can slide along the time axis and the size of it was fixed, when observe the local frequency character of the signal through this window function, which can get a group of Fourier transform, these Fourier transform are character of the time-frequency. But the type of window function has the direct influence for the effect of the time-frequency, it can be shows for the two aspects: first is the type of the window function, it is not difficult to choose a fit window function for a certain signal, but when deal with the overlap signal, this can be difficult; on the other side, the size of the window is also difficult, when analysis the fast signal, the time resolution should be improved, but when analysis the slow signal, the frequency resolution should be improved.

The uncertainty principle defines the constraint relationship between the width and bandwidth of a given signal, that is, the product of the multiplication of the two is greater than a fixed value, and it is impossible to reach infinitely small at the same time, which means that time resolution and frequency resolution are a contradiction.
Figure 1 and Figure 2 show that there is an obvious contradiction between time resolution and frequency resolution. There are two ways to either get rid of the constraints of the window function or try to make the time resolution and frequency resolution meet the requirements for the signal. The requirements of analysis make it achieve the effect of adaptive adjustment.

2.2 Wigner-Ville Distributions

One of the earliest time-frequency representation method is called WVD(Wigner-Ville Distributions, WVD), which has good time-frequency representation methods. WVD has good time-frequency aggregation, high time resolution and frequency resolution, and it is simple to calculate. The WVD of signal $s(t)$ can be defined as,

$$W(t, f) = \int_{-\infty}^{\infty} s\left(t + \frac{\tau}{2}\right) s^*\left(t - \frac{\tau}{2}\right) e^{-j2\pi f\tau} d\tau$$

(2)

Using the signal itself as a window function to perform a STFT can be simulated, which can overcome the shortcoming of the STFT. The window function has some kind of adaptability to the signal. But WVD is not good for the superimposed signal $x(t) = x_1(t) + x_2(t)$, its WVD is

$$WVD_{x_1, x_2}(t, f) = WVD_{x_1}(t, f) + WVD_{x_2}(t, f) + WVD_{x_1, x_2}(t, f) + WVD_{x_2, x_1}(t, f)$$

(3)

where, $WVD_{x_1, x_2}(t, f) = \int_{-\infty}^{\infty} x_1\left(t + \frac{\tau}{2}\right) x_2^*\left(t - \frac{\tau}{2}\right) e^{-j2\pi f\tau} d\tau$, which means when dealing with the superimposed signal, WVD can’t direct reflect the time-frequency’s character, it needs further processing.

2.3 Pseudo Wigner Ville Distribution

In order to solve the bilinear problem of WVD, the WVD is smoothed in the frequency domain as,

$$PWVD_{x}(t, \omega) = \int_{-\infty}^{\infty} h(\tau) s\left(t + \frac{\tau}{2}\right) s^*\left(t - \frac{\tau}{2}\right) e^{-j\omega \tau} d\tau$$

(4)
where the function is defined as $h(t)$. However, PWVD(Pseudo Wigner Ville Distribution, PWVD) can just smooth the direction of variation of $\tau$’s cross terms and reduce the resolution of its.

2.4 SmoothPseudo Wigner Ville Distribution

If add the window and cut from both of the direction of variation of $t$ and $\tau$, if this way the method can cut down the cross term of the two directions to get SPWVD(SmoothPseudo Wigner Ville Distribution, SPWVD),

$$SPWVD_{2}(t, \omega) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u)h(t)\left(s\left(t-u+\frac{\tau}{2}\right)\left(t-u-\frac{\tau}{2}\right)e^{-j\omega t} du \right) d\tau$$

For that the WVD distribution has a high time and frequency resolution, it can be used to describe the signal. However, due to its bilinear, there is a cross term in the middle of the signal term, which generates false signals, which causes trouble to the signal analysis. Therefore, STFT, WVD, PWVD and SPWVD are simulated, and the suppression performance of cross terms by different algorithms is shown in Figure 3 to Figure 5.

Fig. 3. Time-frequency diagram of the four transformations of the 2FSK signal

Fig. 4. Time-frequency diagram of the four transformations of the BPSK signal

Fig. 5. Time-frequency diagram of the four transformations of the LFM signal
2.5 Image Processing

Most image processing methods are defined based on grayscale or binary graphics, so the SPWVD time-frequency image can be transform into the grayscale image. The Figure 6 and Figure 7 show the comparison of the time-frequency image before and after the grayscale change.

![Figure 6. NLFM’s SPWVD time-frequency image](image1)

![Figure 7. NLFM’s time-frequency grayscale change](image2)

In order to make the image can achieve the effect of removing noise and cross terms, the image enhancement technology is used below. The quality of the graphics directly obtained by time-frequency transformation or other methods are not ideal, this is for that the noise and interference signals are often affected in the process of obtaining graphics. It is necessary to improve the quality of images to get the better analyze of images. A common method is to the enhancement of images. The most important part of this technology is highlights features of interest in images, and cut down the features which are not wanted.

Due to the noise and other interference, the grey-scale images obtained by time-frequency technology are always not clearly. One way to enhance graphics is the gray-scale histogram transformation. Its main task is to modify the gray-scale values to make the graphics achieve a uniform effect. The gray level histogram is defined as the correspondence between each gray level (normalized gray level) in the graph and the frequency of the gray level. The probability of the occurrence of any $k$−th gray level is defined as,

$$h_k = \frac{n_k}{N} \quad (6)$$

where, $n_k$ is the number of the $k$−th gray level appear, $N$ is the total of the image gray level. The gray histogram of Figure 8 is shown in Figure 9.

![Figure 8. The diagram of the grayscale histogram](image3)

![Figure 9. Histogram of transformation image](image4)
Observing from the Figure 8, the gray level of the image is mostly concentrated between 0 and 50, and the gray value of this area represents the background part of the image, the gray level of the signal is higher, but its proportion in the gray histogram is smaller. For that the signals (especially some weak signals) can be easily masked on the image. Changing the shape of the histogram is to fill the part with the smaller gray level with the part with the larger gray level to achieve the effect of artificially supplementing the signal energy. In fact, a common method for changing the shape of a histogram is equalization of the histogram. The image and its histogram after histogram transformation are shown in Figure 9 and Figure 10.

The grayscale image after the histogram transformation improves the contrast of the signal, and the signal part can be more clearly displayed on the time-frequency image, which increases the brightness of the signal to a certain extent and enhances the viewability of the signal. However, on the other hand, the noise and residual cross terms are also more apparent in front of the image. In order to improve the contrast of the signal, convert the image into a binary image which was shown in Figure 11. The binarization processing steps are as follows,

1. Sum all pixels on the grayscale image and average them.

\[ \text{thresh} = \frac{1}{N} \sum_{k=1}^{N} h(k) \]  

where, \( N \) is number of points in the grayscale image, \( h(k) \) is the grey value of the image.

2. The gray value in the grayscale image is judged. If the grayscale value is greater than \( \text{thresh} \), the judgment is 1, otherwise the judgment is 0.

After the binarization process, the signal can be clearly displayed, and the cross-term interference left by the SPWVD time-frequency transform is effectively removed.

3 Experiment

Modulation identification of six types of signals LFM, NLFM, BPSK, 2FSK, 2ASK signals, the range of snr is -2dB to 10dB, and the step is 2dB. The length of these five signals is 256, and the sampling frequency is 1000kHz. Each type of signal generates 300 signals in each modulation mode. The flow chart of the simulation experiment is shown in Figure 12.
The main steps are as following: First, input the five kinds of signals, then make SPWVD transform to get their time-frequency image. By processing the time-frequency diagrams of grayscale, equalization, and binarization, time-frequency diagrams of the five signals in Figure 13 after image processing are obtained.

**Fig. 12.** Algorithm flow chart for proposed method.

**Fig. 13.** Time-frequency diagram of the four transformations of the LFM signal.
Then make 300 images for each signal, 70% of the images are test set and else are validation set. Put then into the CNN(Convolutional Neural Networks, CNN). The CNN network which used in this paper is based on the GoogleNet, it was proposed by Christian Szegedy. Before GoogleNet, the AlexNet and VGG are both obtain better training result by increasing the depth of the network, however the increase of the number of the levels can bring many bad effects. In this paper’s method use the inception to get better classification accuracy. The inception model which was used in this paper was shown in Figure 14.

![Inception Model](image)

**Fig. 14.** The architecture of the Inception.

The simulation experiment was based on MATLAB 2019a, after classification in the CNN, the result is shown below. Figure 15 is the confusion matrix of the signal classification results at 0dB. The average classification accuracy can be calculated as,

\[
\text{Accuracy} = \frac{\text{numSignalAccuracy}}{\text{numSignals}}
\]

\[ (8) \]

where, \( \text{numSignalAccuracy} \) is the total number of the signal accuracy, \( \text{numSignal} \) is the number of the kinds of signal. The average classification accuracy can be calculated as, of 5 signals is 90.44%.

![Confusion Matrix](image)

**Fig. 15.** Confusion matrix for signal classification results at 0dB.
In order to analyze the classification performance of the proposed algorithm more intuitively, the average classification accuracy rate is visualized as shown in Figure 16. With the increase of the SNR, the signal can get better and better classification accuracy. In order to analyze the classification effect of the method in this paper, the accuracy curve of the signal recognition and classification results under different signal-to-noise ratios is shown in Table 1.

### Table 1. Recognition and classification results

<table>
<thead>
<tr>
<th>SNR</th>
<th>2ASK</th>
<th>2FSK</th>
<th>BPSK</th>
<th>LFM</th>
<th>NLFM</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2</td>
<td>81.1%</td>
<td>82.2%</td>
<td>90.0%</td>
<td>92.2%</td>
<td>82.2%</td>
<td>85.54%</td>
</tr>
<tr>
<td>0</td>
<td>96.7%</td>
<td>91.1%</td>
<td>77.8%</td>
<td>95.6%</td>
<td>91.1%</td>
<td>90.44%</td>
</tr>
<tr>
<td>2</td>
<td>85.6%</td>
<td>94.4%</td>
<td>94.4%</td>
<td>95.6%</td>
<td>90.0%</td>
<td>92.00%</td>
</tr>
<tr>
<td>4</td>
<td>95.6%</td>
<td>90.0%</td>
<td>95.6%</td>
<td>90.0%</td>
<td>93.3%</td>
<td>92.89%</td>
</tr>
<tr>
<td>6</td>
<td>81.1%</td>
<td>96.7%</td>
<td>98.6%</td>
<td>98.9%</td>
<td>92.2%</td>
<td>93.50%</td>
</tr>
<tr>
<td>8</td>
<td>93.3%</td>
<td>97.75%</td>
<td>87.8%</td>
<td>97.7%</td>
<td>97.8%</td>
<td>94.87%</td>
</tr>
<tr>
<td>10</td>
<td>96.7%</td>
<td>94.4%</td>
<td>98.9%</td>
<td>92%</td>
<td>95.6%</td>
<td>95.56%</td>
</tr>
</tbody>
</table>

### 4 Conclusion

The time-frequency transform of the received signal is analyzed to obtain the corresponding time-frequency map. The image processing can improve the texture details of the time-frequency map. Then use the neural network to classify and recognize the processed time-frequency map. In this method can obtain better classification results. However, the time-frequency map obtained after SPWVD transformation is still subject to the interference of cross terms. In the subsequent research, the time-frequency transformation method of the signal needs to be further improved to achieve the purpose of improving the accuracy of signal classification.

### References


Design of early warning system for reservoir flood level

based on probability analysis

Jiang Wen-bing1, Liu Li2
{jwb00321@sina.com1,yueg00300@163.com2}

(1. Market director, guizhou east century science & technology co., LTD., Guiyang 550081, China;
2. Hunan University of Finance and Economics, Changsha 410205, China.)

Abstract. Due to the poor accuracy of early warning results in the traditional reservoir flood level early warning system, resulting in unstable output voltage, based on this, a reservoir flood level early warning system based on probability analysis is proposed and designed. Its hardware consists of tc401 capacitive sensor, pressure sensor dx130 and water level signal conditioning circuit. Its software design includes database design, water level and flow information setting, water level early warning setting. Through the system hardware design and software design, the design of the reservoir flood water level early warning system is completed. The experimental results show that compared with the traditional early warning system, the designed water level early warning system is more accurate, and the output voltage is more stable.

Keywords: Probability analysis; Reservoir flood; Water level early warning;

1 Introduction

In recent years, large floods have occurred in many river basins in China, which have caused great losses to the national economy and people's lives and property. How to use the latest science and technology to control the loss brought by flood to the minimum is of great practical significance. The key to flood control is to grasp the dynamic state of flood accurately and in real time. In view of these problems, the country has been committed to the research of water resources development and utilization and open source measures for a long time. Among them, flood resource utilization is an effective method. Its purpose is to effectively utilize the unusable flood resources in the flood season. This method has less damage to the ecological environment and no pollution to the green environment. Therefore, it is of great value and potential to study how to transform flood in flood season into available water resources without destroying the original flood control standard of reservoir. For the reservoir, the utilization of flood resources can not only meet the requirements of flood control, but also make full use of flood resources. In order to meet the goal of flood control and profit promotion at the same time, reservoir operation is necessary. In the flood season, when the flood comes, in most cases, in order to ensure the safety of the reservoir, direct flood discharge,
or for the safety of the downstream, first storage and then discharge. Through reasonable reservoir operation, it can be used to cut the peak and hold back the flood, and it can also be used to store the full water to the greatest extent to make profits.

At present, there are few flood warning systems for large-scale basins in China. Most of the existing ones are wireless warning systems for small basins, and most of them are imported from abroad. The wireless early warning system sends the hydrological information to the control center through the wireless transmitter. Literature [1] puts forward the dynamic control method of reservoir operation water level in the dynamic flood prediction period. Firstly, considering that the reserved flood control capacity for the downstream reservoir has the space of overlapping use, the operation water level value in the flood period is optimized; secondly, the model is established by using the prediction and pre discharge method to calculate the upper limit value of the dynamic control area of the operation water level in the flood period; finally, the dynamic change in the flood prediction period is mainly considered, Solve the dynamic control domain based on risk analysis. According to the calculation, the operating water level of the reservoir in flood season can be increased from 347.60 m to 350.50 M. Considering the variation interval of flood forecast period is [3,6] h and the interval of 0.5 h, the dynamic control region of the operating water level in flood season is [350.50,351.40] m. The corresponding power generation benefits can be increased by 19 million KW·h to 25 million KW·h respectively. However, the accuracy of this method for water level early warning is low in the case of severe weather.

In view of the problems of the above system, this paper proposes and designs a reservoir flood level early warning system based on probability analysis.

2 Hardware composition of early warning system of reservoir flood level based on probability analysis

2.1 Tc401 capacitive sensor

The water level sensor needs to be immersed in the reservoir for a long time. The sensors made of common materials are easy to form, corrode and oxidize, which affects the liquid level detection. The inductive digital water level sensor and its system have the advantages of convenient installation and use, strong anti freezing ability, corrosion resistance and high cost performance. Now it has been widely used in the fields of hydrology, water conservancy, hydropower, information prevention engineering, national water resource management, etc. The evaluation of "inductive digital water level sensor" and its system technology by the National Hydrological Standardization Technical Committee provides a key and reliable new technology for the automatic collection and submission of the first line water level data [2]. Therefore, the system adopts the inductive digital water level sensor as the water level data
acquisition equipment.

Tc401 capacitive digital water level sensor (commonly known as electronic water gauge) is a kind of inductive element which uses neural network circuit and mechanical method to locate accurately. The water level signal is processed by digital coding. The design idea of neural network circuit is adopted, and the position inductive element is precisely determined by the method of short-term charging of electrode plate. It has high reliability and anti-interference performance. It is the same kind of sensor in the market at present, and its technical content is also at the leading level. It is a new concept of water level sensor [3]. Its main technical features are high measurement accuracy of 0.5mm, stable and reliable performance, strong anti-interference ability, low energy consumption, adoption of neural simulation theory, from measurement, transformation to transmission, the whole process of digitization, large measurement range, which is suitable for the requirements of flood control and drought resistance of Jianghe reservoir, large temperature range, which is - 20 ~ + 50 ℃; output signal can be connected with other circuits through RS232.

2.2 Pressure sensor DX130

DX130 adopts the design principle of hydrostatics, with the range of 200m, the output voltage range of two-wire system of 0 ~ 5V, and the long-term stability of full degree and zero position of 0.1% FS / year [4]. When the sensor is put into a certain depth of the measured liquid, the pressure formula of the sensor's liquid surface is as follows:

\[ P = p \cdot g \cdot H + P_0 \]  \( \text{(1)} \)

Where: \( P \) is the density of the measured river water; \( P \) is the pressure on the liquid surface of the sensor; \( g \) is the acceleration of gravity; \( P_0 \) is the atmospheric pressure on the liquid surface; \( H \) is the depth of the pressure sensor into the liquid. Because the pressure of the liquid is introduced into the positive pressure chamber of the sensor through the gas conducting stainless steel, and then the atmospheric pressure \( P_0 \) on the liquid level is connected with the negative pressure chamber of the sensor to offset the \( P_0 \) on the front of the sensor, the pressure measured by the sensor is \( p \cdot g \cdot H \), and the liquid level depth \( H \) can be obtained through the measured pressure \( P \). When the ADC interface of the controller is connected with the voltage output of the pressure sensor, the pressure data acquisition and digital processing can be completed, and it can be converted into the water level and height.
information.

2.3 Water Level Signal Conditioning Circuit

Because the pressure sensor works in the low range, in order to improve the measurement accuracy, it is necessary to amplify the output voltage signal of the pressure sensor before analog-to-digital conversion [5]. AD620 is a low-cost, high-precision instrument amplifier. It only needs an external resistance $R_3$ to set the gain. The gain range is 1-10000, and the power consumption is low. The maximum working current is only 1.3mA. The system sets the gain $G = 5$ (i.e. $R_3 = 12.3\, \Omega$ ) [6]. The sixth pin output of AD620 is directly connected with the ADC of the controller. The signal conditioning circuit is shown in Figure 1.

![Fig. 1. Signal conditioning circuit diagram](image)

The output voltage signal of the pressure sensor is connected to the controller through AD620 amplifier for ADC conversion, which improves the accuracy of water level measurement and meets the needs of the system.

3 Software design of reservoir flood level early warning system based on probability analysis

3.1 Database design

Under the unified data format and specification, the database is divided into basic information database and geographic information database according to the data storage mode. Oracle 10g is used as database management platform. The basic information database is divided into flood database and algorithm database from the content [7]. Flood database mainly stores data of reservoir, including:
(1) Historical hydrological, socio-economic and flood loss data, etc;
(2) Real time rainfall and water regime data, including rainfall data of meteorological station, rainfall forecast data and real-time water level and flow data of hydrological station;
(3) Model boundary, parameters, etc. These data can be automatically or manually modified to store new data. The algorithm database includes model calculation and data graphic conversion formula, which are used to support the analysis and calculation of data, as well as the mapping and transformation between data and visual elements. With the help of GIS middleware, a geographic information database is created, which mainly stores vector point data such as rainfall station, water level station, breach; vector line data such as road, linear water system, polder area and embankment; vector plane data such as administrative division, surface water system, land use and flood risk elements. Among them, the elements of flood risk are stored in grid, and the unit domain includes grid coding, maximum inundation depth, maximum inundation duration, maximum flood velocity, flood arrival time and marker domain.

The design of database mainly serves for the probability analysis of flood. The observation series $F_y$ (Time Series) of annual flood frequency is calculated by probability analysis method, and the expression is as follows:

$$F_y = f_y(t) = \left[ f_y(1), f_y(2), \ldots, f_y(n) \right]$$

Where $f_y(t)$ is the number of floods in the $t$-th observation year. Generally, for the multi peak flood caused by continuous rainfall, one flood is counted in the analysis. The $F_y$ series is transformed as follows: if $f_x$ is the total number of $x$ floods per year in the observation period (i.e. the total number of observation years), then there is a series $F$:

$$F = \left[ f_0, f_1, \ldots, f_k \right]$$

That is, in year $n$, there are 0 floods in year $f_0$, 1 flood in year $f_1$ There are $k$ floods in $f_k$ years. According to the calculation results, it is inferred that the probability of no more than $x$ floods per year is:
Let \( p(-) = \max[p(0), p(1), p(2), ...] \), the probability of annual possible flood is \( p(-) \); if \( p(x-1) + p(x) + p(x+1) > 0.6 \), the probability of annual flood between \( x-1 \) and \( x+1 \) is great; the probability of annual flood greater than \( x \) is:

\[
p(x > -) = 1 - \sum_{x=0}^{\infty} p(x)
\]

For different statistical intervals, if there are different \( F \) series, different analysis conclusions can be obtained; for different flood peak standards, if there are different flood frequency series \( F_y \), different flood frequency analysis conclusions can be obtained. The analysis steps are as follows:

Step 1: according to a certain peak discharge value \( QM \), construct the annual flood frequency series \( F \);

Step 2: according to 0 times, 1 time per year, the number of years of \( K \) floods is a series \( F \);

Step 3: according to formula (4), according to the corresponding values of elements \( f_x \) and \( E(x) \) of sequence \( F \), construct statistic \( i_0^2 \), and carry out the goodness of fit test at a given significance level \( T \);

Step 4: if \( F \) series conforms to Poisson distribution, calculate the probability \( p(x) \) of \( x \) floods per year according to formula (5), and then get the number, range and probability of possible floods per year.

3.2 Water Level Flow Information Setting

The water level and flow information setting module defines the maximum, minimum and relative change rate of water level and flow, compares the water level information monitored by the sensor with the warning line, and determines whether the water level of the reservoir is in a safe condition through the control chart, so as to determine whether to alarm. Among them, RS485 communication is mainly used to complete information collection and
RS485 communication is usually used as a relatively economic, high noise suppression, relatively high transmission rate, long transmission distance and wide common mode range communication mode. RS485 standard adopts data transceiver driving bus of balanced transmission and differential reception. The data line shall preferably be composed of twisted pair, and the external shield layer shall be grounded. Because 485 signal is transmitted by differential mode, that is, the voltage difference between a and B is used as signal transmission. When using twisted pair for 485 signal transmission, if there is an external interference source, the interference effect on a and B is the same, the voltage difference remains unchanged, and the interference on 485 signal is minimized. If there is shielding wire, the interference effect of external interference sources will be minimized. Some types of sensors use 485 acquisition communication interface, 651bc184 chip and necessary control circuit to realize the communication between single chip computer and sensors. Through this kind of chip, the corresponding data collected by the single-chip microcomputer on the module can be exchanged with the telemetry terminal of the measuring station, and the data can be reported. 651bc184 is a differential data line transceiver within the scope of sn5176 industry standard, with built-in high-energy transient noise protection device. This design features more than most of the existing devices, significantly improving the reliability of resisting the transient noise on the data synchronous transmission cable, providing a reliable and low-cost direct connection (without insulation transformer) data line interface [8], without any external components. The operating temperature of 651bc184 is - 40 C to 85 C.

Among them, the D and R interfaces are TTL standards, which can be directly connected to the UART port of single chip microcomputer for data transmission and reception, and the re and de interfaces are used to control the bus transmission mode or reception mode. In 485 communication, attention shall be paid to data confusion caused by signal reflection in communication cable, and 120 Ω matching resistance shall be added to realize impedance matching [9]. Due to the complexity of some field conditions, 485 interface uses differential transmission mode, which has the ability of anti common mode interference, but it may cause high common mode voltage between nodes for some reason. When the common mode voltage is more than + 12V or less than - 9V, the limit receiving voltage of 485 communication chip will be exceeded, which will cause 485 communication chip not to work or even burn down. At this time, 485 optical isolation repeater can be used in 485 bus to completely isolate 485 signal and power supply and eliminate the influence of common mode voltage. In order to ensure the reliable operation of single chip in extreme cases, the optical couple isolation device can be added to the interface between 651bc184 and single chip. Due to the simple
connection of 485 communication lines, only four lines a, B, VCC and GND are needed. Through
the unified definition of corresponding pins of 485 communication interface of
module, the 485 acquisition communication interface of telemetry terminal is specified to be
consistent with it, and the unified 485 communication rate, handshake protocol, data format
and encryption mode are customized. It can be realized that the front-end data acquisition of a
new station is compatible, which is convenient for terminal equipment in the future According
to the specific needs, select various types of sensors. In practical application, except for single
point grounding of 485 bus shielding layer, it is better to use galvanized pipe to cover 485
cable when conditions permit. The galvanized pipe is well grounded, which can realize
reliable shielding and lightning protection.

3.3 Water Level Warning

Water level early warning control indicators include water level limit control and water
level change rate, as shown in Figure 2.

![Fig. 2. Water Level Boundary Control Chart](image)

The water level limit control chart is an effective tool for analyzing and judging the
process fluctuation, which can monitor the water level data in real time. Generally, in the early
stage of mine design, the upper and lower water level control limits are specified, and the
historical data are continuously revised [10]. Once the water level exceeds the limit of the
control chart, the alarm indicates that the water level exceeds the safety state. The monitoring
of the relative change rate of water level is based on the historical data to calculate the
maximum change rate of water level \( \delta_{\Delta} \) in the mine, and take \( \delta_{\Delta} \) as the reference to
calculate the relative change rate of water level in unit time:

\[
\delta_i = \frac{H_i - H_{i-1}}{H_{i-1}(T_i - T_{i-1})}
\]  

(6)
Where: $H_n$ is the water level monitoring data collected at n time; $H_{n-1}$ is the water level monitoring data collected at n-1 time; $T_n-T_{n-1}$ is the time interval between N time and the previous time. When $\delta_i > \delta_\Delta$, it indicates that the relative change rate of water level exceeds the reference standard, then the system alarms.

4 System Test Experiment

In order to verify the performance of this system, a comparative experiment with the traditional one is designed. Through comparison, the hypothesis of the experiment is verified. The experimental environment is shown in Table 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating system</td>
<td>Microsoft Windows XP</td>
</tr>
<tr>
<td>Processor</td>
<td>Intel(R)Celeron(R) 2.6GHz</td>
</tr>
<tr>
<td>Internal storage</td>
<td>24.0 GB</td>
</tr>
<tr>
<td>Hard disk</td>
<td>16.0 GB</td>
</tr>
<tr>
<td>Database management software</td>
<td>Microsoft SQL server 2010 R2</td>
</tr>
<tr>
<td>JDK</td>
<td>1.6</td>
</tr>
<tr>
<td>Mathematical software</td>
<td>MATLAB</td>
</tr>
</tbody>
</table>

4.1 Experimental comparative analysis
The experiment in this section will judge the advantages and disadvantages of the two systems by comparing the stability of the two systems. The experimental method is comparative experiment. The experiment verification process is completed under the control of other variables except the experimental variation.

4.2 Experimental result

Two systems are used to test the water level changes of seven rivers. Seven monitoring points (1-7) are set up respectively. The server of the monitoring center is always in operation state, waiting for the TCP/IP connection and communication of the water level terminal. After the water level terminal is connected with the server of the monitoring center, the green communication indicator flashes. See Table 2 and table 3 for water level monitoring data.

<table>
<thead>
<tr>
<th>Table 2 Monitoring and warning data of water level under the system</th>
</tr>
</thead>
<tbody>
<tr>
<td>monitoring site</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3 Water level monitoring and early warning data under traditional systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>monitoring site</td>
</tr>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>Monitoring site</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
</tbody>
</table>

From the contents of Table 2 and Table 3, it can be seen that under the operation of the early warning system in this paper, the water level of the seven monitored rivers is within the preset safety range, that is, there is no drought or flood, nor rapid change, but the water level fluctuates slightly; and in the seven monitoring points, the output voltage of the conditioning circuit and the water level height show a basic linear relationship. Compared with the traditional early warning system, its early warning results are more accurate, and the output voltage is more stable.

In order to further verify the early warning accuracy of this system, the early warning accuracy of this system and the traditional system is compared and analyzed, and the comparison results are shown in Figure 3.
The accuracy of this system

Accuracy of traditional systems

According to figure 3, the accuracy of this system is within 45% ~ 80%, up to 80%, while the accuracy of traditional system is within 25%, up to 25%, indicating that the accuracy of this system is higher than that of traditional system.

5 Concluding remarks

In this paper, a reservoir flood level early-warning system based on probability analysis is designed. Through the contrast experiment with the traditional reservoir flood early-warning system, it is proved that the system designed in this paper can ensure a higher early-warning accuracy and has better stability.

References


The design of high speed multiplex data transmission system based on single chip microcomputer

CUI Rong-rong¹, Ren Yan-yan²
{nide25251@tom.com¹, wudan22@tom.com²}

¹. College of Mathematics and Computer Science, Chifeng University, Chifeng 024000, China;
². College of Mathematics and Computer Science, Chifeng University, Chifeng 024000, China

Abstract: In view of the high packet loss rate in the use of the original high-speed multi-channel data transmission system, the original data transmission system is optimized by using single-chip microcomputer, and the high-speed multi-channel data transmission system based on single-chip microcomputer is designed. The hardware of the main control chip, memory and data channel of the system is designed, and installed in the original system hardware framework to complete the hardware design of the system. According to the requirements of data transmission, data storage format and data source are set, and fuzzy processing calculation is used to complete data preprocessing. Construct data frame structure and complete data transmission. So far, the design of high-speed multi-channel data transmission system based on single chip microcomputer has been completed. By comparing the packet loss rate, it is verified that the designed system can effectively reduce the packet loss rate and improve the integrity of data transmission results.

Key words: Data acquisition; Singlechip; Data transmission; Data comparison;

1 Introduction

With the development of industry and agriculture, data transmission system has been widely used. In order to adapt to this trend, research in related fields is becoming more and more important. In scientific research, the use of data transmission system can obtain a large number of dynamic information, which is one of the important means to obtain scientific data and generate knowledge. No matter in which application field, data processing and transmission will directly affect work efficiency and economic benefits. Data transmission technology is a practical electronic technology. It is widely used in signal detection, signal processing, instrumentation and other fields. In recent years, with the continuous development of digital technology, data transmission technology also presents the development trend of faster speed, more channels and more data [1]. In the process of using the original data transmission system, the system often runs out of control due to the large amount of data information, resulting in the collapse of the data transmission system. In view of the above problems, a high-speed multi-channel data transmission system based on MCU is designed.
Single chip microcomputer is short for single chip microcomputer, which has the characteristics of small size, high integration, fast calculation, economy, etc., so it is widely used in the fields of industrial automation, information collection, home appliance control, etc. In this design, single-chip microcomputer can effectively improve the data load capacity of the system and ensure the normal use of the system performance. Moreover, the single-sided machine is small in size and convenient in installation, which will not cause too much influence on the hardware framework of the system. This paper will be completed by two parts: first, the high-speed multi-channel transmission of information will be realized by single side computer; then, the performance difference between the original system and the designed system will be obtained by system test. Next, this paper will analyze and discuss the hardware design and software design of the high-speed multi-channel data transmission system of single chip microcomputer.

2 The hardware design of high speed multiplex data transmission system based on single chip microcomputer

The hardware design of the high-speed multi-channel data transmission system of single-chip microcomputer consists of three parts: the main chip design, the data channel hardware design and the memory design. In a strict sense, the design of data transmission hardware of single-chip microcomputer should be a system that can automatically detect or patrol the multi-channel data controlled by computer, and can store, process, analyze and calculate the data, and extract the available information from the detected data for display, recording, printing or description [2]. The whole data acquisition and transmission system is controlled by single chip microcomputer. The signal acquisition channel is selected through the control of channel data selection module and A/D conversion circuit module, so as to realize data acquisition and transmission. The system implementation block diagram is shown in Fig. 1.
The above framework is used as the basis and control scheme of hardware framework design to ensure the order and controllability of hardware design and improve the process of system design.

2.1 Design of main control chip

The high-speed multiplex transmission system of single-chip microcomputer mentioned in this paper takes the acquisition and transmission of the relevant parameters of single-chip microcomputer as the main task and the single-chip microcomputer as the core controller. The overall structure of the system is shown in Fig. 2.
There are many kinds and models of single-chip microcomputers available on the market, and the performance of single-chip microcomputers produced by each company is also different. According to the corresponding hardware requirements in this design, c8051f000 series single-chip microcomputers produced by Cygnal company are widely used, and the development tools are relatively perfect, and the product instruction systems of many companies are compatible with this. In addition to the choice of brand, we should consider the performance of SCM. In order to achieve high-speed multi-channel data transmission, the design uses a wide data bus to achieve the performance of the system [3]. The wider the known data bus width, the stronger the function. There are 4-bit computers, 8-bit computers, 16 bit computers and even 32-bit single chip computers. In order to ensure that the design system can run smoothly on the micro computer, 32-bit computer is more suitable. 32-bit computer has high cost performance in price and performance. Considering the internal resources of the single chip microcomputer, the basic units of the single chip microcomputer are as follows: central control unit (CPU), timer / counter, I / O interface, serial communication interface, interrupt response system, internal memory, data bus and address bus for system expansion, etc. [4]. Through the above settings, the design of the main control chip is completed. The specific main control chip is shown below.
The main control chip designed above is introduced to the control host of the original system, which is the hardware control center of this design.

### 2.2 Hardware design of data channel

In this article, the single-chip high-speed multiplexing system is designed to collect 32 data channels, while traditional single-chip microcomputers can collect up to 8 A/D signals. Select, then use the 74LS139 decoder as a chip selection of four 8-channel data selectors, and send the 32-channel data signals to the single-chip microcomputer for storage four times, and then analyze and process them\(^5\). In this way, the data of the 32-channel analog signals can be completely selected and collected by the single-chip microcomputer control, thereby performing high-speed multiplex transmission of the single-chip microcomputer. Data channel selection is shown in Fig. 4.
According to the above structure, the selection and design of the data channel hardware is completed, and it is connected to the main control chip.

2.3 Memory design

The function of data storage in this design system is mainly to temporarily store data, and it has the function of data caching. When the collected data reaches a certain amount, it will be transmitted to the upper computer at one time, which will help improve the data transmission efficiency and the efficiency of storage in the upper computer. All configurations in the memory are transmitted to the nRF905, SIP interface through the SPI interface, which can be set by the SPI instruction. When the nRF905 is in idle mode or shutdown mode, the SPI interface can remain in working state. 256 KB serial electrically erasable programmable read-only memory AT24C256 is used to control the information receiving port of the memory. Fig. 5 is the connection diagram of memory AT24C256 and 89C52P1 port.
It can be seen from the image that this memory only occupies two I/O ports of the microcontroller. CEO, CE1 are address selection inputs, all are grounded, and WC is a write-protected input pin. When connected to ground, write operations are allowed [6]. SCL is a serial clock input pin, which is connected to P1.3 of 89C52. It writes SDA to the memory on its rising edge and reads out data and sends it to SDA on its falling edge. SDA is a bidirectional serial data input and output port. It is connected with P1.4 and exchanges data with the single-chip microcomputer. The figure above is a simplified diagram of the memory design. Through this diagram, the memory is connected to other hardware designed in the article and referenced to the original hardware framework. Then, the optimized hardware framework was used as the development environment for this software design.

3 System software design

In order to realize the use effect of the single-chip microcomputer in the high-speed multi-channel data transmission system, the software design part is divided into three links, which are data acquisition, data preprocessing, and data transmission. Data acquisition and data transmission are the program of the single-chip microcomputer sender and the program of
the upper computer receiver [7]. The program on the sender side mainly includes data collection and transmission, and the program on the receiver side mainly receives and processes the data, and realizes the high-speed data transmission performance of the system.

3.1 Data collection

In view of the problem that the packet loss rate of the original system is too high in use, the big data calculation is used to complete the data collection work and ensure the integrity and effectiveness of the data at the data source [8]. The collected data is stored in the designated database in the form of data table. In order to ensure the integrity of data in the transmission process, the same compression method can be used to expand the processing. The format of data table is set as follows.

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Field</th>
<th>Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>data sources</td>
<td>int</td>
</tr>
<tr>
<td>2</td>
<td>User name</td>
<td>varchar</td>
</tr>
<tr>
<td>3</td>
<td>User number</td>
<td>varchar</td>
</tr>
<tr>
<td>4</td>
<td>Data source No</td>
<td>varchar</td>
</tr>
<tr>
<td></td>
<td>Data</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>acquisition equipment Data</td>
<td>varchar</td>
</tr>
<tr>
<td>6</td>
<td>collection time</td>
<td>varchar</td>
</tr>
<tr>
<td>7</td>
<td>Whether to deal with storage</td>
<td>varchar</td>
</tr>
<tr>
<td>8</td>
<td>storage location</td>
<td>varchar</td>
</tr>
<tr>
<td>9</td>
<td>Usage situation</td>
<td>double</td>
</tr>
<tr>
<td>10</td>
<td>data type</td>
<td>double</td>
</tr>
<tr>
<td>11</td>
<td>Remarks</td>
<td>double</td>
</tr>
</tbody>
</table>

Use the above data table to sort out and store the collected data, so as to ensure the effective processing and transmission of the collected data.

3.2 Data preprocessing

Taking the collected data as the data base of the system design, the data can be integrated
into a variety of transmission formats through fuzzy algorithm.

According to the research on the calculation of multi-channel information transmission in Colleges and universities, the mean fuzzy algorithm combined with preprocessing criteria is used to complete the information processing process. Set the data information to be processed as matrix, name it as $A$, set the dimension of information vector as $B_i (i = 1, 2, ..., n)$, and $n$ as the number of vectors. After the data is initialized, it can be seen that the mean value area is $Z_i$, and then there are:

$$\sum_{i=1}^{A} Z_i = 1$$  \hspace{1cm} (1)

If the fuzzy center is $O$ and the information number is $m$, then there are:

$$O_i = \left(\frac{1}{\sum_{i=1}^{m} B_i}\right)^{1/m} \sum_{i=1}^{m} B_i$$  \hspace{1cm} (2)

The data information processing center is obtained by formula (2), and the data information processing is completed and the processing results are obtained by using this as the reference point. If the processing information is set as $Q$ and $P$ as the frame length of the image information, there are:

$$Q = \left[\frac{1}{P(B_i, t)}\right]^{1/\sum_{i=1}^{m} \left(\frac{1}{P(B_i, t)}\right)^{1/m}}$$  \hspace{1cm} (3)

The collected information is calculated by the above set algorithm until the final value does not drop. This completes the process. In order to improve the efficiency of large-scale data processing, we set the data in advance and complete the convergence from the local to the center.

3.3 Set data frame structure to complete data transmission

After data pre-processing, the processed data is generated into a data frame structure as shown in Table 2. The main task of data transmission is to package the sampled data and set it to an appropriate format to complete the data transmission. Set the standard protocol to complete the communication channel setting during data transmission. The establishment of a standard protocol link usually results in excessive delay, and excessive feedback will increase the complexity of the interconnection between the two [9]. Therefore, in this design, the
high-speed multiplexing of data is realized by high-speed serial transmission between the analog-to-digital converter and the FPGA, and a unidirectional point-to-point serial differential transmission protocol based on SERDES is designed to realize the analog-to-digital on the same circuit board. There are four channels between the converter and the FPGA, and the single-channel 4Gb/s serial transmission rate is set. The clock deviation between the two is not considered. The 8B/10B codec is not selected. The system operating clock is set to 250MHz. Stream mode transmission, with data frames as the basic unit of the protocol. The specific data frame structure is designed as follows.

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Structure</th>
<th>Field length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Header</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>BIT[64:57]</td>
<td>(8 bit MSB first)</td>
</tr>
<tr>
<td></td>
<td>Data1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>BIT[56:45]</td>
<td>(12 bit MSB first)</td>
</tr>
<tr>
<td></td>
<td>Data2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>BIT[44:33]</td>
<td>(12 bit MSB first)</td>
</tr>
<tr>
<td></td>
<td>Data3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>BIT[32:21]</td>
<td>(12 bit MSB first)</td>
</tr>
<tr>
<td></td>
<td>Data4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>BIT[20:9]</td>
<td>(12 bit MSB first)</td>
</tr>
</tbody>
</table>

Table 2: Data frame structure table
With the above settings, the data is encapsulated into frames, so that the receiver can correctly identify the start and end of data transmission, so as to control and detect the data transmission and facilitate the expansion of protocol functions. In response to the requirements of this article, the content of the protocol needs to be defined: synchronization sequence, data frame structure. The synchronization sequence is used to synchronize the system between the high-speed ADC and FPGA. The data frame consists of the frame header (SF), data (D), and frame end (EF). The line code stream for designing a custom protocol is shown below.

**Fig. 6** Schematic diagram of the line code of the custom protocol

The line code stream of the custom protocol is used to complete the efficient multiplexing of data. The above software modules are combined with the hardware designed in this paper. So far, the design of high-speed multi-channel data transmission system based on single chip microcomputer has been completed.
4 Data acquisition system test

Combined with the above hardware design results and software module design, the design of high-speed multi-channel data transmission system based on single chip microcomputer is completed. In order to ensure the effectiveness of the design, the test link of the construction system is used to compare the performance difference between the original encryption and decryption system and the encryption and decryption system designed in this paper.

4.1 Design of system test platform

In order to verify the performance difference between the design system and the original system, the design system and the original transmission system are tested in the same platform. The experimental environment is set up by two parts, namely, hardware environment configuration and software environment configuration.

Hardware environment configuration: Intel CPU series chip, frequency is 6.0GHz, memory capacity is 20.00GB, hard disk is 2TB.

Software environment configuration: The experiment uses SQL software design, the programming environment is SQL2013, and the test environment is configured as a 64-bit Windows 10 operating system.

According to the above hardware and software, the experimental environment is established, and the experimental object is set as the packet loss rate in the transmission process. The set data test sample is shown below.

<table>
<thead>
<tr>
<th>Test sample No</th>
<th>Sample data volume</th>
<th>Data form</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>500</td>
<td>TEXT</td>
</tr>
<tr>
<td>2</td>
<td>1000</td>
<td>TEXT</td>
</tr>
<tr>
<td>3</td>
<td>2500</td>
<td>TEXT</td>
</tr>
<tr>
<td>4</td>
<td>5000</td>
<td>TEXT</td>
</tr>
<tr>
<td>5</td>
<td>10000</td>
<td>TEXT</td>
</tr>
</tbody>
</table>

Using the above system test samples, the calculation process of the packet loss rate of the original system and the designed system is completed.

4.2 Analysis of test results

Through the above design, the system test process is completed, and the specific system test results are as follows.

<table>
<thead>
<tr>
<th>Test sample</th>
<th>Packet loss</th>
<th>Packet loss</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
According to the above experimental results, the packet loss rate of the designed system is significantly lower than that of the original system. The highest packet loss rate of the designed system is 3.04%, and the lowest packet loss rate of the original system is 5.23%. Through the data comparison, we can intuitively feel that the design system in this paper is obviously superior to the original system. At the same time, compared with the original system in the same amount of data, the transmission speed of the designed system is faster. Therefore, the transmission performance of the designed system is better than the original system. In conclusion, the design system is better than the original system.

5 Concluding remarks

The high-speed multi-channel data transmission system introduced in this paper can collect, analyze, process and transmit multi-channel data signals in real time. Compared with the traditional data acquisition system, it not only records the multi-channel acquisition data, but also records the sampling time. It overcomes the shortcomings of the traditional data acquisition system, such as unable to record the sampling time, inconvenient to display and analyze the data, and it is conducive to the user to accurately understand and grasp the status of the transmission process. The hardware and software structure of the system is reasonable. It can collect and transmit parameters, and display the collected data in real time. After simulation debugging, the error of acquisition parameters of the system is within the design requirements, the system is stable and reliable, the operation process is easy to grasp, the acquisition accuracy and speed meet the actual use requirements, with high practical value. It gives full play to the advantages of single chip microcomputer, such as powerful function, high reliability, good flexibility, easy development and easy expansion. On the basis of the system introduced in this paper, users can expand the data processing and analysis part on the single-chip microcomputer according to the actual needs, increase the corresponding output channel, thus increasing more practical application functions. Because this paper does not consider the data security problem in the process of data transmission, in the future research

<table>
<thead>
<tr>
<th>No</th>
<th>rate of original system/%</th>
<th>rate of system designed in this paper/%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.23</td>
<td>2.15</td>
</tr>
<tr>
<td>2</td>
<td>6.45</td>
<td>2.20</td>
</tr>
<tr>
<td>3</td>
<td>6.24</td>
<td>2.56</td>
</tr>
<tr>
<td>4</td>
<td>6.30</td>
<td>3.04</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>2.15</td>
</tr>
</tbody>
</table>
will focus on the data encryption method to ensure the security of data transmission.

Reference


The design of data acquisition system of program automatic shelling based on ARM

Ren Yan-yan\textsuperscript{1}, CUI Rong-rong\textsuperscript{2}

\{wudan22@tom.com\textsuperscript{1}, nide25251@tom.com\textsuperscript{2}\}

\textsuperscript{1,2}College of Mathematics and Computer Science, Chifeng University, Chifeng 024000, China

Abstract. In view of the low efficiency and poor effect of the current program automatic shelling data acquisition system, the design of the program automatic shelling data acquisition system based on arm is proposed. Its hardware consists of ARM (Advanced RISC Machines) processor, ads830 collector and interface conversion module; the software design includes main control software, GPS (Global Positioning System) acquisition data receiving module, FIFO (Flight Inspection Field Office) data cache module and local data processing module. Through the contrast experiment, it is proved that the designed program automatic shelling data acquisition system not only plays an excellent role in shelling effect performance, but also far better than the traditional design in data acquisition efficiency, and higher than the standard line, with high acquisition efficiency and stability.

Keywords: ARM; Program automatic shelling; Data acquisition;

1 Introduction

With the development of computer technology, arm has become one of the focuses of the current IT (Information Technology) industry, showing a huge market demand. The core technology of embedded system is embedded processor and embedded operating system. ARM company's 32-bit RISC (Reduced instruction set chip) processor, with its high speed, low power consumption, low cost, strong function, unique 16 / 32-bit dual instruction set and many other excellent performance, has become the first choice processor in mobile communication, handheld computing, multimedia digital consumption and other embedded solutions [1]. Among all kinds of embedded operating systems, Linux is widely used in data acquisition, instrumentation, measurement and control system, handheld devices and other embedded system applications because of its clear structure and open code.

In recent years, malicious codes (such as viruses, Trojans, worms, etc.) generally use some technologies with strong survivability and high protection strength to hide and transform, so as to avoid the scanning and analysis of detection tools (such as anti-virus software), and the technology of adding shell is a typical representative of them [2-3]. According to statistics, in 2003, nearly 29% of malicious code was cased, which rose to 35%
in 2005 and exceeded 80% in 2007. The development trend of malicious code brings great challenges to the detection tools. After the program is shelled, only the part of the program that is shelled is exposed, which generally does not contain malicious intention. After the content of the original program is compressed and encrypted, the detection tools have been difficult to identify, and the detection success rate is very low. Therefore, how to restore the content of the program and obtain the execution behavior of the program is the focus of malicious code detection technology research. At present, there are two kinds of Shelling procedures: manual shelling and automatic shelling. Manual shelling requires professional reverse engineering experience of analysts, and the process is tedious and energy-consuming [4-5]. Automatic shelling often depends on some existing features, with poor generality, especially for some non-public shell professional analysts who often spend a lot of time to analyze. Moreover, with the development of shelling technology, malicious code not only uses one shell, but also can have multiple shells, which brings great difficulties to the analysis of malicious code. Because the traditional manual and directional shelling methods have obvious defects, such as the lack of universality, the difficulty to keep up with the progress of shelling technology, the development speed of shelling cases, the need to spend a lot of manpower and material resources, so researchers are committed to the research of automatic shelling technology. At present, there have been some new research results in the field of program automatic shelling, such as polyupack, renovo, Ma lwarenorma liza, etc., but these technologies can not evaluate the effectiveness of extracting data. In the face of some highly protective technologies, it is difficult to accurately recover the program content strength, such as stone byte and virtual machine technology. Based on this, it is necessary to optimize and innovate the traditional automatic shelling data acquisition system.

2 Hardware composition of automatic shelling data acquisition system based on ARM

2.1 Selection of ARM Processor

In a multitasking system, the kernel is responsible for managing each task, or allocating CPU to each task. The first step of setting up the hardware development platform of data acquisition system is to make a good choice of ARM core. Various series of arm architecture, such as ARM7, armg, armge, arm10e, securcore, xseale of Intel, StrongArm of hitel, etc., should consider the following main factors when selecting different embedded system applications:

For the selection of mmij unit, if you want to use wince or Linux and other operating systems to reduce the software development time, you need to select ARM chip with MMU (Memory Management Unit) function above arm720t, arm720t, StrongArm, arm92ot,
arm922t, arm946t with MMU function. ARM7TDMI does not have MMU, and does not support Windows CE and most Linux.

The system clock controller determines the processing speed of ARM chip. The processing speed of ARM7 is 0.5mips/mhz, the common main clock of ARM7 is 20mhz-133mhz, the processing speed of arm is 1.1mips/mhz, the common main clock of arm system is 100mhz-233mhz, and the maximum of arm10 can reach 700MHz.

Internal memory capacity, when large capacity memory is not needed, ARM chip with internal memory can be considered.

USB (General serial bus) interface, most arm chips have USB controller, some even have USB host and USB slave controller at the same time.

Among the instructions provided by some chip suppliers, the number of IO (Incoming Orders) usually indicates the maximum possible number of GPIO (General I / O port), but many pins are multiplexed with address line, data line, serial port line and other pins. In this way, we need to calculate the actual number of gpios that can be used in the system design.

The interrupt controller and arm core only provide two interrupt vectors: fast interrupt (FIQ) and standard interrupt (IRQ). However, different semiconductor manufacturers have added their own interrupt controllers in chip design to support hardware interrupts such as serial port, external interrupt and clock interrupt. External interrupt control is an important factor in chip selection. Reasonable external interrupt design can greatly reduce the workload of task scheduling. For example, for Philips saa7750, all gpios can be set to FIQ, and four interrupt modes can be selected: rising edge, falling edge, high level and low level.

LCD (Liquid crystal display) controller, some ARM chip built-in LCD controller, some even built-in 64K color TFT (Thin film transistor) LCD controller. In the design of PDA (Personal digital assistant) and hand-held display and recording equipment, ARM chip with built-in LCD controller, such as 53c2410xis more suitable. As shown in figure 1:
Fig. 1. 53C2410X ARM processor

ADC (Analog to digital converter) and DAC (Damage Assessment Center). Some arm chips have 2-8 channels and 8-12 bit general-purpose ADC, which can be used for battery detection, touch screen and temperature monitoring. Philips saa7750 also has a built-in 16 bit stereo audio ADC and DAC with headphone driver.

Expansion bus: most arm chips have external SDRAM (Synchronous dynamic random access memory) and SRAM (Static random access memory) expansion interfaces. The number of chips that can be expanded by different arm chips is different, that is, the number of chip selection lines. The external data bus has 8 bits, 16 bits or 32 bits.

UART (Universal asynchronous transceiver) and IrDA (Infrared data communication), almost all arm chips have one or three UART interfaces, which can be used to communicate with PC or debug with angel.

DMA (Direct memory access) controller, some arm chips are integrated with DMA, which can exchange data with external devices such as hard disk at high speed, and reduce the CPU (central processing unit) resource occupation during data exchange. ARM series microprocessors provide the best performance in terms of high performance and low power consumption. Arm series microprocessors have the following features: 5-level integer pipeline, higher support for instruction execution efficiency; 32-bit arm instruction set and 16 bit thumb instruction set; support for high-speed AMBA bus interface; full performance MMU (Memory Management Unit), support for Windows CE (chief engineer), Linux, palm 05 and other mainstream embedded operating systems; MPU supports real-time operating system; support for data cache and instruction cache, with higher instruction and Data processing capacity. As shown in figure 2:

Fig. 2. ARM series microprocessors

2.2 ADS830 collector

The ads830 high-speed acquisition chip with a sampling rate of 10ksa / S ~ 60msa / s and a bit of 8 bits is selected from TI company. Analog signal input amplitude of the module is ±
5V, digital interface level is 3.3V, module power supply is +5V single power supply, analog signal input interface adopts two forms, one is pin interface, the other is SMA interface, SMA interface is used in this design. The ADS830 hardware structure is shown in Figure 3.

![Fig. 3. ADS830 Hardware Structure](image)

In the signal conditioning part, the input signal range -5V - +5V is replaced by 1.5V-3.5V to meet the input range of ADC (Analog to digital converter) data acquisition; ad data acquisition adopts single terminal input with a reference voltage of 2V, i.e. the input voltage range is 1.5V-3.5V, and the digital interface adopts 3.3V logic level to facilitate direct connection with FPGA (Field Programmable Gate Array). The output interface of ads830 module is two rows of 5-pin, 10 interfaces in total, which are D8-D1, CLK and N/C (not used). D8-D1 is the digital level of output, which is connected with FPGA / O interface and can transmit the collected data to FPGA. CLK is controlled by CLK clock accessed from external fpgai / O. The ADS830 collector is shown in Figure 4:

![Fig. 4. ADS830 collector](image)
2.3 Interface Switching Module

The output of ads830 high-speed data acquisition module is 10PIN female seat, with two rows of ten pins output, while the available interface of fpgai / O is one row of 16 pins. If the interface conversion is not carried out, it can only be connected with DuPont wire. For high-speed data, the connection of DuPont wire may cause data delay and loss, resulting in inaccurate data. Therefore, the interface conversion module is made to complete ads83 Interface conversion from 0 to fpgai / O. In the process of making interface conversion module, because ads830 is a high-speed data acquisition module, the frequency of data acquisition is high and the change is fast, the distance and width of internal wiring can lead to inaccurate data [6]. The physical connection between ads830 and interface conversion module is roughly designed as follows: ads830e collector on the left side, interface conversion module on the right side, pin connection between P1 interface of interface conversion module and ads830, connection between P2 and fpgai / O of three core development board, transfer the collected data to FPGA, D8-D1 interface of P2 output is used for data reception, and interface CLK is used when FPGA provides ads830 Clock signal.

3 Software design of automatic dehulling data acquisition system based on ARM

3.1 Main control software

The main control software runs on the main control board, which is the core of the whole system. Development environment: take a computer installed Ubuntu 2.04 desktop system as the host development environment, port the main control board of Linux 3.2.0 system through network cable connection, and use NFS to mount the shared directory. Qt4.7 development tool is installed in the host computer. Arm none linumx gnuabi-g + + tool chain is used to compile the program. The program is copied to the main control board through the shared directory and debugged on the main control board. The main control software consists of three subprograms: main program, ad subroutine, parameter subroutine and communication subroutine.

The main program is responsible for completing the logic control and real-time detecting whether there is new parameter input and setting [7], configuration of communication module parameters, calling ad subroutine at the set sampling time, calling communication subroutine at the set sending time, etc;

The parameter subroutine establishes the network connection with the upper computer software, obtains the set parameters in real time, and saves them in the SD card;

The ad subroutine controls the SD card on the main control board;

The communication subroutine receives the data from the communication module,
analyzes the GPS time and completes the time calibration.

3.2 GPS Data Acquisition Module

The difference between GPS receiving module and AIS receiving module is only that the ad_data [7:0] is replaced by CP, SGN and MAG, and the output part of ad clk is removed. GPS data acquisition has been completed by internal ad. the three important data pins output by GPS chip are CP, SGN and MAG signals. GPS intermediate frequency data is sampled and converted at the rising edge of CP and read out at the falling edge of CP. therefore, the falling edge of CP should be judged in the program, and then read out the data. Because the receiving program of GPS data is based on a1s program, the rate cannot exceed 11.2m * 8bit/s. The sampling rate of CP is nearly 16.4m, which is higher than that of AIS. Therefore, the two bits of data should be folded three times in one byte, that is, {sgn1、Mag1、sgn2、mag2、sgn3、MAG3、sgn4、mag4} to represent the first, second, third and fourth sampling points. The state of CP is recorded with the shift register busy reg, which is represented by the busy signal (busy reg). The falling edge trigger signal is generated as the trigger signal of data reception. In this way, there will be a delay of about 10ns in the trigger of K edge drop, but the operation of state machine can completely avoid this disadvantage and complete the data storage.

3.3 FIFO data caching module

FIFO is a first in, first out data cache array. Unlike general memory, its external output has no address line, so its advantages and disadvantages are obvious. Its advantages are convenient for user operation. Its disadvantages are that it can only store and take out data in sequence, and it can't write and read a certain data randomly according to the wishes of its own program [8]. It can complete the data transmission in different clock domain, complete the conversion transmission of data in low frequency domain and high frequency domain; it can complete the data buffer in different bit width, and the read and write clocks can be independent of each other. In FPGA program, FIFO IP core designed by Xilinx is used to complete data cache.

Set Key Parameters:

- Width of FIFO: the number of data bits of a read / write operation of FIFO, which is 8 bits;
- Depth of FIFO: how many n-bit data can be stored in FIFO (n is the width of FIFO), n is 1024;
- The read clock is the UPP read clock, i.e. 111mhz;
- The write clock is 100MHz; it is configured as normal type, and the read and write clock is in asynchronous mode.

FIFO through these configurations and pin connections, when the ad channel data write
flag is set, the data is written into FIFO until 1024 bytes are written. At this time, the full flag signal is set, and the data is sent to the UPP module according to the UPP read clock, so as to complete the data cache and transmission.

3.4 Local Data Processing Module

Linux is a mature free operating system, supporting a variety of hardware platforms, fully compatible with POSIX 1.0 standard, with complete development community support [9], supporting multi-threaded and real-time processing. The local data processing function module can be divided into two processes: basic data processing function module and total data processing function module, as shown in Figure 5.

Fig. 5. Local data processing processes

The basic data processing module is responsible for analyzing the abnormal data of the measurement points and generating corresponding events. After the operation of the module, the message queue, measurement point parameters, shared memory, task parameters, etc. will be initialized and enter the main loop. After a certain delay, read the data and analyze whether
it meets the event conditions. If the conditions are met, upload the data to the data communication module, read the statistical data of the terminal, and compare and analyze whether there is any change. If the data changes, read the historical data record and send it to the remote communication module [10], and then upload it to the main module Station. After all data processing is completed in the whole cycle, check whether the parameters of measurement point, event condition and task cycle have changed. If they change, carry out initialization again, and then continue data processing in the next cycle. The data processing function module of the total plus group is to make statistics on the data of the total plus group and send the data after the total plus to the data storage module for storage. After the operation of the module, the message queue, measurement point parameters, shared memory, total parameters, etc. will be initialized, and enter the main cycle, waiting for a total data statistics cycle. After that, the total data of the current data of the measuring point is calculated by using the summation formula, and the data is analyzed and compared to determine whether it meets the conditions of the difference situation. If it meets the conditions, a differential event will occur, and it will be sent to the remote communication module, and then uploaded to the main station by the module. Then, the current data and historical data records of all measurement points are generated into curve data and sent to the data storage module for storage. After all the processing of this cycle is completed, the parameters of each measuring point, the total plus parameters, the working cycle and the conditions of differential events are detected to determine whether these parameters have changed. If there is any change, restart and initialize, and then carry out the data processing of the next cycle.

4 Testing experiment

In order to verify the performance of the system (shelling effect and collection efficiency), six kinds of commonly used shelling data collection systems were compared and tested. Through comparison, the hypothesis of the experiment is verified. The experimental environment is shown in Table 1.

<table>
<thead>
<tr>
<th>Name</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operating system</td>
<td>Microsoft Windows XP</td>
</tr>
<tr>
<td>Processor</td>
<td>Intel(R)Celeron(R) 2.6GHz</td>
</tr>
</tbody>
</table>
Internal storage 24.0 GB

Hard disk 8.0 GB

Database management software

Microsoft SQL server 2010 R2

Mathematical software MATLAB

4.1 Experimental comparative analysis

In order to verify the effectiveness of the system, six kinds of commonly used shell tools are selected as the test object, and the remote control Trojan x-door is selected as the original program before shell adding. The six most commonly used shell adding tools are upx, pectopact, winupack, aspack, asprotect and mew. The experimental method is comparative experiment. The experiment verification process is completed under the control of other variables except the experimental variation. Figure 6 shows the data transmission terminal of program automatic shelling.

![Automatic dehulling of program data transfer terminal](image)

Fig. 6. Automatic dehulling of program data transfer terminal

Under the data transmission terminal, the remote service is used to collect the data in the program automatic shelling database.

4.2 Experimental result
In the actual test, the shelling effect of the system is compared with that of polyunpack, and the results are shown in Table 2.

Table 2 Comparison of system shelling and polyunpack shelling in this paper

<table>
<thead>
<tr>
<th>Shelling Tool</th>
<th>PloyUnpack</th>
<th>Dynamic dehulling</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPX</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>PECompact</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>WinUPack</td>
<td>Part</td>
<td>Yes</td>
</tr>
<tr>
<td>ASPack</td>
<td>Part</td>
<td>Yes</td>
</tr>
<tr>
<td>ASProtect</td>
<td>Yes</td>
<td>Part</td>
</tr>
<tr>
<td>MEW</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Among them, yes indicates correct shelling, no indicates no shelling, part indicates incorrect entry point identification, but part of the original program is obtained.

From Table 2, it can be concluded that polyunpack has not completely shelled except for the correct shelling of upx and asprotec. The shelling effect of the system technology proposed in this paper is better than polyunpack. At the same time, through the comparison of shelling time, the average time of shelling implementation in this paper is less than polyunpack. The reasons are as follows: first, polyunpack's single step comparison object is the generated decompilation result, while it is difficult to ensure that the binary executable is completely correct in the decompilation, and the shelling program often has the phenomenon of multiple shells. Compared with the system designed in this paper, polyunpack's processing effect on multiple shells is relatively poor; Second, polyunpack is based on single-step execution. Before each instruction is executed, it is checked and processed. The execution time of polyunpack is tens to tens of times of that of the original program, so the actual execution time of polyunpack is longer than that of the system in this paper.

In the actual test, the data collection efficiency of this system is compared with that of polyunpack, and the result is shown in Figure 7.
According to figure 7, the design system in this paper is far better than the traditional design in data collection efficiency, and higher than the standard line, with high collection efficiency and stability.

5 Concluding remarks

In this paper, a program auto shelling data acquisition system based on arm is designed. By comparing with six traditional auto shelling data acquisition systems, it is proved that the system designed in this paper has high shelling effect, high efficiency and better stability in data acquisition.

References


Design of distributed storage system of digital media information based on Metadata

ZHUXiao-Jun¹,LI Cai-hong²
{wz20160401@163.com¹,lgf147852@tom.com²}

(¹NanTong Normal College, Nantong 226010, China; ²Changchun Institute of Technology, Changchun 130021, China)

Abstract. In view of the poor load balancing performance of traditional distributed storage system of digital media information, a distributed storage system of digital media information based on metadata is designed. The hardware configuration of the system is server module and content distribution module. The server module is mainly composed of database server, data acquisition device server, SCADA server and routing table server. The content distribution module is mainly composed of multiple content distributors. The software of the system consists of central node module, task scheduling module, network module and client module. Through the combination of hardware and software, the distributed storage of digital media information is realized. In order to prove that the load balancing performance of the system is better, the traditional system is compared with the system. The experimental results show that the load balancing performance of the system is better than the traditional system.

Keywords: Metadata; Digital media information; Distributed storage system; Load balancing performance;

1 Introduction

The fifth information revolution in the 20th century refers to the digital revolution, which is marked by the widespread application of computers and the achievement of digital media represented by electronic technology and network technology, thus affecting the development of various industries in social life. For more than half a century, the application of digital art has become the mainstream of digital media art in entertainment culture and commercial display. Today, digital media art is widely used and flourishing. Digital media has fundamentally changed the social environment, economy and culture of people's lives. As an interdisciplinary subject of culture, art and technology, space design has also been affected unprecedentedly [2]. In this context, the creative industry economy has given birth to a new form of digital media art represented by digital animation, digital film and television, online advertising, online games. This kind of digital media art, with information science and digital technology as the main means, mass communication theory as the basis and modern art concept as the core, applies the information art form to the fields of military industry, culture, art, education, medicine and commerce, which is also a new media form of interdisciplinary integration of modern science and technology and art.
With the continuous development of digital society, the creative industry represented by digital media art animation and game has become a key industry in many countries. The development of digital media technology and the continuous integration of media provide new technical means for the content and form of art design, and form a new art form. Its application has become the most dynamic and creative emerging industry in the cultural and creative industry, and has formed a certain industrial model. People use digital media and information technology to integrate text, image, voice, image and other information, and generate a series of related products, technologies and services.

Twenty-one In the 21st century, driven by the context of "globalization", the era of digital information is coming. With the rapid development of information technology, the number of digital media information is increasing, the area of digital media information is expanding and the speed of transmission is speeding up. People's life has entered the era of digital media information explosion. Digital media information has penetrated into people's daily life and become the most common in our life a part. At the same time, the development of the network provides a wider range and real-time way for the dissemination of digital media information. In reference [3], a distributed data storage architecture for data storage and management of power transmission and transformation projects is proposed. Based on the metadata model, the architecture refines the three types of data, i.e. engineering geographic information, three-dimensional design model and document data of power transmission and transformation projects, and according to the different data storage modes, A distributed storage architecture is designed to deal with all kinds of data, but the load balancing performance of the storage system is poor.

In view of the problems of the above methods, this paper designs a digital media information distributed storage system based on metadata, which solves the problems of the traditional system.

2 Design a distributed storage system of digital media information based on Metadata

2.1 Hardware design

The hardware configuration of the metadata based digital media information distributed storage system is server module and content distribution module [4].

2.1.1 Server module

The server module is mainly composed of database server, data acquisition device server, SCADA server and routing table server. The access flow of digital media information data is as follows:

1. Access the local routing table server and query the routing table according to the IP
address;

2. Start from the forwarding server of the client, enter the next hop according to the routing table results, and continue to step 1 until reaching the target server or the query fails. If it fails, the unreachable information will be returned; otherwise, the corresponding operation processing flow will be executed. In the process of processing, if the client requires the query or processing result to be an intermediate value, the intermediate result data will be returned after corresponding calculation and processing at the target server;

3. After the client displays the corresponding results or integrates the intermediate results, continue to perform the follow-up tasks [5].

2.1.2 Content distribution module

The content distribution module is mainly composed of multiple content distributors, which can manage all activities of all user nodes in its domain. The content distributor only stores the basic information of the digital media information file and the index information of the digital media information file segments, not the actual content of the digital media information file resources. When a user node uploads a file, the content distributor splits the uploaded file, distributes it to the user node under its jurisdiction, and records the basic information and index information of the file locally; when a user node queries a file, the content distributor finds the corresponding information of the query file by querying the file information stored in the whole upper P2P network, and returns it to the user section. When a user node requests to download a file, the content distributor that receives the download request requests the user for file segmentation according to the local partition index information, and reconstructs the original file and transmits it to the user node. In the upper P2P network, the content distributor will send the local file information and the partition index information to its neighbor nodes for redundant storage, so as to ensure the fault tolerance and invulnerability of the system [6]. The content distributor maintains the network topology of the system by sending Ping messages to each other and receiving heartbeat information from user nodes.

The content distribution module uses the message trigger mechanism, and the content distributor calls the corresponding function module according to the received message type. The functional partition of the content distributor is shown in Figure 1.
According to the figure, the content distributor is composed of resource query function, file upload function, file download function, information fusion redundant storage function and network extended flutter maintenance function.

2.2 Software design

The software structure of the distributed storage system of digital media information based on metadata includes central node module, task scheduling module, network module and client module.

2.2.1 Central Node Module

The central node module is designed based on metadata. The main tasks of the central node module include processing the location information requests of clients, assigning new cache tasks and prefetch tasks to cache nodes, or assigning new storage tasks to persistent nodes, maintaining all the metadata of the cluster, managing all the read cache nodes in the cluster, writing and prefetch cache nodes and persistent node management node failures, etc [7]. The main classes related to the central node are shown in Figure 2.
Each central node in the central node module corresponds to a DBPs < center instance. Running its run() function service, it starts. It uses three service ports, one for processing the client's request, one for communicating with the data node, and the other for synchronization between the master and the standby. The events that nodes need to deal with are mainly divided into two categories: one is network events in digital media information, the other is timing events in digital media information.

Among them, network events are driven by messages. When the central node reads and parses the messages from the socket, it processes them according to the source and type of messages. These events include processing client's request, receiving cache node's heartbeat information, error report information and metadata update information, and processing cache node's data location request. For example, the network part of the node adopts the I / O multiplexer and non blocking I / O reactor mode. DBPs < center defines the callback function

```
#include <stdio.h>

typedef void (*TimerFunction)(struct TimerEntity);

struct TimerEntity {
    TimerFunction timerFunction;
    int cycle;
    int cycleCount;
    struct timer_t timer;
    short inPeriodic;
};

struct Timer {
    TimerEntity timerEntity;
    int timerID;
    struct timer_t timer;
    short inPeriodic;
};

struct TaskScheduler {
    char *centerHandle;
    DBPsCenter *m_center;
    void *m_data;
    void *taskScheduler;
    int m_taskCount;
    DBPsCenter *m_center;
};

struct DBPsCenter {
    short m_numMaster;
    char *m_center;
    void *m_data;
    void *taskScheduler;
    int m_taskCount;
};
```

Fig. 2. Major classes associated with the central node
onreadactionforclient to handle network events. The variable mainreactor is used to listen for newly initiated connections. Other reactor instances stored in the reactor pool run in multiple threads, mainly responsible for the central node and the storage section Communication between point and client [8].

Timer events are executed by timers, as shown in class timer in Figure 2. Timer only uses a real timer, and all timer events registered by nodes are realized by counters. For example, timer counts every 1ms, and timer events with a period of 1s will be executed when the count value reaches 1000, and each timer event is represented by a timerentity It records whether the functions and events to be executed at a fixed time are periodic events, and whether they have a long time or an execution cycle. The timing events mainly include the routine inspection procedures and cleaning procedures of the central node. For example, it checks whether the data node times out, does not contact with the central node, carries out corresponding processing, checks whether there is a timeout client connection, closes them, and regularly tracks the state time of the source object last written to the disk, regularly clears the invalid location information and node information, and also includes the timeout connection from the data node.

DBPS Center uses a circular buffer clientConnectionCircleBuffer to handle the Conn timeout of the client. Each buffer slot stores the client ConnectionAgent with corresponding time to the timeout. The pointer lastSlot always points to the slot with the longest time to the timeout. The adjacent slots are separated by a unit of time, so the previous slot of the lastSlot is the current timeout connection Section agent collection. Timing event will make lastslot move forward one slot in each unit time and close the connection in the slot. Each client connection will move itself from the last stored slot to the slot indicated by the current lastslot when it is active, so as to update the timeout.

2.2.2 Task scheduling module

Task scheduling module mainly includes task scheduler, whose tasks include allocating appropriate read cache nodes and write and prefetch cache nodes for client requests, allocating persistent nodes for new traceability objects or agents, selecting appropriate nodes to replace the tasks of the original nodes in case of node failure, etc. After receiving the client's request, the central node will parse the message, get the request type and parameters, and then call the corresponding task allocation function of taskscheduler according to the request type to get the allocation result and return it to the client. The whole process is shown in Figure 3.
Taskscheduler defines the task scheduling function used by the central node when receiving the client location information request. According to the operation type and operation object type of the client request, it can be divided into: assigning the traceable object or agent in the query request to the read cache node where the data is stored in the uncached time period, creating or modifying / deleting the traceable object or agent in the request. Configure the prefetch node that is responsible for processing the request, assign the persistent node that is responsible for managing its data to the newly added traceability object or agent [9]. According to the metadata maintained in metadatamanager, taskscheduler uses load balancing strategy to allocate tasks. In addition, it provides a node list parameter minorserver

Fig. 3. Client request task scheduling
in all task allocation functions. The list contains the nodes with which the client has recently failed to connect. It enables the client to participate in the selection of nodes and avoid getting invalid scheduling nodes again. As a result, the task scheduler avoids the nodes in the list when selecting service nodes for clients.

First, call getReadLocationForEntity to get several time periods and their corresponding cache nodes, then check the time periods that have not been cached and assign them appropriate service nodes. TaskScheduler segmenting the non buffering time segment according to the unit time sizePerSecond of the traceable object, then selecting multiple alternative nodes by function nth_optimalReadServer, and selecting the highest priority node to cache the data in a certain time period. It will call the MetaDataManager distributeReadServerForE after each assignment task is completed. The entity function updates the metadata.

2.2.3 Design network module

The network module consists of four main components. One is the I/O event source to be monitored. In network events, it mainly refers to the socket file. The other is the multiplexer, such as select, poll, epoll. The monitored event source needs to be registered with these multiplexers, and it is informed of the events to be monitored (readable, writable or with errors) and the handling to be called after the event occurs. Program, the third is the manager that deals with the whole event driven process, which provides the user with the interface to register the event, and registers the event to the I/O multiplexer. It manages all event sources, circulates the execution of waiting and processing programs, and the fourth is the event processing function that contains the application logic. Reactor will associate the event source with its corresponding processing function, and adjust when the event occurs. Use it.

The reactor is the manager. The reactorepoll is the I/O multiplexer using epoll. The reactorevent corresponds to an event source, and records the file descriptors related to the event source, the event types to be monitored and the event handling functions. It is registered to the reactorepoll through the reactor. The reactorepoll will create an epoll "event" according to the reactorevent and open it in the next event cycle. To start listening, reactor contains a reactorepoll instance. It runs the event listening function of reactorepoll. When an event occurs, it obtains the event source related to it, reactorevent, and calls its event processing function. The event processing function is registered by the creator of reactorevent, usually serveragent or connectionagent, etc.

Serveragent represents a server agent, clientagent represents a client agent, and connectionagent represents a connection. The connectionagent instance corresponds to a successfully created connection, records the corresponding socket, and contains the event
source created for it, reactorevent, which provides the event handling functions of socket readability, writability and error to reactorevent, through which the user interacts with the opposite end, and provides the function of sending messages to the opposite end and the user registration callback function to the user Function. The callback function describes the processing that the user needs to do when different events occur, such as when the connection is successfully established, when the data is readable, when the message is sent, etc. it reads and writes the data of the socket through the buffer of the application layer, so that the user only needs to read and write the buffer, and does not need to interact with the kernel [10].

In the network module, there is communication between the client and each node. They send request, response or report information to each other. The message format is shown in Figure 4.

![Message format](image)

**Fig. 4. Message format**

The message header contains three fields, type indicates the type of message, and the node will parse the text part of the message body according to the message type after receiving the message; length indicates the number of bytes occupied by the text of the message body, and the node will judge whether the message received is complete according to it; option indicates optional parameter, in which the node can store some important information or simple message information Set text to null. For example, when a node returns error information for a request, it can store the error code in it. Text is the main part of the message, which is the byte order encoded by protobuf. After the node receives the message, it uses the deserialization function of protobuf to parse the text according to the message type, so as to read the content. The format of all message bodies needs to be in advance Defined in the protobuf file, the encoding and compression algorithm of protobuf reduces the message volume.
2.2.4 Design client module

The client module designs the API including crud function according to the user's demand. The user creates the instance of entity, activity or agent by executing the constructor of class entity, activity and agent, and sets their properties through the member function, including the object name and version number of entity, the agent it belongs to, the time when it begins to exist, the start and end time of activity, and the sending time Starting from its agent and action name, agent namespace and name, etc., by setting the relationship between them, create an association for them, and finally use the commit function to submit the new instance to the system. The API related to the creation operation includes the API shown in Table 1 in addition to the constructor and member function of the entity / activity / agent class.

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Create operation API</th>
<th>API function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>void setRelationship(ProvObj* obj, ProvObj* obj, rel_t type)</td>
<td>Set the relationship between two objects, including wasused, wasattributedto, etc</td>
</tr>
<tr>
<td>2</td>
<td>int commit(ProvObj* obj)</td>
<td>Write objects to DBPs system</td>
</tr>
</tbody>
</table>

3 Simulation experiment design and result analysis
3.1 Design comparison experiment

The distributed storage system of digital media information based on metadata is used in the experiment of distributed storage of digital media information. Set up the experimental platform of the system, and build spark on Hadoop HDFS, which is jointly stored by HDFS and HBase, coordinated by zookeeper, and spark is responsible for the calculation part of the experiment. The experimental cluster is configured on 7 machines, and the operating system uses Ubuntu 14.04 LTS. Table 2 shows the specific configuration and version. You can expand the nodes according to the actual application. The configuration of all nodes of each machine in the cluster is shown in Table 3.

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Name</th>
<th>To configure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hadoop version</td>
<td>Hadoop-2.6.0</td>
</tr>
<tr>
<td>2</td>
<td>Spark version</td>
<td>Spark-1.5.0</td>
</tr>
<tr>
<td>3</td>
<td>HBase version</td>
<td>HBase-1.1.4</td>
</tr>
</tbody>
</table>
In order to ensure the effectiveness of the experiment, the traditional digital media information distributed storage system is compared with the metadata based digital media information distributed storage system designed in this paper. Compare the load balancing...
performance of the distributed storage system of digital media information. The judgment of load balancing performance is based on the fluctuation of load value of each system. The greater the fluctuation is, the better the load balancing performance is.

3.2 Analysis of experimental results

The experimental results of load balancing performance between the traditional digital media information distributed storage system and the metadata based digital media information distributed storage system designed in this paper are shown in Figure 5.

![Fig. 5. Experimental results of load balancing performance](image)

According to the experimental results of load balancing performance in Figure 5, the load balancing performance of the digital media information distributed storage system based on metadata is better than that of the traditional digital media information distributed storage system.

4 Concluding remarks

Due to the poor load balancing performance of traditional distributed storage system of digital media information, this paper designs a distributed storage system of digital media information based on metadata. Through the combination of hardware and software, the distributed storage of digital media information is realized. The distributed storage system of digital media information based on metadata can improve the load balance performance, which is of great significance for the distributed storage of digital media information.

References


Data storage system of personalized multimedia network teaching resources based on Hypertext

WANG Fei-fei¹, LIANG Yan²
{WANGfeiF06@163.com¹, LiangYan3263@163.com²}
¹,²Jilin International Studies University, Changchun 130000, China;

Abstract: The feedback effect of traditional multimedia network teaching resource data storage system is not good, which leads to the incomplete data when students query or download teaching resources. Therefore, a personalized multimedia network teaching resource data storage system based on hypertext is designed. On the basis of the original system hardware, the memory with larger capacity is selected, and the interface of the memory with larger capacity is redesigned; at the same time, the drive circuit of the memory is designed to ensure the smooth operation of the data. In the aspect of software design, by establishing the queuing model of teaching resource storage process, redefining the data storage program, and based on hypertext management storage system database, the high feedback storage system design is realized. The experimental results show that the feedback effect of the system is better than that of the traditional data storage system. Therefore, the system can be applied to the current personalized multimedia network teaching.

Keywords: Hypertext; Personalized Multimedia Network Teaching; Resource Data Storage System

1 Introduction

The feedback ability of traditional teaching resource data storage system is weak, so in order to ensure teaching resources and teaching level, improve students' learning interest and learning ability, a personalized multimedia network teaching resource data storage system based on hypertext is designed [1]. Hypertext is a form of information management technology or electronic document. One of its characteristics is an anti-traditional way of recording information in linear order. It imitates human's associative memory thinking and stores and records the information of teaching resources in a network structure [2].

In the design of multimedia teaching resource storage system, the non-linear technology
of hypertext and hypermedia breaks the linear rule of organizing multimedia information \cite{3-4}. According to people's habitual way of thinking, it can build a database feedback mode with strong individualized teaching resource data in the form required by students, and realize data query and download with stronger feedback for students. Provide perfect learning tools.

2 Hardware Design of Personalized Multimedia Network Teaching Resource Data Storage System

2.1 Selection and Interface Design of Large Capacity Memory

At present, there are two kinds of memory used in embedded storage system: ROM (read-only memory) and RAM (random access memory). Both of them belong to semiconductor memory, ROM can still save data when the system stops power supply, and the data in RAM will be lost after the system is powered off. Considering the importance of personalized multimedia network teaching resources, the most suitable memory for data resource storage system is Flash ROM.

The mainstream flash ROM includes NAND FLASH and NOR FLASH, both of which are fast and safe storage media without data loss due to power failure. NorFlash is characterized by faster reading, smaller capacity and slower writing and erasure, and is suitable for storage of programs and code. Nand Flash's storage unit is serially connected to external pins, and its internally integrated multiplexed ports are used for time-sharing of commands, addresses, and data. It is characterized by fast writing and erasure, large storage capacity, suitable for a system or electronic device that needs a lot of data storage. According to the requirement of rich and diversified teaching resources, the designed data storage system should choose the Nand Flash chip with large data storage capacity, quick response and parallel communication as the large capacity memory of the system. In terms of storage capacity, Nand Flash has 512KB-16GB of different capacity levels to choose from, is divided into MLC (MultiLevel Cell) architecture and SLC (SingleLevel Cell) architecture from the storage cell structure type. The NAND FLASH of the MCL framework can store multi-bit information in each storage unit, and the different data can be accurately stored in the storage unit through the precise control of different potential voltages in the cell. Each unit of the SCL framework, Nand Flash, can store 1-bit information. The advantage of the architecture is that the storage technology is relatively simple and the storage function is relatively fast.

Based on the specific requirements of the storage system, the storage data capacity needs to reach gb level, but the high temperature Nand Flash that can be purchased on the market at present is only the LHD S 1 GANand Flash chip that qingdao zhiteng company began to
represent at the end of 2017, as shown in figure 1 below:

**Figure 1** Physical diagram of large-capacity storage

The large-capacity storage in the above picture, its essence is Samsung's K9F8G08UOM type Nand FLASH, but it is encapsulated in the outside and made high temperature gold plating treatment, the maximum working temperature is 175 °C, which meets the temperature requirement of the system hardware under the multimedia long time teaching. K9F8G08UOM is constructed using SCL with a 1 GB of Nand Flash chip with 1 8GB memory chip (8GB for 1G byte) inside, each Chip contains 1 layer (Plane) and each layer contains 4096 storage blocks (Block). The following figure 2 shows the storage array organization diagram of K9F8G08UOM.
Fig.2 Structure of memory array

Where Block is the minimum unit for the Nand Flash to perform the erasure operation. Each block contains 128 pages (page), each page has (2 k64) bytes storage space, and the page is also the minimum unit for nand flash to program (write in). Therefore, the overall storage capacity of K9F8G08UOM is 8448 Mb, and the conventional storage capacity is 8GB, that is 1 GB.

2.2 Design of memory drive circuits

K9F8G08UOM type Nand Flash has eight command, address, data reuse of the I/O parallel bus interface, can transfer both the instruction of the teaching resources read and write, can also execute the erasure instruction, and can transmit the address of each storage unit, and the read and write data is also transmitted through these eight pins. The unique communication connection of Nand FLASH makes it different from that of Nor FLASH, which requires too many address pins, greatly improves the storage capacity of the memory, simplifies the hardware connection, and liberates the application resources of the main control chip. Its pin function is shown in Table 1 below:

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Pin Name</th>
<th>Pin Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I/O0-I/O7</td>
<td>I/O for command, address, data</td>
</tr>
<tr>
<td>2</td>
<td>CLE</td>
<td>Command latch enable</td>
</tr>
<tr>
<td>3</td>
<td>ALE</td>
<td>Address latch enable</td>
</tr>
<tr>
<td>4</td>
<td>CE</td>
<td>Chip Select Enable</td>
</tr>
<tr>
<td>5</td>
<td>RE</td>
<td>Read enable</td>
</tr>
<tr>
<td>6</td>
<td>WE</td>
<td>Write enable</td>
</tr>
<tr>
<td>7</td>
<td>WP</td>
<td>Write protected</td>
</tr>
<tr>
<td>8</td>
<td>R/B</td>
<td>Ready / busy output</td>
</tr>
</tbody>
</table>

K9F8G08UOM type Nand Flash is powered by 2.7V-3.6V DC voltage. Except for the power supply and the ground, the other angle P is connected to TMS320F28335, of which ALE, CLE are connected to the AO and A1 address lines of TMS320F28335; Read enable end RE, write enable end WE and F28335 DSP external read enable and external write enable respectively. The off-chip storage expansion Zone Zone Zone 7 of the chip selector connection DSP, the R/B pin selects pull up and connects a GPIO port of the DSP, and the write protection side also selects pull up, 8 data lines connected to main control chip data bus. A schematic diagram of the connection is shown in Figure 3 below:
Figure 3 Schematic diagram of the connection

It can be seen from Figure 3 that the design of memory driving circuit ensures the continuous driving ability of the storage system and the smooth operation of data storage, thus realizing the hardware design of personalized multimedia network teaching resources and data storage system.

3 Software Design of Teaching Resource Data Storage System Based on Hypertext

Based on the features of hypertext, the software of data storage system is designed to ensure the compatibility between storage system software and hardware.

3.1 Establishing the Queuing Model of Teaching Resource Store Process

The instructional resource storage queuing model, also called the service model, is used to sort the data of the supertext resource storage order. As we all know, service organization and service object constitute service model together. Service object is operator's command time, and storage system's service time to operator is random. Figure 4 is a schematic diagram of a simple queuing model:

![Figure 4 Queuing Model](image)

According to the above figure, the interface between the user and the database is stored
procedure, the user can access the database by calling the stored procedure of different functions, and by establishing the queuing model, the configuration of the stored procedure of the database as the service window is analyzed, so as to arrange the calling program more reasonably to realize the purpose of personalized multimedia network teaching.

This paper analyzes the user request, storage process and response process, and summarizes the scheduling characteristics of the model according to the characteristics of the queuing model. Take three indicators as an example to describe the input process of the storage system. Suppose for a certain period of time T, As n student users continue to use the teaching resource storage system to query teaching materials, the total number of proposed user sources is infinite, so exist \( n \rightarrow \infty \); When the users of the teaching resource data storage system are the \( f(x) \) service objects of the queuing model, by default, they come in random order, appear individually and independently of each other; because the users come randomly, appear individually and independently of each other, The sequence formed by the arrival of users can be compared to the input stream \( x_n \). It is assumed that the service window serving time for each customer and the interval between users' arrival are all negative exponential distribution. In \((0, T)\) time, the number of users arriving is a Poisson process.

Then set the sorting rules of the queuing model. When the user makes an operation request, the storage system first responds to the earliest request. If the stored procedure has been occupied, the user will enter the waiting stage. Therefore, the queuing rule of the system is first come first serve waiting. Suppose that in the storage system, a task is completed by a stored procedure, which completes the request according to the passing of a parameter. According to the above analysis and feature collection, the storage system queuing model can be described as follows:

\[
f(x_n) = \sum_{i=1}^{m} k_i x_n - \sigma_a / q(T) \tag{1}
\]

In the formula: \( k_i \) represents the search feature of the \( i \)th user; \( \sigma_a \) represents the constraints of the \( a \) user; \( m \) is the number of services. After the user puts forward the query operation request, some relevant parameters can be obtained by adding code to the stored procedure to report the execution progress, as shown in Table 2 below:

**Table 2 Relevant parameters**
Average column length | Average waiting column length | Average length of stay | Average waiting time
---|---|---|---
27 | 26 | 0.964 | 0.965
298 | 297 | 0.997 | 0.996
568 | 567 | 0.998 | 0.998

It is assumed that customers arrive according to the Poisson distribution with parameter \( g \). Service time is \( T \). The average service rate is \( \tau \), make \( n(T) = L \) the queue length at \( L \) time \( T \), the probability that \( q_n(T) = q \{ n(T) = L \} \) represents the queue length of \( n(T) = L \) is \( q_n \). According to the probability theory, we can get the balance equations of the storage system:

\[
\begin{align*}
q'(T) &= -(g + \tau)q_n(T) + gq_{n-b}(T) + \tau q_{n+b}(T) \\
q'_b(T) &= -gq_b(T) + gq_e(T)
\end{align*}
\]

In the formula: \( b \) represents random task requests among \( n \) operation requests. When \( T \to \infty \), Theoretically, it can be proved that when the storage system is stably distributed. By substituting formula (2) into formula (1), a complete queuing model of teaching resource storage can be obtained.

3.2 Storage function driver technology setting data storage program

According to the NAND flash storage array structure and sorting model, the driver and control program are designed to match the hardware of the storage system to realize the erasure and read-write control of NAND flash. Because there is no NAND flash Hardware Manager in TMS320F28335, it is necessary to develop the corresponding driver of NAND flash to realize the setting of data storage program.

It is known that NAND flash has a variety of data storage methods, including full page writing, single byte writing and other storage methods. At the same time, NAND flash has 8 I/O parallel bus interfaces for command, address and data multiplexing. When using single byte write storage mode, each time data storage is carried out, 8 data lines need to carry out write command transmission and addressing operation. The memory system designed in this paper has a memory space of 1024 bytes per page. If a single byte is used to write the memory, the operation cycle required to write a full page of data is relatively long, and the overall working
efficiency of the system is low. Therefore, the system uses the storage mode of whole page writing to store data, so as to achieve the purpose of relatively fast data storage [7-9].

When storing, the size of each page of the memory is 1024 bytes, and the data structure of each page is designed as follows: the first to the fourth words contain knowledge points and time information, and the later storage space is used to store the notes information of the multi-channel A / D sampling channel of the system to the sensor output. At the same time, in consideration of high-temperature chip cost, the data stored in NAND flash is read out to the on-chip RAM of DSP in turn by using queuing model, and the data in the middle buffer RAM is packaged and read out in turn by using the system development environment CCS. Before the operation of NAND flash, the corresponding operation instructions should be transmitted first, and the corresponding block or page should be addressed. Known operation instructions of LHDs 1 GA NAND flash are shown in Table 3.

<table>
<thead>
<tr>
<th>Table 3 Operation instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Features</td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>Read</td>
</tr>
<tr>
<td>Page</td>
</tr>
<tr>
<td>programming</td>
</tr>
<tr>
<td>Block erase</td>
</tr>
<tr>
<td>Reset</td>
</tr>
</tbody>
</table>

NAND flash has a special addressing cycle. It usually takes five cycles to locate a byte in the storage space. Taking LHDs 1 GA NAND flash as an example, the address of the first two cycles is called column address, in which the significant bit is ao-a12, which meets the intra page addressing of lhds 1ganand flash (211 = 2048, in which A12 is not used)[10-12]. The last three address cycles are called row addresses, which are used for inter page addressing and inter block addressing within blocks. The effective bits are a13-a30. The invalid bits of NAND flash address cycle should be set low, and its addressing cycle is shown in Table 4 below.

<table>
<thead>
<tr>
<th>Table 4 Addressing cycle table</th>
</tr>
</thead>
<tbody>
<tr>
<td>Add.</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>IO0</td>
</tr>
<tr>
<td>IO1</td>
</tr>
<tr>
<td>IO2</td>
</tr>
<tr>
<td>IO3</td>
</tr>
<tr>
<td>IO4</td>
</tr>
<tr>
<td>IO5</td>
</tr>
<tr>
<td>IO6</td>
</tr>
</tbody>
</table>
For the specific operation of NAND flash, first compare the operation instructions in the table, and send the first cycle operation instructions through the command latch enable end cle of NAND flash, and then transmit the address of the storage block or page required for the operation through the address latch enable end of NAND flash. After the operation is completed or the data transmission is completed, the second cycle operation instruction is sent through the command latch enable end cle. At the same time, through the ready / busy status bit R / B to determine whether the operation is completed or not, to achieve the storage system storage program settings.

3.3 Hypertext Management Storage System Database

Hypertext Technology is a kind of information management technology. It uses nodes (also known as knowledge points or information blocks) as the basic unit, and organizes nodes into a mesh structure with chains. If the teaching resource data node includes not only text, but also multimedia information such as sound, animation, graphics, image, it is called hypermedia. The hypertext structure is shown in Figure 5:

![Schematic diagram of the hypertext structure](image)

Figure 5 Schematic diagram of the hypertext structure

As can be seen from the figure, hypertext structure is composed of several internally interconnected text blocks (or other types of information, so it is generally called information blocks)[13-15]. Each node has several pointers to or from other nodes, which are called links. If each text block in Figure 5 is regarded as a node, and each chain (regardless of its specific starting position in the text block) is regarded as a directed edge, then the directed graph as shown in Figure 6 is obtained, which is the topological structure of hypertext.
Manage storage system database according to hypertext topology. Suppose \( \{s_1, s_2, s_3, s_4\} \) is an independent knowledge state, and all the knowledge states are known, that is, the set that does not conflict with the premise relationship. The knowledge space of the knowledge domain \( W \) is composed of empty set \( \emptyset \) and full set \( W \). The Hasse diagram of the space is shown in Figure 7 below:

According to the above analysis, the database planning algorithm is set to control the feedback mode of the storage system database. The calculation equation of the algorithm is:

\[
h(\mathcal{G}) = \frac{k + \sqrt{1 - k^2}}{1 + \lambda^{-\mu(\mathcal{G} - k)}} \quad (3)
\]

Where: \( \mathcal{G} \) represents the standard value of system feedback capability; \( k \) represents the logical relationship of teaching materials; \( \lambda \) is the general value of database feedback; \( \mu \) is the feedback coefficient. At this point, the management of the storage system database is completed, and the design of personalized multimedia network teaching resource data storage system based on hypertext is realized.
4 Testing experiment

The data storage system based on hypertext is verified and tested by Windows operating system, and the system is compared with the data storage system designed by traditional methods to analyze the functional characteristics of the hypertext storage system. Open the multimedia network teaching system, as shown in figure 8 below, upload a large number of teaching resource data to the system, and use it as the basic data source center to design the data storage system.

![System login interface](image)

**Figure 8** System login interface

Login the system to obtain experimental basic data, using two methods, design personalized multimedia network teaching resources data storage system. 100 students were randomly selected and divided into two experimental groups, group A and group B, of which group A was the test participants of the designed system and group B was the test participants of the traditional system. The results of the experiment are shown in Table 5 and Table 6 below:

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Feedback result</th>
<th>Serial number</th>
<th>Feedback result</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>✓</td>
<td>M11</td>
<td>✓</td>
</tr>
<tr>
<td>M2</td>
<td>✓</td>
<td>M12</td>
<td>✓</td>
</tr>
<tr>
<td>M3</td>
<td>✓</td>
<td>M13</td>
<td>×</td>
</tr>
<tr>
<td>M4</td>
<td>✓</td>
<td>M14</td>
<td>✓</td>
</tr>
<tr>
<td>M5</td>
<td>✓</td>
<td>M15</td>
<td>✓</td>
</tr>
<tr>
<td>M6</td>
<td>✓</td>
<td>M16</td>
<td>✓</td>
</tr>
<tr>
<td>M7</td>
<td>✓</td>
<td>M17</td>
<td>✓</td>
</tr>
<tr>
<td>M8</td>
<td>✓</td>
<td>M18</td>
<td>✓</td>
</tr>
<tr>
<td>M9</td>
<td>✓</td>
<td>M19</td>
<td>✓</td>
</tr>
</tbody>
</table>
According to the data in the above table, under the same experimental test conditions, the hypertext based data storage system has only one prompt error in its feedback, while the traditional method design data storage system has nine prompt errors in its feedback results. It can be seen that hypertext-based storage systems have better feedback.

5 Concluding remarks
The method of designing the data storage system of teaching resources, according to the functional characteristics of hypertext technology, a storage system with better compatibility of software and hardware and stronger feedback ability is designed to ensure that the data can be accurately feedback the data search results through hypertext topology results. However, the design of the storage system is only for the current stage of hardware design, if there is a better performance of hardware, teaching resources data storage system can be re-designed.

6 Fund projects
The article is a phased research achievement of the National Computer Basic Education Research Project "Research on the Reform of Computer Practice Teaching System for Liberal Arts Majors in the Perspective of Innovation and Entrepreneurship"(No: 2019-AFCEC-042)

References


Abnormal Data Mining Method in Environmental Monitoring Data of Animal Husbandry Farm

XU Xiao-hua¹, CHEN Chang-xi²*
{yg69030@163.com¹,xuxh69850@163.com²*}
(¹.Tianjin Agricultural University, Tianjin 300384, China; ².Tianjin Agricultural University, Tianjin 300384, China)

Abstract: In order to solve the problem of low accuracy of traditional anomaly data mining methods, this paper proposes an anomaly data mining method in the environmental monitoring data of livestock farms. Through collecting the environmental monitoring data of animal husbandry by sensors, after getting the environmental monitoring data, the environmental monitoring data is preprocessed, and the data after preprocessing is mined to complete the design of abnormal data mining method in the environmental monitoring data of animal husbandry. Compared with the traditional methods of outlier data mining, the experimental results show that the proposed outlier data mining method has higher mining accuracy.

Key words: Animal husbandry; Environmental monitoring; Abnormal data; Mining;

1 Introduction

With the adjustment of agricultural structure and the support of relevant national policies, agriculture, forestry, animal husbandry and fishery have developed rapidly, especially animal husbandry accounts for 1/3 of the overall output value. According to the data of the National Bureau of statistics, animal husbandry is maintained at an average value of about 30% every year, which not only increases the economic income of farmers, but also promotes the circulation of the national economy and increases the GDP [1-3]. Although large-scale breeding has gradually replaced the traditional extensive breeding mode, this trend still cannot change the extensive management mode in the breeding process. The ventilation, sewage, heating and other facilities are not complete, or even not, resulting in the poor growth environment of livestock, which is easy to cause animal disease, death, and poor quality of animal products [4-6].

With the concept of "healthy breeding", some domestic farms have improved their understanding of breeding management, strengthened supervision and ensured product quality. However, imperfect breeding management is still common, resulting in frequent occurrence of animal diseases. Traditional breeding, extensive breeding, and sky dependent breeding still account for a considerable proportion [7]. With the emergence of new animal husbandry production organization forms, the traditional family and single family farming forms decrease
year by year. It is of great practical significance to use high technology to carry out environmental monitoring on livestock farms for improving the quality of livestock products and reducing the occurrence of various animal diseases. By mining the abnormal data in the environmental monitoring data of the animal husbandry farm, we can find the abnormal environment in the animal husbandry farm, so as to deal with it in time. However, in the traditional animal husbandry environmental monitoring data, the method of abnormal data mining has the disadvantage of low mining accuracy, which needs further research. On this basis, a method of mining abnormal data in the environmental monitoring data of livestock and poultry farms is proposed.

2 Method of abnormal data mining in environmental monitoring data of livestock farm

2.1 Data collection

First of all, the environmental monitoring data of livestock farms are collected by sensors. The sensor is responsible for collecting data, and then transmitting it to the microcontroller, which then transmits it. Figure 1 shows the flow chart of the sensor.

![Processing flow of acquisition sensor](image)

Figure 1 Processing flow of acquisition sensor

As can be seen from Figure 1, the upper computer can control the acquisition of data by sending instructions. The upper computer sends instructions, and MCU receives the collection
information instructions from the monitoring end, then packs the data and sends the data to the upper computer system. After obtaining the environmental monitoring data of animal husbandry farm, the pretreatment is carried out [8].

2.2 Data preprocessing

Data preprocessing includes data cleaning, data integration, data conversion, data discretization, data reduction and data reduction. The data preprocessing process is shown in Figure 2.

![Flow chart of data processing](image)

**Figure 2** Flow chart of data processing

In the integration stage of data cleaning and data, first, simple data filtering and cleaning are carried out, errors and duplicate data are removed, and then relatively "clean" data is obtained, and then incomplete data is merged or completed, then data is standardized and normalized, and the final data values are all between [0,1]. So it provides a good foundation for the abnormal data mining of environmental monitoring data of livestock and poultry farms. [9].

The data reduction algorithm can well deal with attribute reduction under the condition of dynamic increase of environmental monitoring data in animal husbandry farm. The reduction effect is obvious, and the time is very short. The processing structure is shown in Figure 3.
In Figure 3, the environmental monitoring data is reduced. The specific implementation process is as follows:

According to the rough set theory, mutual information equality can be used as the termination condition to find the relative reduction of knowledge. Based on this, a new calculation formula of attribute importance is proposed in combination with mutual information theory. This formula does not calculate the importance of data unrelated to attribute reduction, reduces the calculation time of attribute importance, and thus reduces the time of attribute reduction. The formula is described as follows:

$$
\text{sig}(a, B; C) = H(B \cup \{a\}) - H(B) \\
= - \sum_{i=1}^{n} p(X_i) \log 2^{p(X_i)} + \sum_{i=1}^{n} p(Y_i) \log 2^{p(Y_i)}
$$

(1)

In formula (1), \( p(X_i) \) represents the probability distribution of each object in \( B \cup \{a\} \), \( p(Y_i) \) represents the probability distribution of each object in \( B \) (where \( p(X_i) = \frac{\text{card}(X_i)}{\text{card}(U)} \)), \( a = \{b | b \in \{C-B\} \wedge \{b\} \cup c_y \} \) and \( c_y \) represent the elements in the difference matrix. The attribute importance is calculated by formula (1). The specific steps of the algorithm are as follows:

Input: livestock farm information \( S = (U, C, V, f) \) and new object set (new livestock farm environmental monitoring data) \( \{x_{s1}, x_{s2}, \ldots, x_{sn}\} \); output: a attribute reduction \( B \).
calculate attribute importance, if the conditions are met, the algorithm ends.

For the processing of missing value, Lagrange interpolation method is used: take out the five data before and after the missing value respectively. If the data before and after the missing value is empty or nonexistent, directly round off the data. According to the ten data, form a group, and use Lagrange difference formula to interpolate the data. The Lagrangian difference formula is shown in formula (2):

$$L_n(x) = \sum_{i=0}^{n} l_i(x) y_i$$
$$l_i(x) = \prod_{j=0,j\neq i}^{n} \frac{x-x_j}{x_i-x_j} \quad (2)$$

Where, $x$ represents the $x$ data, $y_i$ represents the sample value, and $l_i(x)$ represents the Lagrangian polynomial.

After that, the data is normalized. If the data values of different dimensions are different or distributed unevenly, the training time will be longer and the data processing effect will be affected. Normalization processing in order to solve this problem, it will be used in the abnormal data mining of environmental monitoring data of animal husbandry farm, and the environmental monitoring data will be normalized using formula (3):

$$NL = \frac{L - \min(x)}{\max(x) - \min(x)} \quad (3)$$

In formula (3), $L$ represents the original value before data normalization, $\max(x)$ represents the maximum value of all values under the dimension before data normalization, $\min(x)$ represents the minimum value of all values under the dimension before data normalization, and $NL$ represents the normalized characteristic value. Through the above process, realize the pretreatment of the environmental monitoring data of the livestock farm [10].

2.3 Data mining

Data mining is done by clustering and fast computing algorithm. In clustering algorithm, outliers are usually regarded as leaf nodes in clustering feature tree, and the density of leaf nodes is very low, usually set a fixed threshold. If the number of sample data objects of one leaf node is less than the set threshold or the number of samples in this leaf node is the least compared with other leaf nodes, then it can be considered that it is very possible that the leaf
node is an outlier, and then the leaf node is added to the outlier data set. If the set threshold changes, so that the sample object that has been included in the abnormal data set does not meet the condition of being an abnormal point, it should be reclassified into the clustering feature tree. First, cluster the original data set, and calculate the size of the local anomaly factor, and it is also related to the density of its surrounding neighbors.

The density based lof algorithm mainly involves the calculation of \( k \)−distance, \( k \)−distance neighborhood, reachable distance and reachable density of data objects. The \( k \)−distance of data object \( p \) can determine which data objects in the data set can constitute \( p \)'s neighbors. To a certain extent, it measures the density of the area around \( p \). If the \( k \)−distance of \( p \) is larger, it indicates that the area around \( p \) is sparse; on the contrary, if the \( k \)−distance of \( p \) is smaller, it indicates that the area around \( p \) is dense. Figure 4 vividly explains the \( k \)−distance of \( p \) and the neighbor object of \( p \).

![Figure 4 Neighborhood of \( k \)-distance and \( k \)-distance of \( p \)](image)

In Figure 4, a total of six data objects are set, which are \( p_1, p_2, p_3, p_4, p_5 \) and \( p_6 \) respectively. At the same time, the distance from each data object to \( p \) is marked. Suppose \( n \) represents the total number of objects in the sample data set, there are \( k−1 \)
objects in the circle, and the remaining \( n-k-1 \) objects are distributed outside the circle, then the distance between the object \( p_k \) and the object \( p_i \), which is just on the circumference boundary, is the \( k \) distance between the object \( p_1 \) and the object \( p_i \).

The data object located on and within the circle boundary is defined as the neighbor object of object \( p_i \), for example, in Figure 4 (\( p_2, p_3, p_4 \)), it is the neighbor object of object \( p \).

According to Figure 4, the following definitions are given:

Defines the \( k \) distance of data object \( p \). For any positive integer \( k \) in the data set, the \( k \) distance of object \( p \) is defined as the Euclidean distance between \( p \) and an object \( o \), where object \( o \) belongs to data set \( D \) and the following conditions need to be met simultaneously: there are at least \( k \) data objects \( o' \in D \setminus \{ p \} \) and \( d(p,o') \leq d(p,o) \);

there are at most \( k-1 \) data objects \( o' \in D \setminus \{ p \} \) and \( (p,o') < d(p,o) \). When calculating the \( k \) distance of object \( p \), you can first calculate the distance of all objects reaching \( p \) respectively, and then select \( k \) minimum distances. The maximum distance of \( k \) objects already selected is defined as the \( k \) distance of \( p \).

Defines the \( k \) distance neighborhood of data object \( p \). The definition of \( k \) distance of given object \( p \) in data set is the collection of all data objects within its \( k \) distance neighborhood whose distance from object \( p \) does not exceed \( k \) distance, namely:

\[
N_k \text{dist}(p) = \{ q \in D \setminus \{ p \} \mid d(p,q) \leq k \}\)  

\( (4) \)

Defines the reachable distance of data object \( p \). For a given natural number \( k \), object \( p \)
is defined relative to object \( o \) as:

\[
reach - dist_k(p,o) = \max\{k - \text{dist}\tan ce(o), d(p,o)\} \tag{5}
\]

Defines the local reachable density of data object \( P \). For the given positive number \( k \), the reciprocal of the average reachable distance between the object \( P \) and its neighborhood relative to its distance \( k - \) is defined as the local reachable density, and the calculation method is as follows:

\[
lrd_k(p) = \frac{1}{\sum\limits_{q \in N_k(p)} reach - dis_k(p,q)} \tag{6}
\]

In formula (6), \( lrd_k(p) \) represents the local reachable density, and \( \sum\limits_{q \in N_k(p)} reach - dis_k(p,q) \) \( \frac{1}{\left| N_k(p) \right|} \) represents the average reachable distance between the object \( p \) and its neighborhood relative to its \( k - \) distance. Define the local outlier of data object \( p \). For a given positive number \( k \) in the data set, the calculation method of the local outlier factor \( LOF_k(p) \) of the object \( p \) is as follows:

\[
LOF_k(p) = \frac{\sum\limits_{q \in N_k(p)} lrd_k(q)}{N_k(p)} \tag{7}
\]

The local outlier indicates the abnormal degree of an object. If the larger the local outlier is, the greater the abnormal degree of the object is. Otherwise, the smaller. Because the local exception factor of each object of lof algorithm is related to the environment of the object, in the dynamic incremental database environment, new data objects often affect the local exception factor of some objects in the original data set, so it is necessary to call lof algorithm to recalculate the local exception factor for all data objects. However, the new data object will
only affect the local abnormal factors of some data objects, but not all data objects. Therefore, it is improved. If the data in the database is updated, only the affected data objects need to be recalculated, which greatly improves the speed of abnormal data mining.

For a given sample data set $D$, for a known object $q$ in $D$, whenever an object $p$ is added, the $k$-distance neighborhood of $q$ changes due to the addition of object $q$, which also changes the reachable density and local outlier of $q$, and then a series of changes of $q$ will cause a series of changes. Here we define the affected object set: In a given sample object $D$, due to the addition of new object $p$, the local outlier of data in set $D$ changes, then the set of data is called the affected object set.

When adding object $p$ to data set $D$, it is necessary to recalculate $k$-distance neighborhood and local outlier factors for the affected object set, and update the neighborhood in the following three cases. If $q$ is the affected object and there are $d(q, p = \|k - dis(q)\|$, then $p$ can be directly added to the neighborhood of $q$. If $q$ is the affected object, and there are $d(q, p < \|k - dis(q)\|$, and at most $k - 2$ data $o$ make $d(q, o < \|k - dis(q)\| exist at the same time, then $p$ can also be directly added to the neighborhood of $g$. If $q$ is the affected object and $d(q, p < \|k - dis(q)\|, k - 1 data are met at the same time. Make $d(q, o < \|k - dis(q)\|$, then $p$ can be added to the neighborhood of object $q$, and the object farthest from $q$ can be deleted. According to the above definition, the abnormal data in the environmental monitoring data of animal husbandry farm are clustered.

Cluster the data in the original data set: first, cluster the original data set, because according to the definition of the abnormal data, the abnormal data is some data that accounts for a small part of the data set, so after clustering the original data, the calculation of the normal data that accounts for a large part of the data can be eliminated, so the normal data cluster $N$ and abnormal data $U$ are formed by clustering, continue processing until the end of clustering, forming data clusters $N_1, N_2, \cdots, N_i$ and $U_1, U_2, \cdots, U_i$.

Clustering new data: when dealing with abnormal data mining in the dynamic
incremental data environment, the newly added data are usually normal data, only a small amount of abnormal data. The algorithm avoids the high complexity of re-clustering when new data arrive. Make full use of the normal cluster \( N_i \), calculate the center object \( O_i \) in each cluster \( N_i \), then calculate the distance from each newly added object to the center object \( O_i \), and record the cluster when the minimum distance is obtained. If the minimum distance calculated is less than the set radius value \( \epsilon_i \), the object can be directly added to the cluster, if the distance can not meet the conditions for adding any cluster, a new abnormal cluster is formed. Add all new data objects to a normal or abnormal cluster.

Calculate the local outlier factors for the data in the abnormal cluster: after the clustering, the initial abnormal cluster \( U_1, U_2, \ldots, U_i \) and the newly added abnormal cluster \( U'_1, U'_2, \ldots, U'_i \) are formed. Because the amount of data in these abnormal clusters is not very large, the data can be directly calculated as \( k \)-distance neighborhood, local reachable density and local outlier factors.

In order to improve the speed of abnormal data mining in environmental monitoring data, the lof algorithm and DBSCAN algorithm are combined to improve the algorithm. First of all, DBSCAN algorithm is used to cluster the original data set. The dense data objects are subtracted from a large number of data sets, and the sparse data objects that are not clustered are retained to form abnormal clustering. On the basis of clustering, the data in the cluster is judged, and the abnormal data is added to the abnormal cluster. Finally, lof algorithm is used to calculate the local abnormal factors for the data objects in the abnormal cluster.

After clustering the new data objects, we can judge which data objects are dense and which are sparse, and add the sparse data objects to the new abnormal data set \( U' \). Then, on the basis of the above improved algorithm clustering, we calculate the average of the cluster centroid and the distance from all data objects to the centroid of the cluster, and then calculate the distance between each data object in the cluster and the cluster centroid. If the distance is less than the average value, the data object is considered normal, otherwise, it is considered abnormal data object, and the abnormal data object is added to the new abnormal data set port. Because the main research object of lof algorithm is the data object with sparse distribution in the data set, so the data object with sparse distribution in the original data set and the data object with sparse distribution in the new data set need to use lof algorithm to calculate the
local abnormal factor. Because the sparse data objects in the original dataset (the data objects in $U$) have calculated the local exception factor with the lof algorithm in advance, when the new data objects arrive, the data objects in the new exception dataset $U'$ and the exception dataset $U$ are affected by the data objects in the new exception dataset $U'$, resulting in the change of the local exception factor. Therefore, the data object in $U$ can be recalculated with the local exception factor and sorted according to the size of the local exception factor, so the former data object is the real exception data object.

At this point, the design of abnormal data mining method in environmental monitoring data of livestock farm has been completed.

3 Experiment

In order to verify whether the anomaly data mining method in the monitoring data of animal husbandry farm has higher mining accuracy, the paper compares the proposed anomaly data mining method with the traditional anomaly data mining method.

3.1 Experimental process

The experimental operation environment is shown in Table 1.

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Items</th>
<th>Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ComputerCPU</td>
<td>Pentium(R) Dual-Core <a href="mailto:T4300@2.10GHz">T4300@2.10GHz</a></td>
</tr>
<tr>
<td>2</td>
<td>RAM</td>
<td>2.0GB</td>
</tr>
<tr>
<td>3</td>
<td>Hard disk</td>
<td>320G</td>
</tr>
<tr>
<td>4</td>
<td>Software operating</td>
<td>Windows 7,VC++6.0</td>
</tr>
</tbody>
</table>

environment |

The experiment was carried out in the above experimental environment. First of all, to ensure the smooth progress of the experiment, we use the shuttle data set, which contains 43500 data objects. Now take 3000 data objects as the initial data set, and take 10, 30, 100, 200, 500 as the number of newly added data objects. In the case of $Esp = 18.245, Minpts = 20$, when the new data objects arrive, the running time result of the algorithm used is shown in Table 2.

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Number of new objects</th>
<th>Algorithm running time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>0.174s</td>
</tr>
</tbody>
</table>
It can be seen from table 2 that with the increase of new data objects, the calculation time of the algorithm does not increase significantly. The results show that the algorithm is efficient and meets the calculation requirements.

On this basis, in view of the five groups of animal husbandry environmental monitoring data with different amount of abnormal data (10, 20, 30, 40, 50), the mining accuracy of this paper's abnormal data mining method and traditional abnormal data mining method is compared.

3.2 Analysis of experimental results

Compare and analyze the mining accuracy of the proposed anomaly data mining method and the traditional anomaly data mining method. The comparison results are shown in Figure 5.

![Figure 5](image)

**Figure 5** Comparison results of mining accuracy

As shown in Figure 5, the maximum difference between the number of abnormal data mined by traditional exception data mining method and the actual number of abnormal data is 10; the maximum difference between the number of abnormal data mined by the proposed exception data mining method is 2. Through comparison, it is found that the proposed anomaly data mining method has higher mining accuracy.
4 Concluding

In view of the low accuracy of the traditional methods of outlier data mining, a method of outlier data mining in the environmental monitoring data of animal husbandry is proposed. Through the comparative experiment, compared with the traditional abnormal data mining methods, the experimental results show that the proposed abnormal data mining method has a higher mining accuracy, hoping that it can provide a certain reference value for the research of abnormal data mining in animal husbandry environmental monitoring data.

5 Fund projects
Post Scientist of National Broiler Industry Technology System (CARS-41-G25)
The Science and Technology Planning Program of Tianjin Science and Technology Bureau, China (18ZXRHNC00080)
Information Post of Tianjin Dairy Industry Technology System Innovation Team(ITTCRS2020023)

Reference
Design of multi-agent based decision system for building spatial planning

LIU Lin¹, YU Jia Jie²

{yjl59951000@163.com¹, yjj901229@163.com²}

(1. Department of architecture and art, ChongQing JianZhu College, ChongQing 400000, China; 2. Department of architecture and art, ChongQing JianZhu College, ChongQing 400000, China)

Abstract: In view of the low efficiency of traditional system planning, a multi-agent based decision-making system for building space planning is proposed. The spatial characteristics of traditional villages in southeastern ChongQing are analyzed and the overall structure is designed. B/S three-layer architecture is adopted, MapObject functional components are selected, and web browsers are set up to facilitate information retrieval. According to the actual situation, the Web server is divided into application hierarchy or scale scale, server structure and hardware type. Design the spatial decision-making information system by using various data and information, and process the business and basic information of the enterprise according to the information processing system of public participation. Two kinds of distributed database systems are designed to accommodate different databases for different purposes and improve the performance of the whole system. Under the natural environment, agricultural production and living conditions, design the spatial planning decision plan. According to the experimental results, the maximum planning efficiency of the system can reach 97%, providing a guarantee for building safety construction.

Keywords: Multi-agent; Architectural space planning; Decision making; Public participation in

1 Introduction

In practice, many business models have been formed by experts in the field of architectural spatial planning and decision-making, which urgently needs a platform for unified organization, management and sharing. At the same time, a large number of planning and implementation evaluation tools have been accumulated in the process of building spatial planning and decision-making information construction, which adopt different technology development systems and depend on different operational support environments. In order to enable relevant users of planning business to operate and use directly on the integrated and unified decision-making and management platform of architectural space planning, it is necessary to establish a unified environment supporting the operation of architectural space planning and business tools, build up a group of architectural space planning and decision-making tools, and make them play a greater application value[1].

Architectural spatial planning decision-making system is a complex system, whose
spatial evolution is characterized by a large number of dynamic and adaptive micro-behavior subjects (enterprises, residents, farmers and various organizations, etc.) as well as the nonlinear interaction between the behavior subjects and the environment, producing discontinuous architectural spatial decision-making behaviors. Because of its static characteristics, the traditional decision-making system of architectural spatial planning can not reflect the interaction process of many spatial time accumulation and spatial dimension in urban planning decision-making. The more difficult the spatial decision-making behavior of urban planning participants is to be effectively expressed, the more difficult it is to deal with the planning results, and there are certain restrictions on the acceptable level. Oriented to the needs of urban planning, and relevant theories of multi-agent and intelligent science and technology, build can clearly express the city planning to participate in decision-making behavior subject space-time dominant feature space for urban planning decision system, and based on the build system, carried out the architectural space planning decision system and analysis of applied research, auxiliary to recognize and solve encountered in the process of architectural planning related unstructured problems, so as to enhance the scientific nature and rationality of architectural space planning decision-making, in order to promote the sustainable development of city decision-making to provide advanced technical method and reference.

2 The overall architecture

Chongqing southeastern area of chongqing is located in the basin edge mountains where the two mountains of dalou mountain and wuling mountain meet in the southeast of sichuan basin. The four provinces of Chongqing, Hubei, Hunan and Guizhou are merged, including Qianjiang Development Zone and four minority autonomous counties, including Shizhu, Pengshui, Youyang and Xiushan, covering an area of about 16900 square kilometers, accounting for 20.5% of the total area of the city. There are 28 ethnic minorities including miao, tujia, meng, hui, qiyegui, dong, zang, yi, hani, zhuang, man and yao. At the end of 2007, the registered population of southeast chongqing was 3,549,400, of which 86.96% were agricultural.

Architectural space planning decision system based on the planning space, database, the spatial planning oriented visual modeling, spatial planning tool group building, multi-dimensional comprehensive decision support model and on the basis of participatory planning decision support technology, computer technology, geographic information system technology, the remote sensing technology, communication technology, network technology, space planning and the theory of system science and comprehensive decision method is applied in the spatial planning and management of the integrated system, implements the effective integration of spatial planning tool and space planning decision-making model
library building, service and management for the decision of spatial planning provides a strong support, it also provides a set of reliable and efficient analysis tools and decision analysis means for spatial planning[4-6].

The overall system framework is divided into five layers from bottom to top: base layer, architecture layer, core data layer, operation support layer and application layer, as shown in Figure 1.

Figure 1 System overall structure diagram

(1) Basic layer: including basic supporting environment, hardware infrastructure, software infrastructure and network infrastructure. By establishing the building space maintenance database, the functions of storage, management, analysis and processing of various building maintenance information can be realized. It is used for the management and planning of building space planning and maintenance to provide basic data support[7].

(2) Architecture layer: including standard system and exchange system, using scientific methods to realize quantitative analysis and calculation of building space status, providing data support for land suitability evaluation and space utilization planning.

(3) Core data layer: including data metadata database, basic database, professional database and management database, which provides scientific basis for the suitability evaluation of small-space buildings. Through this system, the appropriate grade of land in each block of buildings can be determined objectively and accurately, providing technical support for further governance work[8].

(4) Operation support layer: including model library and toolbox, it can realize the
optimization of building utilization in small space, formulate the optimal utilization mode of building in small space, and provide spatial information support for decision-makers in order to obtain the best decision keeping scheme.

(5) Application layer: including planning results management, planning decision analysis, participatory decision support and Shared services. This layer mainly has a good user experience interface, which can realize system functions through simple operation.

3 System hardware architecture

Based on the regional characteristics of southeastern chongqing, this paper studies the rural settlement landscape and considers the following three feasibility points for the establishment of the research object.

First, the urbanization level in southeastern chongqing is very low, and the rural population accounts for about 90% of the total population, which lays a necessary material foundation for the study of rural settlement landscape.

Second, the southeast of chongqing is a typical mountainous region, which is a typical "karst" landform. It has special topographic and geomorphic characteristics, abundant natural and cultural resources, complex engineering construction, unique landscape and other characteristics, and high landscape sensitivity, which is an important condition for the study.

Thirdly, the minority population in the southeast of chongqing accounts for about 70% of the total population, among which tujia and miao are the main minority in the southeast of chongqing. The rural settlement landscape in the southeast of chongqing has obvious regional characteristics and ethnic characteristics. This provides necessary conditions for the study of rural settlement landscape based on the regional characteristics of Southeast Chongqing.

The hardware architecture of the system adopts a three-tier architecture based on B/S, as shown in Figure 2.
The browser side is developed with multi-agent technology, and its main module is the public participation function module developed based on Java Applet -- PPGeoTool module, including the geography operation submodule and information feedback submodule. The submodule of geographic operations realizes the basic geographic operations of the client public, including attribute query, location query, spatial analysis, map browsing, map editing, display output, etc. The information feedback sub-module is used to realize the information interaction between the public and the system, including the inquiry of the information related to the planning project, the input and preservation of the information related to the public participation in decision-making and the public resume.

On the server side, MapObject functional components provided by ESRI are used to develop corresponding functional modules in VB language, including sub-modules of public participation information processing system and sub-modules of spatial decision-making information system. The sub-module of public participation information processing system is used to realize the functions of statistical analysis, spatial analysis, model analysis, information database building based on data model, and information display output. The sub-module of spatial decision information system is used to realize the spatial analysis function of traditional geographic information system and the auxiliary decision function of decision support system. On the basis of traditional geographic information database, public participation information database and rule database, the spatial analysis and decision-making of the system must introduce new analysis models and methods into the ten thousand square data of urban planning projects.
3.1 Browser

A web browser is an application that retrieves and displays information about the world wide web. These information resources can be web pages, pictures, videos or other content, they are marked by the uniform resource identifier, hyperlinks in the information resources to facilitate the user to browse relevant information.

The browser is accessed through a client program, known as a web browser, because it allows users to roam based on hypertext links without having to make purposeful queries. The browser is software that displays the contents of HTML files on a web server or file system and lets users interact with those files. Web browsers primarily interact with and retrieve web pages from web servers over the HTTP protocol, which are specified by urls, usually in HTML format, and specified by MIME in the HTTP protocol.

3.2 Web server

Network server is the core part of computer LAN, its efficiency directly affects the whole network operating system. Web servers have run the main function of the network operating system, control and coordination of all the work the computer in the network, the maximum to meet the requirements of customers and make a response and handling, storage, and management in the network sharing resources, to monitor and control network activity, to actual network management, resource allocation system, to understand and adjust the system running state and shut down/start some resources, etc. The Web server structure is shown in Figure 3.

![Figure 3 Web server](image)

The server type can be divided according to the actual situation:
(1) classification according to application level or scale
Entry-level server: the lowest level server, mainly used for office file and print services.
Workgroup server: suitable for smaller network, suitable for providing Web, mail and
other services for smes.

Departmental server: mid-range server, suitable for data center, Web site and other applications of medium-sized enterprises.

Enterprise server: high-end server, with super data processing capacity, suitable as a large network database server.

(2) Partition by server structure

Desktop server: also known as tower server, this is the most traditional structure, with good scalability.

Rack-mounted servers: rack-mounted servers are installed in a standard 19-inch cabinet and are available in 1U (1U= 1.75in), 2U, 4U, and 6U sizes, depending on height.

Blade server: a high-availability, high-density, low-cost server platform designed for specific application industries and high-density computing environments, where each "blade" is essentially a system motherboard.

Machine cabinet server: the machine cabinet is machine cabinet, in the server need to install many module components.

(3) Divide by hardware type

Dedicated server: specially designed advanced server, using special operating system (such as UNIX, MVS, VMS, etc.), mainly used for database services and Internet business, generally by professional companies to provide a full set of software and hardware systems and full service.

PC server: a server with Intel or Motorola dedicated processors as the core, compatible with a variety of network operating systems and network applications, performance can reach the level of mid-range RISC servers.

3.3 Spatial decision information system

Spatial decision information system is a comprehensive use of a variety of data, information, knowledge, artificial intelligence and model technology to assist high-level decision to solve semi-structured or unstructured decision problems. It is a human-computer interactive information system based on computer processing. In this system, the latest achievements in management, mathematics, database and computer science are fully applied. The structure of spatial decision information system is shown in Figure 4.
Spatial decision support system is composed of spatial decision support, spatial database and other interdependent and interactive elements, and completes the processing, analysis and decision-making of spatial data as an organic whole. Its main behavior is a spatial decision support, and spatial decision support is the application of spatial analysis of all kinds of means to deal with spatial data transformation, in order to extract implicit in the certain facts and relationship of spatial data, and directly expressed in the form of graphics and text, and to provide a variety of applications in real world scientific and rational decision support. Because the method of spatial analysis directly integrates the spatial positioning ability of data and can make full use of the current characteristics of data. Therefore, the decision support provided by it will be more in line with the objective reality and therefore more reasonable.

3.4 Public participation in information processing systems

Public participation in information processing systems refers to computer-based processing systems. It is composed of input, output and processing, or hardware (including CPU, memory, input and output devices, etc.), system software (including operating system, utilities, database management system, etc.), applications and databases. An information processing system is an information conversion mechanism with a set of conversion rules. The structure of the public participation information processing system is shown in figure 5.
Figure 5 Public participation information processing system structure

Public participation information processing system is a very complex system, the design, construction, operation and maintenance of the system need a lot of costs, so it needs to be analyzed and studied from the point of view of systems engineering. System software is divided into two parts: program and database, which are equally important to information processing system. A good information processing system must have a good man-machine communication interface. The technology of developing information processing system is still developing, and the information processing system that has been applied also needs to be updated constantly. Public participation information processing system an information system that utilizes modern information processing technology to process business transactions and basic information of an enterprise in order to improve the efficiency and automation of business processing.

3.5 Distributed database system

There are two types of distributed database systems: one is physically distributed, but logically centralized. This kind of distributed database is only suitable for a single, small unit or department. Another kind of distributed database system is physically and logically distributed, which is called federated distributed database system. Because the seed database systems that make up the Federation are relatively "autonomous", such systems can accommodate various databases for different purposes, and are suitable for large-scale database integration. The structure of the distributed database system is shown in figure 6.
Distributed database system includes distributed database management system and distributed database. In the distributed database system, an application can be transparent to the database operation, data in the database in different local stored in the database, managed by different database management systems, running on different machines, supported by different operating systems, are different communication network connection together. Compared with the centralized database system, the distributed database system is scalable, and the reliability of the system can be improved by adding appropriate data redundancy. In a centralized database, minimizing redundancy is one of the system goals because redundant data wastes storage space and tends to cause inconsistencies between copies. In order to ensure the consistency of data, the system has to pay a certain maintenance cost. The goal of reducing redundancy is achieved through data sharing. However, in the distributed database, redundant data is expected to be added, and multiple copies of the same data are stored in different locations. The reasons are as follows: (1) improve the reliability and availability of the system when a site has a failure, the system can operate on the same copy of another site, and the whole system will not be paralyzed due to a single failure. (2) improve system performance the system can select the nearest data copy to the user for operation according to the distance, reduce the communication cost and improve the performance of the whole system.

4 Software part design

There are two factors influencing the site selection of traditional villages in southeastern chongqing, namely, natural environment factors and agricultural production and living needs.

(1) Natural environmental factors

The settlement's location is largely influenced by the local natural environment, it is because the self-sufficient small-scale peasant economy has the economic capacity and technical conditions is limited, can only make full use of natural conditions, through the
reasonable conditions of site selection for good farming homes built environment and adapt to the local natural environment, satisfy the building ventilation, lighting, heat, cold basic living needs.

(2) Agricultural production and living needs

Agricultural production is the most basic economic activities of the villagers, but also the daily source of livelihood. Since ancient times, villagers have relied on agricultural production to provide for their daily needs. With the development of the society, the villagers depend on the output of agricultural production to meet the needs of food and clothing, and trade the surplus agricultural products to obtain economic sources. It can be said that agricultural production is the most important issue related to the development of settlements and the livelihood of villagers. Therefore, landform, soil conditions, climate characteristics and irrigation water resources and other factors related to agricultural production become the necessary conditions for settlement site selection.

According to the spatial characteristics of traditional villages in the southeast of Chongqing, the design and planning decision plan is made. Using multi-agent decision technology, an abstract view of an agent is presented, as shown in Figure 7.

![Figure 7 Multi-agent decision abstract view](image)

In this block diagram, you can see that agents produce action output to affect their environment. In an environment of moderate complexity, an agent cannot fully control its own environment, but can only partially control it at best, that is, influence the environment. From the agent's point of view, this means that executing the same action twice in the same environment can have completely different effects.

System functions mainly include three parts: system maintenance management, spatial planning model/tools and decision support management. Among them, the system maintenance management module supports some maintenance management functions of system operation, including database maintenance, data source management, data source dictionary management, user rights management, system log management, etc. Model/tool integrated spatial scenario evolution analysis and dynamic simulation, spatial intelligent zoning, spatial planning implementation evaluation, spatial utilization constraint identification
and dynamic potential evaluation, and other model tools integrated management applications.
Decision support management includes decision support system index setting, comprehensive
decision making, decision support for participatory group planning, and outcome release.
Main research contents of multi-agent system:

1) Communication between multiple agents

The communication between agents is the basis of the interaction and cooperation
between agents. The communication between agents involves the understanding and
generation of physical mode and communication language. If agents are heterogeneous, how
to translate different knowledge into a unified communication language of mutual
understanding is also an important problem. At present, there are two common communication
language design methods: process method and declaration method. The idea of process
method is that communication can be simulated by the exchange of process instructions, the
design process requires the information of the receiver, and the communication process is
one-way, while many information exchanges of agents should be two-way, so the process
method is not applicable to the communication between agents.

2) Coordination and collaboration of multiple agents

Multi-agent coordination means that multiple agents with different goals make reasonable
arrangements for their goals and resources to coordinate their behaviors and achieve their
goals to the maximum extent. Multi-agent collaboration means that multiple agents work
together to achieve a common goal by coordinating their behaviors. Multi-agent systems can
be thought of as open, distributed environments in which one agent sometimes needs to work
with other agents to construct complex programs, or to accomplish tasks that it cannot
accomplish alone.

For multi-agent systems with common goals, the existing negotiation methods mainly
include contract network protocol. Agents are dynamically assigned to the roles of manager
and collaborator. An agent receives a new task and becomes a manager, responsible for the
assignment of tasks. The other agents are collaborators, bidding on the current task, expressing
capabilities and intentions for the task. The manager assigns tasks to the most suitable bidder
based on the commitment of all bidders. Based on the work model, the formal definition of
joint intention, social commitment and rational behavior is proposed to describe or constrain
the cooperative behavior of intelligent bodies. In the collaborative process of multi-agent
system, the thought of decision making and learning always permeates. Multi-agent
interaction and cooperation based on game theory has a complete theoretical system and
axioms of derivation. The countermeasures take the equilibrium point as the goal of the
cooperation, so the convergence and stability of the cooperative process of intelligent bodies
are introduced into the cooperative research of intelligent bodies. Many of the theories in
game theory can be used in the framework of multi-agent cooperation.

(3) Multi-agent conflict resolution

In the multi-agent system, each agent has autonomy and will act according to its own
knowledge, ability and goal during the problem solving process. For some shared resources,
shared conflicts often occur, and agents sometimes have different goals. Especially for
multi-agent systems, it is not possible to design or implement an agent and then build it to
match the goals of other potential agents. Due to the high degree of autonomy and flexibility
of agents, their understanding of the environment is different, and their acquisition of global
knowledge is often not comprehensive. Therefore, for multi-agent systems, dynamic conflict
management is an inevitable requirement.

At present, the main method of conflict resolution in multi-agent systems is negotiation.
Negotiation techniques include refactoring, limiting mediation, and arbitration. Negotiation
technology is usually based on game theory, which assumes that agents have complete global
knowledge and choose their own behaviors according to the principle of maximizing utility,
and the utility matrix of agents is shared knowledge. However, the knowledge of agents is
often not complete, and their utility is not shared but private. In order to simulate problems in
the real world, the assumption of conflicts is usually avoided by establishing social rules. But
social rules and standards impede the flexibility and adaptability of multi-agent systems. If
individual agents simulate human intelligence, then many agent systems simulate human
society. The conflicts in human society are solved by the social rules observed by the group,
and the design of corresponding multi-agent system should also include the content of social
rules.

The permission to use the system is controlled by the IP address through the hardware,
and the specific process is shown in figure 8.
5 The experiment

In order to verify the rationality of the design of building space planning decision system based on multi-agent, the experimental verification analysis is carried out under the environment of MATLAB simulation tool, Microsoft Windows XP operating system, Intel (R) Celeron (R) 2.6GHz processor and 24 GB memory. A 10-point scale was used for evaluation, so as to determine the planning rationality of the experimental target in the decision-making system of building space planning, as shown in table 1.

<table>
<thead>
<tr>
<th>Serial number</th>
<th>The degree of planning</th>
<th>Weight value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unreasonable</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>Between very reasonable and moderately</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>Moderate and reasonable</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>Between average and moderate reasonable</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>general</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Very reasonable</td>
<td>0</td>
</tr>
</tbody>
</table>

According to the weight values shown in table 1, the optimal planning effect of the traditional system and the multi-agent system on the building space was compared and analyzed respectively. In this process, two experimental conditions were set, namely natural
environmental factors and agricultural production and living needs. The comparison results are shown in Figure 9.

![Figure 9](image)

**Figure 9** Comparative analysis on the efficiency of trajectory planning of the two systems

It can be seen from figure 9 that under the natural environment, the efficiency of the two systems is maintained at above 75%. Under the conditions of agricultural production and living needs, the efficiency of traditional system planning is lower than 60%. However, the planning efficiency of multi-agent based system is still high, which can be maintained at over 75%. It can be seen from this that the design rationality of the decision system of building space planning based on multiple agents is reasonable.

### 6 Conclusion

With the development of information technology, the planning assistance decision-making service will also change. In order to further meet the service needs of the public, the required elements will also increase, the data sources used will also gradually become multi-source, and the system function modules will also increase and improve. In terms of comprehensive research, the following aspects need to be further studied:

1. The analysis quality of auxiliary data service products has been improved. With the existing data conditions and technical routes, a large part of the city details cannot be identified, and for the areas with close building spacing or lush trees, the information at the bottom of buildings cannot be accurately expressed. Therefore, the future development direction of the system will be to add the road street view collection information.

2. The efficiency of relevant algorithms is improved. In some analyses involving a large number of models and terrain data, such as sunshine analysis, the execution efficiency is low and users need to wait for a long time. Therefore, relevant algorithms need to be optimized to
improve data processing capacity and provide a good user experience.

7 Fund projects
Chongqing Municipal Education Commission Science and Technology Research Project: Spatial Features and Planning of Traditional Villages in Southeast Chongqing Research on design strategy (KJQN201904301).

Reference


Research on power control system of inspection robot based on wavelet transform

GUAN Hong-mei1, LIU Yue2
{dawda985@163.com1, tosn614@163.com2}

(1. Department of Automation, College of Humanities Information Changchun University of Technology, Changchun 130122, China; 2. Department of Automation, College of Humanities Information Changchun University of Technology, Changchun 130122, China)

Abstract: Due to the poor stability of the traditional power control system, inspection control errors will occur during the inspection process. Therefore, a new type of power supply control system for inspection robots was designed based on wavelet transform, and the drive hardware circuit and control hardware circuit of inspection robots were redesigned in hardware design to ensure the stability of hardware operation. In the software part, by extracting and checking the structure and functional characteristics of the robot, setting the system control parameters based on the wavelet transform, and configuring the system communication parameters, in order to enhance the stability of the system operation and realize the complete design of power control. The test results show that, compared with the traditional power control system, the designed system has better stability performance and stronger patrol robot control ability. The designed power control system can be applied to the patrol robot control.

Keywords: Wavelet transform; Patrol robot; Power control system

1 Introduction

There are many shortcomings in the traditional manual inspection method. High labor intensity, low work efficiency, scattered detection quality, and single means; under severe weather conditions such as thunderstorms, manual inspections have greater safety risks and cannot be inspected in time; traditional video surveillance systems are restricted by various conditions. There is a large blind spot in surveillance, and it is difficult to truly meet the requirements of all-round coverage of video surveillance. At the same time, due to the complexity of the system, the large number of cameras, and the large workload of installation and wiring, the failure rate is high and maintenance is difficult. With the promotion of smart substation and unattended substation modes, higher requirements are placed on the inspection of substations, and a more advanced and intelligent inspection mode is urgently needed.

With the rapid development of special robot technology in recent years, inspection robot technology that replaces manual inspection has become the focus of attention at home and abroad. Checking robots involves various disciplines and needs to be studied in combination with actual conditions. Only by combining the previous experience and actual investigation can the design of the main body of the transmission line robot and the obstacle avoidance strategy be completed, and the corresponding control system be designed to achieve. The line...
is easy to meet the emergency, so it is necessary to monitor the data and status of the robot in real time during the inspection of the robot, and transmit the image data to the ground, which can be used to remotely control the robot through the ground, and the remote operating system evaluates the patrol of the transmission line Environment, find the inspection status of the transmission line, perform simple processing, etc. [1]. However, the traditional design method has insufficient control stability, so it is necessary to propose a new design scheme to correct the traditional design method. In the redesign, the design of the inspection robot power control system is based on the wavelet transform theory as the basic design idea. The hardware circuit of the power supply control system improves the stability of the control system through continuous wavelet transform, thereby achieving the reliability of the inspection robot. This article involves the system to provide a powerful guarantee for improving the inspection work of the inspection robot.

2 Design of hardware circuit of power control system of patrol robot

2.1 Design of driving circuit for inspection robot

The driving circuit of the circuit patrol motion control system mainly consists of three parts:(2) the obstacle removal system that controls the robot to remove high-voltage line obstacles;(3) robotic arm system to control the movement of robot manipulator. The routing system controlling the robot walking is mainly used to control the forward and backward functions of the robot on the line. The driving control circuit is shown in figure 1.

![Fig. 1. Inspection robot drive circuit.](image)

The main control chip TMS320F28335 sends the control signal of high and low level to the circuit. When the GPI081 pin sends a low voltage to the circuit, $R_1$ and $R_2$ divide the voltage, $R_3$ divide the voltage of 1V, the transistor $L_1$ realizes conduction, the relay circuit conducts conduction, and the 24V voltage is connected to the stepper motor driver. Then the main control chip is used to input the PWM signal and the high and low level signal to drive
the stepper motor forward and backward, so as to realize the forward and backward control of the robot. When GPIO81 pin sends a high level (3.3v) to the circuit, $R_1$ and $R_2$ divide the voltage, $R_3$ divide the 0.34v voltage, the transistor $L_1$ turns off, and the stepper motor stops [2].

In order to reduce the control error of the line inspection robot, the single power supply mode is adopted, and the 24V battery is used as the power control system hardware of the whole robot. However, since multiple chips need to be powered, 24V to 5V power supply circuit composed of LM2576S-5.0 voltage regulator chip is adopted here to solve this problem. The circuit shown in figure 2 is a schematic diagram of 5V voltage stabilizing circuit.

![Fig. 2. Regulating circuit.](image)

The voltage regulator chip used in the voltage regulator circuit is LM2576S-5.0, which belongs to the step-down switching voltage regulator, has very small voltage adjustment rate and current adjustment rate, and has the load driving capacity of 3A, which can replace the general three-terminal linear voltage regulator, fully reduce the area of the heat sink, and even do not use the heat sink under some application conditions. At the same time, the chip is relatively simple and has fewer peripheral components, built-in frequency compensation circuit and fixed frequency oscillator, as well as built-in overcurrent protection circuit and overheat protection circuit. Under the conditions of specified input voltage and output load, the input voltage of LM2576S-5.0 reaches the fixed voltage output of 36V and 5V, and the error range of the output voltage is ±14%. The oscillation frequency error range of the oscillator is 10%. The typical standby current is 50 mA, thus realizing the hardware design of the driving circuit of the patrol robot [3].

### 2.2 Electric control circuit design of inspection robot

Based on the hardware design of the driving circuit of the inspection robot, the control circuit hardware of the power control system of the inspection robot is designed. Known inspection robot movement, mainly rely on the power control system board send action instruction, so as to complete the corresponding action, at the same time control system board needs to be real-time display robot servo system signal acquisition system, the working state of the robot were collected including robots working voltage, working current, robot power. The block diagram of the control circuit hardware design of the power control system is
shown in figure 3 below.
Since the power control system of the inspection robot is the key control center of the control drive system, the main control chip of the power control system of the inspection robot based on wavelet transform is MSP430F437.

According to the process described in the figure 3, the control port circuit of the robot control board is designed. In the control panel of the known robot, the button used for the signal sending button is a waterproof self-locking button with LED lamp, and the button port circuit is shown in figure 4 [4].

When the self-locking button is pressed, the first pin and the third pin in the figure are connected to the first pin through a diode. When the self-locking button is released, the first
pin and the third pin are disconnected. Two control chip pins with 1K electricity connected to the first pin through a diode are set as high level. Then the hardware of the numerical display circuit of the control board of the inspection robot is designed. The design of the circuit is for real-time robot working status in the process of movement, through the timely display of robot inspection work, to achieve timely adjustment of the control system. The digital tube display circuit is shown in figure 5.

Fig. 5. Digital tube display circuit.

According to the design of drive circuit hardware and control circuit hardware, it is convenient for the application of wavelet transform technology in power control system.

3 Design control system software based on wavelet transform

On the basis of the circuit design, the control system software is redesigned according to the wavelet transform theory and application technology.

3.1 Extract the dynamic characteristics of the inspection robot

The inspection robot consists of many parts, including support columns, inspection platform, lift structure, beam, etc. Because the structure parameters of the robot during the inspection are not fixed, it is necessary to analyze the dynamic characteristics of the inspection robot in order to optimize the design of the power control system software. Figure 6 is the schematic diagram of the inspection robot in the visual direction.
According to the parameter information in the figure 6, the inertial dynamic value of the robot is calculated as follows:

\[ F_z = \frac{C_z^2 H_z}{12} - \frac{c_z^2 - h_z}{12} \]  

(1)

Where: \( F_z \) represents the inertial dynamic value of size \( \xi \); \( C_z \) represents the size of the robot base; \( c_z \) represents the top size of the robot inspection device; \( H_z \) represents the horizontal height of the robot; \( h_z \) represents the distance from the base of the robot to the ground.

It is known that under different working intensities, robots will have different motion modes due to different working properties and working environments, and the deformation occurring in the plane XOY [5] has the greatest impact on their working characteristics. When the robot is stationary, its static deflection is basically unchanged, but when starting and braking, the deflection of the robot column increases with the increase of acceleration, and the deflection of the column has a great relationship with the horizontal acceleration. Therefore, the deflection of the robot column is calculated as:

\[ \sigma = \sigma_x + \sigma_y + \sigma_q \]  

(2)

Where: \( \sigma_x \) represents the deflection of the column tip caused by gravity of each mass element; \( \sigma_y \) represents the deflection of the column tip caused by the inertial force of each mass element; \( \sigma_q \) represents the deflection at the top of the column caused by the inertial force evenly distributed by the column itself. According to the results obtained in formula (1) and formula (2), the dynamic characteristics of the inspection robot are obtained as follows:

\[ \dot{\kappa} = \sum \frac{\phi_i \gamma_i \phi_{ij} \sigma}{F_i} \left( \ddot{\chi}_i - \dot{\chi}_i \right) - \frac{\sigma}{6EI} \]  

(3)

Where: \( \dot{\kappa} \) represents the dynamic characteristic value of the robot; \( \chi_i, \dot{\chi}_i \) and \( \ddot{\chi}_i \) represent the position, velocity, and acceleration of the \( i \)-th point of the robot, respectively.
represent the node coordinates of the \( i \) mobile joints of the inspection robot; \( Q_i \) represents the load capacity value of the robot; \( \varepsilon \) represents the bending moment of each mass element to the axis of the column; \( V \) is the acceleration of gravity; \( \dot{h} \) represents the height of the robot; \( c \) represents the acceleration in the walking direction; \( EI \) represents the force modulus [6]. According to the above formula, the dynamic characteristics of the inspection robot are extracted.

3.2 The system control parameters are set based on wavelet

According to the obtained dynamic characteristics of the robot, the system control parameters based on wavelet transform are set. The setting process of such parameters is shown in figure 7 below.

![System control parameter setting process](image)

According to the figure 7, the calculation condition of the wavelet function is set as follows: \( \tau(\nu) \in \mathcal{L}_2(\mathbb{R}) \), its Fourier transform is \( \tau(\nu) \), and it satisfies \( \tau(0) = 0 \), then \( \tau(\nu) \) can generate the family of functions through expansion and translation \( \{\tau_{\alpha,\beta}\} \), and the relation is:

\[
\tau_{\alpha,\beta}(\nu) = |\beta|^{-\frac{1}{2}} \tau \left( \frac{\nu - \beta}{\alpha} \right)
\]

Where \( \alpha \) represents the scale factor, \( \beta \) represents the translation factor, and \( \alpha, \beta \in \mathbb{R} \), \( \alpha \neq 0 \). \( \tau \) represents the base wavelet or the mother wavelet [7]. According to the continuous wavelet transform results of the square integrable function obtained from the above equation, the transformation equation of the integrable function \( h(\nu) \in \mathcal{L}_2(\mathbb{R}) \) is:

\[
\int_{-\infty}^{\infty} h(\nu) \psi' \left( \frac{\nu - \beta}{\alpha} \right) d\nu = [\hat{h}, \tau_{\alpha,\beta}]
\]
Where: \( \psi(a,b) \) represents the continuous wavelet transform coefficient; \( \psi^* \) is the conjugate. \( a > 0, b \) can be positive or negative in practice. Since each parameter is a continuous variable, the control structure of the system by parameters is set according to the linear, time-shift and scale characteristics of the continuous wavelet, as shown in figure 8 below.

![Fig. 8. Parameter control structure diagram.](image)

In the figure 8, D represents the original signal, and X1, X2 and X3 represent the parameters of surface data control structure in the three stages. Y1, Y2 and Y3 represent the hidden control structure parameters of the three stages, and set the control parameters [8].

### 3.3 Configuration of communication parameters to achieve the power control system design

To complete the communication between control system software, but also to configure the communication parameters. According to the communication control mode of the power control system of the inspection robot, the communication protocol code of the control algorithm is adjusted, and the adjusted parameter values are shown in table 1 below [9].

<table>
<thead>
<tr>
<th>Parameter name</th>
<th>Dynamic value</th>
<th>Static value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Communication control</td>
<td>2074</td>
<td>2000</td>
</tr>
<tr>
<td>Frequency setting</td>
<td>2085</td>
<td>2001</td>
</tr>
<tr>
<td>Inverter status</td>
<td>3863</td>
<td>30001</td>
</tr>
<tr>
<td>Operating frequency</td>
<td>3485</td>
<td>3001</td>
</tr>
<tr>
<td>Set frequency</td>
<td>3095</td>
<td>3003</td>
</tr>
<tr>
<td>The output voltage</td>
<td>2374</td>
<td>2001</td>
</tr>
<tr>
<td>Output current</td>
<td>2099</td>
<td>2000</td>
</tr>
</tbody>
</table>

According to the communication change parameters in table 1 above, the demo program of the power control system is set, and the schematic diagram of the delay program is shown in figure 9 below.
According to the delay program set in the figure 9, the response sensitivity of the inspection robot is controlled to ensure that all movement of the robot can meet the requirements of inspection. Thus, the electric control system of the inspection robot based on wavelet transform is designed [10].

4 Experimental analysis

In order to verify the effectiveness of the research in this paper, a comparative test experiment was put forward. The designed power control system and the traditional designed power control system were put into the use of patrol robots. According to the use effect of patrol robots, specific experimental test conclusions were drawn.

Two inspection robots with the same model and configuration were selected as experimental test objects, and two power control systems were loaded into the robot. The physical picture of the inspection robot is selected, as shown in Figure 10.
The experimental test platform is mainly composed of the following parts: the experimental support is used to simulate the power line, and the robot can detect and cross over the simple obstacles on the top; the pressure measuring unit is set up to measure the pressure output of the clamping and pressing mechanism when the robot goes over obstacles. At the same time, remote sensing connection will be established between the inspection robot, key control box and computer to ensure the continuity of the test system. After completing the above operations, the test capability of the experimental test system is tested. The following table 2 is the test system operation parameter values obtained by the test.

**Table 2.** Test the parameters of the system running smoothly.

<table>
<thead>
<tr>
<th>Testing time</th>
<th>Stability parameter</th>
<th>deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1h</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>2h</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>3h</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>4h</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>5h</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>6h</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>7h</td>
<td>0.98</td>
<td>0.01</td>
</tr>
<tr>
<td>8h</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>9h</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>10h</td>
<td>0.98</td>
<td>0.01</td>
</tr>
<tr>
<td>11h</td>
<td>0.98</td>
<td>0.01</td>
</tr>
<tr>
<td>12h</td>
<td>0.98</td>
<td>0.01</td>
</tr>
<tr>
<td>13h</td>
<td>0.98</td>
<td>0.01</td>
</tr>
<tr>
<td>14h</td>
<td>0.97</td>
<td>0.02</td>
</tr>
<tr>
<td>15h</td>
<td>0.97</td>
<td>0.02</td>
</tr>
<tr>
<td>16h</td>
<td>0.97</td>
<td>0.02</td>
</tr>
<tr>
<td>17h</td>
<td>0.97</td>
<td>0.02</td>
</tr>
<tr>
<td>18h</td>
<td>0.96</td>
<td>0.03</td>
</tr>
</tbody>
</table>
According to the above table, within 24 hours, the deviation between the stability index of the test system and the required value of standard stability is within 0.05, so it can be known that the test result of the system is reliable and the experiment can be started.

In this experiment, the test results of the inspection robot control system based on wavelet transform were taken as the experimental group. The test results of the electric power control system of the patrol robot designed under the traditional design method were taken as the control group, as shown in figure 11 below, that is, the experimental test comparison results.

![Experimental test comparison results.](image)

According to the test results in the figure 11, the stability of the power control system designed based on wavelet transform decreased by 0.5% in the third stage, and the calculated average stability was about 95.13%. However, the stability of the power control system designed by the traditional method began to decline in the first stage, and by the third stage, the stability degree was already around 80%, and the decline degree was about 15.86%. It can be seen that the designed power control system is more suitable for the daily control requirements of patrol robot.

**5 Conclusion**

The design of electric control system based on wavelet transform, the drawbacks of the existing in the traditional control system as the breakthrough point, take full advantage of wavelet transform concept, to better drive hardware circuit and communication circuit of hardware support, has realized the speed inspection robot more moderate, stability, strong electrical control system design, for use in the inspection robot in the future development, to
provide reliable technical support. But the design of the control system, in the hardware circuit part did not explain in detail, driving circuit hardware, as well as all the hardware circuits in the communication circuit, the future research can be aimed at this problem, make detailed data description.

References

Research on Fault Location Method for Wind Turbine Line in Power System

LIU Yue¹, GUAN Hong-mei²
{tosn614@163.com¹, dawda985@163.com²}

(1. Department of Automation, College of Humanities Information Changchun University of Technology, Changchun 130122, China; 2. Department of Automation, College of Humanities Information Changchun University of Technology, Changchun 130122, China)

Abstract: In the conventional fault location method of the wind turbine line, when multiple points are positioned in parallel, a location fault or a location error often occurs. Therefore, a fault location method for power system wind turbine is proposed. Identify the fault range of the power system wind turbine line, match the fault data within the controllable range, and correct the matching data to deviate from the execution range. The fault location of the wind turbine line of the power system is realized through the transformation. The simulation experiment is designed. The results show that the design method can achieve accurate positioning under multi-point fault location, and solve the problems in traditional methods.

Keywords: Power system; Wind turbine; Line fault; Fault location.

1 Introduction

Wind turbines are important components in the power system. The composition of the wind turbines is complex, and often in complex and changeable operating conditions. As a result, the wind turbine's line is one of the components with the highest failure rate, and its failure causes the longest downtime. Therefore, it is of great research interest to realize accurate and timely fault location of this component.[1-2]

When the extreme points of the signal from the wind turbine are not evenly distributed, the line resolver will cause "overshoot" and "undershoot" phenomena, resulting in aliasing in the power system. Starting from solving the problem of distribution of line fault points, by adding Gaussian white noise to the wind turbine to make the distribution of line fault points uniform, in order to reduce the impact of modal aliasing[3]. Human experience intervention, greatly reducing the adaptability of line fault location methods. Based on this, many researchers have proposed a variety of different methods for parameter tuning. To a certain extent, the blindness of artificial selection of parameters is avoided, but the selection of noise intervals in the fault location method still needs to be set manually, and there is a long time for pre-processing to select parameters.

The magnitude of line fault location deviation has a significant impact on the effect of
wind turbines. When the noise is too small, the line fault location cannot solve the problem of power system aliasing. For this reason, a fault location method for power system wind turbine lines is proposed.

2 Fault location method for power system wind turbine

2.1 Identify the fault range of the power system wind turbine line

According to the characteristics of the power system, the search procedure of the wind turbine is set using CS technology to determine the effective range of the line fault location. Specify the line fault judgment step factor. This factor is the parameter for judging the search area. When the judging step factor is too small, the search program is controlled to start a regional search. When the step factor is too large, the line parameter can control the search program. Expand local search in local area [4]. In order to ensure the searching balance of the search program in the whole area and the local area, a balance adaptive adjustment parameter is added to make the search program perform distributed synchronous search.

The calculation formula for this step size factor is:

$$A = l_{\text{max}} - \frac{Q(l_{\text{max}} - l_{\text{min}})}{\varepsilon Z} \frac{H_1 - H_2}{H_1}$$

(1)

In equation (1): $A$ represents the $Q$th search step size, where $Q$ represents the current number of times; $l_{\text{max}}$ represents the maximum upper limit of the step size factor; $l_{\text{min}}$ represents the minimum lower limit of the step size factor [5]; $\varepsilon$ represents the introduced adjustment parameter; $Z$ represents the maximum total number of iterations; $H_1$ represents the optimal function at the initial fitness; $H_2$ represents the optimal function suitable for this iteration. At the same time, under the search control of this parameter, the probability of finding the approximate location of the line fault of the wind turbine is calculated. In CS technology, the discovery probability value is generally between 0.2-0.3, and the probability under this parameter is set to $P$. When the probability is always at a large value, the convergence speed of defect search can be accelerated. When the value of the probability $P$ is found to be small, the convergence speed at this time is reduced, so the optimal fitness is set by the ratio of the optimal fitness and the previous solution:

$$F = \begin{cases} 1 & f_i > f_{i-1} \\ 0 & f_i \leq f_{i-1} \end{cases}$$

(2)

In equation (2): $f_i$ represents the optimal fitness of the previous generation; $f_{i-1}$ represents the fitness of the $i$th solution under the action of $l_z$ [6]. According to the above calculation formula, freely adjusting the fitness of the discovery probability, the calculation result of the obtained discovery probability is as follows:
In equation (3): $P'_{i}$ represents the high-quality probability of the solution after iteration; $H_{i}$ represents the optimal fitness for the $i$th solution; $e$ represents the total number of solutions; $P_{\text{min}}$ and $P_{\text{max}}$ represent the upper and lower limits of the discovery probability respectively. Combine the above two sets of formulas to get a search program, implement a defect search instruction for the power system, and obtain the approximate area range of the wind turbine line fault, as shown in Figure 1 below.

![Fig. 1. Defining scope of wind turbine line faults.](image)

The recognition result in the figure above is the target range of wind turbine line faults. However, this target range not only contains defective nodes, but also contains a small number of normal nodes. Therefore, the data of the defective nodes need to be matched and processed.

### 2.2 Fault data matching processing

After obtaining the defined fault range of the wind turbine line, the main characteristics of the original wind turbine line are detected and matched with the characteristic points of the fault of the wind turbine line to be matched. Set the characteristic point of the line of the original wind turbine as $a_{i}$, and the characteristic point of the line fault of the wind turbine to be matched as $b_{j}$, then:

\[
\begin{align*}
\mathbf{a}_{i} &= \begin{pmatrix} x_{i}^{y} & y_{i}^{y} & z_{i}^{y} \end{pmatrix} \\
\mathbf{b}_{j} &= \begin{pmatrix} x_{j}^{y} & y_{j}^{y} & z_{j}^{y} \end{pmatrix}
\end{align*}
\]

(4)
In equation (4), \( i = 0, 1, 2, \ldots, n \), there are \( n \) pairs of feature points in total; \( w \) represents the point cloud set corresponding to \( a_i \), and \( v \) represents the point cloud set corresponding to \( b_i \). Taking point cloud \( W \) as a reference, rigid transformation is performed on point cloud \( V \), and \( V \) is transformed into the coordinate system of \( W \), and the matrix is as follows:

\[
W_i = \begin{pmatrix}
\alpha_{00} & \alpha_{01} & \alpha_{02} \\
\alpha_{10} & \alpha_{11} & \alpha_{12} \\
\alpha_{20} & \alpha_{21} & \alpha_{22}
\end{pmatrix} \begin{pmatrix}
\beta_0 \\
\beta_1 \\
\beta_2
\end{pmatrix}
\]  

(5)

The definition of each variable is the same as the above formula. Due to the certain mapping relationship between the fault points in the wind turbine line and the matched three-dimensional point cloud, the matching points in the three-dimensional space of the wind turbine line fault are obtained, as shown in Figure 2.

As shown in Figure 2, the characteristic points in the wind turbine line are matched with the three-dimensional spatial fault points of the wind turbine line, but because of the matching process, it may be disturbed by random factors [8]. To this end, in the neighborhood of feature points, reconstruction of inter-point interpolation is required to replace defective points with missing phases. Extend the X direction or Y direction of the position to be matched after matching. When two points are found to be defective, solve the three-dimensional space where the two points are located. Use the points obtained by interpolation to replace the original defect points and calculate. The process is as follows:

![Fig. 2. Schematic diagram of line fault points matching for wind turbines](image-url)
\[
\begin{align*}
\nu'_j &= \frac{(\nu^{\text{proj}}_j + \nu^{\text{proj}}_k)}{2} \\
\nu'_l &= \frac{(\nu^{\text{inter}}_j + \nu^{\text{inter}}_k)}{2}
\end{align*}
\]

Where \((\nu^{\text{proj}}_j, \nu^{\text{proj}}_k)\) represents the point where interpolation replaces \(\nu_j\), and \((\nu^{\text{inter}}_j, \nu^{\text{inter}}_k)\)
represents the point where interpolation replaces \(\nu_l\).

The specific replacement diagram is as follows:

(a) Defective point replacement process
As shown in Figure 3, the nodes in the Figure 3(a) are used to represent nodes. The filled squares indicate defects or faults at the location, the upper part indicates that the original wind turbine line contains characteristic points, and the lower part indicates the wind turbine to be matched corresponding defect fault points in the line. Figure 3(b) shows the cross-section lines obtained after matching in the wind turbine line. So far, the fault data matching process is completed, and the fault data deviation correction is implemented on this basis.

2.3 Error data deviation correction

In the process of matching fault data of wind turbine units, due to the influence of random factors on the fault location of each line, a certain deviation occurs. In order to accurately locate the fault point of the line, the fault data deviation needs to be corrected. The fault data deviation is detected by using the wind turbine circuit diagram. On the edge of the wind turbine circuit fault, the mark is left blank, as shown in Figure 4:
According to the actual situation, if the distance between the two color scales remains unchanged, there is no deviation between the two marks, otherwise the fault data deviation correction is implemented [9]. To this end, a fault data deviation correction model is established and the fault data deviation is revised. The specific structure of the revision process is shown in Figure 5:

Fig. 5 Schematic diagram of error data deviation correction

As shown in Figure 5, the master-slave structure of the fault data deviation correction model is used to control the deviation generated in the wind turbine line. The control effect is divided into two parts: the mechanical deviation of the wind turbine and the deviation of the line fault. Among them, the correction of the mechanical deviation of the wind turbine must be completed by setting the relevant parameters and the correction of the state of the wind turbine equipment during operation. The line fault deviation revision uses the unit controller, photoelectric encoder, motor drive and other related equipment to modify the line fault deviation in real time, and control and display it. Both parts can complete the fault data deviation correction through the CAN bus, so that the wind turbine line fault is guaranteed. The photoelectric encoder installed on the line main shaft in the wind turbine equipment runs simultaneously with the operation of the wind turbine control unit. The specific physical diagram is shown in Figure 6.
2.4 Power system wind turbine line fault location

After the fault data deviation is corrected, the fault current travelling wave head arrival time is calibrated. In the wind turbine line, the traveling wave signal is decomposed into components with different characteristic scales, and the instantaneous maximum value is obtained for the decomposed component to calibrate the wave head arrival time. When the wind turbine line fault is in the fault transient state, the current signal will generate high-frequency components due to the fault. The fault signal is decomposed by EMD, and the highest frequency component in the decomposed signal is Hilbert transformed to obtain the fault signal. Instantaneous frequency graph. Solve the maximum value of the instantaneous frequency to remove the interference wave head [10]. In the time-frequency diagram, when the fault traveling wave head appears as the first maximum value, the time when the fault traveling wave head reaches the detection point can be determined. In addition to locating the fault area of the wind turbine line, in addition to knowing when the fault traveling wave head reaches the detection point, it is also necessary to calculate the traveling wave velocity. When the frequency-dependent interference of the relevant parameters of the wind turbine line is not considered, the formula for calculating the traveling wave velocity v of a fault is as follows:

$$ v = \frac{1}{\sqrt{NV}} \quad (7) $$

Among them, $N$ is the inductance value of the unit length wind turbine line, and $V$ is the capacitance value of the unit length wind turbine line. Theoretically, it is feasible to calculate the traveling wave velocity of a fault using a formula, but in practical applications, because the wind turbine unit line is usually constructed across regions, the inductance, capacitance, and resistance of the line will change as the line grows. The wave velocity formula is as follows.

$$ J = \frac{\nu}{\sqrt{\frac{1}{2} \left[ \omega^2 N - J' G + \sqrt{G^2 + \omega^2 C^2} \right]}} \quad (8) $$

In equation (8), $G$ is the conductance value of the unit wind turbine unit line, and $\omega$ is the angular frequency. The figure is a schematic diagram of the wind turbine line, and the specific
principle of measuring the traveling wave velocity of a fault is explained in detail.

![Diagram of wind turbine](attachment:image.png)

**Fig. 7** Schematic diagram of the wind turbine

As shown in Figure 7, the length of the line is 1, detection points A and B are placed at both ends of the line, and detection points C are set at 1/2. In practical applications, the detection points are not placed at equal intervals, and the clocks of the detection points at different intervals are implemented by tracking. After a fault occurs, the fault occurrence section is judged according to the phase separation phase difference of the fault current.

Assume that the fault occurs in the BC section, and the time $t_B$ and $t_C$ at which the fault traveling wave reaches the detection points B and C respectively. Calculate the transmission time $t_{BC}$ on the line in the BC section. Simply calculate the fault traveling wave wave speed. Simply calculate the fault traveling wave velocity, and similarly calculate the AC traveling fault wave velocity. After the line fails, the information collected by the fault information collector at each monitoring point is sent back to the monitoring center. After receiving the data, the monitoring host obtains the arrival time of the wave head from the wavelet scale transformation according to the fault identification result, analyzes the section where the fault point is based on the principle of phase-separated current and phase difference, and calculates the traveling wave velocity of the fault to locate the fault area of the motor unit circuit. At this point, the fault location of the wind turbine in the power system is completed.

### 3 Experiment

In order to verify the fault location method of the power system wind turbine line designed in this paper, experiments are designed to verify whether the location information and obstacle source are accurate in the event of multiple faults in the wind turbine line.

#### 3.1 Experiment preparation

The wind power unit of the power system is used as the experimental research object, including 1 breakpoint failure point, 10 operating failure points and 5 positioning points. And experimental simulation diagram are shown in **Figure 8**.
The coordinates of the fault point and each set anchor point are shown in the following Table 1:

<table>
<thead>
<tr>
<th>Experiment library number</th>
<th>X-axis</th>
<th>Y-axis</th>
<th>Occurrences / times</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>34</td>
<td>114</td>
<td>2</td>
</tr>
<tr>
<td>P2</td>
<td>25</td>
<td>53</td>
<td>2</td>
</tr>
<tr>
<td>P3</td>
<td>64</td>
<td>158</td>
<td>2</td>
</tr>
<tr>
<td>P4</td>
<td>57</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>P5</td>
<td>142</td>
<td>113</td>
<td>1</td>
</tr>
<tr>
<td>P6</td>
<td>153</td>
<td>58</td>
<td>1</td>
</tr>
<tr>
<td>P7</td>
<td>176</td>
<td>114</td>
<td>1</td>
</tr>
<tr>
<td>P8</td>
<td>134</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>P9</td>
<td>176</td>
<td>174</td>
<td>1</td>
</tr>
<tr>
<td>P10</td>
<td>134</td>
<td>25</td>
<td>2</td>
</tr>
</tbody>
</table>

Under the above experimental conditions, the traditional method and the method in this paper were used to locate and analyze the experimental results.

### 3.2 Experimental results and analysis

Under the above experimental conditions, two methods were used to locate the data, and the test point results were obtained. The experimental results are shown in Table 2:

<table>
<thead>
<tr>
<th>Experiment number</th>
<th>Experiment set point</th>
<th>Experimental results of the method of this paper</th>
<th>Experimental results of traditional method</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1, 2, 9, 10</td>
<td>1, 2, 9, 10</td>
<td>1, 9, 10</td>
</tr>
<tr>
<td>P2</td>
<td>4, 5, 6</td>
<td>4, 5, 6</td>
<td>4, 5, 6</td>
</tr>
<tr>
<td>P3</td>
<td>3, 7, 8</td>
<td>3, 7, 8</td>
<td>3, 4, 8</td>
</tr>
<tr>
<td>P4</td>
<td>11, 15, 16</td>
<td>11, 15, 16</td>
<td>11, 13, 9</td>
</tr>
<tr>
<td>P5</td>
<td>8, 5, 10</td>
<td>8, 5, 10</td>
<td>8, 5, 10</td>
</tr>
</tbody>
</table>
As can be seen from Table 2, compared with the traditional method, the positioning results of this article are based on the experimentally set point coordinates. The basic positioning points of this method are correct, and the experimental results appear marked. The smaller the distance between the wind turbines, the more accurate the positioning results. Omissions and errors in the number of anchor points make the method in this paper more applicable.

4 Conclusion

This paper designs a fault location method for power system line faults, and solves the problem of inaccurate or missing multi-points in conventional fault location methods for power system line faults. It is hoped that the research in this paper can provide help for fault line maintenance of power units.

In future research, we will focus on improving the efficiency of power system line fault location and detection, and further improve the proposed algorithm.

References

Design of medical database information query system based on Android

ZHU Yan-Chen¹, Du Jian-Qiang ²*, Li Zhi-Biao³, Zhong Guo-Yue⁴, Deng Chun-Sheng⁵
{ddd24616563@163.com¹, zhuyc126@163.com²*, dingqin201703@163.com³, zyj201707@163.com⁴, hy20160401@163.com⁵}
¹,²,³,⁵School of Computer, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, China; ⁴Research Center for Traditional Chinese Medicine Resources and National Medicine, Jiangxi University of Traditional Chinese Medicine, Nanchang 330004, China;

Abstract. Aiming at the problem that the traditional national medicine database information query system is easy to produce data redundancy in the actual query, which leads to the long query time of the query system, an Android based national medicine database information query system is designed. BrainLink single chip microcomputer is selected as the processing core, and the hardware part is designed by analyzing the controller, development board and circuit connection part. The software part designs the web architecture, uses the app program to interact with the user, calculates the redundant weight value of the system data, according to the size of the weight value, uses the soap serialization envelope object to describe, and completes the implementation of the query system on Android. The experimental results show that: compared with the traditional query system, the query time of the national medicine database information query system based on Android is the shortest.

Keywords: Android; National medicine database; Information Service; WEB architecture;

1 Introduction

Ethnic medicine refers to the medicine used by ethnic minorities in China, guided by the theory and practice of traditional medicine of the ethnic group. It is an important part of the traditional medicine treasure house. Its function and value in disease prevention and health protection have attracted wide attention at home and abroad. Therefore, the modernization research of national medicine is not only related to the development of national medicine itself, but also has great significance for enriching the treasure house of traditional Chinese medicine resources, improving the level of effective and comprehensive utilization of resources, and realizing the sustainable utilization of traditional Chinese medicine resources [1]. There are many kinds of ethnic medicine. There are some reports that there are about 8000 kinds of ethnic medicine in China, accounting for 85% of Chinese herbal medicine resources.
According to published ethnic medicine monographs and published literature, there are about 5500 kinds of ethnic medicine. As an industrial resource, ethnic medicine has gone through a long history of accumulation and systematic excavation. After the founding of the people's Republic of China, the party and government attached great importance to the inheritance and development of national medicine. From the "national health work plan for ethnic minorities" approved by the culture and Education Committee of the Government Council in December 1951 to the guiding spirit of the 18th National Congress of the Communist Party of China on supporting the development of traditional Chinese medicine and ethnic medicine, it is clear to vigorously develop ethnic medicine. Due to the large quantity, variety and wide distribution of ethnic medicine, it is difficult to integrate and query the information of ethnic medicine resources.

With the development of mobile Internet, the original web application of B / S architecture is gradually transplanted to the mobile platform. Under the net platform, generally through the web service technology provided by Microsoft to provide interface services to the outside world, it is convenient to realize service mobility. In this paper, a national medical information query system is developed by using the existing web service interface and Android platform to meet the needs of national medical researchers to obtain national medical information [2-3].

At present, in the aspect of hospitals, the trend of information management has begun to take shape. Some large hospitals and enterprises have begun to invest in the development. It is urgent to design a national medical data query system. In order to realize information sharing and reduce physical labor, the research of this topic has a very practical social significance.

2 Hardware design of information query system of national medicine database

2.1 Controller and development board

BrainLink single chip computer is selected as the processing core of the controller, which receives and processes the index data related to the query system [4]. The port of the single chip microcomputer is allocated to realize the control function of the single chip microcomputer. Before the allocation, the brainLink single chip microcomputer has four groups of I / O ports (P0, P1, P2 and P3) according to the I / O drive of the single chip microcomputer. The pull-up resistors are respectively connected at the eight pins of the I / O port to ensure that all the internal output transistors of the ports are in the cut-off state and the lower transistors are in the open state. Control two pull-up resistors in parallel at port P0 to ensure the output of "0" and "1" processing instructions. The connection diagram of single chip microcomputer when it is used as the output port is shown in the figure below:
As shown in Figure 1 above, group P0 port is connected to the circuit driving the LCD display, group P1 port is used to store the input national medicine data, group P2 port controls the LCD display driving signal of the query system, group P3 port is used to query the change of system information, and the specific allocation of I/O port is shown in the table below:

<table>
<thead>
<tr>
<th>Serial number</th>
<th>I/O port</th>
<th>Allocation function</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Group Q0</td>
<td>Provide data signal for LCD display</td>
</tr>
<tr>
<td>2</td>
<td>Q1.0 pin</td>
<td>Receive display output</td>
</tr>
<tr>
<td>3</td>
<td>Q1.1 pin</td>
<td>Receive data signal output</td>
</tr>
<tr>
<td>4</td>
<td>Q 1.2 - Q 1.4 pin</td>
<td>Provide driving signal for LCD display</td>
</tr>
<tr>
<td>5</td>
<td>Q 1.5 pin</td>
<td>Driver chip controller</td>
</tr>
<tr>
<td>6</td>
<td>Group P2</td>
<td>Drive control chip</td>
</tr>
<tr>
<td>7</td>
<td>Q 3.2 - Q 3.4 pin</td>
<td>Receive key input</td>
</tr>
<tr>
<td>8</td>
<td>Q 3.5 pin</td>
<td>Receive controller output</td>
</tr>
</tbody>
</table>

According to the port function shown in the above table, connect each port of the chip, design the control circuit of the controller, the left side of the single chip is mainly connected.
with the power circuit, the right side is connected with the clock circuit and the reset circuit. In
the actual work of single-chip microcomputer, it is defined as a machine cycle when the single-chip microcomputer accesses the memory from Rom. the internal oscillator of single-chip microcomputer is used to store the query data of a machine cycle. The xtal1 and xtal2 ports of the oscillator are used as the input / output ports of the oscillator. The xtal1 port of the oscillator uses internal and external clock mode to connect a quartz crystal, and then the external is equipped with a capacitor to form a parallel resonance circuit, so that the internal oscillation circuit generates self-excited oscillation [5].

ARM processor is selected as the processor of development board, and its 32-bit reduced instruction set processor architecture is used to be compatible with multiple 8-bit and 16 bit devices. The chip uses S3C2440 microprocessor, which is a 16 / 32-bit RISC embedded microprocessor integrated with ARM920T core. It supports 400MHz main frequency, up to 522mhz. Using an LCD display, connect the Ethernet RJ-45 interface. Connect the three serial ports, USB host, USB slave B and PWM control buzzer respectively, and connect the 34 pin GPIO pin of the expansion interface of the development board. Finally, it is connected to the bus interface of 40 pin system. The actual development board is shown in the figure below:

![Circuit board actually connected](image)

As shown in the figure above, the RF module of the development board uses the open ISM frequency band with an effective working range of 433 / 868 / 915MHz, uses the frequency modulator integrated inside nRF905 chip to realize the conversion of different query instructions, and uses the receiver with demodulator on chip to receive the frequency of different conversion instructions [6]. Using power amplifier, different database can be called with the same instruction. The 32 byte data of the control command is received and sent by the regulator, the transmission rate is adjusted to 50kbps, and the external 433MHz antenna is set to receive the data of the query command.

A s003nrf905 wireless data transmission module is set in the control and development
board. The working frequency range of the data transmission module is adjusted from 422.4mhz to 473.5mhz. 512 communication channels are built in to meet the control mode of multi-point communication and frequency modulation grouping. The SMA interface is used to connect the external antenna to enhance the function of sending and receiving the query command signal.

2.2 Circuit connection

Adjust the reset circuit of the single chip microcomputer system to the level switch reset mode, so that when the query system is connected to the power supply, the capacitor charging is in the short circuit state, and adjust the reset pin to connect to the high level. After the power supply is stable, the reset pin is grounded through the resistance, so that the capacitor can isolate the DC level. Redefine the pin function, as shown in Figure 3:

![Fig. 3. MCU pin](image)

As shown in the figure above, the pin function of the single-chip microcomputer is redefined, the bus bit of the crystal oscillator circuit is located on the same layer as the chip, and the "ground" network is used to surround it, so as to avoid laying the ground under the crystal oscillator. A 0.1 μ f decoupling ceramic capacitor is connected at each power pin to eliminate the high-frequency noise caused by the switch of IO port. Each pair of VCC and GND pins of the single-chip microcomputer are equipped with a capacitor with a short lead, and 64K ROM and 8K RAM are integrated internally to complete the design of the hardware part of the query system [7].

3 Software design of information query system for national medicine database

3.1 Design Web Architecture

The presentation layer (UI) in the web architecture is divided into web management end and Android client. The web management end is developed by asp.net 4.5, using IE browser to realize the interface with users, Android client is developed by Android technology, and using
app program to interact with users. The design business logic layer (BLL) is responsible for key business processing and data transmission. The operation and logical judgment of the database are processed in this layer. The processing of national medicine information and open web service interface of the system are processed in this layer [8]. The data access layer (DAL) is responsible for database access, mainly providing data for the business logic layer. The overall architecture design is shown in the following figure:

According to the requirements of the system, the system software is divided into two parts, one is web server, including system management module, role management module, user information management module, national medicine information management module, web service interface management module and other major functional modules [9]. The other part is Android client, including user registration module, personal information management module, query module by Chinese name of ethnic medicine, query module by ethnic medicine ethnic language and other sub query modules, as shown in the following figure:
According to the function module shown in the figure above, design the database of the system software part, complete the design of the information query module of the national medicine database, and then divide the database function of the web end to complete the design of the web architecture.

SQL Server 2016 database is adopted in the database of web end, and multiple data tables are designed according to the functional modules, including passport information table, material attribute table, cultural attribute table, chemical composition table, intellectual property table, literature source information table, germplasm resource table, resource image description table, sharing utilization, user description, etc. [10]. Each data table is associated with the same field name, mainly forming a data relationship with the drug ID as a foreign key, and the main database table relationship structure, as shown in the following figure:

![Database Table Relation Structure](image)

Fig. 6. Database table relation structure

Based on the above design, complete the design of the query system Web architecture, then calculate the data redundancy of the system in the web architecture, and complete the software design of the query system.

3.2 Calculate system data redundancy weight value

When using big data technology to calculate the weight value of various information in the national medicine database, the off-line calculation method is used to calculate the weight value of the national medicine information name in the national medicine database according to the analytic hierarchy process by using the Hadoop calculation mode. Firstly, a data redundancy judgment matrix is constructed, which is expressed as:

$$ C = (v_{ij})_{H \times V} $$  \hspace{1cm} (1)$$

Where, $H$ is the number of horizontal and vertical lines in the matrix, and $v_{ij}$ is the degree coefficient, which is generally taken as 1-10. Therefore, for the above formula (1), the degree of redundancy influence can be expressed as:
Normalize each column of data in judgment matrix $C$ to get matrix $C'$, add each row of $C'$ to get vector $c$, and use sum product method to get the maximum eigenvalue of vector, and get:

$$v_p = \frac{1}{V_p} (i, j = 1, 2, ..., n) \quad (2)$$

$$\alpha_{\text{max}} = \frac{1}{m} \sum_{j=1}^{n} \frac{W \times \psi^j}{V_p} \quad (3)$$

Where $\alpha_{\text{max}}$ is the maximum eigenvalue and $R$ is the coefficient of the sum product method. Using the random consistency ratio $WR$ to judge the consistency of the largest characteristic root, the following results are obtained:

$$WR = \frac{WJ}{RJ} \quad (4)$$

Among them, $WJ$ is the consistency index and $RJ$ is the average random consistency index. Therefore, the final weight value of each information can be calculated according to the consistency index:

$$WJ = \frac{\alpha_{\text{max}} - m}{m - 1} \quad (5)$$

In the above formula, $m$ is the order of judgment matrix. The formula (3) and (5) are combined to synthesize the random consistency table to get: when matrix $C$ meets $WJ < 0.1$, the normalized eigenvector can be used as the weight vector, and the final result of formula (2) is the weight value of the weight value index of the final information. Repeat the above calculation process, and finally calculate all the weight values in the information base. Then, according to the size of the weight value, the database with large import weight value is arranged.

When importing the query database into Android port, first use PC port to import the jar package of the query system, ksoap2-android-assembly-3.6.2-jar-with-dependencies, then find the namespace of web service, the parameter method called and the URL in the WSDL file. Then generate the soap request information that calls the web service method, which is described by the soap serialization envelope object.

4 Experiment

4.1 Experimental preparation

The WEB server of the system uses the .NET Framework framework of Microsoft Visual
Studio 2015, and the development language uses C#. In order to make the system have a better interactive interface and user experience, Javascript program is also used in the development process. First log in and log in to the web server to change the personal password. When entering the password, there will be a password strength prompt.

Then add the user window. The administrator can manually add other administrators to manage different national medicine data. The administrator can set the permissions of other users, or initialize passwords for other users. Because the user password system uses MD5 encryption, add new user / initialization user passwords, and input data according to the data provided by the members of the research group. The input data needs to be encoded when it is stored in the database, because the national language and characters cannot be queried and retrieved normally in the SQL Server database. In order to ensure that the Android client can retrieve the data correctly, the input national medicine data needs to be encoded.

After coding and importing the national medicine data, use soap WSDL to communicate with web service to link web service server and Android client. After the preparation of the experiment, we use two traditional national medicine database information query systems and Android based national medicine database information query system to experiment, and compare the response speed of three systems.

4.2 Analysis of experimental results

Based on the above preparations, take the bytes needed for each query of ethnic medicine information as independent variables, and finally the speed of the three query systems when actually querying ethnic medicine data information, as shown in the following figure:

(a) Experimental results of traditional query system 1
Experimental results of traditional query system 2

Fig. 7. Experimental results of three query systems

It can be seen from the above three experimental figures that for 1000-6000 bytes, the processing time of the standard query system increases by 0.05s with the increase of one thousand bytes, and the processing time of the traditional query system 1 increases unsteadily with the increase of bytes, resulting in data redundancy in the processing process. The traditional query system 2 increases with the increase of processing bytes, and the standard system The average difference of processing time is about 0.08s, and the data redundancy produced by the system is less. The final processing time of the national medicine database information query system based on Android is the same as the standard query time of the system, which basically does not produce system data redundancy. Compared with the two traditional query systems, it is more suitable for practical application.

5 Concluding remarks

This project develops a national medicine data information query system based on
Android platform by using a variety of computer technologies. After the system is completed, it will be deployed in the mobile intelligent terminal, which is convenient for users to carry around, which not only increases the way of national medicine knowledge dissemination, but also improves the efficiency of knowledge dissemination. This system provides convenient and quick learning tools for medical learners at home and abroad, provides good help for researchers to acquire professional knowledge of ethnic medicine, and provides powerful help for the dissemination, research and sustainable development of ethnic medicine. In the process of medical database information query, the problem of medical database information security is not considered, which leads to the decline of data quality. In the future, we will focus on the research of medical database information encryption technology to ensure the information security of medical database.

6 Fundprojects

National key research and development plan (2019YFC1712301), Science and Technology Project of Jiangxi Provincial Department of Education (GJJ150862) Tibet Autonomous Region Education Department College Young Teacher Innovation Support Plan Project (QC2015-67)

References


Measurement method of enterprise information asymmetry based on probability analysis

LIU Li¹, JIANG Wen-bing²
{yueg00300@163.com¹, jwb00321@sina.com²}
(¹Hunan University of Finance and Economics, Changsha 410205, China; ²market director, guizhou east century science & technology co., LTD., Guiyang 550081, China)

Abstract. Aiming at the problem of low sensitivity in the traditional measurement method of enterprise information asymmetry degree, a probabilistic analysis based measurement method of enterprise information asymmetry degree is designed. By analyzing the causes and countermeasures of asymmetric information of enterprises, the proxy variables are determined, the coherence of proxy variables is calculated by probability analysis, and the cost of adverse selection of enterprises is decomposed by covariance decomposition method, which is combined with the coherence of proxy variables to calculate the degree of asymmetric information of enterprises. The experimental results show that compared with the traditional measurement methods, the method based on probability analysis is more sensitive and suitable for practical projects.

Keywords: Probability analysis; Information asymmetry; Measurement method;

1 Introduction

The information asymmetry theory is a kind of classical theory, which means that in the market economic activities, all kinds of people have different understandings of relevant information. Those who have enough information are often in a favorable position, while those who have little information are at a disadvantage[1]. According to the theory, the seller knows the goods better than the buyer in the market. The party with large amount of information benefits from the market by transmitting reliable information to the party with small amount of information; the buyer and the seller with small amount of information try to obtain information from the other party; market signal can make up for information asymmetry to some extent[2].

In the enterprise, prone to two extremes, asymmetric information is a kind of adverse selection, due to the advance of information asymmetry, managers and other internal company personnel use may be bad for investors information favorable contract has been signed, while investors due to informational disadvantage in against their position, which makes the process
of market transactions from the desire of the lack of information\(^3\). Influence the decision of investors and damage the interests of investors. The other is moral hazard, usually after the conclusion of the transaction contract, the work results depend on both the investment efforts of the agent and various objective factors not determined by the subjective will, and these two factors cannot completely distinguish or distinguish the cost of very high agent moral hazard problem\(^4\).

For in the above case, adopt measure models or methods to measure the enterprise information asymmetry degree, according to the calculated result, the analysis of different degree of information asymmetry of adverse selection and moral hazard problems, from the angle of practical problems, through effective management means to reduce the company and the results of the capital market information asymmetry problem. Improving the governance mechanism of diversified companies, improving the relationship between the company and external investors, and achieving effective cooperation with the company's internal scientific management are the keys to give full play to the strategic advantages of diversified operations and ensure smooth financing channels.

Most of the previous methods for measuring the degree of information asymmetry of enterprises were based on the VAR model, and there was a problem of low sensitivity. Therefore, probabilistic analysis technology was adopted to design a method for measuring the degree of information asymmetry of enterprises based on probabilistic analysis, so as to solve the problems existing in the traditional methods.

2 Measurement method of enterprise information asymmetry based on probability analysis

2.1 Identify the proxy variable for enterprise information asymmetry

The agent variable of information asymmetry is the variable of information transaction, or the reason of information transaction, or the result of information transaction, which has certain causality. It is mainly the internal variables of the enterprise market, such as price change, bid-ask spread, turnover rate, internal ownership and enterprise growth\(^5\).

Before determining the proxy variable of enterprise information asymmetry, analyze the above variables and measure the value of the variables in measuring the degree of enterprise information asymmetry. According to the actual asymmetric information generation and response to determine the proxy variables.

Through the actual investigation of enterprises and markets, it is found that when the participants of enterprises and the market are faced with information asymmetry, they usually increase the sellers or decrease the buyers to avoid the losses caused by the exchanges with information traders. Therefore, the bid-ask spread is taken as one of the proxy variables. Due
to the serious social informatization at the present stage, the information transactions in enterprises are becoming larger and larger, and the volume of transactions involved is also becoming larger and larger. If there is no information transaction in the market, the enterprise holds a diversified portfolio and trades according to a certain proportion of the number of shares outstanding by other individuals; If a company believes that it has private information about the stock market, the information trader's trade will disrupt the stock market. Thus, turnover is also one of the proxy variables; at the same time, the internal personnel of the growth enterprise have an advantage over the general external personnel in understanding the investment opportunities of the enterprise and the future cash flow generated by the existing assets of the enterprise. Therefore, the growth of enterprises and the investment opportunities of enterprises can be used as the proxy variables of information asymmetry.

From the perspective of the value of enterprise information itself, the price of information itself, the number of information transactions, the demand of enterprises or markets for information and the information provided by enterprises or markets are regarded as the proxy variables to measure the information asymmetry of enterprises. The coherence between enterprise information's own variables is shown in Figure 1.

![Fig. 1. Correlation between enterprise information's own variables](image)

Probability analysis is used to calculate the probability distribution of the above basic variables, and the coherence of each variable to the measurement of enterprise information asymmetry degree is calculated.
2.2 The coherence calculation of proxy variables based on probability analysis

In the calculation of the coherence of proxy variables, multiple variable parameters are involved. Before the coherence calculation, the data of the above variables are randomly sampled to obtain the characteristic values of sample mean, variance, mathematical expectation, skewness coefficient and kurtosis coefficient, which to some extent reflect the distribution characteristics of proxy variables\(^6\). The random sampling process is shown in Figure 2.

![Flow chart of random sampling of proxy data](image)

Fig. 2. Flow chart of random sampling of proxy data

Using the sampled data, the mathematical expectation, variance, skewness coefficient and kurtosis coefficient are calculated to reflect the distribution characteristics of the proxy variables. The calculation formula is as follows:

\[
\hat{\epsilon}^2 = \frac{\sum (u - \bar{u})}{n} \quad (1)
\]
\[ E(i) = \sum_{n=1}^{N} u_n p_i \] (2)

\[ h_i = E \left( \frac{u_n p_i}{e^3} \right) \] (3)

\[ g_i = \frac{E \left[ (h_i - e)^4 \right]}{E \left( e^2 \right)^2} \] (4)

In the formula, \( e_i^2 \) represents the variance of the \( i \)th variable, \( \bar{u} \) represents the mean of sample data, \( u \) represents the sample data, \( n \) represents the number of sample data, \( E(i) \) represents the mathematical expectation of the \( i \)th variable, \( p_i \) represents the probability distribution of the \( i \)th variable, \( h_i \) represents the skewness coefficient of the \( i \)th variable, and \( g_i \) represents the kurtosis coefficient\(^7\). The above eigenvalues can only roughly reflect the characteristics of the distribution. In order to better calculate the coherence of the proxy variables, the distribution function of the proxy variables is obtained by histogram, and a probability distribution model is fitted by combining the eigenvalues. The coherence of the proxy variables is calculated by using the fitted probability distribution model.

The obtained eigenvalues are arranged in order of size, and the sequence of obtained variables is called empirical distribution\(^8\). Then these values are assigned to a series of intervals, where the interval division is estimated by Struges formula according to the sample size \( Q \):

\[ W = 1 + 3.31 \log Q \] (5)

After the upper limit \( a \) and lower limit \( b \) of the data are known, and the number of intervals \( W \) is selected, the interval interval can be calculated:

\[ \eta = \frac{a - b}{W} \] (6)

Sometimes the value of \( Q \) is not necessarily strict in order to make the interval interval
The calculation results can be adjusted properly. After the interval is divided, the median value of each interval is taken as the representative value of all data in the interval, the frequency, frequency and cumulative frequency of each interval are calculated, and the histogram is drawn\(^9\). The relation between frequency and cumulative frequency histogram is equivalent to the relation between probability density function and probability distribution function.

The normal distribution probability density curve is used to fit the above probability distribution model\(^{10}\). The probability density function is as follows:

\[
\beta_i(x) = \frac{1}{e\sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{x-\mu}{e} \right)^2 \right]
\]  

(7)

In the formula, \(x\) and \(\mu\) represent the distribution parameters. It is assumed that the sample data of the proxy variable is \(u_1, u_2, \ldots, u_n\), and the probability distribution model after fitting is \(F_i(u)\). The normal distribution is shown in Figure 3.

According to the sample interval and the size of the sample data, it is divided into groups
of data without intersection, then the coherence of the proxy variable is as follows:

\[ \gamma_i = \sum_{z=1}^{Z} \left( \beta_i - n u_z \right) \frac{1}{n u_z} \]  \hfill (8)

In the formula, \( u_z \) represents the sample data of the \( z \)th interval. Through the above process, the coherence of proxy variables in the measurement of enterprise information asymmetry is calculated, which is combined with the decomposed adverse selection cost to calculate the degree of enterprise information asymmetry.

### 2.3 Decompose the adverse selection cost

According to the microstructure theory, the asset price of an enterprise is affected by instruction processing cost, survival cost and adverse selection cost, among which the adverse selection cost reflects the degree of information asymmetry of the market. By decomposing short-term price changes, various costs are separated out. If asymmetric information is the only factor that affects the spread between bid and offer prices, the transaction will reflect the information transmitted by the exchange. A buy with a buy quote results in a sustained increase in the buy and sell quote, reflecting the information transmitted by the buy exchange. If liquidity costs are the source of displacement in the spread, then the quote adjustment reflects an inventory balancing transaction. After the buy quote is sold, the buy quote does not decline, but the price change is not permanent. Covariance method was used to decompose bid-ask spread into transaction cost, revenue cost, expense cost and spread.

The method of covariance decomposition is used to decompose the adverse selection cost. The covariance matrix of the actual factors in the cost of adverse selection is decomposed to achieve the purpose of decompressing the cost of adverse selection. The covariance is expressed as:

\[ s(n) = \sum_{i=1}^{N} \phi_i v_i \]  \hfill (9)

In the formula, \( \phi_i \) represents the non-negative eigenvalue, \( v_i \) represents the corresponding eigenvector, and \( i \) represents the number of costs after the inverse selection cost decomposition.

The decomposition of adverse selection cost is realized by the method of differential solution, and the parameters of transaction cost, revenue cost and expenditure cost after decomposition are used in the calculation of subsequent information asymmetry degree.
2.4 Calculate the degree of information asymmetry

Assuming that in a typical enterprise, there are many suppliers and demanders, both of which have certain information, the final transaction price is:

\[ d = d + \chi \| \bar{d} - d \| \quad (10) \]

In the formula, \( d \) represents the lowest transaction price acceptable to the supplier, \( \bar{d} \) represents the highest transaction price willing to pay by the demand side, and \( \chi (0 \leq \chi \leq 1) \) represents the degree of information available to the supplier in the pricing process. Therefore, \( \chi \| \bar{d} - d \| \) reflects the surplus captured by the supplier in the process of reaching the transaction price.

Next, the corresponding formula is used to show the degree of information acquired in the pricing process of the supply side and the demand side. Describe the "fair" transaction price of \( r(x) \) under the given conditions of individual basic characteristics ii, and it always meets: \( d \leq r(x) \leq \bar{d} \), so \( (\bar{d} - r(x)) \) represents the expected surplus of the demand side in the transaction price reaching process; \( (r(x) - d) \) represents the expected surplus of the supplier. Which side can "grab" more of the surplus will depend on the level of information they possess and the bargaining power based on it. With these remaining definitions, the formula is reformulated as:

\[ d = r(x) + [d - r(x)] + \chi [\bar{d} - r(x)] - \chi [d - r(x)] \quad (11) \]

Formula (11) shows that the supplier can increase the transaction price by taking part of the surplus expected by the demander, and the surplus size is \( \chi [\bar{d} - r(x)] \); In the same way, the demand side can reduce the transaction price by taking part of the surplus of the supply side, and the surplus size is \( (1 - \chi) [r(x) - d] \). The surplus captured by the supplier depends on the degree of information possessed by the supplier \( \chi \) and the total expected surplus of the demander \( \bar{d} - r(x) \). Similarly, the surplus obtained by the demand side
depends on the information level of the patient \((1 - \chi)\) and the total expected remaining
\(r(x) - d\) of the supply side, while the demand side can also lower the transaction price
through the information level it has acquired. Therefore, in formula 11, it is mainly composed
of three parts. The first part \(r(x)\) represents the "fair" transaction price given individual
characteristics, which is called the benchmark price. The second part \(\chi[d - r(x)]\) reflects
the surplus captured by the supplier through the degree of information. The third part
\((1 - \chi)d\) is the surplus obtained by the demand side through the degree of
information. Net surplus \(F_x = \chi[d - r(x)] - (1 - \chi)[r(x) - d]\) can be used to
describe the degree of information asymmetry in the formation of transaction prices.

The above process is an ideal state, measuring the degree of the asymmetry of the
structure of the enterprise in the process of actual measurement, there are some interference is
proxy variables and adverse selection costs, from the above process, the interference by a
probability distribution, has the characteristics of normal distribution, therefore, derive the
probability density function of complex interference item \(\xi_i\) as follows:

\[
F(\xi_i) = \frac{\exp(\gamma_i)}{\epsilon_i} + \frac{\exp(\nu_i)}{\phi_i} \quad (12)
\]

For the measurement of information asymmetry degree of enterprises, it is mainly
measured by the surplus obtained by the information degree mastered by the supply side and
the demand side. Therefore, the net surplus in the bargaining process is calculated as follows:

\[
F_x = \xi_i\left[\chi d - r(x) + (1 - \chi)d\right] \quad (13)
\]

The net surplus obtained through the above process is the degree of enterprise
information asymmetry, so the measurement method of enterprise information asymmetry
based on probability analysis has been designed.

3 Simulation experiment and analysis
3.1 Set experimental constraints

Because it is difficult to obtain real enterprise information, MATLAB software is used to
simulate the experimental data. Prior to this, in order to prevent external factors from interfering with the experimental results, experimental constraints were formulated to ensure the effectiveness of the experiment.

According to the hypothesis of "information diversification", the level of diversification has a negative effect on the degree of information asymmetry, that is, the higher the level of diversification, the lower the degree of internal and external information asymmetry. The correlation between the operating departments of diversified enterprises and the estimation bias is consistent. The higher the level of non-correlated diversification is, the lower the degree of internal and external information asymmetry is. Therefore, constraint 1 is proposed. Information diversification has no influence on the degree of internal and external information asymmetry, and internal and external information is unified.

According to the hypothesis of "information transparency", the level of information transparency will have a positive effect on the degree of information asymmetry, that is, the higher the level of information transparency, the higher the degree of internal and external information asymmetry should be. However, when there is a high degree of correlation between different departments of an enterprise, the accuracy of reflecting the operating conditions of the department through the overall cash flow will also affect the transparency of information. Therefore, constraint condition 2 is put forward. Information transparency has no influence on the degree of internal and external information asymmetry of enterprises, and enterprise information transparency is unified.

3.2 Sensitivity test results and analysis

Under the above constraints, we use MATLAB software to extract the sample information of enterprises, and use the design budget method and the traditional budget method to measure the sensitivity of enterprise information asymmetry. The experimental results are as follows:
According to the results in the observation figure, as shown in figure a, the sensitivity obtained by the traditional measurement method is between 0.2 and 0.6 in the early stage, and
basically stable between 0.2 and 0.4 in the later stage. As shown in figure 2, the sensitivity obtained by using the designed measurement method is relatively stable, always between 0.6 and 0.8. Compared with the two, the designed measurement method is more sensitive than the traditional one, indicating that the designed measurement method based on probability analysis is better than the traditional one.

4 Conclusion

The degree of information asymmetry of an enterprise has an important impact on the management and development of an enterprise. Traditional measurement methods cannot meet the current measurement of enterprise information asymmetry degree. Therefore, a probabilistic analysis based measurement method of enterprise information asymmetry degree is designed. Through the design comparison experiment, it is proved that the design measurement method solves the problems existing in the traditional measurement method, which is of great significance for the future development of enterprises.

Reference

[10] Yuqi, M., Ruipeng, H., Yanyun, Z.: The study of ownership, industry difference and exit mechanism
Predicting Sales of Cross-chain Discrete Manufacturing Products Based on LSTM

Zhicheng Wang¹, Ye Tao², Xiangqian Ding¹, Ruichun Hou¹, Hui Zhang¹
{wangzhicheng@stu.ouc.edu.cn, ye.tao@qust.edu.cn, hrch4968@sina.com}

¹ College of Information Science and Engineering, Ocean University of China, Qingdao, Shandong, China,
² School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, China

Abstract. In the mass customization product value chain, the sales forecast in the marketing chain is important. Traditional forecasting methods only consider the previous sales volume, which is a single value chain factor. In fact, with the rapid development of online media, the factors affecting the sales of manufacturing products have become increasingly complex. For example, the attention and evaluation of online social media is an important influence on sales forecasts. In addition, other influencing factors such as stocks also have a certain impact on product sales. Therefore, it is possible to integrate cross-domain information to improve the accuracy of prediction results. In this paper, we propose an LSTM model with the portal's attention index, and involve factors from multiple domains, such as the network media, market environment and original product sales volume. Experiment results show that using the proposed cross-domain approach obtains more accurate prediction results than other mainstream models.

Keywords: value chain, cross-chain prediction, LSTM.

1 Introduction

In modern mass customization manufacturing, product sales forecasting is one of the important issues in business intelligence. Accurately predicting product sales trends can help companies better grasp market demand, adjust output and optimize inventory. In the modern market, the product life cycle has become a "pencil type" with extremely short-lived peaks. It is no longer a parabola-like slow change, but straight up and down. Product life cycles are getting shorter and shorter, so sales forecasting is becoming more important. This makes the senior management of the enterprise must formulate flexible marketing and inventory allocation strategies for the rapidly changing market, and the forecast of sales volume will be an important reference for the senior management of the enterprise to formulate corporate strategies.

Most forecasting methods are built upon time series models with single input and have been applied in some specific actual scenarios and has a good forecasting effect. Traditional time series forecasting methods fit historical time trend curves by establishing appropriate mathematical models, but the data they rely on are simple and often face lag problems.
With the continuous improvement of the Industrial Internet system, it is possible to obtain multidimensional data from multiple value chains. For example, with the development of the Internet, the attention of various brands on the Internet has had a significant impact on sales. Many consumers can post their own evaluation of a certain product on social media, and the impact of consumer evaluation on sales is becoming more and more important. Therefore, many scholars [1]-[3] find that using cross-domain data to predict sales can get better results through in-depth research.

This article conducts an experiment on the time series of sales volume of a manufacturing industry, and collects four types of cross-value chain data: sales data, user evaluation, Internet attention index, and corporate stock price fluctuations of 18 different manufacturers. According to the actual situation, the data are optimized for the influence of de-seasonal factors. Different from the data of a single sales value chain used in previous studies, this article uses four types of cross-value chain data to predict the sales volume of a manufacturing industry through the LSTM model. The experiment compares the impact of cross-value chain data and single value chain data on the accuracy of sales volume prediction, the effect of de-seasonal data and seasonality data on the accuracy of sales volume prediction, and the impact of using LSTM model and LSSVR model on sales volume prediction. The experimental results show that compared with the past, the experimental results have been greatly improved.

The rest of this study is organized as follows. Related work about predicting sales is introduced in Section 2. Section 3 presents methodologies utilized in this study and the proposed framework. Section 4 depicts the results of this study. Conclusions are delivered in Section 5.

2 Related work

There are many traditional prediction methods for time series data. F. M. De and X. Yao [4], [5] used the traditional method to forecast the load of the power grid, and used the sales volume of the previous day as the input to forecast the sales volume data of the next day, and achieved a certain forecast effect. However, only the data of previous day is used, and the prediction result of the previous data input as input for the time series model is more error. D. Trigg and A. Leach [6], [7] used the exponential smoothing method to predict the discrete signals and achieved a certain prediction effect. Exponential smoothing is based on previous predictions and the percentage of prediction errors. Because of the simplicity, transparency, and ability to capture various time series data patterns. In recent years, exponential smoothing has become one of the most popular and practical methods in time series prediction. However, for the sales forecast of manufacturing products, it is clear that the index smoothing model cannot accurately reflect the market and predict sales volume. Therefore, the method of index smoothing is not applicable to the sales forecast of manufacturing products in modern volatile market.

In recent years, with the development of artificial intelligence algorithms [8], BPNN [9] and SVR [10] algorithms have also become common methods for predicting time series data. In previous related studies, some scholars used the BPNN [11], [12] algorithm to predict time series data. In some areas, such as automobile sales forecasting, certain results have been achieved. But for the BPNN algorithm, it requires a large amount of data, and the prediction result of small samples is not ideal. For small sample predictions, the SVR [13], [14] algorithm performs better. Therefore, in recent years, many scholars have used SVR algorithms to predict time series data and have achieved some results. However, the SVR algorithm has poor practicability and poor generalization ability.
In recent years, with the development of online social media, the value chain of online social media has become an important factor affecting sales. Previously, some scholars have also done a lot of related work to prove that the promotion and evaluation of social media can affect the sales of manufactured products\cite{15}\-\cite{17}. At the same time, some scholars have also pointed out that market macro performance, such as stock information, industrial indexes and other indicators, will also affect the sales of manufacturing products\cite{18}, \cite{19}.

In these studies, LIU and PAI\cite{20} used the LSSVR model to model car sales, taking past car sales, Twitter reviews, and the Dow Jones Industrial Index as inputs, and predicting car sales for the next month through training. However, in this article, the Dow Jones Industrial Index as a macro value that represents the overall market does not reflect the performance of different companies and manufacturers in the capital market. Therefore, there is still room for improvement in the forecast results. This article will focus on comparing the LSTM prediction model with LIU’s LSSVR model, emphasizing that the prediction accuracy of prediction models across multiple value chains is greatly improved compared with the prediction results of single value chain prediction models.

### 3 Predicting based method

#### 3.1 Basic framework

We use two types of data sets for prediction. The first type is the total monthly sales of home appliances from different manufacturers, and the second type is independent variable data, which includes Weibo users' sentiment on the corresponding manufacturers' products, and the Baidu index of the corresponding manufacturers.

![Figure 1: The proposed monthly total sales predicting framework](image)

We performed experiments in two groups, one using raw data and the other using moving average (MA) to remove seasonal factors. The data across multiple chains are then used as input to be trained in the LSTM model and the LSSVR model, and the prediction results are obtained. The algorithm flow chart in this paper is shown in Fig.1.
3.2 Deasonalization method

If the value of the time series at time $t$ can be explained by a linear combination of time $t$ and $i$ error terms before time $t$, then a moving average MA($q$) model can be used to predict the value at time $t$.

$$ y_t = \varepsilon_t + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i} + \mu $$

In Eq. (2): $y_t$ represents the value of the predicted time series at time $t$; $\varepsilon_t$ and $\varepsilon_{t-i}$ represent the random errors at time $t$ and $t-i$ respectively; $q$ represents the order of the model; $\beta_i$ is the regression parameter; $\mu$ represents a constant term. In the following sections, we will use the MA algorithm to remove the seasonal factor operation.

In the following sections, we will use the MA algorithm to remove the seasonal factor operation.

3.3 LSTM (Long Short-Term Memory) method

In 1997, Hochreiter first proposed Long Short-Term Memory (LSTM), which is a special recurrent neural network (RNN), which can effectively solve the problem of gradient disappearance or gradient explosion of RNN, and can learn long-term dependencies. Compared with RNN, LSTM is designed for the controller of the neural cell, which can judge whether the information is useful. The Cell control unit is shown in the figure.

![Cell structure in LSTM model](image)

**Fig. 2.** Cell structure in LSTM model

The control unit in the LSTM model consists of a memory unit $C$ for recording the current state and three gates that control the flow of data. These three gates are composed of input gate
\( i \), output gate \( o \), and forget gate \( f \). At time \( t \), after the data enters the control unit, through calculation, the LSTM can choose to remember or forget some information, output the control information, and pass these status information to the next time \( t + 1 \).

The forget gate is the information that determines what we will discard from the state of the cell. The gate reads \( h_{t-1} \) and \( x_t \) and outputs a value in the range 0-1 to the number in state \( C_t \) of the previous output. 1 means "completely reserved", 0 means "completely discarded". The calculation method is shown as Eq.(3). Where \( f_t \) represents the forget gate information at time \( t \).

\[
f_t = \sigma \cdot \left( w_f \cdot [h_{t-1}, x_t] + b_f \right)
\]

The input gate determines how much new information is added to the cell state. The calculation method is shown as Eqs.(4-5). Among them, \( i_t \) indicates the input gate information at time \( t \), and \( C_t \) indicates the update of the memory unit, which indicates how much information is forgotten and which of the current input information needs to be updated into the current memory unit.

\[
i_t = \sigma \cdot \left( w_t \cdot [h_{t-1}, x_t] + b_i \right)
\]

\[
C_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(w_c - [h_{t-1}, x_t] + b_c)
\]

The output gate determines what value to output. The calculation method is shown as Eqs.(6-7). Among them, \( o_t \) indicates that the door information is output at time \( t \), which is used to control the update of the information to achieve the purpose of increasing and decreasing the information. \( h_t \) produces the current output result, and the output gate decides which information is finally output.

\[
O_t = \sigma \cdot \left( w_o \cdot [h_{t-1}, x_t] + b_o \right)
\]

\[
h_t = o_t \cdot \tanh(C_t)
\]

In the above formula, \( \sigma \) represents the sigmoid function, and \( w_f, w_t, w_o, b_f, b_i \) and \( b_o \) represent the weights and offset values of the three gates, respectively.

### 3.4 LSSVR (least square support vector regression) method

The least-square support vector regression (LSSVR) model is one of the modified techniques which can reduce the computation load of the original support vector regression model. Suykens et al. Proposed the least squares support vector regression (LSSVR) model based on the standard SVR. LSSVR changed the inequality constraints of the SVR to equality constraints, which made the solution of the original convex quadratic programming problem into a solution of linear equations, which greatly reduced the calculation difficulty and calculation time.

Given a set of training samples \( \{x_i, y_i\}_{i=1}^M \), \( x_i \in R^p \) and \( y_i \in R \) represent the input and expected output of the model, respectively, and \( M \) represents the number of samples. \( f(x) \) represents the mapping relationship between model input \( x \) and output \( y \). The calculation formula is as follows:
\[
f(x) = x^Tw + b
\]  
(8)

In Eq.(8), \(w\) and \(b\) are weight vectors and offset terms, respectively. In order to solve for the appropriate \(w\) and \(b\), the loss function of LSSVR needs to be calculated:

\[
\min \frac{1}{2} ||w||^2 + \frac{1}{2} \sum_{i=1}^{M} e_i^2
\]  
(9)

subject to \(x_i^Tw + b - y_i = e_i\)  
(10)

Finally, the equation is solved. The prediction function \(f(x)\) of LSSVR is expressed as:

\[
f(x) = \sum_{i=1}^{M} \alpha_i x_i^Tx + b
\]  
(11)

4 Experiments and discussions

4.1 Data set description

We use two types of data sets for prediction. The first type is the total monthly sales of home appliances from different manufacturers, and the second type is independent variable data, which includes Weibo users' sentiment on the corresponding manufacturers' products, and the Baidu index of the corresponding manufacturers' products. And the stock price of the corresponding manufacturers.

The data of manufacturers' home appliances from 2012.1-2018.6 was provided by related companies. There are a total of 18 different product models. Related data involves confidentiality. The product data of Weibo users from 2012.1-2017.12 was found out by the crawler code, and the Weibo were collected by keywords, that is, "buyer electricity". Through text analysis, distinguish and count whether the user's evaluation of the product is positive or negative.2012.1-2017.12 Baidu index data is the number of times the manufacturer was searched and followed in the Baidu search engine. The stock data of 2012.1-1017.12 is the percentage change of stocks of corresponding manufacturers for the month.

4.2 Data preprocessing and model building

In this study, a time series model and a multivariate regression model were performed to predict total monthly sales of white goods. This article will predict the overall sales of white goods for two-time series models, LSSVR and LSTM. In addition, the non-seasonal de-seasonalization procedure obtained from the monthly white goods sales data applies to the monthly white goods sales data, popularity scores, Baidu index, and stock market value. The three independent variable data are arranged and combined, and 16 LSTM / LSSVR models are listed to use different data combinations to predict the total monthly sales volume and study it.
Table 1. Data types and corresponding codes

<table>
<thead>
<tr>
<th>LSSVR/RNN models</th>
<th>Data used</th>
<th>LSSVR/RNN models</th>
<th>Data used</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM/LSSVR 1</td>
<td>Y</td>
<td>MA-LSTM/MA-LSSVR 1</td>
<td>DY</td>
</tr>
<tr>
<td>LSTM/LSSVR 2</td>
<td>X1, Y</td>
<td>MA-LSTM/MA-LSSVR 2</td>
<td>DX1, DY</td>
</tr>
<tr>
<td>LSTM/LSSVR 3</td>
<td>X2, Y</td>
<td>MA-LSTM/MA-LSSVR 3</td>
<td>DX2, DY</td>
</tr>
<tr>
<td>LSTM/LSSVR 4</td>
<td>X3, Y</td>
<td>MA-LSTM/MA-LSSVR 4</td>
<td>DX3, DY</td>
</tr>
<tr>
<td>LSTM/LSSVR 5</td>
<td>(X1, X2), Y</td>
<td>MA-LSTM/MA-LSSVR 5</td>
<td>(DX1, DX2), DY</td>
</tr>
<tr>
<td>LSTM/LSSVR 6</td>
<td>(X1, X3), Y</td>
<td>MA-LSTM/MA-LSSVR 6</td>
<td>(DX1, DX3), DY</td>
</tr>
<tr>
<td>LSTM/LSSVR 7</td>
<td>(X2, X3), Y</td>
<td>MA-LSTM/MA-LSSVR 7</td>
<td>(DX2, DX3), DY</td>
</tr>
<tr>
<td>LSTM/LSSVR 8</td>
<td>(X1, X2, X3), Y</td>
<td>MA-LSTM/MA-LSSVR 8</td>
<td>(DX1, DX2, DX3), DY</td>
</tr>
</tbody>
</table>

Table 2. LSSVR\LSTM models and data used for predciting monthly total sales

<table>
<thead>
<tr>
<th>Data codes</th>
<th>Data descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>Sentiment scores of Weibo</td>
</tr>
<tr>
<td>X2</td>
<td>Baidu index</td>
</tr>
<tr>
<td>X3</td>
<td>Percent change in stocks Market values</td>
</tr>
<tr>
<td>DX1</td>
<td>Deseasonalized sentiment scores of Weibo</td>
</tr>
<tr>
<td>DX2</td>
<td>Deseasonalized Baidu index</td>
</tr>
<tr>
<td>DX3</td>
<td>Deseasonalized Percent change in stocks Market values</td>
</tr>
<tr>
<td>Y</td>
<td>Monthly total sales</td>
</tr>
<tr>
<td>DY</td>
<td>Deseasonalized Monthly total sales</td>
</tr>
</tbody>
</table>

In the flowchart, we divide the independent variables into two cases, one is the original data, and the other is fed into the MA seasonal factor algorithm, and then the independent variables and sales are combined and fed into the LSTM / LSSVR model for training.

4.3 The results

In this section, the results of predicting monthly total sales by time series models and multivariate regression models with various data types are illustrated. The predicting performance is measured by MAPE, WAPE[21] and NMAE[22] are shown as Eqs. (12-14).

\[
MAPE(\%) = \frac{100}{N} \sum_{t=1}^{N} \left| \frac{Y_t - F_t}{Y_t} \right|
\]  

\[
WAPE(\%) = 100 \frac{\sum_{t=1}^{N} \left| Y_t - F_t \right|}{\sum_{t=1}^{N} Y_t}
\]  

\[
NMAE = \frac{1}{Y_h - Y_l} \left[ \frac{1}{N} \sum_{t=1}^{N} |Y_t - F_t| \right]
\]

where \( N \) is the number of predicting periods, \( Y_t \) is the actual value at period \( t \), and \( F_t \) is the forecasting value at period \( t \), \( Y_h \) is highest actual value, and \( Y_l \) is the lowest actual value.
Comparison of results using seasonal data

Table 3. MAPE, WAPE and NMAE values by using LSTM/LSSVR models with seasonal data of independent variables and original monthly total sales to predict monthly total sales.

<table>
<thead>
<tr>
<th>LSTM Models</th>
<th>MAPE</th>
<th>WAPE</th>
<th>NMAE</th>
<th>LSSVR Models</th>
<th>MAPE</th>
<th>WAPE</th>
<th>NMAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM1</td>
<td>12.36%</td>
<td>24.10%</td>
<td>0.030</td>
<td>LSSVR1</td>
<td>24.10%</td>
<td>16.59%</td>
<td>0.050</td>
</tr>
<tr>
<td>LSTM2</td>
<td>12.23%</td>
<td>21.18%</td>
<td>0.046</td>
<td>LSSVR2</td>
<td>15.12%</td>
<td>15.50%</td>
<td>0.046</td>
</tr>
<tr>
<td>LSTM3</td>
<td>11.83%</td>
<td>21.51%</td>
<td>0.026</td>
<td>LSSVR3</td>
<td>15.64%</td>
<td>15.67%</td>
<td>0.047</td>
</tr>
<tr>
<td>LSTM4</td>
<td>9.38%</td>
<td>23.64%</td>
<td>0.019</td>
<td>LSSVR4</td>
<td>16.56%</td>
<td>16.52%</td>
<td>0.050</td>
</tr>
<tr>
<td>LSTM5</td>
<td>11.08%</td>
<td>16.59%</td>
<td>0.024</td>
<td>LSSVR5</td>
<td>7.43%</td>
<td>7.43%</td>
<td>0.017</td>
</tr>
<tr>
<td>LSTM6</td>
<td>11.08%</td>
<td>16.59%</td>
<td>0.024</td>
<td>LSSVR6</td>
<td>7.43%</td>
<td>7.43%</td>
<td>0.017</td>
</tr>
<tr>
<td>LSTM7</td>
<td>8.43%</td>
<td>15.12%</td>
<td>0.018</td>
<td>LSSVR7</td>
<td>5.71%</td>
<td>5.71%</td>
<td>0.018</td>
</tr>
<tr>
<td>LSTM8</td>
<td>8.43%</td>
<td>15.12%</td>
<td>0.018</td>
<td>LSSVR8</td>
<td>5.71%</td>
<td>5.71%</td>
<td>0.018</td>
</tr>
</tbody>
</table>

It can be seen from Table 3 that when using data without seasonal factors for prediction, when the number of independent data combinations increases, the values of WAPE and NMAE decrease significantly. It can be seen that for the sales forecast of this type of product, the more independent variables, the smaller the error percentage of the predicted sales. For the three independent variables of Weibo's product evaluation, the Baidu index, and corporate stock percentage change, the most relevant to the sales forecast is the stock percentage change.

As can be seen from Table 3, comparing the LSTM model with the LSSVR model, the values of MAPE and NMAE predicted by the LSTM model are significantly smaller, indicating that the error percentage of the LSTM model predicted sales is smaller. At the same time, the NMAE value predicted by the LSTM model is also smaller, indicating that the LSTM is more accurate in predicting extreme values in the prediction. It can be seen that for this type of product sales forecast, the prediction effect of the LSTM model is significantly better than the prediction effect of the LSSVR model.

In this paper, five types of data, real sales, LSTM1, LSSVR1, LSTM8, and LSSVR8, are selected for comparison as shown in Fig. 3.
As can be seen from Table 4, when using the data without seasonal factors for prediction, as the number of independent data combinations increases, the values of WAPE and NMAE decrease significantly. It can be seen that for the sales forecast of this type of product, the more independent variables, the smaller the error percentage of the predicted sales. For the three
independent variables of Weibo's product evaluation, the Baidu index, and corporate stock percentage change, the most relevant to the sales forecast is the stock percentage change.

Comparing the MA-LSTM model with the MA-LSSVR model from Table 4, the values of MAPE and NMAE predicted by the LSTM model are significantly smaller, indicating that the error percentage of the LSTM model predicted sales are smaller; meanwhile, the NMAE values predicted by the LSTM model are also smaller, indicating that LSTM is more accurate in predicting extreme values in prediction. It can be seen that for this type of product sales forecast, the prediction effect of the LSTM model is significantly better than the prediction effect of the LSSVR model.

We compare the MA-LSTM model and the LSTM model from Tables 3 and Tables 4. In general, the prediction result using the data without seasonal influence is better than the prediction result using the data without seasonal influence. Comparing Weibo Review, Baidu Index, and stock fluctuations independently of seasonal factors to remove the impact of seasonal factors on forecasting results, it can be found that Weibo Review and Baidu Index's data removing seasonal factors affect the forecast accuracy. And the stock data has almost no improvement after removing the influence of seasonal factors.

In this paper, five types of data, real sales, MA-LSTM1, MA-LSSVR1, MA-LSTM8, and MA-LSSVR8, are selected for comparison as shown in Fig. 4.

![Fig. 4. Actual and predicted deseasonal monthly total sales of LSTM/LSSVR models.](image)

Experimental results were evaluated using the mean percentage error (MAPE). The experimental results show that the prediction error of the cross-value chain data is 42.15% lower than the prediction result of the single value chain data; the prediction result of the de-seasonal factor data is 15.25% lower than the prediction result of the data without seasonal factor; the prediction result of LSTM model is 52.15% lower than that of LSSVR model.

5 Conclusions

This research proposes a new algorithm MA-LSTM, which includes a time series prediction model and multivariate regression technology to predict monthly sales. De-
seasonalization algorithms are used to process different types of data. Numerical results show that using mixed multivariate regression data to predict sales can obtain more accurate prediction results than univariate models. The excellent prediction performance can be concluded as follows. First, using a mix of data that includes portal attention, social media ratings, and stock market value can improve forecast accuracy. Secondly, quantifying the impact of seasonal factors on the attention value of social networking sites and social media evaluation data has significantly improved the accuracy of sales forecasting. The de-seasonal factor of stock data does not greatly improve the accuracy of the sales forecast. It may be because the stock market is more rational, and the correlation between stock market fluctuations and seasonal factors is not great.

In this data set, compared with the prediction of the sales volume of the LSSVR algorithm of PAI and LIU, the MA-LSTM algorithm proposed in this paper is significantly better than the LSSVR algorithm. The reason may be that the LSSVR algorithm has advantages for small sample predictions, but has a poor prediction effect for the 18 products and large data sets of 90 months in this article. Therefore, for data sets with a large amount of data, the MA-LSTM algorithm is significantly better than the LSSVR algorithm.

For future research, since the identification of Weibo keywords has a significant impact on the search results of Weibo and the accuracy of predictions, more systematic techniques for selecting the correct keywords from Weibo maybe future research directions. Another possible direction for future research is to use other social media data, such as the number of product introductions related to video sites, to predict sales. Finally, the collection of geographic information on Weibo may be an important issue for future research to improve Weibo analysis, and shopping preferences of people in different regions may be significantly different.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1702902, and in part by the Shandong Province Colleges and Universities Young Talents Initiation Program under Grant 2019KJN047.

References

Prediction model of spatial distribution pattern of building based on Neural Network

YU Jie1, LIU Lin2

{yjjy901229@163.com, yjjy9951000@163.com}

1. Department of architecture and art, Chongqing JianZhu College, Chongqing 400000, China;
2. Department of architecture and art, Chongqing JianZhu College, Chongqing 400000, China

Abstract. The traditional prediction model of building spatial distribution pattern change has the problem of poor prediction accuracy, so a prediction model of building spatial distribution pattern change based on neural network is designed. Based on the analysis of the spatial characteristics of traditional villages in Southeast Chongqing, this paper uses the neural network method to predict the change of the spatial distribution pattern of buildings, so as to complete the construction of the prediction model of the change of the spatial distribution pattern of buildings, and puts forward the planning and design strategies of traditional villages in Southeast Chongqing for the effective planning of traditional villages in Southeast Chongqing. The experimental results show that the prediction model based on neural network is more accurate than the traditional model, and can meet the needs of the prediction of the change of the spatial distribution pattern of buildings.

Keywords: Neural network; Architecture; Spatial distribution pattern; Change; Forecast;

1 Introduction

The spatial pattern of traditional villages reflects the formation and evolution of villages, as well as the allocation of natural and social resources. It is a concentrated reflection of its characteristics, which can directly and comprehensively reflect the historical evolution and value characteristics of the whole village. It is the focus of traditional village protection. However, in the protection process of traditional villages in Chongqing, there are not only a large number of problems, scattered distribution, remote region, backward development, but also the phenomenon of ethnic cultural differences, ineffective protection and management, which leads to the poor accuracy of the traditional prediction model of the change of the spatial distribution pattern of buildings. Therefore, it is necessary to predict the change of spatial distribution pattern of buildings in Chongqing. This study focuses on the southeast of Chongqing, which is located in the basin edge mountainous area where the two mountain systems of Dalushan and Wuling Mountain meet in the southeast of Sichuan Basin, mainly including Qianjiang District, Shizhu, Pengshui, Youyang, Xiushan and "one district, five counties" of Wulong county. Southeast Chongqing, as one of the few minority gathering areas in the country, is the key area for China to implement the strategy of targeted poverty
alleviation and relocation. Alpine ecological migration is an important measure to solve poverty and protect ecology in combination with the reality and influencing factors of poverty in Chongqing. In recent years, it is highly valued by the state and the government. However, with the rapid rise of rural construction and the policy support of poverty alleviation through relocation, ecological migration has made great achievements. At the same time, there are a series of problems. In order to pursue the scale effect of rapid rural construction, the spatial planning of new immigrant villages often applies the urbanization construction mode. This "homogenization" construction can not be applied to the new ethnic minority immigrants with regional cultural characteristics. In villages, the blind application of model construction will only bring about a series of problems, such as spatial alienation, cultural fault, employment difficulties and psychological maladjustment caused by immigrants' loss of land survival skills after they are far away from their homes.

Literature [1] put forward the spatial distribution pattern and combination method of typical natural secondary forest species in Zhejiang Province. In this study, four 1-hm2 stands were established in coniferous and broad-leaved mixed forest and broad-leaved mixed forest, including two mixed forest (No.1 and No.2), two broad-leaved forest (No.3 and No.4). The distribution and relationship of dominant species in different life stages were studied, and the combination and distribution of species were detected. The species with DBH ≥ 5cm were identified. The density of Pinus massoniana in zone 1 was lower than that in zone 2. In this study, the spatial distribution of dominant tree species and communities of different tree species were analyzed by spatial point model analysis. However, the prediction accuracy of this method is low.

In view of the problems of the above methods, this paper starts from the analysis of the spatial characteristics of traditional villages in Southeast Chongqing, designs a prediction model based on neural network for the change of building spatial distribution pattern, and puts forward the planning and design strategies of traditional villages in Southeast Chongqing according to the existing problems, so as to complete the prediction of building spatial pattern change based on neural network.

Among them, neural network is an algorithm mathematical model that imitates the behavior characteristics of animal neural network and carries out distributed parallel information processing. This kind of network relies on the complexity of the system, through adjusting the relationship between a large number of internal nodes, so as to achieve the purpose of processing information.

2 Prediction model of spatial distribution pattern of building based on Neural Network
2.1 Spatial characteristics of traditional villages in Southeast Chongqing

The settlement space of Tujia Nationality in Southeast Chongqing is affected by natural environment, clan system, livelihood and other aspects. According to the distribution of terrain and geographical environment, it is mainly divided into mountain village settlement and town settlement, which correspond to clan settlement and trade settlement respectively. Clan settlement is called a community formed by farming together in the same family unit, an autonomous group with common living habits and certain rules and regulations, a self-sufficient natural economy within the clan, a closed and autonomous group social organization established in strict accordance with traditional standards, from the whole clan's life organization to spiritual organization and social organization, the patriarchal system is firmly printed. The spatial form of clan settlement is hierarchical, and different levels of public space will appear in the level of social attribute [2]. The higher the level, the more popular the places gather, and the larger the activity space correspondingly. Although the hierarchy will be accompanied by the central spatial form, the Tujia settlements in Southeast Chongqing are not central due to the limited mountain conditions, and most of them are distributed in a decentralized layout, closed and stable.

Most of the clan settlements are staggered with Han and Miao villages. Most of the villages are built on the hillside near the mountain and the water. This is because the plain dam in Wuling mountain area is very precious and there are many mountains and few land, so most of the flat land is left for farming. As shown in the figure below:
Fig. 1. Settlement distribution map

Most of the clan rural settlements are the combination of mountain buildings and alleys according to local conditions. Their development form is to form buildings first, and then to form streets and alleys. The whole growth process is based on nature and integrated into nature, so on the whole, the form presents an occasional interest. This is because the clan settlement is affected by the terrain and environment, and less restricted by the etiquette system. The stacked Tujia courtyard appears random and flexible in space. One of these two forms of settlements is called "market", which is an open-air trading and trading mode that specifies a certain area at a certain time node; the other is that on the flat dam, a single house gradually develops into a market. Such trade settlements generally have water systems passing through, and the houses expand along the street. The general layout of Tujia traditional town settlements in Southeast Chongqing is mainly in linear form. Due to the local topography and residents' sense of settlement, such settlement area will not be too large and will not continue to expand. If the buildings that violate the rules and the streets and lanes are regarded as positive and negative spaces respectively, the buildings that form the relationship between the front and back of the streets and lanes are negative spaces, and the streets that are full of twists and turns are positive spaces. If they are extracted in the form of the relationship at the bottom of the figure, they will show the form of natural beauty, as shown in the following figure:

Fig. 2. Spatial characteristics of some traditional villages in Southeast Chongqing
2.2 Construction of prediction model for the change of spatial distribution pattern of buildings

From the perspective of public space, the sun dam space of traditional courtyard carries the function of activity, but the model rural construction loses the public space of communication, as shown in the following figure:

**Fig. 3.** New rural planning destroys the original spatial pattern

In the practical work of predicting the change of spatial distribution pattern of buildings, the multi model combination with mutual error correction function should be a means to obtain the ideal spatial analysis effect. The prediction process of building spatial distribution pattern change is shown in the following figure:
In the study of evolution of urban residential spatial distribution pattern, neural network method is used to predict the change of architectural spatial distribution pattern, and the linear distance between the regional center point and the residential space center point is solved by demand [4]. If the coordinate of demand point is \((x, y)\) and the coordinate of residential space center is \((k, l)\), the algorithm steps are as follows:

Step 1: randomly assign \(m\) living space;

Step 2: according to the principle of the nearest distance, find out the ownership relationship between the residential space and the demand point [5], expressed by binary variable \(a_{ik}\), and find out the total distance;

Step 3: compare the total distance calculated this time with the total distance calculated last time. If the difference between the two values is less than the error limit [6], stop the calculation. The calculation process is as follows:
Fig. 5. Calculation process of total building distance

Step 4: according to the following formula, solve the location of each living space in the next step to study the overall trend and difference of spatial correlation of observation variables in the whole study area, and the formula is as follows:

\[ I = \frac{s \cdot t(x_i - x^o)}{\sum o} \]  

(1)

In formula (1), \( n \) is the number of spatial units in the study area, \( t \) is the observation time, \( x_i \) and \( x^o \) are the observed values of architectural spatial units, \( \sum o \) is the spatial weight matrix, and \( S \) is the sum of spatial weight matrix.

Step 5: on this basis, reveal the heterogeneity between the attributes of different spatial units in the study area [7], and the calculation formula is:

\[ I = z_i \sum o / g \]  

(2)

In formula (2), \( z_i \sum o \) represents the standardized value of the observed value of the spatial element, and \( g \) is the spatial weight.

Step six: neural network is used to analyze the concentration degree of observation values in local space, which can further measure the spatial distribution of hot spots and cold spots

\[ G(o) = \frac{\sum q / x}{j} \]  

(3)

In formula (3), \( \sum q / x \) represents the space distribution calculation parameters of hot spot area and cold spot area, and \( j \) represents the original value of building space distribution.

According to the above process, the change prediction of building spatial distribution pattern [8] is completed, and the whole prediction model is shown in the following figure:
3 Planning and design strategy of traditional villages in Southeast Chongqing

On the basis of the above prediction of the spatial distribution pattern of buildings, in order to effectively plan the traditional villages in Southeast Chongqing, the planning and design strategies of the traditional villages in Southeast Chongqing are put forward. The characteristics of the southeast mountainous area of Chongqing make Tujia people have deep special feelings for the mountain. They advocate nature and think that the mountain is the spiritual existence in their mind. The mountain not only provides them with material means of production, but also the sustenance of their spiritual belief. Therefore, the planning of Tujia NEW village should respect the nature and develop spatial layout according to the mountain situation. The layout of commercial and residential mixed functions is shown in Figure 7:
In order to adapt to Tujia's deep feelings for mountains and create a familiar living space, the project divides the site into half slope area, central flat dam area and along block area according to the terrain characteristics, and launches different building combination layout for different terrain areas. First of all, along the hillside of Banpo District, the qiqikou setback building is arranged, and the setback courtyard group has a good view line and ecological environment; secondly, the central Pingba district is arranged with a combination of courtyard manual workshop group [9], which is a courtyard street space composed of a unit mode; there are also the commercial building group near the street, which uses the compound architecture It encloses a linear commercial street and market space. Through the building layout mode that matches the mountainous terrain, the rich and changeable building forms and levels are formed, and the Tujia mountain villages with aesthetic and regional characteristics are created, which is conducive to the combination of rural tourism and new rural areas, on the basis of building a beautiful village environment, improving the income of farmers, transforming living conditions, and thus the rural construction goal of "beauty in the farmhouse" is achieved.

According to the needs of different types of houses, three kinds of single building modules are formed in the planning, and the single group Village growing spatial pattern is adopted. The settlement pattern of Tujia traditional towns is flexible and changeable. The growing spatial texture is concave and convex, beautiful and conducive to creating a variety of living spaces. In order to restore the noumenon characteristics of Tujia villages and avoid "barracks style" The destruction of the traditional rural space by the urban pattern, this spatial planning is carried out on the basis of adjusting measures to local conditions, using the spatial design pattern of "single group Village" to continue the traditional growing texture. This method refers to the basic unit of Tujia residential "room" and the composition mode of unit assembly yard. This mode determines the building unit module according to the needs of different house types, and forms a dense, flexible and changeable village as a whole through a certain regular
combination of collages. The spatial mode of "single group Village" starts from the basic unit, and considers the integration of building base boundary and courtyard. According to the needs of housing type in the resettlement area, two or four basic units of different forms are determined, and then the unit evolves into a large group with spatial changes. The orderly growth of the group can develop into a continuous and changeable village space, which can be based on different items. The purpose of this paper is to develop the spatial layout flexibly and to be operable. The growing spatial texture of "single group courtyard" is a continuation of the traditional Tujia village form and a certain logical village spatial growth mode, which is of reference significance to the new village spatial planning. This kind of space technique is helpful to build a living place with traditional urban intention and living atmosphere, and bring the traditional rural aesthetic feeling and comfortable street dimension to the new village residents to become a harmonious and beautiful living space, so as to achieve the design goal of "the beauty of farmhouse". The planning scheme of traditional villages in Southeast Chongqing is as follows:
Reasonable functional layout can optimize industrial structure, effectively promote the development of industry and efficient use of space. Due to the severe employment test faced by ecological migrants, how to use space to create employment needs to be considered from the perspective of space function. Most of the resettlement areas lack of effective use of space, so it is difficult to meet the basic needs of the way of life and production of migrants. In order to achieve the design goal of "rich in rural areas" and realize the needs of residents' employment in the land through the study of spatial functions, the following are the specific design methods of functional spatial layout, mainly including the composite functional mode and functional space placement, as shown in the following figure:

(a) Project Space Streamline
The above figure shows the boundary reorganization process of the building function group. The farmer's market is located in the north of the commercial market. It is dominated by the e-commerce farmer's market, which can not only provide the place for farmers' trade, but also provide the employment space for the industrial transformation. The farmer's market can meet the purchase needs of residents' daily life. The e-commerce market realizes the local employment of some residents through the industrial transformation. As lijiaxi project is located in Qianjiang high tech Industrial Park, the surrounding area has the development advantages of Internet industry. Immigrants can take advantage of the industrial transformation and improve their skills and employment ability. The government can vigorously develop the Internet of things industry by virtue of the industrial advantages, establish and improve the circulation network, use the Internet, the Internet of things and other traditional industries to upgrade, and help migrant residents expand employment and participate in innovation Industry and so on. In terms of spatial planning, the large skylight shed formed by building enclosure can become the agricultural trade market, and the Internet of things industry can be distributed in the commercial buildings on the street according to the spatial advantages. In this way, the mixed industry model will also make full use of the space, and combine the agricultural trade market and the e-commerce space. The mixed industry model is a space with dual effects, and effectively realize the transformation and prosperity of
immigrants The design goal of the rich in the farm. So as to complete the planning and design of traditional villages in Southeast Chongqing.

4 Experimental comparison

In order to verify the validity of the prediction model based on neural network, the experimental comparison is made. In order to ensure the preciseness of the experiment, the designed model is compared with the traditional model, and the prediction accuracy of the two models is compared.

The building data of a certain place from 2011 to 2012 is taken as the experimental object, and the basic experimental data are shown in the table below:

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Sub option</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Village environment</td>
<td>Mountains and rivers, geology and geomorphology</td>
</tr>
<tr>
<td>2</td>
<td>Traditional village pattern</td>
<td>Village, pattern mechanism, important public space</td>
</tr>
<tr>
<td>3</td>
<td>Traditional building</td>
<td>Area, function, use situation</td>
</tr>
<tr>
<td>4</td>
<td>Historical environment</td>
<td>Production, fire protection, defense</td>
</tr>
<tr>
<td>5</td>
<td>Intangible Cultural</td>
<td>Production lifestyle</td>
</tr>
</tbody>
</table>

Take the above data as the experimental objects of the two models.

4.1 Experimental platform

NS2 is used to design the experimental platform, as shown in the figure below:
The experimental data are recorded and the corresponding experimental results are output.

4.2 Analysis of experimental results

The comparison results of prediction accuracy between the traditional model and the designed prediction model based on neural network are shown in the following figure:

Analysis of the above comparison results shows that the prediction accuracy of the designed model is higher than that of the traditional model, and the prediction accuracy of the traditional model is gradually declining. However, the overall prediction accuracy of the design model is high and the prediction is relatively stable, which can prove the validity of the design model based on neural network.
5 Concluding remarks

In this paper, a prediction model of building spatial distribution pattern change based on neural network is designed, and the validity of this design model is verified by experiments. However, there are still some deficiencies in the design model, which is limited in the research scope and depth of typical buildings, and can not fully cover all the traditional villages in Chongqing area, and the village research can not be fully in-depth, so it may cause the loss and deviation of village value indicators. These insufficient factors are the direction for further research in the future. In the future research, with the enrichment of data and in-depth research, it is hoped that the prediction model of building spatial distribution pattern change will be gradually improved, which will provide reference for the prediction of building spatial distribution pattern change in the whole southwest region and even the whole country.

6 Fund projects

Chongqing Municipal Education Commission Science and Technology Research Project: Spatial Features and Planning of Traditional Villages in Southeast Chongqing Research on design strategy (KJQN201904301).

References


196-202 (2018)


Sentiment Prediction for User Comments on Home Appliances Products

Ye Tao¹, Cao Shi¹*, Canhui Xu¹, Ruichun Hou², Zhifang Xu³

Qingdao University of Science and Technology¹, Qingdao, China
Ocean University of China², Qingdao, China
Haier Technology Co., Ltd.³, Qingdao, China

Abstract. Subjective information in products reviews play vital role in home appliances manufacturing industry. Generally, the comments are trusted worthy since we assume the customers will not make false ones. But in fact, there are cases that ratings and comments are not matched for some products. This paper proposed an approach to detect improper ratings by classifying and predicting the corresponding sentiment expressions in text reviews. To evaluate the effectiveness of the proposed method, we conducted experiments on a dataset which consists of customer reviews in 14 models of home appliances made by Haier. Results show that the sentiment polarity of the reviews can be predicted accurately, and the proposed method can be applied to detect and prevent a product from false rating.

Keywords: Sentiment Analysis, Product Review, Continuous Vector Model

1 Introduction

With the explosion of Chinese e-commerce, customer reviews have become a key factor in sentiment analysis, such as opinion mining and product recommendation. Reviews not only support customer decision making, but also enable manufacturers to know their customers preferences, and implement continuous iterative improvement [1].

In traditional home appliances manufacturing industry, communication between end-users and organizations often includes too much transferring and waiting, thus a simple feedback information exchange may become a complicated process. Haier has established an innovation platform that creates connection among user community, designers, suppliers and manufactures [2]. It collects the requests, challenges and recommend solutions that are posted and commented by end-users to certain products. Extraction subjective information in these large amount of text materials provides accurate understanding of sentiment expressions and may further bring tremendous business opportunities [3].

Most online shopping sites allow users to give rating (e.g. from 1 to 5) to indicate the degree of satisfaction for the product. In addition, users are often required to post comments regarding the products representing their opinion on different aspects of products. Typical consumer reviews contain valuable information about the quality of products, as well as additional recommendations. It has been studied that both ratings and comments carry different weight in decision making [4]. However, ratings and comments are not always emotionally consistent. For example, a user gives 5 stars to a product, but leaves an obvious negative comment corresponding to it. Sentiment analysis can be helpful to solve these unfair

* Corresponding author at: School of Information Science & Technology, Qingdao University of Science and Technology (LaoShan Campus), No.99, SongLing Road, LaoShan District, QingDao city, China, 266061. Mobile: +86 159 6420 3220.
judgments which are solely calculated by counting the number of likes/dislikes clicks, and therefore prevent a product from falsely gaining/losing its popularity [5].

In this paper, we focus on sentiment prediction on consumer reviews/comments in the domain of home appliance products. A web crawler is used to obtain the comments for product items of a certain category. Automatic sentiment classification approaches are explored and applied to help both manufactures and users analyzing the volume of information from the increasing amount of comments.

The rest of the paper is organized as follows. Section 2 gave a brief review of the state-of-the-art research in natural language processing, opinion mining, and sentiment classification. Section 3 introduced the proposed algorithms. Section 4 presented the datasets and discussed the experiments results, and Section 5 concluded the paper.

2 Related Works

One of the basic sentiment analysis is the binary classification problem, i.e. to classify a given review’s polarity as positive or negative. It plays a vital role in opinion mining for social network [6], document summarization [4], product reviews [7], and etc. Usually, this can be done by using either lexicon-based [8] or machine learning approaches [9]. The lexicon-based approach measures the polarity from lexicons derived from various resources which contains the typical semantic words or phrases [10]. Machine learning approaches apply a series of traditional machine learning algorithms on features extracted with vector space model (VSM) [11]. For example, Pang and et al. [12] demonstrated standard supervised methods on movies review data, and illustrated that support vector machine (SVM) provided more accuracy as compared to Naive Bayes and maximum entropy classification (ME).

Due to the complexity and ambiguity of sentiment expression, simple words and phrases cannot express the sentiment polarity accurately. Most traditional methods suffer from the high dimensionality and high sparsity problems, and therefore does not show good enough performances in sentiment classification. Recently, document representation methods such as Doc2vec[13] where the model is trained to represent a sentence or even a paragraph, have been introduced to solve sentiment classification problems from document level, and it shows higher performance than bag-of-words approaches [14]. It is also noted that deep learning technology has been shown to be effective in a wide range of natural language processing tasks [15].

In this work, continuous vectors [16] are exploited to represent online comments on appliances products from customers, and the SVM is used as the classifier system to increase the performance.

3 Continuous Vector Model for Comments Representation

Words or sentences representation is an important procedure in sentiment analysis. The principle underlies continuous vectors representation is that human beings can easily perceive the exact meaning of words in a review, whereas computer does not. Thus, effective comments representation models are required to distinguish different words.
Generally, a comment in text format can be divided into $N$ words, i.e. $w_1, w_2, \cdots, w_N$. Each $w_i$ is projected to different vectors, and the position in vectors of individual word is unique. These column vectors get together to form the word matrix, as is showed in Fig. 1.

![Fig. 1 Word matrix for comment representation.](image)

A neural network (NN) is employed to automatically predict a target word $w_i$ from its neighbors $w_{i-2}$, $w_{i-1}$, $w_{i+1}$, $w_{i+2}$. As shown in Fig. 2a, this NN consists of three layers: input layer, hidden layer and output layer. In practice, the word matrix that consists of all input candidates are used as inputs, and the column vector into which $w_i$ is projected becomes the output.

Define the input vector as:

$$\begin{bmatrix} x_1, x_2, \cdots, x_K \end{bmatrix}^T$$

which is a $K \times 1$ column vector determined by word. And the customer comment containing several words can be denoted as:

$$c = [w_1, w_2, L, w_N]^T.$$

The word matrix that contains all input vectors can be represented as:

$$[\cdots, v_j, v_{j+1}, \cdots]_{K \times M}, \quad 1 \leq j \leq M,$$

where $M$ is the given dimension of the vector. The weights connect input vector, hidden layer, and the output layer are defined respectively as:

$$\Omega_{K \times V} = [\omega_j^i]_{K \times V}, \quad \Omega'_{V \times K} = [\omega'_j]_{V \times K},$$

where $V$ is the number of the nodes in hidden layer. The input of a node in the hidden layer is calculated using the following formula:

$$\sum_{j=1}^{V} \omega_j^i x_j.$$

And outputs of hidden layer are worked out using sigmoid function:
\[ h_j = \frac{1}{1 + \exp(-\sum_{j=1}^{V} \omega_j x_i)}, \]

where \( h_j \) represent the whole hidden layer:

\[ H = [h_1, h_2, \cdots, h_V]^{T}. \]

The softmax function is used to get final output vector:

\[ y_l = \text{exp}\left(\sum_{j=1}^{V} \omega_j h_j\right) / \sum_{m=1}^{K_m} \left[\text{exp}\left(\sum_{j=1}^{V} \omega_j' h_j\right)\right], \]

where \( 1 \leq l \leq K \), and the output layer can be represented by:

\[ [y_1, y_2, \cdots, y_K]^{T}. \]

The cross-entropy function is exploited to evaluate error, and the stochastic gradient descent algorithm and back propagation algorithm is adopted to update \( \Omega_{K \times V} \) and \( \Omega_{V \times K}' \). Considering potential contribution of the entire comment, the comment ID information is added into NN, as shown in Fig.2b. Similar training process can be applied.

![Neural Network for continuous bag of word and the Comment ID.](image)

**Fig. 2** Neural network (NN) for continuous bag of word and the Comment ID.

### 4 Experiments and Discussions

We have collected 84000 items of online customer reviews from 14 models of refrigerators manufactured by Haier. The source of the product review data is crawled from www.jd.com which is a popular E-commerce website in China. As is shown in **Table.1**, all comments are semi-automatically mapped into three polarities, *i.e.* positive, neutral and negative, based on the corresponding ratings. Comments with 5 and 4.5 stars are classified as positive items, those with stars between 2.5 to 4.5 are mapped to neutral category, and the rest are considered as negative comments. To obtain a balance dataset, 2000 items are taken from
each polarity, and all these selected comment items are mixed together in a random order. It is helpful to balance positive and negative documents so that we focus on the efficiency of proposed framework rather than other factors which may affect the performance of classifiers.

The utility of *Doc2vec* was used to calculate the vector representations of comments by implementing the proposed NN. The polarity is used as the comment ID to learn the representation of each review. To test classification performance, all vectors are sent to SVM classifier, and results are evaluated based on the testing features.

**Table 1** The size and distribution of the dataset.

<table>
<thead>
<tr>
<th>No.</th>
<th>Type</th>
<th>Rating</th>
<th>Preprocessing</th>
<th>Polarity</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BC-93TMPF</td>
<td>5 Stars</td>
<td>Positive</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 Stars</td>
<td>Neutral</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 Stars</td>
<td>Neutral</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Stars</td>
<td>Negative</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Star</td>
<td>Negative</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>BCD-642WDVMU1</td>
<td>5 Stars</td>
<td>Positive</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>4 Stars</td>
<td>Neutral</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3 Stars</td>
<td>Neutral</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2 Stars</td>
<td>Negative</td>
<td>2000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 Star</td>
<td>Negative</td>
<td>2000</td>
<td></td>
</tr>
</tbody>
</table>

Among the 2000 comments items of each type of refrigerators, training set and validation set are divided with a ratio of 1:1. Commonly used indicators in sentiment analysis are used to evaluate the effectiveness of our proposed method, and the prediction accuracy can be defined as follows:

\[
\text{accuracy} = \frac{\text{tp} + \text{tne} + \text{tne}}{\text{tp} + \text{tne} + \text{tne} + \text{fpo} + \text{fne} + \text{fne}},
\]

where *tp*, *tne*, *tne*, *fpo*, *fne*, *fne* denote "true positive", "true neutral", "true negative", "false positive", "false neutral" and "false negative", respectively.

**Fig. 3** shows the accuracy of the classified reviews with different length of comment vectors. The sapphire dots are worked out according to Equation (11). The black line is the curve fitting of accuracy. It is supposed that the error of accuracy obeys Gaussian distribution, and then the lower and upper bounds are ±3%. It demonstrates that almost all sapphire dots fall into the area closed by two red dot lines.

Apparently, the accuracy increases fast when vector dimension increases from 20 to 100. That means the prediction accuracy improves as the vector dimension increases. However, when vector dimension is greater than 100, the accuracy does not show appreciable development.

The experimental results show that our proposed sentiment prediction model and the classifier succeed to classify more than 75% of the reviews correctly and can be effective to correct those mis-marked comments/reviews.
5 Conclusion

In this paper, we presented a study on classifying the product reviews using continuous vectors from balanced review datasets. SVM is applied as the classifier. Based on the experiment results, we believe that the proposed method can be helpful to identify and detect those fake reviews which are probably mismarked by users. We are still in the process of investigating the possibilities to integrate the sentiment analysis with other key modules of manufacturing industry, to enhance the performance of the value chain.

Acknowledgments. This research is supported by National Key Research and Development Plan (No. 2017YFB1400903).

References

Design of distance learning process monitoring system
based on 5G communication network

YU Bing-yan¹, XU Wei-Dong²

{YuBingyan3812@163.com¹, XuWeidong3812@163.com²}

(¹Jiangxi institute of applied science and technology, Nanchang 330100, china;
²Jiangxi institute of applied science and technology, Nanchang 330100, china)

Abstract: In view of the low stability of the conventional monitoring system, a remote learning process monitoring system based on 5G communication network is designed. Analyze the requirements of monitoring system, optimize flash memory circuit, select dts6677 data concentrator, use 5G communication network to transmit monitoring data and instruction information, use sigmoid function to calculate the error between expected value and output value, optimize function module, and realize the design of remote learning process monitoring system under 5G communication network. The simulation experiment tests the performance of the system. The test results show that applying 5G communication network to the design of the remote learning process monitoring system can effectively improve the stability of the monitoring system and optimize the system performance.

Keywords: 5G communication network; remote learning; monitoring system

1 Introduction

5G mobile network is the same as the early 2G, 3G and 4G mobile networks. 5G network is a digital cellular network. In this kind of network, the service areas covered by suppliers are divided into many small geographic areas called cellular. Analog signals representing sound and image are digitized in mobile phones, converted by analog-to-digital converters and transmitted as bitstreams. The main advantage of 5G network is that the data transmission rate is much higher than the previous cellular network, up to 10Gbit / s, faster than the current wired Internet, and 100 times faster than the previous 4G LTE cellular network (¹). Another advantage is lower network latency, less than 1 millisecond, compared to 30-70 milliseconds for 4G. 5G's performance objectives are high data rate, reduced latency, energy saving, cost reduction, increased system capacity and large-scale device connectivity. The first phase of the 5G specification in release-15 was to accommodate early commercial deployments. The second phase of release-16 will be completed in April 2020 and submitted to ITU as a candidate for imt-2020 technology. ITU imt-2020 requires a speed of up to 20 Gbit / s, which
can realize wide channel bandwidth and large capacity MIMO. The monitoring system of
distance learning process is designed by 5G communication network.

2 System design analysis

In distance learning, if the whole process of teaching is controlled by students, only a few
learners with rich knowledge and experience can achieve good learning results. For most
learners, self-control of learning pace, the number and difficulty of exercises often lead to
disappointing results. So if the initiative of control is transferred from computer to learners,
the interaction control will be unbalanced, which will eventually lead to the disappearance of
real interaction. Therefore, learning monitoring mechanism should be introduced in the design
of network courses to seek a balance between learners and computers, and give full play to the
teaching interaction function of network resources to maximize learning. Psychological
research shows that learning process self-control can be divided into two categories:
monitoring component and control component [2]. The former is the conscious evaluation of
the learning state being experienced, while the latter is the domination and domination of their
own learning process. Learning control is based on monitoring, consciously choosing learning
methods and actively adjusting learning process. In distance learning, the monitoring and
control of learning process interact and both are in dynamic process.

In the design process of learning behavior monitoring system, "monitoring the learning
behavior of learners" is the core, so we should make clear which learning behaviors of
students should be monitored, whether the behaviors are effective, whether they have an
impact on the learning results of students, and the results of these learning behaviors should be
quantifiable and measurable, and consider how to collect these learning behaviors, what
technical route to adopt, and whether the acquired behaviors are real and effective. After
obtaining these behavior data, statistical analysis shall be carried out. Through in-depth
analysis and mining of online learning behavior data, find out the internal relationship between
online learning behavior and learning effect, and realize the formative assessment of distance
learners. This fundamentally solves the problem that teachers evaluate students by subjective
experience in the network teaching environment, so as to truly evaluate students objectively,
fairly and fairly. This system starts from the monitoring of learning behavior to monitor
learning behavior, but monitoring and evaluation are inseparable, so "monitoring and
evaluation" work together to promote learners' learning [3]. The remote learning process
monitoring system designed by us introduces the monitoring, feedback, evaluation and
interaction mechanism designed by human, and has the functions of recording learners' learning situation, timely feedback learning effect and perfect communication mechanism. The
system involves the following core issues: first, establish a perfect learning behavior database,
track the learning process of learners, and collect the data related to learning behavior in this process. Whether the data collection is complete, correct and effective affects the effectiveness of monitoring and evaluation to a great extent. Therefore, behavior data collection must ensure the comprehensiveness and accuracy of data collection. The second is to use a stable technical route to scientifically collect, sort, extract, process and analyze the data in the learning behavior database, and provide a visual operation interface to realize the real-time statistical analysis of students' online learning behavior. The third is to find out the relationship between learning behavior and learning effect according to the collected historical data of learning behavior and the learning achievements that students have achieved, and to carry out formative assessment on learners to achieve personalized learning support services.

3 Hardware design of remote learning process monitoring system

According to the above analysis results, it is optimized on the basis of the original hardware structure.

3.1 Design of Flash Memory Circuit

Flash memory can keep data for a long time without current supply. It is a kind of non-volatile memory, which is fundamentally different from common memory\(^4\). The memory module is mainly composed of flash memory chip and FPGA minimum working system, and its circuit principle is as follows:
As designed in Figure 1, the circuit part of the memory module is set as a dual backup system to improve its viability and reliability. Receive the data and instruction information through FPGA, write and erase flash according to the system command, and feed back the storage status to the standard acquisition module to ensure the normal operation of the system in the next step.

3.2 Design of signal acquisition and processing equipment

In the system hardware of the monitoring system, the signal acquisition equipment is the basic hardware to carry out the monitoring work. The hardware includes the signal concentrator, acquisition terminal and communication terminal, and the signal concentrator and acquisition terminal are distributed in different monitoring units. The two connect through the communication terminal to realize the transmission and reception of the monitoring signal. The signal concentrator connects the monitoring terminal, computer, communicator and sensor through the network signal, in addition, the collected signal is transmitted to the data monitoring and processing unit of the next stage, which has the functions of signal acquisition, instruction transmission, password communication and time recording. Therefore, the power
module, processing module, communication module and monitoring module are connected with them at the same time, that is, the signal concentrator becomes the data signal transmission and receiving hub of each system hardware [5]. In this design, the selected data concentrator model is dts6677 LCD, and its parameters are shown in the table below:

<table>
<thead>
<tr>
<th>Table 1  Concentrator parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scope of application</td>
</tr>
<tr>
<td>major function</td>
</tr>
<tr>
<td>Accuracy grade</td>
</tr>
<tr>
<td>Communication protocol</td>
</tr>
<tr>
<td>Reference voltage</td>
</tr>
<tr>
<td>Reference current</td>
</tr>
<tr>
<td>Total power consumption</td>
</tr>
<tr>
<td>Outline size</td>
</tr>
<tr>
<td>Installation dimension</td>
</tr>
</tbody>
</table>

The collection of power supply data in the monitoring system mainly depends on the circuit of ATT7022CU. The circuit uses a special power metering chip with high precision to measure various parameters and read and convert data through A/D converter. The collected data is displayed through the main display board of the touch screen. The model of the main display board selected here is V30U2417H, which is connected with the drive control chip, power switch, circuit switch, etc. So far, the optimization design of the hardware part of the system has been completed.

4 Software design of remote learning process monitoring system

With the support of the above hardware structure, the software part of the remote learning process monitoring system is designed by using the 5G communication network.

4.1 Overall functional optimization

In order to ensure the security of the system and the consistency of the data, the data storage layer is used to monitor the storage of the data, and the related data of the fault point is transferred to the processing layer. The monitoring results are docked with the data of other parts through the docking layer, as follows:
With the support of the above system architecture, the 5G communication network is applied to the monitoring system.

4.2 Optimization of Monitoring Signal Processing Based on 5G Communication Network

Processing the monitoring signal through each hidden layer, transmitted to the output layer, will not reach the expected value of the output value of the reverse transmission, after the output value is infinitely close to the expected value, the output of the monitoring signal is realized.

If the sample data set of the input layer is set to \( \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \), its input data vector is:

\[
x = \{x_1, x_2, \ldots, x_n\}
\]

(1)

The input and output values of the hidden layer are calculated as follows:
Among them, sigmoid function has the properties of single increase and inverse function single increase, which is used for the output of hidden neurons. When the input value is \( x \), the output value of the neural network is calculated as follows:

\[
\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (2)
\]

Where, sigmoid function has the properties of single increase and inverse function single increase, which is used for the output of hidden neurons.

When the input value is \( x \), the output value of the neural network is calculated as follows:

\[
w_{r}^{i} (t+1) = w_{r}^{i} (t) + \Delta w_{r}^{i} (t) \quad (3)
\]

Where, \( c^r \) represents the two nodes connected, \( i \) represents the number of layers where the value is located, \( w_{r}^{i} \) represents the weight of the \( i \)-th layer, and \( t = 0 \), \( \Delta w_{r}^{i} \) represents the weight modification.

At that time, \( r = 1, 2, \ldots, n_{-1} \), we get:

\[
\Delta w_{r}^{i} = \mu \cdot \frac{\partial e}{\partial w_{r}^{i}} \quad (4)
\]

Where, \( \mu \) represents the learning step, and \( \frac{\partial e}{\partial w_{r}^{i}} \) represents the derivative of the error with respect to the weight.

The derivative value of the error of the output layer and the middle layer with respect to the weight is different, and the calculation process of the derivative value of the error of the output layer with respect to the weight is as follows:

\[
\alpha_{r}^{i} = y_{r} (1 - y_{r}) (d_{r} - y_{r}) \quad (5)
\]

Where, \( d_{r} \) represents the expected output value, and \( r = 1, 2, \ldots, \alpha \),

\[
y = [y_{1}, y_{2}, \ldots, y_{\alpha}]^{T}.
\]

Through the above calculation, the error between the expected value and the output value can be obtained, and the relationship between the learning step and the convergence rate can be found. So far, the processing optimization of monitoring signal based on 5G communication network is completed, and the system function module is designed on this basis.

### 4.3 Functional Module Design

The system has the following function modules: (1) course learning. Browse the learning course content according to the chapter directory of the course. The system sets record points
based on sections. After each section, the learner selects "complete learning" to record the learning time, learning path and other information [6]. (2) Class notes. This module is similar to the learning notes in traditional classroom learning. Learners can sort out and record the key points and difficulties in the learning process at any time, learning resource information related to knowledge points, or understanding of a knowledge point, etc. The system stores the learning note information in the database one by one, and each record is associated with the corresponding chapter, which is convenient for the learners to consult later. (3) Course interaction. Provide a public discussion area for all learners to discuss topics related to the course content, express problems or opinions, such as different understanding of a difficult knowledge point or the application of learning methods and learning strategies of the course. (4) Online exam. After learning a certain chapter, learners can take an online test to test their learning effect. The online test module supports the online test of learners. After submitting the test paper, the system can automatically score, give the test results, and record the results of each test [7]. The specific situation of its function module is as follows:

---

After logging in as a teacher, there are two functional modules: information management and user management. The function modules of information management are as follows: (1) Chapter management. This module realizes the organization, editing and modification of online course content by teachers. The content of online courses is pre-selected by teachers. After logging in the system, online editing is carried out based on chapters or sections to add, delete or modify the text and picture content of courses. (2) Course introduction. It is used to
describe the relevant information of the online course, such as the introduction of course content, relevant teaching materials, teachers and learning methods. (3) Class notes. Teachers can view learners' learning notes. Teachers can find and display the information recorded in the learning notes module here. Input the student number of the learner to be viewed, then the information of the learner's learning notes can be displayed in the form of a directory, with the title and time of each learning note written by the learner as the directory order, and the specific learning note content can be viewed. By checking the learning notes, the teacher can get a general understanding of the learners' understanding of the course content and some questions in the learning process, so as to give appropriate guidance and help [8]. To some extent, the content and time record of learning notes reflect the learners' emotion, attitude, method and other information of learning the course, which is also the reference for teachers to master the learners' learning situation. (4) Course interaction. Teachers can check the opinions and opinions expressed by learners here. By looking at the topics discussed between learners, teachers can understand the overall mastery of the course content and find out the problems, such as which questions are common problems, which need to be guided as a whole, which questions are the problems of individual learners, which need to be guided individually, etc. Similarly, this module provides a reference for teachers to understand the learners' feelings, attitudes, methods and other information of learning the course. (5) Online exam. This module supports teachers to organize online examination questions, check the examination situation of learners, etc. The test results of learners after learning a chapter are the important basis for teachers to master the learning effect of learners. Teachers can edit the test questions of each chapter online, check the test situation of learners, and adjust the difficulty of the test questions in time according to the test results, so as to obtain better learning effect. (6) Study time. In this module, teachers can query the learning time of each chapter. Enter the student ID of the learner to be viewed to display the time that the learner has completed each chapter and the time spent completing each chapter in a list. The learning time function module is the core module of the whole learning monitoring system and the main basis for monitoring learners' learning activities [9]. Through this module, teachers can understand whether the allocation of learning time is reasonable. Combined with the online examination and curriculum interaction of learners, teachers can objectively reflect whether the learning behavior of learners conforms to the learning law, how the learning efficiency and learning effect are, and what aspects should be given guidance and help.

User management module includes two parts: learner management and administrator management: (1) learner management. Used to find, add, modify or delete the learner account information. The login account of learners in the learning monitoring system is preset by the
teacher, that is, only the designated learners can log in to use the learning system, and other learners can't log in by themselves or wait for the approval of the administrator after registration, so as to avoid the use of irrelevant personnel at will, display a large number of useless data information, and disturb the monitoring management of teachers\textsuperscript{10}. (2) Administrator management. It is used by teachers to manage their administrator identity. You can also find, add, modify or delete administrator account information\textsuperscript{11}. When there are teachers and teaching assistants or multiple teaching teachers in a course, you can set administrator authority in this module.

5 System Function Testing

In order to verify the performance of the above system, a control experiment is proposed to simulate the running process of the designed system and get the test results.

5.1 Preparation process

The operating system of the simulation experimental computer is windows 2018 or Windows XP, 8g memory, hard disk of 500GB or more, tms320c6678 debugging version, 5V, 3A DC stabilized power supply is used for the hardware. The relay is used to control the power of the equipment in operation state, so as to ensure that the CPU power of the monitoring equipment is kept between 50W and 80W.

With the support of 40s68s37n03 chip shown in Figure 4, eight external charging posts are connected with Arduino single chip microcomputer, ieee802.15 data transmission module, ARM processor and CC2530 device in sequence. Each of the two posts corresponds to one system element, which is connected from top to bottom, left to right. C18, R11, C8 and U5 are four same directional device control switches, among which C18 and R11 switch directly control the connection and closing state of 40s68s37n03 chip, C8 and U5 are connected with external resistance, and whether the resistance is connected to the circuit is controlled by the same connection or closing state.
Fig. 4 Smart controller monitor circuit charging chip

Under the premise of ensuring the normal operation of the 40S68S37N03 chip, the C18, R11, C8, U5 switch is closed at the same time, and the external resistor is directly connected to the charging circuit, the experimental detection of the monitoring effect of the remote learning process under the 5G communication network is started.

5.2 Interpretation of result

The long-distance learning process monitoring system and conventional monitoring system under 5G communication network are respectively connected with the experimental acquisition device, which are the experimental group and the control group. Adjust the relay several times to control the electric power between 65W and 70W, and ensure the operation frequency of the monitor between 35Hz and 50Hz. After each experimental monitoring operation, the equipment is allowed to stand for 2 minutes to ensure the reliability of the test results. Seven connection ports of the optical access network are selected for four experiments, and the following data results are obtained.

<table>
<thead>
<tr>
<th>Port</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.96</td>
<td>0.95</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td>A2</td>
<td>0.94</td>
<td>0.95</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>A3</td>
<td>0.95</td>
<td>0.96</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td>A4</td>
<td>0.94</td>
<td>0.95</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>A5</td>
<td>0.97</td>
<td>0.94</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>A6</td>
<td>0.94</td>
<td>0.94</td>
<td>0.96</td>
<td>0.94</td>
</tr>
<tr>
<td>A7</td>
<td>0.95</td>
<td>0.95</td>
<td>0.95</td>
<td>0.94</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Port</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>0.91</td>
<td>0.85</td>
<td>0.84</td>
<td>0.88</td>
</tr>
<tr>
<td>A2</td>
<td>0.89</td>
<td>0.87</td>
<td>0.9</td>
<td>0.85</td>
</tr>
<tr>
<td>A3</td>
<td>0.9</td>
<td>0.91</td>
<td>0.88</td>
<td>0.85</td>
</tr>
<tr>
<td>A4</td>
<td>0.91</td>
<td>0.91</td>
<td>0.89</td>
<td>0.87</td>
</tr>
<tr>
<td>A5</td>
<td>0.85</td>
<td>0.83</td>
<td>0.87</td>
<td>0.85</td>
</tr>
<tr>
<td>A6</td>
<td>0.88</td>
<td>0.84</td>
<td>0.9</td>
<td>0.91</td>
</tr>
<tr>
<td>A7</td>
<td>0.9</td>
<td>0.85</td>
<td>0.87</td>
<td>0.88</td>
</tr>
</tbody>
</table>

Comparing the above two groups of results, under the premise of stable electric frequency voltage, the stability of the designed automatic monitoring system's received signal
reaches 0.94, and the maximum value reaches 0.97, the average stability result is 94.89%;
while in the control group's four experiments, the signal reception stability fluctuates greatly,
the maximum value is 0.91, the minimum value is 0.83, and the average stability test result is
87.64%. In contrast, the stability of the designed automatic monitoring system is 7.25% higher
than that of the conventional monitoring system. It can be seen that the designed system is
more stable and more in line with the actual monitoring requirements.

6 Conclusion

5G communication network is applied to the remote learning process monitoring system
to provide users with better service and better user experience.

References

Automation and Applications, 2019,38 (05): 79-81 + 91
Communications Technology, 2019,52 (02): 404-408
[4] CHEN Bo, CHEN Jianlong. Knowledge Discovery in Gesar Studies Based on Word Vector and
Distant Learning [J]. Journal of Tibet University, 2019,34 (02): 122-130
[5] LI Lvtui, YANG Yongxing, LIU Xu. Experience and Enlightenment of establishing remote learning
demonstration sites at Weizhou Island meteorological station [J]. Journal of Meteorological Research and
Application, 2018,39 (02): 138-141
[6] MENG Xianyu. Knowledge "Block" Combination Teaching Method: Design and Development of
Adult Distance Learning Textbooks [J]. Journal of Beijing Radio and TV University, 2019,24 (03): 55-62
[7] LI Shuang, LI Rongqin, YU Chen. Evaluation Model of Distance Student Engagement: Based on
LMS Data [J]. Education Research, 2018,24 (01): 91-102
[8] XIONG Sichang, ZHOU Yuntao. Design of Motion Images Monitoring System Based on Embedded
Linux [J]. Computer Measurement & Control, 2018,26 (04): 82-84 + 96
2018,21 (01): 143-146
Design of Education Management System Based on Wireless Self-organizing Network

LIU Xiao-Ling¹, LI Jun²
{LiuXiaoLing3812@163.com¹, LIjun3812@163.com²}

(1. Institute of Education and Physical Education, Jiangxi Institute of Applied Science and Technology, Nanchang 330100, China;
2. Institute of Education and Physical Education, Jiangxi Institute of Applied Science and Technology, Nanchang 330100, China)

Abstract: In view of the current imperfect construction of education management system in most colleges and universities, the design of education management system based on wireless self-organizing network architecture is proposed. On the basis of the original hardware structure, human-computer interaction design is added; S3C2440 processing chip and EPCS16SI16N storage chip are selected to ensure the normal operation of the system; data transmission bandwidth is reasonably allocated, and the wireless self-organizing network is used to optimize the login interface to realize the design of education management system based on the wireless self-organizing network architecture. The test results show that the original system can not meet the needs of users for knowledge competition and test management, and the designed system improves this defect, and its performance is obviously better than the original system.

Keywords: Wireless Self-organizing Network; Dynamic Routing; Educational Management;

1 Introduction

The traditional wireless cellular communication network needs the support of fixed network equipment, such as base station, for data forwarding and user service control[1]. However, wireless ad hoc network does not need the support of fixed equipment. Each node, i.e. the user terminal, organizes its own network. When communicating, other user nodes forward the data. This kind of network breaks through the geographical limitation of the traditional wireless cellular network, and can be deployed more quickly, conveniently and efficiently. It is suitable for some urgent communication needs, such as the battlefield individual communication system. In reference [2], aiming at the problem of resource optimization in self-organized heterogeneous networks, considering the energy efficiency and the stability of network connections, a scheme of educational resource allocation is proposed. The channel is included into the available channel spectrum pool, and the stability of network connection is guaranteed by the education user selection mechanism based on Morse potential.
energy. In order to improve the energy efficiency of the network, the power is allocated to the education users. In the selection process, the users are ranked according to the Morse potential energy value, and the users with lower potential energy value are preferred to provide services. The results show that the proposed scheme can save the data transmission power loss of education platform, ensure the stability of network connection, and optimize the energy efficiency. Reference [3] designed and built a wireless self-organizing network education management test platform system. By developing the underlying network card driver and routing algorithm protocol module, ordinary PC becomes an independent self-organizing network terminal. At the same time, through monitoring, statistics of data transmission between multiple educational terminals to obtain the measured performance indicators of the network. At present, the platform has implemented the test support for a variety of routing algorithm protocols.

However, wireless ad hoc network also has the disadvantages of limited network bandwidth, poor support for real-time services, and low security. Wireless ad hoc network is a multi-hop mobile peer-to-peer network composed of dozens to hundreds of nodes, which adopts wireless communication and dynamic networking. Its purpose is to transmit multimedia information flow with QoS requirements through dynamic routing and mobile management technology. Usually nodes have a continuous energy supply. It is applied to the design of education management system to optimize the system performance.

2 Hardware design of education management system based on wireless self-organizing network architecture

2.1 Design of human-computer interface

Through the human-computer interface, the communication between users and the system can be realized, and different operations can be performed. In the education management, the form of touch screen is used to select the input of the selected information according to the coordinate position of touch screen or press touch screen. Using touch human-computer interaction function in education management system makes management more convenient and visible. The industrial LCD LQ104VLDG52 is used to realize the display function, which has the advantages of high resolution and high data transmission speed. Its LCD driver module mainly uses programmable CPLD chip to realize the function of data bus and address bus in data exchange between LCD and driver module, so as to ensure high-speed and reliable data transmission. In the driver module, the RAM chip is used to realize the data buffer function in the process of data transmission, so as to prevent the data loss of education management data in the process of transmission.

The combination of industrial LCD and driver module can be used as an external device
to read and write relevant data. The timing diagram of the writing process of the driver module and the LCD is shown in Fig. 1.

**Fig. 1.** Sequence diagram of the writing process of the LCD screen and the driver module

In Figure 1, t1 represents the minimum effective time from the chip selection signal CS and address signal A0 of the drive module to the data line D0-D15 of the write process; t2 is the minimum time from data line to write signal WR during data writing; t3 represents the minimum time from WR, RD line to the end of CS, A0; t4 is the minimum time from the end of WR to the effective time of busy; t5 represents the maximum time of busy. The circuit connection structure of each part of human-computer interaction is shown in Figure 2.

**Fig. 2.** Circuit connection diagram of LCD screen

The three outputs of cal programmable logic make up the address line of LCD and driver module. After logic operation, the input data is connected to the output. The change of output
signal will not affect the input signal and effectively isolate the signal.

2.2 Data Processing Chip Selection

The S3C2440 processor produced by SamSung company based on ARM920T processor core has 5-level pipeline and Harvard structure, and its main frequency reaches 400MHz. It is a 32-bit microcontroller with high cost performance for communication field and system control, which has the following characteristics: (1) the standard working frequency is 400MHz, and the operation rate reaches 450mps. (2) Built in memory management unit and instruction data cache space. (3) On chip oscillator with two PLLs. (4) The chip contains two 16 bit timing/counters with three channels. (5) 117 external common I/O and 24 external interrupt sources. (6) External bus interface EBI supports SDRAM, nor flash and NAND flash. (7) Integrated LCD controller, support 64K color TFT, and a DMA channel dedicated to memory refresh. (8) Four universal synchronous different/step serial transceivers USART, serial peripheral interface SPI. (9) Sdhost supports 1.1 multimedia protocol. (10) Support 10/100M Ethernet interface. According to their own needs, users can choose the hardware devices suitable for the characteristics of the core plate and the bottom plate for welding. The physical figure of core board is shown in Figure 3.

![Fig. 3. S3C2440 Core Plate](image)

As shown in Figure 3, ① is a Samsung S3C2440 microprocessor chip with the main frequency reaching 400MHz, ② is the SDRAM memory pin interface. Users can customize suitable memory chips according to their needs, or weld them by themselves. ③ is the reserved NAND flash interface pin. The core board is surrounded by address bus, data bus, JTAG interface pins, GPIO pins of each group and touch screen pin interfaces. These interfaces and backplanes can correspond to the matching backplane slots one by one.

2.3 Circuit design

Due to the power supply for some precision devices, many devices are sensitive to mutation and rising edge and other variables, so the burr ripple should be as small as possible,
the load mutation should be as stable as possible, the heating should be as small as possible, the volume should be as small as possible, and the protection function should be complete. The devices used in this system are all small power devices, so using power frequency transformer and linear voltage regulator can completely meet the technical requirements of the system. As the basis of hardware system, power supply can only work normally, and subsequent work can be carried out. See Figure 4 for details of power supply design.

**Fig. 4. Power Hardware Circuit Diagram**

As shown in Figure 4, 3.3V voltage is used, while 5V voltage is generally provided. Here, a power adapter is used to convert 220V alternating current into 12V DC voltage. 12V voltage is converted into 5V voltage through lm2596s-5v chip, and then 5V voltage is connected to lm10851-3.3v chip to reduce voltage to 3.3V.

JTAG is a customized PCB and IC test protocol initiated by several international well-known electronic manufacturers. It is an international standard test protocol, mainly used for chip internal test, and also supports online programming of advanced logic devices such as FPGA and DSP [6]. Al standard protocol has four main signals: TMS for mode selection, TCK for clock, TDI for data input, and TDO for data output as shown in Figure 5.

**Fig. 5. JTAG hardware circuit diagram**
FPGA based on static memory will lose data information after power failure, so the system needs to reload the configuration file after each power up. In an independent system, the configuration file is usually included in the external flash memory. After power on, FPGA configures the configuration file by loading it into the internal memory. Based on the above reasons, it is necessary to select a memory to store the power on program. The system selects epcs16si16n as the power on program memory chip, as shown in Figure 6.

![Fig. 6. Circuit diagram of power-up program memory chip](image)

Based on the above, the software part of education management system based on wireless self-organizing network architecture is designed.

3 Software design of education management system based on wireless self-organizing network architecture

3.1 Distribution of data transmission bandwidth

In the application process of education management system, there are many people online at the same time. In order to ensure the normal operation of the system, determine the priority of the transmission rate, and allocate the network bandwidth resources to the appropriate users.

Set $\mu^f$ is used to represent the $\mu$ local user group in network $f$. The users of the local user group are registered users. For the terminals used by users to access the system, the transmission priority is determined by the decision factors of different education management services [8]. The priority decision factors are compared to form a matrix of $X \times X$, namely:

$$X = \begin{bmatrix}
\alpha_{11} & \alpha_{12} & \cdots & \alpha_{1x} \\
\alpha_{21} & \alpha_{22} & \cdots & \alpha_{2x} \\
\cdots & \cdots & \cdots & \cdots \\
\alpha_{x1} & \alpha_{x2} & \cdots & \alpha_{xx}
\end{bmatrix}$$

(1)

Where $\alpha$ represents different decision factors. The weighting vector of each decision
factor is $\hat{c} = [\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_s]$, which is the eigenvector corresponding to the maximum
eigenvalue of matrix $\chi$. According to this vector, the weighting value of the decision factor
can be obtained. In the education management system, the decision factors include throughput,
transmission rate, security and cost. Each decision factor compares with each other according
to the importance of the influence on the management service, obtains the weighted value of
the decision factor, judges the priority of the transmission rate according to the weighted
value, and forms the priority decision rules. Based on this, the network bandwidth resources
are allocated to different user terminals.

Then the optimal bandwidth allocation matrix is.

$$G = \frac{W_{\infty}(z, c)}{i(z, c)} \quad (2)$$

In the formula, $z$ represents the user terminal, $\nu$ represents the network transmission rate,
$i(z, c)$ represents the spectral efficiency of the $n$ bandwidth resource in the network $f$ to
the user terminal $z$ of the distance education management service. Due to the limited
bandwidth resources and the limitation of system capacity, the optimal broadband allocation
matrix is constrained by certain conditions, then the constraint conditions of the bandwidth
allocation matrix are.

$$G_{\min} \leq \sum_{f=1}^{m} \sum_{n=1}^{N} G \leq G_{\max} \quad (3)$$

Where $G_{\min}$ and $G_{\max}$ represent the actual maximum and minimum bandwidth
resources of the user terminal supporting the education management service, and meet the
needs of the actual education management service through the above constraints.

3.2 Design of Login Interface Based on Wireless Self-organizing Network Architecture

The functions of wireless ad hoc network management include performance
management, fault diagnosis, service discovery and topology discovery. If we adopt the
network management model based on policy and agent, the above network management
functions are driven by policy. Performance collection can be active or passive, such as event
response, on-demand request, etc\cite{9}. The performance analyzer is responsible for processing
the basic performance data obtained by the performance collector, and the performance
monitoring agent monitors the network nodes through the performance data base and the
performance event scheduler\cite{10}. The fault agent scheduler is responsible for generating, releasing, receiving and destroying the fault detection agent and fault diagnosis agent. Service discovery mainly includes service registration and service query. Topology discovery mainly includes topology report, topology collection, topology timing, etc. It is applied to the application design of login interface, as follows.

The education management subsystem will need to include all links and processes of teaching activities, and its management scope also includes management organization, administrative part, teaching management personnel and teachers, students, other teaching resources and courseware library, etc\cite{11}. A good education management subsystem can create a good learning environment and a personalized and intelligent learning atmosphere for students, so a good education management subsystem should include the following functions.

1. Student registration: provide student registration status to facilitate the management of students in the later stage, and ensure the access rights of students, and the access rights of students to teaching resources and courseware library.
2. Student status management: this function is mainly for the management of personal information, grades, course selection, examination, graduation and other aspects of students to ensure the integrity and timeliness of student status.
3. Course management: it mainly includes the establishment and adjustment of major setting, discipline and training plan, and at the same time, it is necessary to ensure the access of visitors within the allowed range, including the course selection of students, etc\cite{12}.
4. Management of teaching managers and teachers: to ensure the management of information about teaching managers and teachers, qualification of teaching managers and teachers, assessment and teaching of teaching managers and teachers.
5. Payment function: including the complete realization of the functions of examination, inquiry and statistics of students' tuition fees and examination fees.
6. Administrative document management function: it is convenient for teaching units to release documents related to teaching, including curriculum adjustment, course adjustment, examination information and other functions, and it is also necessary to ensure that these information can be modified and inquired.
7. System setting and maintenance functions: mainly realize the maintenance of the security and stability of the whole system, as well as the backup of relevant data and important data information, the integrity and consistency of data and important data.

User authorization and authentication is an important part of the system, which plays a key role in the security of the whole system\cite{13}. We set the browsing mode as the mode and any user with authority can log in to the system. The user needs to enter the user name and password, and the user authorization and authentication module will review and authenticate the user's qualification to check and determine whether the user is a legal user. If the
authentication is passed, the user group of the user should be found according to the user name to determine the permissions of the user and the operations that can be performed, and then return to the corresponding page to display the access permissions of the user. In order to ensure the realization of the system security requirements, the password and user authority values are transmitted in an encrypted way [14]. In order to facilitate flexible authorization, the system uses the way of connecting with the background database to store the corresponding user permissions and corresponding files one by one, so when the user logs in and issues a request for an operation, the system automatically compares the user information with the records saved in the database [15]. If the conditions are met, the requested operation can be allowed, otherwise the access request will be denied. In the login window of the client terminal, the validity of some data and important data is verified first, and then the validity of the user's identity is verified. If it passes the verification, the page corresponding to the user's permission will be returned. Otherwise, the error information will be displayed and the login window will be returned. Of course, the login here is divided into manager login and general user login. The manager is the operator who maintains the normal operation of the system, system security and timely update of information, and is one of the key points of the whole system, the source code of the manager login module is as follows.

```html
<!DOCTYPE html>
<html>
<head>
</head>
<body>

<?
$dir = 'username'  
$dir = 'password'
$username = trim('username'),  
$password = trim('password')
$sql = 'select * from admin where password="$password" and username="$username"'

$conn = mysql_connect() or die('Could not connect: ' . mysql_error());
$sql = mysql_query($sql) or die('SQL Error: ' . mysql_error());
$row = mysql_fetch_array($sql, MYSQL_ASSOC);
if ($row['username'] == $username and $row['password'] == $password) ....

</body>
</html>
```

Fig. 7. Administrator login module source code

In the process of user login, firstly, the user name and password entered by the client terminal are submitted to the server. Secondly, the existence of the user is searched in the background database and the matching of the password is compared. The existence of the user and the matching of the password must be satisfied at the same time. So far, the design of education management system based on wireless self-organizing network architecture has
been completed.

4 Management system function test

In order to verify the effectiveness of the education management system based on the wireless self-organizing network architecture, design experiments, test the basic functions of each module of the system, whether it can achieve the design requirements smoothly, and compare the results with the original system, complete the experiment test.

4.1 Testing process

In order to meet the needs of the experiment, the hardware environment of the test needs to use Intel HD 4000 or above graphics card, Intel Core i3 2.4 GHz or above processor, use more than 4GB memory, and the hard disk space is at least 20G or above. The software environment shall at least meet the requirements of Windows 7/8/10 64-bit or Mac OSX Sierra (10.12.6 +), and the browser version above IE6.0 shall be used. In the above experimental environment, using the Pingdom Tools, using its network and web page detection function, the module function detection of the education management system is completed. The specific operation interface is as follows.

Fig. 8. Operating interface of Pingdom Tools

In the operation process as shown in the figure above, input the relevant information data of a certain college, test whether the relevant functions of user management meet the expectations, and compare with the test results of the original management system to get the results and analyze them.

4.2 Analysis of test results

The designed system management test results obtained by Pingdom Tools simulation are compared with the management test results of the original system, as shown in the following figure.
In order to ensure the preciseness and reliability of the test, several simulations have been completed during the test. Through the above tests, it can be found that the original system can meet the user's needs in user management, teaching material management, score query management, online classroom management, course selection management, video playback management and system authority management, but in knowledge competition management and test. If the test results of these two items are blank, it indicates that the original system has problems in these two items and cannot meet the current user needs. The designed
management system makes up for this defect and improves the requirements of knowledge competition management and test management. It can be seen that the designed education management system based on wireless self-organizing network architecture can effectively improve some deficiencies of the original system and make the education management work more convenient and efficient.

5 Conclusions

In view of the problems of low efficiency and imperfect interaction in the construction of educational administration management system in most universities, this paper proposes a kind of educational administration management system based on wireless self-organizing network structure. Based on the original hardware structure, human-computer interaction design is added. Reasonable allocation of data transmission bandwidth, and use of wireless self-organizing network to optimize the login interface, realize the design of the educational administration management system based on wireless self-organizing network architecture. It has been proved that the application of wireless self-organizing network in the educational administration management system of colleges and universities can optimize the performance of the system and contribute to the educational administration management of colleges and universities.

References

Research on the management of multimedia online education based on the virtual architecture of network function

TANG Li Li¹,TANG Li²

¹²Jiangxi Institute of Applied Science and Technology, Nanchang 330100, China;

Abstract: The traditional multimedia online education management method can not effectively recommend the learning content that students need, which leads to poor learning effect and no obvious improvement of students' performance after learning. Therefore, a multimedia online education management method based on the virtual structure of network function is designed. Firstly, the online education resource learning recommendation model is constructed, the student attribute eigenvalues are extracted, and then the multimedia online education database is generated. Finally, the multimedia online education management system is established to complete the multimedia online education management based on the network function virtualization architecture. The experimental results show that the students' learning performance is higher than that of the traditional management method after learning the multimedia online education management method based on the network function virtualization architecture, which proves that the design of the multimedia online education management method based on the network function virtualization architecture can improve the students' learning effect.

Keywords: network function virtualization architecture; multimedia; online education; management; evaluation;

1 Introduction

With the development of Internet technology, the education mode has been upgraded gradually. In recent years, multimedia online education has developed rapidly. The characteristics and advantages of multimedia online education are two-way interaction, that is, through information technology to achieve the interaction and exchange between human and human, which can not only strengthen the communication between teachers and students, but also encourage and promote students' self-study and cooperation between students. Modern information technology provides multimedia online education with learning resources and learning environment that are more conducive to students' construction, multimedia online education breaks the traditional education model of "school attendance", overcomes the limitations of traditional education in time, space, age and economic conditions, pushes
education to the front of the educated, while students pull education resources to their front through the network, according to the needs of colleges and universities, majors, courses and teachers, make their own teaching plans and arrangements to make teaching and learning more flexible and personalized. The position of multimedia online education in higher education system has risen to the main force of mass education and life-long education, becoming the first choice for governments to realize life-long education, and the development of educational technology characterized by knowledge and intelligence, which is a new trend in the development of education in the world. Multimedia online education with various media as the means of communication has input a continuous stream of human blood and knowledge resources for the development of human society. This kind of education method is favored by governments and enterprises all over the world because of its flexible means, continuous tracking of modern scientific and technological achievements, and direct transformation of those cutting-edge science and technology with application prospects into productivity. The traditional online education management method does not fully consider the content that users need to learn, and the performance after management has not been significantly improved, therefore, it is of great significance to design a multimedia online education management method based on the virtual structure of network functions to help analyze the learning mode of online course users and develop the most appropriate teaching evaluation system.

2 Online education resource learning recommendation model

The design of multimedia online education management method based on network function virtualization architecture [1] mainly provides personalized courses for learners to improve the effect of education management. The mechanism of multimedia online education management is to judge the current students' learning needs through the virtual architecture of network functions, and to provide appropriate personalized learning resources and learning support services with the help of recommendation technology. According to the rules of online learning and personalized learning needs of users, the collection, analysis and judgment of individual characteristics of learners are the premise of the implementation of personalized learning system recommendation mechanism. In this paper, the user learning style is tested by using the style scale, and then the test results are recorded in the user model database. Learning resource database is mainly used to store learning related materials, including courseware, video, audio and other electronic resources. Teaching strategy database is used to store some process learning activity templates designed by teachers, such as collaborative learning, inquiry learning, project learning, etc. During the operation of the system, we need to use the recommendation technology to extract the required information from the database, and then through the analysis and transformation of the program algorithm, the content of the
resources that meet the current learners' individual needs will be presented in turn, so as to achieve the dynamic recommendation effect of resource personalization. The overall architecture is shown in Figure 1.

![Fig.1. Personalized learning architecture based on network functional virtualization architecture](image)

When learners enter the website for the first time, they need to fill in basic personal information to complete the registration, mainly including name, gender, login name, login password, professional background and other basic information. At the same time, under the guidance of the system, learners need to use the learning style test module to evaluate their own learning style type. The user application process is shown in Figure 2.

![User application process](image)
Fig.2. User Application Process

The learner modeling component processes the information and stores it in the user model database. In the process of learning, we can calculate and record the course resources learned by learners in real time by recorder, and update the learning record data at any time; At the same time, the resource recommendation mechanism needs to extract the current user's personal information from the user model database to determine the type of learning style, and use the recommendation algorithm to extract the dynamic recommendation of resources matching the user's learning style from the resource database and the teaching strategy database. The recommended process of collaborative filtering is as follows:

Step 1: fill in the missing value of the evaluation score matrix of the user knowledge object according to the requirements of the content filtering algorithm [3], and set the initialization score of the knowledge unit without evaluation to 0;

Step 2: build a user model based on the four dimensions of user's basic information, learning style, cognitive level and learning record, and calculate the similarity of the target user. The calculation formula is:

\[ f = \sum_c d \frac{m}{g} \]

In formula (1), \( f \) represents the user basic information, \( \sum_c d \) represents the user similarity calculation parameter, \( g \) represents the information for the website, \( m \) represents the learning basic content.

Step 3: Calculate the average score for all users as the cluster center. Calculate the distance between the current target user and the cluster center, and find out all other virtual user sets in the class to which the target user belongs.

Step 4: calculate the evaluation score of each knowledge object by calling the algorithm of knowledge content characteristic attribute value matrix. The calculation formula is:

\[ b = m \frac{s}{\sqrt{c/i}} \]

In formula (2), \( b \) represents the user to learn the knowledge content, \( m \) represents the user to measure the score calculation factor, \( s \) represents the knowledge content characteristic attribute, \( \sqrt{c/i} \) represents the knowledge object appraisal index.

Step 5: judge whether the target user conducts knowledge unit evaluation, and if the
result is yes, switch to the next step;

Step 6: calculate the similarity of all users in the target user according to the basic characteristics of the user (background information);

Step 7: calculate the similarity between the target user and all users in the website according to the score of knowledge unit evaluation;

Step 8: calculate the neighbor user set of the current target user according to the similarity ranking;

Step 9: according to the similarity ranking calculation, recommend the knowledge object list to the target user.

Learners use the auxiliary support service tools provided by the website (such as learning notebook, player, communication tools, etc.) for online personalized learning. At the end of learning, learners can learn about self-study through the unit test module provided by the website, and their test scores will also be recorded in the user model database in real time as the basis for judging learners' cognitive level when dynamically recommending resources. Recommendation technology is the core of realizing intelligent effect of personalized learning system, and it also embodies the recommendation idea of the system.

3 Realization of multimedia online education management in virtualization architecture

3.1 Student attribute eigenvalue extraction

On the basis of the above-mentioned online education resource learning recommendation model, for the multimedia online education management of the virtualization architecture, first extract the student attribute eigenvalue [4]. The traditional method of extracting the characteristic value of students' attributes is mainly based on the examination of students' achievements and whether the teachers have completed the teaching tasks. This evaluation method is relatively simple and does not fully consider the participation of the teaching subject, so it cannot accurately evaluate the teaching effect, let alone be used as the evaluation method of online courses. Online course is different from traditional classroom. Students who participate in online course leave a lot of data. By analyzing the user data of online course, the definition of student attribute is put forward, that is, the nature or behavior characteristics of students themselves or objectively. In short, it is the objective performance of students in a specific aspect. Based on the analysis of online course user data, this study summarizes the following student attributes and gives them characteristic values, which can be used in online learning modeling of teaching evaluation. It mainly includes the following aspects:

First, the student's gender, if the student's gender is male, the eigenvalue is 1; if the student's gender is female, the eigenvalue is 2;
Second, classroom activity, its classification and characteristic values are shown in Table 1.

### Table 1. Class Activity Characteristics

<table>
<thead>
<tr>
<th>Activity</th>
<th>Classification basis</th>
<th>Eigenvalues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly active</td>
<td>Class participation &gt; 200% 100%</td>
<td>1</td>
</tr>
<tr>
<td>More active</td>
<td>&lt;class participation &lt;200% 20%</td>
<td>2</td>
</tr>
<tr>
<td>Moderately active</td>
<td>&lt;class participation &lt;100% 0%</td>
<td>3</td>
</tr>
<tr>
<td>Low activity</td>
<td>&lt;Class participation &lt;20% Class</td>
<td>4</td>
</tr>
<tr>
<td>Inactive</td>
<td>participation = 0%</td>
<td>5</td>
</tr>
</tbody>
</table>

Third: Forum activity, its classification and eigenvalues are shown in Table 2.

### Table 2. Characteristics of Forum Activity

<table>
<thead>
<tr>
<th>Activity</th>
<th>Classification basis</th>
<th>Eigenvalues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highly active</td>
<td>Forum response rate &gt; 200% 100%</td>
<td>1</td>
</tr>
<tr>
<td>More active</td>
<td>&lt;forum response rate &lt;200% 20%</td>
<td>2</td>
</tr>
<tr>
<td>Moderately active</td>
<td>&lt;forum response rate &lt;100% 0%</td>
<td>3</td>
</tr>
<tr>
<td>Low activity</td>
<td>&lt;forum response rate &lt;20%</td>
<td>4</td>
</tr>
<tr>
<td>Inactive</td>
<td>Forum response rate = 0%</td>
<td>5</td>
</tr>
</tbody>
</table>

Fourth: job quality, its classification and eigenvalues are shown in Table 3.

### Table 3. Job Quality Characteristics

<table>
<thead>
<tr>
<th>Operating quality</th>
<th>Classification basis</th>
<th>Eigenvalues</th>
</tr>
</thead>
<tbody>
<tr>
<td>high quality</td>
<td>Timely price increase, excellent rate &gt; 90%</td>
<td>1</td>
</tr>
<tr>
<td>Higher quality</td>
<td>70% &lt;excellent rate &lt;90%</td>
<td>2</td>
</tr>
<tr>
<td>Medium quality</td>
<td>20% &lt;excellent rate &lt;70%</td>
<td>3</td>
</tr>
<tr>
<td>Low quality</td>
<td>0% &lt;excellent rate &lt;20%</td>
<td>4</td>
</tr>
<tr>
<td>No quality</td>
<td>Excellent rate = 0%</td>
<td>5</td>
</tr>
</tbody>
</table>

Fifth: the degree of autonomous learning. The degree of autonomous learning mainly determines the positive degree of student's pre-study and the degree of participation in review after class [5].

One of the after-class questions with the classroom activity and forum activity part of the coincidence, so no longer set properties. According to the above process, the extraction of eigenvalues of students'attributes is completed, which provides the basic significance for the multimedia online education management of virtualization architecture.

### 3.2 Multimedia online education database generation

In order to improve the speed of data response, multimedia online education database is designed. The database of this design is mainly composed of three parts. The first part is the
standardization of data processing, the second part is data mining analysis and integration [6], the third part is the database. The core of the middle tier is the second part, which is also the core of online education management, as shown in Figure 3.

In the face of rapidly increasing complex data, online education database uses network function virtualization architecture for modern data management, supporting all data types [7], such as files, pictures, videos, blogs, click streams and geospatial data, and the network function virtualization architecture is used to store the data in the data center, keep the data updated in real time, realize data sharing, analysis, discovery, integration and optimization, and improve the data value. Using the advantage of load balance, we can effectively and transparently expand the bandwidth of network devices and servers, increase the throughput of multimedia online education, strengthen the network data processing capacity of the platform, and improve the flexibility and availability of services. In the face of a large number of concurrent access or data flow of users, it can be shared to multiple devices for processing, reducing the waiting response time of teachers and learners; at the same time, parallel processing is done, and the processing results are summarized and returned to the online
interactive platform, which greatly improves the processing capacity of the platform system. Offline data is the information resource of various databases accessed by users, and it is the user access information and behavior information collected from the server, the client [8], and the proxy server. The big data technology is used to process the data and clear the unnecessary data. The clustering and classification algorithms are used to analyze the pattern of the processed data. The sample data resources are established to prepare for the data flow mining and analysis. Online data is due to the dynamic and large flow characteristics of data flow. When data flow mining is realized, the network function virtualization architecture is used to mine association rules, classification and clustering with less memory and faster processing speed. The structure of user registration module data table and the meaning of fields are shown in Table 4.

Table 4. Structure of User Registration Module Data Sheet and Significance of Fields

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Field Name</th>
<th>Chinese description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ID</td>
<td>Primary key</td>
</tr>
<tr>
<td>2</td>
<td>YH</td>
<td>username</td>
</tr>
<tr>
<td>3</td>
<td>XM</td>
<td>actual name</td>
</tr>
<tr>
<td>4</td>
<td>GK</td>
<td>Do you want to make your name</td>
</tr>
<tr>
<td>5</td>
<td>XB</td>
<td>public gender</td>
</tr>
<tr>
<td>6</td>
<td>MM</td>
<td>password</td>
</tr>
<tr>
<td>7</td>
<td>SFZH</td>
<td>identification number</td>
</tr>
<tr>
<td>8</td>
<td>EMAIL</td>
<td>E-MAIL</td>
</tr>
<tr>
<td>9</td>
<td>CSRQ</td>
<td>date of birth</td>
</tr>
<tr>
<td>10</td>
<td>LXDH</td>
<td>contact number</td>
</tr>
<tr>
<td>11</td>
<td>LXDD</td>
<td>contact address</td>
</tr>
<tr>
<td>12</td>
<td>YB</td>
<td>Postal code</td>
</tr>
<tr>
<td>13</td>
<td>WHCD</td>
<td>Educational level</td>
</tr>
<tr>
<td>14</td>
<td>RZDW</td>
<td>Working unit</td>
</tr>
<tr>
<td>15</td>
<td>ZW</td>
<td>Occupation</td>
</tr>
<tr>
<td>16</td>
<td>JKZK</td>
<td>Health status</td>
</tr>
<tr>
<td>17</td>
<td>GZSJ</td>
<td>time of participation in work</td>
</tr>
<tr>
<td>18</td>
<td>BYXX</td>
<td>Graduated school</td>
</tr>
<tr>
<td>19</td>
<td>SXZY</td>
<td>Majors studied</td>
</tr>
<tr>
<td>20</td>
<td>BZ</td>
<td>Remark</td>
</tr>
<tr>
<td>21</td>
<td>CJRQ</td>
<td>Creation date</td>
</tr>
<tr>
<td>22</td>
<td>ZT</td>
<td>status</td>
</tr>
</tbody>
</table>
Integration data is to take offline data as the reference of sample database, analyze online data, timely and effective feedback results, and update resource analysis results in time with the passage of time and the change of users' demand for information resources. Through the data mining process [9], we can filter, analyze and integrate data, establish multi-resource classification results, make decisions according to different needs of users, and form indexes to provide convenience for users to access and use services. The integration of data is mainly to prepare for the integration of users, analyze the similarity of users' information resources, classify similar users and allocate similar information resources. According to the four dimensions of service requirements of online learners, including basic information, learning style, learning satisfaction and learning perception, the customized service, personalized service and accurate service can be realized, which is convenient for users to extract the resources they need. Finally, according to the different authorization, the resources needed by users are standardized and uploaded to the server.

3.3 Multimedia Online Education Management System

By analyzing the attribute data of students in online courses, we can build a reasonable teaching management system, guide teachers to focus on the students who are not active and have poor learning autonomy, so as to help teachers to teach pertinently. Because there are many attributes of students, the eigenvalues also show diversified distribution characteristics, and the artificial sample analysis is time-consuming, laborious and low accuracy, this paper uses the network function virtualization architecture to build the multimedia online teaching management system. The specific method is to use a large number of student data as samples, extract the eigenvalue data and use the eigenvalue data as the learning data of the neural network. After the training of the network function virtualization architecture, it can automatically output the criteria for students' learning attention, so as to help build a reasonable teaching management system [10]. Because the model is a typical multi-parameter nonlinear model, the RBF network with good nonlinear approximation performance is selected, that is, radial basis function network. The network has three layers of structure: input layer, hidden layer and output layer. Combined with the analysis of the previous section, the eigenvalues of students' gender, classroom activity, forum activity, homework quality and autonomous learning degree are taken as the input of RBF network, and the attention degree of students is taken as the output of RBF network. The input neuron and output neuron of the established network are 5 and 1 respectively, and the network topology is shown in Figure 4.
It can be seen from Figure 4 that the network topology of the multimedia online teaching management system is mainly completed by courseware activities, independent learning, forum activities, etc. on the platform of three-tier architecture, the function module of customer management is realized by using programming technology, function business: online review management module, course dynamic management module, resource download management module, system login interface module Performance management and attendance management module, user password modification interface module, student information management module, teacher information management module, online teaching management module, etc. According to the above definition, the multimedia online education management based on network function virtualization architecture is completed.

4 Experiment

In order to verify the effectiveness of a multimedia online education management method based on network function virtualization architecture, and to ensure the rigor of the experiment, compare the traditional method with the management effect of the design method.
4.1 Experimental data

Randomly selected experimental data from a multimedia online education management platform, randomly selected 10 students as the experimental objects, and compared the results of students after two multimedia online education management methods.

In order to record the students' learning situation in real time, the experimental platform is set up, as shown in figure 5.

![Experimental Platform Diagram]

**Fig.5.** Experimental Platform

In the above experimental platform, the display shows the experimental situation, and the server provides network support for the experiment.

4.2 Analysis of experimental results

The traditional multimedia online teaching management method is compared with the multimedia online education management method based on the network function virtualization architecture designed in this paper. The comparison results of the learning achievements after the management are shown in Table 5.
### Table 5. Experimental results

<table>
<thead>
<tr>
<th>Student serial number</th>
<th>Traditional management methods (learning scores / points after management)</th>
<th>The design management method (learning score / point after management)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>58</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>45</td>
<td>98</td>
</tr>
<tr>
<td>3</td>
<td>78</td>
<td>96</td>
</tr>
<tr>
<td>4</td>
<td>58</td>
<td>95</td>
</tr>
<tr>
<td>5</td>
<td>63</td>
<td>93</td>
</tr>
<tr>
<td>6</td>
<td>67</td>
<td>98</td>
</tr>
<tr>
<td>7</td>
<td>75</td>
<td>97</td>
</tr>
<tr>
<td>8</td>
<td>77</td>
<td>96</td>
</tr>
<tr>
<td>9</td>
<td>80</td>
<td>93</td>
</tr>
<tr>
<td>10</td>
<td>75</td>
<td>94</td>
</tr>
</tbody>
</table>

Analysis of the above results shows that among the 10 students, the learning performance of the multimedia online education management method based on the network function virtualization architecture is significantly higher than that of the traditional management method. In the comparison of the results of the second student, the difference between the two management methods is the biggest, 53 points and 13 points. Therefore, through the above experiments, it can be proved that the multimedia online education management method based on the network function virtualization architecture can improve students' learning performance, meet the needs of multimedia online education management, and provide some help for other multimedia online education management.

### 5 Conclusion

In the current big data environment, the development mode of the education industry has changed dramatically. With online teaching, students can contact experts and scholars around the world, expand the scope of knowledge and expand their thinking. At present, the application of big data in the education industry is not comprehensive, and the attention of teachers and students is far from enough. Online teaching is only limited to homework and preview, which is a waste of educational resources. At the same time, after the traditional management method of multimedia online education, the learning achievement is low. Therefore, a management method of multimedia online education based on the virtual structure of network function is designed. The experiment shows that the students' performance after the design method management is higher than that after the traditional method management. Because the teaching management system can automatically analyze the
learning state of students according to the data information of students, so that students can be
classified according to the degree of attention they should receive. Teachers can carry out
targeted teaching according to this classification, so as to improve the teaching effect.

References
[1] Huang Aihua. Evolution, research and transformation of the teaching mode in higher education [J].
Standards" in the U.S. and the Evaluation Index System of Excellent Online Open Course in China [J].
[10] Zhou Zhicheng. The Dilemma in the Reform of Academic Evaluation System in Colleges and
Universities and the Deduction of Dynamic Equilibrium Model of Academic Utility [J]. Fudan Education
Access control method of multimedia distance education resource database based on Web Platform

LIANG Yan¹, WANG Fei-fei²

{LiangYan3263@163.com¹, WANGfeiF06@163.com²}

¹,²Jilin International Studies University, Changchun 130000, China;

Abstract. In view of the shortcomings of the traditional access control method of multimedia distance education resource database, which is slow to access due to the poor control effect, this paper proposes a web-based access control method of multimedia distance education resource database. According to the web platform to establish access control model, on this basis, through the role division, user management, role management, authority management, user role assignment management and role authority assignment management, the design of access control method of multimedia distance education resource database based on Web platform has been completed. Through the comparative experiment, compared with the traditional access control method of multimedia distance education resource database, the experimental results show that the access speed of the proposed access control method of multimedia distance education resource database based on Web platform is faster.

Keywords: Web platform; Database; Access control;

1 Introduction

With the development of Internet, resource sharing becomes more and more simple, and resource sharing platforms emerge in endlessly. However, the database security problems brought by resource sharing for resource sharing platform are increasingly serious. Access control technology is an effective way to protect database resources [1-3]. Multimedia distance education resource database is a typical resource sharing platform. The access control of database can restrict the access ability of users, and effectively prevent the invasion of illegal users or the misoperation of legitimate users from bringing adverse consequences to the database [4-6]. Security research has always been a hot topic in the field of computer science and technology. Most scholars focus on how to improve and optimize the authorization access control strategy, for example, from the traditional access control such as independent access control DAC and mandatory access control MAC to role-based access control strategy.
The current authorization access control can be roughly divided into two categories: database level security management, such as operating database level security management, database level security management, etc.; the other is application level security management, which mainly depends on the specific database [7]. Such authorization access control module is usually composed of user authorization database or data file, user authentication and authorization hard coding in application database. Every time a new application is established, the authorization access control part has to be redesigned; such mechanism is not conducive to the control under distributed environment, reduces the efficiency of application database design and development and database compatibility Sex and expansibility.

The rapid development of network technology has given birth to new network applications. Web based applications and clients are penetrating into traditional desktop applications. The increasingly rich network applications have put forward new challenges and requirements to the security control technology. At the same time, with the rapid development of Web services technology, software tends to exist in the form of services. Authorized access control technology based on Web Services provides a new idea for security design. Web service is recognized as an advantage technology to solve the integration of heterogeneous databases in the network environment. It provides good integration performance of heterogeneous database and cross platform. On this basis, combined with RBAC and web services technology, the access control method of multimedia distance learning resource base based on Web platform is designed. This method makes the authorized access control module easier to adapt to the security requirements of heterogeneous database, and easier to configure and maintain.

2 Access Control Method of Multimedia Remote Teaching Resource Database Based on Web Platform

2.1 Establish access control model

Based on the web platform, RBAC (role-based access control) is introduced into the concept of grouping to realize active authorization. It is not like RBAC, users can have access control rights to resources after granting roles to users, nor simply allow users belonging to a certain group to have full control over resources of their own group, but group them according to colleges and universities. Because resources and users of multimedia remote teaching resource database are from colleges and Universities, when colleges and universities are
grouped, the number of The database also groups users [8]. When permissions are assigned to the role, the permissions of users in this role will be limited to the group in which they belong, and the access to resources by users will be limited to the resources in this group.

The user permission information in the group and role-based access control model will be transferred through sessions. There are three kinds of sets in the model: user set \( u \), role set \( R \) and group set \( G \), as shown in Figure 1.

![Access Control Model Diagram Based on Grouping and Role](image)

**Fig. 1.** Access Control Model Diagram Based on Grouping and Role

When the user is authorized to enter the database, the information interaction between the session and the user \( u \) is carried out through the user session set. The grouping and role information of the active user stored in the session set are the activated grouping and role information, and the authority judgment can be effectively carried out in the database [9]. The group assignment between the group information of user \( u \) and group \( G \) uniquely determines the scope of user activity. The role assignment existing between the role information of user \( u \) and role \( r \) uniquely determines the user operation permission. Users authorized by groups and roles access resources, that is, access resources within their scope of activity according to their own permissions.

First, in the model, we build the relationship among groups, roles and users. According to the above-mentioned access control model based on grouping and role, combined with the actual situation of multimedia distance education resource database, the designed relationship between grouping, role and user is listed as follows. Because grouping is based on colleges or units, it also involves the subordination relationship between grouping and colleges. The database has four roles: database administrator, general administrator, teacher and student; each group has three roles: general administrator, teacher and student; the database administrator does not belong to any group (it can also be considered that the database administrator should not be assigned any role).
administrator belongs to all groups), the management group can be added and deleted, and any information (user information, courseware information) of any group can be seen. Information, courseware entity, resource statistics query report, etc.; Roles and users are one to many relationships. A user can only have one role, but a role can correspond to multiple users. After users register, they are assigned roles first, and then assigned groups; each user's corresponding roles and groups are one to one relationships; groups and colleges are one to many relationships, and each group has one or more colleges or units, one to many relationships. After a user belongs to a certain group, one or more institutions or units in the group should be redistributed. Ordinary administrators and teachers can assign one or more institutions under their jurisdiction; students can only assign one institution; Therefore, the relationship between users and institutions can be divided into two types according to different roles: the relationship between ordinary administrators and teachers and institutions is one to many; the relationship between students and institutions is one to one.

Secondly, in the model, the user status is designed as follows: the huge user group will inevitably bring the complexity of user account management. Even if the access control of users is limited by groups and roles, it cannot guarantee to find and solve problems in time when the database is running.

In the design of access control system, after user authentication and before authority judgment, the management of user status is added to enhance database security. Users can have three statuses in the database: the pending status of newly registered users, the enabled status with legal permissions, and the disabled status caused by illegal or misoperation. In the design of user status, setting all newly registered users does not immediately have access to the database, but needs to pass the administrator's audit. The administrator views the user information in detail during the audit process to ensure that the user is trustworthy in the database. At the same time, the administrator can also adjust the user's grouping and role information to assign reasonable rights to the user the limit is [10]. When users pass the audit, they can become legal users of the database. If the user's behavior in the database endangers the database security or information security, resulting in database exceptions, the administrator can immediately freeze the user's access rights by disabling the account, and retain the user's information in the database, and re-enable it after eliminating the user's threat. The function of user status is embodied by storing the corresponding status fields in the
database. Therefore, the basic identity information of each user logging in the database, in addition to his / her user ID, password, has his / her user status word.

Finally, in the model, the permission string is designed as follows: throughout the permission management of the whole database, due to the complexity of resources and user information, plus retrieval, statistics and other resource services, if users are grouped and divided into roles, more permission points need to be controlled. If the database users are grouped or the role changes, it will inevitably cause a lot of changes in permission points, and the frequent setting of permission system will have a lot of work. Therefore, in the design of user management module of the database, all the parts that need to be controlled by permission will be summarized and listed one by one in the way of permission limit points. In the database, the method of permission string is used, and "1" is used to represent With this permission, "0" means no permission, and a string corresponds to the permission range of a role to simplify the permission control process. As shown in Figure 2, it means that the user can browse resources, view resource statistics results, retrieve resources, download resources and publish feedback on resources.

![Permission String Diagram](image)

**Fig. 2.** Permission string diagram

The permission corresponding to each function page of the database will occupy one of the permission strings. When a user accesses the page, he can determine whether to allow the user to access the resource by reading the character bit of the page corresponding to the permission string in the user session information as 1 or 0. 1 indicates the legal user, and 0 indicates the illegal access. The emergence of the permission string also provides great
convenience for the role to assign permissions. The database administrator only needs to select the corresponding string for each role on the permission string setting page to complete the authorization of the role, which greatly simplifies the authorization process. Due to the existence of user groups, the assigned permission string can only take effect within the scope of the group to which the user belongs. Administrators don't need to worry about group access of users after permission assignment. After the establishment of the model, based on the web platform, the access to the multimedia remote teaching resource database is controlled, and the specific content is described as follows.

2.2 Role division control

Role classification is only a general division of roles, any information processing database is to complete certain task functions, so any role should be set according to the specific task requirements of the database, and the role should always perform certain tasks in the database. For this purpose, symbols are introduced to represent: \( A = \{a_1, a_2, \cdots, a_n\} \) role set. Among them, \( a_n \) is the specific role of multimedia distance education resource database.

\( T = \{t_1, t_2, \cdots, t_n\} \) task set. Among them, \( t_n \) represents a specific task in the database. Task set for role \( A^T \subseteq 2^T \). In the multimedia distance education resource database, \( t_i \) represents the entry of multimedia distance education resource item, \( t_j \) represents the modification of multimedia distance education item, and \( t_k \) represents the audit of database. If \( r^T_i = \{t_i, t_j\} \), it means that \( r_i \) is the role of resource recorder in the multimedia distance education resource database; if \( r^T_j = \{t_k\} \), it means that \( r_j \) is the role of resource auditor in the multimedia distance education resource database. The new role is added to the database by using the addrole (role) algorithm. The premise is that there are no new roles in the current role set. Enter the name of the newly added role; if the role is added successfully, the role ID will be returned; otherwise, null will be returned.

The specific implementation process is as follows:
Step 1: if there is a role with the same name as the newly added role in the current role set, you will be prompted "role already exists" and null will be returned, otherwise continue;

Step 2: add a new role in the current role set;

Step 3: update role information table roleInfo, that is, add related records in role information. RoleInfo;

Step 4: return the system generated role ID.

2.3 User management control

The users of the multimedia remote teaching resource database based on the web platform are registered through the network registration. The user objects are teachers, students, educational administrators, school administrators, teaching and scientific researchers and other different categories, which are filled in by the user when registering. The user classification actually belongs to the basic role type in the system. The filling in of user registration application is only the basis for the system to assign roles, which does not mean that users automatically have roles. Users should also fill in the name, gender, unit, discipline and other information when registering. The purpose of the system background user management module is to manage the user's content and design the user's add, delete, modify and other operations. The trust value of the user is calculated by formula (1):

\[
\text{TrV}=\text{initTrV}+\frac{\alpha \cdot N_{OB} + (-\beta) \cdot N_{RB} + (-\gamma) \cdot N_{DB}}{N_{OB} + N_{RB} + N_{DB}}
\]

(1)

In formula (1), \(\text{TrV}\) represents the user's trust value, \(\text{initTrV}\) represents the user's initial trust value, \(\alpha\), \(\beta\) and \(\gamma\) represent the trust evaluation weight, and \(N_{OB}\), \(N_{RB}\) and \(N_{DB}\) represent the total number of user behavior records with the behavior modes of \(OB\), \(RB\) and \(DB\) respectively. The control of user management is realized through the following steps.

Add user (user) algorithm is used to control user add operation. If there is no new user in the current user set, add a new user to the multimedia remote education resource database,
enter the new user name user; if the user is added successfully, the user ID will be returned, otherwise null will be returned.

The specific steps are as follows:

Step 1: if there is a user with the same name as the newly added user in the current user set, a prompt "user already exists" will be prompted, and null will be returned, otherwise continue;

Step 2: add new users to the current user set;

Step 3: update user information table userinfo, i.e. add relevant records in user information table userInf o;

Step 4: new users are mapped to empty sessions;

Step 5: return the system generated user ID.

The user addition in this database is added to the user information table userinfo after the user registration.

The deleteuser (user) algorithm is used to control user deletion. The premise is that the user to be deleted must belong to the current user set, and a user is deleted from RBAC system. Enter the name of the user to be deleted user; the output user deletion success returns true, otherwise returns false. The specific implementation steps are as follows:

Step 1: if there is no user to be deleted in the current user set, the prompt "user does not exist" will be prompted, and return false, otherwise continue;

Step 2: if the user is associated with an active session, delete the session (force close or exit)

Step 3: update the user assignment relationship (UA), delete all user assignment relationships of roles, that is, delete the user role assignment information table userroleinf. Relevant records in;

Step 4: delete the user from the current user set;

Step 5: update user information table userinfo, that is, delete relevant records in user information table userinfo;
Step 6: returns true.

Deleting a user usually includes the following situations: the user logs off actively, the user is forced to log off due to illegal operations, and the user is automatically deleted if he/she does not log on to the system within 90 consecutive days.

2.4 Role management control

The role of multimedia remote teaching resource database includes two basic role types: teacher and student. By inheriting the role of teacher, we can further divide the role types, such as education administrator, resource auditor, school administrator and teaching and scientific research personnel. These different roles have different permission sets. They have the rights to add, delete and modify roles. The content of roles in the database will change with the actual situation. Therefore, the purpose of role management module is to manage the content of roles and adjust the role information of users in time. Therefore, the operation of adding, deleting and modifying user roles is designed.

Using addrole (role) algorithm to control role management operation. If there is no new role in the current role set, add the new role to the multimedia remote education resource database, input the name of the new role, and output the role ID if the role is added successfully, otherwise null will be returned. The specific implementation process is as follows:

Step 1: if there is a role with the same name as the newly added role in the current role set, a prompt "role already exists" will be given, and null will be returned, otherwise continue;

Step 2: add a new role in the current role set;

Step 3: update role information table roleinfo, that is, add related records in role information table roleinfo;

Step 4: return the system generated role ID.

Use the deleterole (role) algorithm to delete the role (provided that the role to be deleted must belong to the current role set). Enter the name of the role to be deleted. If the output role is deleted successfully, return true, otherwise return false. The specific process is as follows:
Step 1: if there is no role to be deleted in the current role set, it will prompt "role does not exist", and return false, otherwise continue;

Step 2: if the role is associated with an active session, delete the session (force close or close after exit);

Step 3: update the user assignment relationship (UA), delete all user assignment relationships of roles, that is, delete the relevant records in the user role assignment information table userroie;

Step 4: update the permission assignment relationship (PA) and delete all the permission assignment relationships of the role, that is, delete the relevant records in the role permission assignment information table rolepermissioninfo;

Step 5: update roleInfo, i.e. delete relevant records in roleInfo;

Step 6: delete the role from the current role set;

Step 7: returns true.

2.5 Authority Management Control

The purpose is to manage the content of authority, that is, to manage the operation of object objects, such as the addition, deletion and modification of authority. The following is a list of some algorithms that implement permission operations in the permission management module.

Use registerpermission to control permissions. For the registered legal permission (associate the resource object with the corresponding operation), the precondition is that the new operation is not associated with the object and the object supports the current operation. Enter the object resource object name and operation name operation. If the output registration succeeds, the permission ID will be returned. Otherwise, null will be returned. The specific implementation steps are as follows:

Step 1: if the precondition is met or not, return false, otherwise continue;

Step 2: whether there is an object in the current object object set. If there is no object, execute addObject (object) and continue;
Step 3: add the current operation item in the object ACL, and the new item is not associated with any role;

Step 4: return the system generated permission ID.

Unregisterpermission (object, operation) algorithm is used to deregister the specified permission, provided that the operation to be deleted has been associated with the object.

Input object resource object name and operation name operation; output logoff success returns true, otherwise false. The specific implementation process is as follows:

Step 1: if the precondition is met or not, return false, otherwise continue;

Step 2: delete all related operation items in the ACL of the specified object;

Step 3: returns true.

2.6 User - Role Assignment Management Control

User role assignment management is mainly based on the functional requirements of the user role assignment information table userpermissionInfo for related operations.

The assignuser (user, role) algorithm is used to assign a role to a user, provided that the user belongs to the current user set, the role belongs to the current role set, and there is no assignment relationship between the user and the role. Enter the user name and role name role; if the output assignment succeeds, true will be returned; otherwise, false will be returned. The specific implementation process is as follows:

Step 1: if the precondition is met or not, return false, otherwise continue;

Step 2: update the user assignment relationship (UA) and add the user assignment relationship of the role, that is, add the relevant records in the user role assignment information table userroleInfo;

Step 3: returns true.

Use the deassignuser (user, role) algorithm to cancel the user to role assignment. The premise is that the user belongs to the current user set, the role belongs to the current role set, and the user and the role have an allocation relationship. Enter the user name use and role
name role; the output returns true if the assignment is cancelled successfully, otherwise false. The specific implementation process is as follows:

Step 1: whether the preconditions are met or not. If not, return false. Otherwise, continue;

Step 2: if the user and role are associated in an active session, delete the session (force close or close after exit);

Step 3: update the user assignment relationship (UA), delete the user assignment relationship of the role, that is, delete the relevant records in the user role assignment information table userroleinfo;

Step 4: returns true.

The user subset of the current role is obtained by using the assigned use: (role) algorithm. If the role belongs to the current role set, enter the role name role and output all user sets assigned to the specified role. The specific implementation process is as follows:

Step 1: whether the preconditions are met or not, if not, return null, otherwise continue;

Step 2: all users assigned to the specified role form a set;

Step 3: returns a pointer to hold the collection.

We use the assigned roles (user) algorithm to get a subset of the roles assigned by the current user, provided that the user belongs to the current user set. Enter the user name user, and output all the role collections assigned to the specified user. The specific implementation process is as follows:

Step 1: whether the preconditions are met or not, if not, return null, otherwise continue;

Step 2: all roles assigned to the specified user form a set;

Step 3: returns a pointer to hold the collection.

2.7 Role - Authority Assignment Management Control

The grant permission (permission, role) algorithm is used to assign permissions to roles. The precondition is that the execution role belongs to the current role set and the permission belongs to legal permission. Enter the permission name, authorization role name and role; if
the output assignment succeeds, true will be returned; otherwise, false will be returned. The specific implementation process is as follows:

   Step 1: if the precondition is met or not, return false, otherwise continue;

   Step 2: update the permission assignment relationship (PA), and add a new permission assignment relationship to the role, that is, in the role permission assignment information table rolepermissionif. Add relevant records in;

   Step 3: returns true.

Revokepermission (permission, role) algorithm is used to revoke the relevant permissions assigned to the role, provided that the executing role belongs to the current role set and the permissions have been assigned to the specified role. Enter the permission name to be revoked, and the authorization role name role; if the revocation is successful, true will be returned; otherwise, false will be returned. The specific implementation process is as follows:

   Step 1: if the precondition is met or not, return false, otherwise continue;

   Step 2: update the permission assignment relationship ((PA), and add a new permission assignment relationship for the role, that is, in the role permission assignment information table rolepermissionif. Delete relevant records in;

   Step 3: returns true.

Through the above content, based on the web platform, the access control of multimedia remote teaching resource database is realized.

3 Experiment

In order to verify whether the proposed control method has better control effect (faster access speed), a web-based access control method for multimedia distance education resource database is used.

3.1 Experimental process

First, the access control method is tested. The test environment is shown in Table 1.

Table 1 Test environment
<table>
<thead>
<tr>
<th>Types of section</th>
<th>Hardware environment</th>
<th>Software Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Client computer</td>
<td>Windows XP</td>
<td>CPU 2.0GHz</td>
</tr>
<tr>
<td></td>
<td>Internet Explorer 5.0 and above</td>
<td>256MB memory</td>
</tr>
<tr>
<td>Server computer</td>
<td>CPU 2.4GHz</td>
<td>Microsoft SQL Server 2005</td>
</tr>
<tr>
<td></td>
<td>1GB memory</td>
<td>IIS 6.0</td>
</tr>
</tbody>
</table>

In the above experimental environment, the access control method is tested. To test the logical relationship, we only need to test the functional logic of each module in the database which needs access control. Take the user status test as an example. There are three user statuses: enabled, disabled, and to be approved, corresponding to the status words "1", "0", "0".

The test process of user status is as follows. In each step of the process, you need to check the field changes of the test user's status word in the data table to complete the test of user status.

1. Register new user, expected status word: 0;
2. Audit the user as administrator "failed", expected status word: 0;
3. Approve the user as an administrator with expected status word: 1;
4. Modify the user grouping information as the user, expected status word: -1;
5. Confirm the group information modified by the user as an administrator. The expected status word is: 1;
6. Delete the user's group as administrator, expected status word: -1;
7. Re select the group as administrator for the user, expected status word: 1;
8. The user is actively disabled as an administrator, with the expected status word: -1;
9. Log off the user as the user, and the user no longer exists.
After testing, the user status word will switch between "1", "- 1" and "0" according to the operation of each step in the process, and meet the expected results. All the user status functions are realized. The modules of input authentication, user login, user registration, user audit, user modification, user status, my resources, user deletion, group management, role management, resource access are tested. In the test process, combine the design and implementation conditions to design various use cases for various possible permission exceptions. Repeat each test case for 2-3 times. If each test result is the same and meets the expected result, the realization of this function meets the design requirements. Otherwise, it cannot pass the test and does not meet the design requirements. At this time, the program shall be modified and tested again Try until the design requirements are met. After that, the paper compares the speed of access under the control of web-based access control method and traditional access control method.

3.2 Analysis of experimental results

The comparison results of database access speed under the control of web platform based multimedia distance education resource database access control method, traditional multimedia distance education resource database access control method 1 and traditional multimedia distance education resource database access control method 2 are shown in Figure 3.

![Figure 3: Access speed comparison results](image-url)
As shown in Figure 3, the average response time of the three access control methods increases with the number of concurrent access users. Among them, the average response time of traditional access control method 1 can be as long as 20ms when the number of concurrent access users is different; the average response time of traditional access control method 2 can be as long as 16ms when the number of concurrent access users is different; and the average response time of the proposed access control method is as short as 7ms. Through the analysis, it is found that the proposed access control method is based on the web platform, which greatly reduces the average response time and significantly improves the access speed.

4 Concluding remarks

In view of the shortcomings of the traditional access control method of multimedia distance education resource database, which is caused by the poor control effect and slow access speed, this paper proposes a web-based access control method of multimedia distance education resource database. Through the comparative experiment, compared with the traditional access control method of multimedia distance education resource database, the experimental results show that the access speed under the control of the proposed control method is faster, hoping that it can provide a certain reference value for the research of access control of multimedia distance education resource database.

5 Fund projects

This article is the phased research result of the demonstration virtual simulation experiment teaching project “Business Interdisciplinary Virtual Simulation Training” of Universities in Jilin Province (NO: JI JIAO GAO [2019]8)

References


Design of Online Education Resource Scheduling System Based on Multi-level SDN Architecture

TANG Li¹, TANG Li Li²
{TangLi3812@163.com¹, TangLiLi3812@163.com²}

(1. Jiangxi Institute of Applied Science and Technology, Nanchang 330100, China;
2. Jiangxi Institute of Applied Science and Technology, Nanchang 330100, China)

Abstract: In order to improve the time efficiency of online education resource scheduling system in resource scheduling, an online education resource scheduling system based on a multi-level SDN architecture is designed. The hardware part of the system designed this time mainly includes a controller and an expansion board; the software of the system in part, the online education resource scheduling based on the multi-level SDN architecture is realized through the construction of the online education resource database and the extraction of the associated dimension features of the resource information flow. The experimental results show that the online education resource scheduling based on the multi-level SDN architecture is designed. The system spends less time on resource scheduling than traditional systems, which proves that the system designed this time is more superior.

Keywords: multi-level SDN architecture; online education; resources; scheduling; equilibrium;

1 Introduction

In the process of global education information, streaming media technology has been widely used in the field of education. Modern distance education system based on network provides abundant information resources, friendly interaction performance and excellent openness. The transmission of multimedia files on the network has become an important technical problem in modern distance education system. The small size of streaming media, less resource consumption, short waiting time, multiple media types, low purchase cost and various compression rates provide effective solutions for modern distance education system. With the development of online education, the scale of online education data has become larger and larger, and the real-time requirement of online education data processing is higher and higher. The traditional online education resource scheduling system has a poor real-time performance in resource scheduling. Educational resource scheduling system. SDN architecture is a new software-based network architecture. Layer and forwarding layer, as well as centralized network state control, to achieve transparent multilevel control network strategy of the underlying infrastructure. The programmability of the network provides automatic management and unified control functions for the network, and effectively solves the current problems. The problems of poor flexibility, limited expansion of resources and diversification
of service types in networking. The system designed in this paper completes the design of
online educational resource scheduling system based on multi-layer SDN architecture from
hardware and software aspects. Practice has proved that the real-time system based on multi-
layer SDN architecture designed in this paper has better real-time performance than the
traditional system.

2 Online Education Resource Scheduling System Framework

The system designed this time is mainly divided into three layers, the system layer, the
application layer, and the data layer. It also includes an administrator login interface, teaching
material management interface, and member management interface. The main function of the
administrator login interface is to enter the user name and password, judgment, can modify, add
and delete the content of the system; teaching material management interface, which mainly
manages the content of tutoring materials, all materials can be displayed in full, and can also be
displayed by type; member management interface, administrators can delete, modify and lock
users, and can retrieve passwords for users who have forgotten their passwords. The framework
of the online education resource scheduling system based on the multi-layer SDN architecture
designed this time is shown in Figure 1:

![Framework of an online education resource scheduling system based on multi-layer SDN](image)

**Fig.1.** The framework of an online education resource scheduling system based on multi-layer SDN.
3 Hardware design of online education resource scheduling system

3.1 Controller design

The controller hardware can be divided into two parts: one is the USB control chip, and the other is the JTAG main control chip. The controller uses the USB control chip to implement the USB protocol to communicate with the PC. At the same time, by operating the JTAG main control chip, the data output on the JTAG bus conforms to the IEEE1149.1 standard, so as to achieve the purpose of controlling the JTAG bus by a PC. The EZ-USBFX2 chip CY7C68013 produced by the company is a USB interface control chip, and the CY7C68013 chip includes an enhanced 8051 processing Device, a serial port engine (SIE), a USB2.0 transceiver, 8.5kB on-chip RAM, 4kB FIFO Memory, and a universal programmable interface (gpif). Its gpif can be connected with any asic or dsp, it also supports all current universal bus standards. The USB interface chip of this system is used as an intermediary for data transmission, and it can be completed simultaneously with a portable computer. The exchange of control information and the execution of the control process. Transmission mode: Among the 4 transmission modes defined by the USB protocol, batch transmission and isochronous transmission are faster. However, batch transmission has error checking, which can ensure data transmission correctness. In this solution, because of the high requirements for the accuracy of the data, the batch transmission method is used. The 8051 running code is downloaded from the PC terminal. This way the system software modification and function upgrade are more flexible, and the external connection is also eliminated. rom, making the circuit more concise and reliable.

Select the company's boundary scan chip. The internal structure of the chip includes queue management module, host module, serial module, event manager, counter, command management and read-write bus.

3.2 Expansion board design

The w5100 network expansion board is used to provide communication functions for the innovative education network ecosystem. The expansion board integrates a 10/100 Ethernet controller to meet the system design requirements.

This chip supports hardware TCP/IP protocol, embedded 10 Base/100 BaseTX Ethernet physical layer, and supports auto answer, ADSL link, 4 independent ports, internal 16KB memory as TX/RX cache, working The voltage is 3.3V, and the I/O port can withstand 5V voltage. At the same time, it contains a variety of indicator lights, where L represents the programming indicator; LINK represents the network has been linked, this indicator is flashing when sending and receiving innovative education network data; RX will flash when the
3.3 processor design

ADSP-21161N is a powerful 32-bit floating-point DSP chip recently launched by ADI of the United States. It uses a super Harvard structure, has multiple internal buses, high-speed computing units, large-capacity memory, and flexible and diverse external interfaces. The core operating frequency can reach 100MHz, and the external bus operating frequency can reach 50MHz. Since it includes two sets of processing units inside, each set uses a three-stage pipeline structure for processing, so the operation processing speed can reach 600MIPS to achieve the DSP. The function of low operating frequency and high processing capacity can reduce power consumption. Large-capacity internal dual-port SRAM, the capacity can reach 1Mbit, divided into two storage areas, and one cycle can complete the access of instruction code and operands at the same time, and can be arbitrarily set. It can be 16-bit, 32-bit or 48-bit word width, which is convenient for different applications. The host (HOST) and multi-processor interface do not need external circuits, relying on the on-chip bus arbitration logic and DMA controller support, it can be easily Forms a tightly coupled parallel bus/shared memory parallel system. The on-chip SDRAM controller can directly manage SDRAM, and multiple DSPs can coordinate and use SDRAM well, so that an integrated processing system. Two sets of two-way high-speed LINK data transmission, each LINK port is supported by an independent DMA controller and send/receive data FIFO, which can perform high-speed data transmission up to 100MB/s, greatly improving Parallel processing capabilities can be used to form a loosely coupled distributed parallel system. In addition, there are communication ports such as SPI interface, programmable I/O pins (FLAG), and synchronous serial ports.

4 software design of online education resource scheduling system

4.1 Construction of Online Education Resources Database

Construct an online education resource scheduling database to reduce the time for online education resource scheduling. The distance education system includes 8 entity classes, student entity class, teacher entity class, system administrator entity class, course subject entity class, role entity class, test questions. The entity class and the problem entity class are as follows: First, the student entity class, attributes: student number, user name, password, real name, class, mailbox, landline, address, mobile phone, etc. The student entity attribute diagram is shown in Figure 2:
Second, the teacher entity class, attributes: work number, user name, password, real name, class, mailbox, fixed phone, address, mobile phone, etc. The attributes of the teacher entity are shown in Figure 3:

Third, the system administrator entity class, attributes: number, user name, password, real name, mailbox, fixed phone, address, mobile phone, etc. The system administrator's implementation attribute diagram is shown in Figure 4:
Fourth: Course subject entity class, attributes: number, course subject name, etc. The course subject entity attributes are shown below:

Fifth: role entity class, attributes: number, role name, etc. The role entity attribute diagram is shown below:

Sixth, the entity type of the test question, attributes: number, title, answer option a, answer option b, answer option c, answer option d, correct answer. The test entity attribute diagram is
shown in Figure 7:

![Test entity attribute map](image)

**Fig. 7. Test entity attribute map**

Seventh, the problem entity class, attributes: number, question title, answer analysis, etc. The problem entity attribute diagram is shown in Figure 8:

![Problem entity attribute map](image)

**Fig. 8. Problem entity attribute map**

The online education resource scheduling system stores education resources in accordance with the above database storage format to improve the speed of online education resource scheduling. The business flow chart of this database is shown in Figure 9:
4.2 Relevant dimension feature extraction of resource information flow

The traditional method uses the auto-correlation matching method of resource information for resource scheduling. When the interference in the data transmission channel is large and the prior data of the resource information flow is lacking, the balance of resource scheduling is not good, and the accuracy of registration is not good. In order to overcome the shortcomings of the traditional method, Gao designed an online education resource scheduling system based on a multi-level SDN architecture. The filter extracts the correlation dimension characteristics of the resource information flow, and firstly gives the function of information resource scheduling of open network large databases under cloud computing. The consumption is expressed as:

\[ FK = G \times G \frac{x}{M} \] (1)

In formula (1), \( FK \) represents a subspace of the features of the associated dimension, \( G \) represents resource filtering parameters, \( \frac{x}{M} \) represents power consumption parameters.

Under the condition of resource load balancing, the cloud resource information is screened to achieve information fusion, and the association dimension single frequency
characteristics of the open network large database information resources are obtained as follows:

\[ g_i = s_d j \prod f \tag{2} \]

In formula (2), \( g_i \) represents real vector which the information resources of the open network large database, \( s_d j \) is the information fusion scale in the correlation dimension feature extraction process, \( \prod f \) is resource information flow.

Based on this, the correlation dimension features of the resource information flow are extracted. The graph extraction process is shown in Figure 10:

![Diagram](image)

**Fig. 10.** Association dimension feature extraction process of resource information flow

The calculation formula is:
In formula (3), \( |\mathbf{s}| = \sum_{a} d / gh \) represents the time delay of resource scheduling, \( d \) is the scale factor, \( gh \) is the spectral registration coefficient, \( d \) is state transition parameters for resource scheduling.

As the time delay increases, the focus performance of the open network large database resource scheduling process improves. It can be seen that by extracting the associated dimension features of the resource information flow, the open network large database resource scheduling problem can be transformed into a multivariate unknown. Parametric channel adaptive equalization problem.

### 4.3 Online Education Resource Scheduling

According to the result of the feature extraction of the above resource information flow, resource load balancing control and channel adaptive equalization design are used to implement the open network large database resource scheduling. Under different interference intensities, the resource load balancing control function \( T \) for resource scheduling is obtained as:

\[
T = \frac{\mathbf{a}^*}{S \ast \mathbf{d}} \quad (4)
\]

In formula (4), \( T \) is the scale equilibrium coefficient, \( S \ast \mathbf{d} \) is scheduling the distribution elements for the open network large database resources, \( \mathbf{a}^* \) is evenly distributed time window.

Based on the above calculations, the upper and lower thresholds of the resource scheduling of the open network large database are:

\[
F = \frac{|d|}{S \ast \mathbf{A}} \quad (5)
\]

In formula (5), \( F \) represents the information flow sampling window delay, \( |d| \) represents the channel equalization coefficient, \( S \ast \mathbf{A} \) is continuous function vector.

Complete the online education resource scheduling based on the multi-layer SDN architecture according to the above process. The overall process is shown in Figure 11:
5 Experimental comparison

In order to verify the effectiveness of the multi-level SDN architecture-based online education resource scheduling system designed this time, an experimental comparison was performed. To ensure the rigor of the experiment, the traditional system was compared with the designed system, and the online education of the two systems was compared resource scheduling time.

5.1 Experimental platform construction

Set up a simulation platform based on Mininet. Mininet is a lightweight software-defined network test platform that can simulate a complete network host, link and switch on the same computer. The simulation platform supports custom complex topologies and can be on the same computer a complete network host, link and switch are simulated on the plane. The simulation platform supports custom complex topology, and the experimental platform is shown in Figure 12:
Use the above experimental platform for experiments. During the experiment, an online education platform was selected as the experimental object, and 7 experiments were performed, and the scheduling time of online education resources was compared with the 7 experimental processes.

5.2 Analysis of experimental results

The comparison result of traditional online education resource scheduling system based on multi-layer SDN architecture and traditional system online education resource scheduling time is shown in Figure 13:
Based on the analysis of the above experimental comparison results, it can be seen that the online education resource scheduling time of the system designed this time is shorter, and the overall scheduling time is less than that of the traditional scheduling system. The traditional online education resource scheduling system has a longer scheduling time, which is longer than that of the system designed this time. Therefore, the above experiments can prove that the system designed this time can reduce the scheduling time of online education resources and improve the real-time performance of online education resource scheduling system.

6 Conclusion

An online education resource scheduling system based on a multi-level SDN architecture is designed to solve the problem of long time for online education resource scheduling in traditional systems. It proves that the designed online education resource scheduling system improves the information registration ability of resource scheduling. The overhead is small, the technical indicators are superior, and have high application value. There are still some imperfections in this design. If you design the parallel scheduling algorithm of the teaching resource access scheduling system, then combine linear programming and numerical values. Analyzing the theory, the effect will be better. We hope that the online education resource scheduling system based on the multi-layer SDN architecture designed this time can provide some help for other online education resource scheduling.

References

European Alliance for Innovation

EAI is a non-profit organization with free membership and the largest open professional society for advancing research careers through community collaboration and fair recognition. Members benefit from finding feedback and mentorship for their work and they are guaranteed to be evaluated fairly, transparently, and objectively through community.

ISBN: 978-1-63190-271-0
ISSN: 2593-7642
http://eudl.eu/series/CORE| www.eai.eu