The Design of an Intelligent Monitoring System for Human Action

Xin Liang[1,2], Mingfeng Lu[1,2(✉)], Tairan Chen[1], Zhengliang Wu[3], and Fangzhou Yuan[1,2]

[1] School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
lumingfeng@bit.edu.cn
[2] Beijing Key Laboratory of Fractional Signals and Systems, Beijing 100081, China
[3] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

Abstract. Now the monitoring equipment such as cameras has been widely used in social life. In order to solve the problem that the current monitoring equipment relies on manual screening for the recognition of abnormal human action and is not time-efficient and automatic, an intelligent monitoring system for human action is designed in this paper. The system uses object detection, classification and interactive recognition algorithm in deep learning, combines 3D coordinate system transformation and attention mechanism model. It can recognize the local human hand actions, head pose and a variety of global human interaction actions in the current environment in real time and automatically, and judge whether they are abnormal or special actions. The system has high accuracy and high speed, and has been tested successfully in laboratory environment with good effect. It can also reduce labor costs, improve the efficiency of security monitoring, and provide help for solving urban security issues.

Keywords: Action recognition · Intelligent monitoring · Deep learning · Security issues

1 Introduction

In recent years, with the continuous development of economy and society, the deployment scope of surveillance cameras has become more and more intensive, covering all corners of the city and playing an important role in security and public security management. Security guards can use surveillance video captured and stored by cameras to detect dangerous action. The police can also use it as evidence in solving a case. However, this traditional method cannot give real-time and automatic warning of abnormal actions. It relies on repeated manual screening, which is troublesome and has no real-time capability. In order to solve this problem, this paper designs an intelligent monitoring system for human action, which can alarm the abnormal actions in the environment in real time and capture, deal with and record them in time. It

S. Shi et al. (Eds.): AICON 2020, LNICST 356, pp. 558–570, 2021.
https://doi.org/10.1007/978-3-030-69066-3_49

realizes the purpose of reducing labor cost, improving safety monitoring efficiency and reducing the probability of occurrence of hazards.

Now intelligent monitoring systems are already being used in transportation and agriculture, often to monitor the speed of cars on roads and the growth of crops. The main purpose of the intelligent monitoring system designed in this paper is to monitor abnormal actions of people in the environment, and its core technology is human action recognition. This technology obtains the human action data through the sensor, and intelligently recognizes the human action. Usually, human action signals can be characterized by images, motion sensors and environmental sensors. Different specific signals reflect different actions. The intelligent monitoring system mainly uses the human action recognition method of images, which is more convenient to obtain data and cheaper to buy equipment compared with the sensor method.

Human action recognition based on images is one of the basic problems in the field of computer vision and can be divided into three types according to different modeling methods. The first one is end-to-end human action recognition [1]. This method inputs the original image sequence information, extracts multiple features of space and time dimensions to construct the classifier, and finally outputs the action types of human in the image. The disadvantage is that it only works for single-player action analysis in small scenarios. The second one is the recognition of human skeleton pose, which is mostly used for multi-person pose estimation. This method generally estimates the skeleton pose of human first, and then classifies human actions, which can be divided into two ways: top-down and bottom-up. The top-down approach transforms the multi-person into the single-person pose estimation problem. It obtains k coordinates of key points of the human body by directly regressing the coordinates, or by calculating the expectation of the thermal diagram of each key point and taking the position with the highest probability. Then it uses the coordinate distribution for classification prediction. CPM [2] and HRNet [3] networks belong to this method. The disadvantage of the top-down approach is that the performance is related to the detection network and the running time increases with the number of people in the image. The bottom-up approach is to first detect the key points and then group them to get multiple body poses. The speed of this method can be realized in real time, and it is suitable for deployment in mobile terminal. OpenPose [4] and HigherHRNet [5] networks belong to this method. The disadvantage of action recognition based on human skeleton pose is that only the motion information of skeleton is taken into account without the image information, so the detailed interaction action can't be described. The third one is instance-centered human interaction recognition. This method defines the problem as a structure for human and object interaction, namely <human, verb, object> . It uses the target detection algorithm to detect people and objects respectively, and then learns the interaction between people and objects in the form of topological nodes or thermal maps by referring to the graph convolution method or the attention mechanism of NLP domain. Finally, it analyzes human action. InteractNet [6] and VSGNet [7] network belong to this method. Compared with the previous two methods, this method has higher recognition accuracy and is widely used in practical systems.

According to the different application scenarios, types of actions and functions realized, the intelligent monitoring system will be divided into three parts, namely Hand action recognition, Head pose estimation and Human-object interaction

detection. The data processed is the frame of the video shot by the camera. Hand action recognition mainly uses SSD [8] and OpenPose method to locate hand, and then uses SqueezeNet to classify and finally output hand action types [9]. Head pose estimation mainly uses SSD to locate the head, and then uses facial landmark-based classical method and image-based deep learning method [10] to identify the head angle. Human-object interaction detection uses two methods. One is to use Faster R-CNN [11] to identify the position of people and objects, and then use iCAN [12] network to estimate the interactive relationship. The other is the network based on YOLO structure improvement. It will be the location identification and classification at the same time.

2 Methods

2.1 Hand Action Recognition

We hope to protect the information security of special places such as confidentiality room, and real-time identification and alarm of abnormal actions of people within the monitoring scope to prevent information leakage. In general, the action of obtaining information, such as making phone calls, taking photos, operating computers and so on, is done by hand, so it is very important to obtain information about hand action. Hand action recognition mainly includes two steps: hand location and action classification. The location will use SSD and OpenPose methods.

SSD networks use a single deep neural network to detect objects. It applies the small convolution kernel to the feature map, predicts the type score and offset of a set of default boundary boxes, and generates the prediction of different proportion with different proportion of feature map to improve the detection accuracy. The SSD network generates several initial bounding boxes with different aspect ratios, and then returns to the correct truth value. In this way, the target position can be calculated once to improve the detection speed. SSD provides a unified framework for training and prediction, and it has high accuracy, high speed and good performance in detecting small targets.

OpenPose is action recognition based on human skeleton pose. It generates Part Confidence Maps for skeleton key point regression and Part Affinity Fields between skeleton key points according to the input image. By using the key points generated by CNN and the confidence mapping of the connection, the original problem is divided into several maximum power matching problems of bipartite graph to solve the problem that the key points are combined into human skeleton. Then the hand is located through the human skeleton. OpenPose is one of the most popular open source pose estimation algorithms with high precision and low computational complexity.

After the hand position is located, this area needs to be extracted and put into the classifier to detect the action type. SqeezeNet was chosen as the classifier because it can achieve AlexNet's accuracy in classification and reduce the size of the model. At last, SoftMax was used to calculate the probability of multiple classification problems, and the type with the highest probability was selected as the result output.

2.2 Head Pose Estimation

The direction of people's attention is mainly reflected by the direction of people's visual Angle, that is, the facial direction. When people focus their attention on dangerous areas for a long time, such as high-voltage cables and rivers without fences, the intelligent monitoring system can capture such abnormal action in real time and give early warning, which can be used as a prediction to reduce the risk of dangerous events.

The head action analysis algorithm consists of two parts: face target detection and head pose estimation. The location of the face is the location of the head. Since the face detection is also the recognition detection of small targets, the SSD algorithm mentioned in Sect. 2.1 is also used to obtain the face image. Then input it into the head pose estimation algorithm to estimate the head yaw, pitch and roll. In the head pose estimation part, landmark-based classical method and image-based deep learning method were respectively used to estimate the head pose. Since the coordinate system of the head pose estimation is the plane of the face, the judgment of attention needs to be converted into the world coordinate system. The system uses perspective transformation method to modify and finally get the angle range of human attention.

Classical method can be simulated through PerspectivenPoint(PnP) to solve the problem. N 3D points of the known object and their 2D projections in the image can be used to calculate the pose relationship between the object and the camera in the world coordinate system by using the internal and external parameters of the camera. In the process of 2D-3D matching and pose estimation, it involves the world coordinates system representing the 3D coordinate system of facial features, the camera coordinates system centering on the camera, and the image coordinate system using the internal parameters of the camera to project 3D points to the image plane. See Fig. 1.

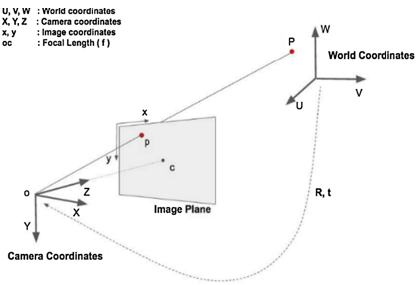

Fig. 1. Relationship among world coordinate system, camera coordinate system and image coordinate system

In order to calculate the 3D pose of the head, the system needs to obtain the 2D coordinates of 68 characteristic positions of the face, such as the tip of nose, canthus, chin and mouth corner. At the same time, the universal 3D face model is used to provide the coordinates of 3D face feature points to be matched. In order to better calculate the head pose, we use EPnP [13] algorithm. EPnP iteratively uses a set of virtual control points to represent the feature points in the world coordinate system as the weighted sum of the control points, rather than directly solving the depth of the

feature points. It uses the internal parameters of the camera to convert the coordinates of the reference point into the control point through formula transformation, and then solves the translation and rotation matrix [14], and converts the rotation matrix into euler angle. So we get the angle and the position of the head in world coordinates.

For the facial image-based deep learning method, we use FSA-Net network, which is a method to estimate the head pose with a single image. The FSA-Net uses two different branches to extract the features of the input image and conduct feature fusion at each stage. The fusion module first combines the two feature maps by element multiplication, then applies C 1×1 convolution to transform the feature map into C channel, and then uses average pooling to scale the feature map to make the size equal to the original image. Finally, we obtained the feature graph U_k of stage K. The feature graph U_k is a spatial grid, which is input into the attention mechanism module to obtain the weight of the feature graph, after which the more important parts of the feature can be highlighted.

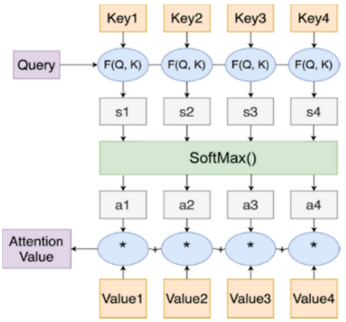

Fig. 2. Computational process of attention mechanism

Attention mechanism refers to a data processing method embedded in neural network which is inspired by human vision to selectively focus attention on important information and ignore irrelevant information. It can give different weights to the extracted features and improve the quality of the network. Attention mechanisms can be divided into four categories according to input types and processing methods, including item-based for processing sequences and location-based for processing feature graphs. Each category can be further divided into differentiable soft attention and non-differentiable hard attention [15]. The module in this system is a location-based soft attention mechanism. Figure 2 shows the computational process of attention mechanism. It inputs the similarity function F(Q,K) into the position Key_i in the feature and Query_i in the query result to calculate the weight s_i of the position, and gets the final attention weight value through the normalization processing of SoftMax function. It then weights the feature vector and the weight value to obtain the final attention value, and identifies the key parts of the image data with a higher weight coefficient. Because soft attention can be differentiated, it can participate in the process of learning and training and learn attention by itself.

After obtaining the feature map and its attention weight map, the fine-grained structure mapping is carried out to obtain the importance weighted feature. Fine-grained structure can focus attention on features that have a greater impact on facial posture, and reflects the spatial relationship between features. The robustness of attention mechanism can be improved by calculating weights such as learnable convolutional layer and unlearnable variance in spatial position. The weighted feature set was input into the feature aggregation method and the head pose was obtained by stepwise regression.

2.3 Human-Object Interaction Detection

When a scene requires detection of full-body action, or of multiple actions of the same person, these two approaches fail to meet the requirements. Therefore, we designed the detection of global human-object interaction action. The system abstracts the action into the interaction between human and objects, namely <human, verb, object>. The key is how to identify the interaction action between two targets. This system will use two different methods to realize human-object interaction detection, one is an iCAN network that combines the attention mechanism with the instance-centered and the other is an improved method based on YOLO structure that combines target detection and action recognition simultaneously.

ICAN network includes two parts, target detection and interactive recognition. After detecting the bounding box, class and probability values of people and objects, iCAN simultaneously input them as intermediate quantities along with the original image into the interactive detection network to detect interactive detections, and finally output the action type. The target detection algorithm adopts the Faster R-CNN network with high accuracy and uses COCO [16] data set for pre-training. The region proposal algorithm is mainly used to assist the location of objects. RPN is a full convolutional network that simultaneously predicts the boundary position and score of each object, and it can tell the whole network where to focus. Because it shares full image convolution feature with detection network, it can consume almost no computing resources.

Faster R-CNN network can be divided into four main steps. The first step is to take the VGG16 model as the backbone network structure, extract the feature map with the same size as the original image, and provide input for the region proposal network and classification network respectively. In the second step, RPN generates the candidate boundary box with the extracted feature map, and gets the more accurate position through regression. The third step is to pool the feature map and candidate areas after the above two steps, extract the features of local locations, and prepare for the fourth step. In the fourth step, the object type is classified by the full connection layer, and the boundary frame is regressed again to obtain more accurate boundary frame position and object type.

Figure 3 shows the overall structure of the iCAN network. Where b_h is the detected human bounding box, b_o is the object bounding box, s_h^a is the action prediction score based on human body, s_o^a is the action prediction score based on object, s_{sp}^a is the spatial relationship score between human and object, x_{inst}^h or x_{inst}^o is the appearance characteristics of a single person or object, and $x_{context}^h$ or $x_{context}^o$ is the context characteristics based on attention diagram. In the interactive detection part, iCAN use the attention module to evaluate all pairs of people and object boundary boxes to predict the interactive action score. This module is the attention mechanism module in Sect. 2.2. By learning to highlight key areas in the image dynamically, it can selectively summarize and identify features related to human interaction action. The image feature (x_{inst}^h or x_{inst}^o) and the convolution image of the whole image are mapped to 512 dimensional space, and the similarity is calculated by dot product. Then softmax function is used to get the instance-centered feature map. The overall image is weighted by the product of the feature map and combined with $x_{context}^h$ and $x_{context}^o$ extracted by the full connection layer for subsequent score calculation.

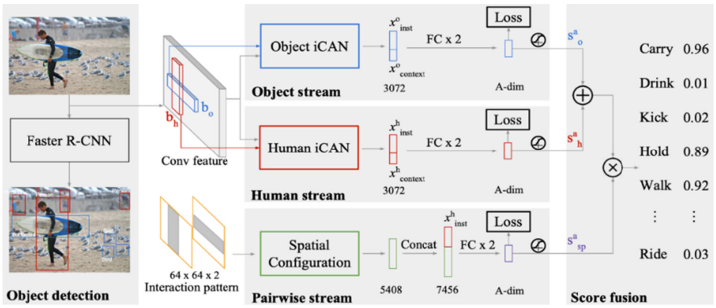

Fig. 3. Overall structure of the iCAN network

In order to solve the problem of ambiguity generated by different interaction relations of the same person and object, such as making phone calls and playing mobile phones, iCAN uses binary images to represent spatial interaction, that is, the images in the boundary box are 1, and the images in other positions are 0, and it uses the generated spatial features to judge together with the appearance features of the human body. The final action score of iCAN is calculated as follows:

$$s_{h,o}^a = s_h \cdot s_o \cdot \left(s_h^a + s_o^a\right) \cdot s_{sp}^a. \tag{1}$$

Where s_h and s_o are the confidence detected by a single object. For some action that do not include objects, such as walking, only human body branches are used to calculate $s_h \cdot s_h^a$.

An improved network based on YOLO can conduct target detection and action recognition at the same time. Its structure consists of three parts. The first part is feature extraction network. Inspired by YOLOv2 it is composed of 24 convolution layer, other convolution layer using batch normalization, the last layer using leaky rectified as linear activation layer. The second part is the two branches of target detection and

interaction detection. The input is extracted feature. The target detection network has the same structure as YOLO [17], and the output is the classification probability of the target and the corresponding detection box. The interaction detection network uses non-maximum suppression processing, and introduces an additional target detection branch to improve the positioning accuracy. The output is the probability of the interaction and the positioning boundary box of human and objects. The third part merges the results of the previous two branches and outputs the final result.

In the interaction detection branch, the input image is divided into N × N grids, and the grid where the interaction action center is located is responsible for detecting the interaction relationship. This step can provide the detection speed. Like YOLO, the network uses regression to predict bounding boxes. The design of the <human, verb, object> of human action in this system leads to the need to predict the boundary frame of human and object respectively. The loss function is calculated in the same way as the YOLO network. The network detection speed is fast and can meet the real-time requirement. But there is only one set of person-object matches, so it is impossible to identify a person's multiple action.

3 Experimental and Results

In order to verify the reliability, effectiveness and practicability of the system, an Ezio CS-C6HC-3B2WFR camera will be used to shoot video and intercept the action frame from the video for testing and evaluation. The training platform is the server with the Tesla V100 graphics card, and the reasoning platform is the computer with GTX1060 graphics card. The tests were divided into hand action recognition, head pose estimation and human-object interaction detection. The head pose estimation was tested on the 300 W-LP [18] data set using the classical method and the FSA-Net network. Human-object interaction detection was tested on V-COCO [19] and HICO-Det [20] data sets for iCAN networks and improved networks based on YOLO.

Hand action recognition captures nearly 8,000 action frames from the video captured by the camera and cuts them into 224 × 224 pixels as a data set. The classifier uses mixed samples for training. We used the Adam optimizer to train 100 epochs in 30 min until the loss stopped falling. At this time, the accuracy of the model on the training set was almost 100%. In order to make the hand action recognition of the system more reliable, we retain three most likely action types in the result section and give the confidence of each type. The result is a system that recognizes seven actions, including making a phone call, using the phone (other ways to use the phone besides making a phone call), opening a door, holding a file, using a keyboard, using a mouse, and others, and empty. And a threshold can be set to determine whether the abnormal action and alarm. Figure 4 is the test result of hand action. Both SSD and OpenPose target detection methods can correctly identify and classify hands. There is no significant difference in accuracy between them, but SSD is superior to OpenPose in recognition speed.

Fig. 4. Results of hand action test

Head pose estimation uses the pre-trained SSD face detection weight model to detect the face region and output the boundary box. The recognition results of SSD face detection are shown in Fig. 5 (a). The classical pose estimation method uses dlib library to extract and label the key points of 68 individual faces, and then matches the positions with the general 3D face model. The results are shown in Fig. 5 (b). Then it uses OpenCV's solvePnP function for head pose estimation. The input is characteristic point coordinates and camera internal parameters in world coordinate system and image coordinate system. The output is rotation translation vector between world coordinate system and camera. The head pose estimation results of the test figure are shown in Fig. 5 (c). The output angle value is [5.59276, 0.42750707, 4.0778937], which basically meets the expectation. The pose estimation of the FSA-Net network uses TensorFlow to describe the model. The 300 W-LP data set is a face data set that has been aligned with 68 key points and is expanded from the 300 W data set. Among them, 101,144 samples were used as the training set, and the remaining 21,306 samples were used as the test set. The Adam optimizer is used, and the final model weight is obtained by iterating 100 times. The same picture was used for testing, and the output angle value was [5.02749, 0.573317, 3.7282182]. It can be seen that the results of the deep learning method and the classic method are basically consistent and in line with expectations.

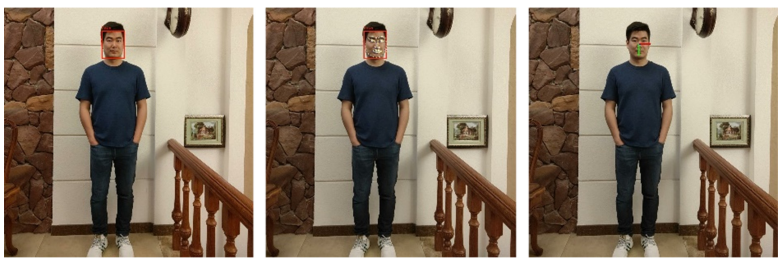

Fig. 5. (a) SSD face detection effect (b) Face feature point extraction effect (c) Head pose estimation result

The evaluation index of head pose estimation adopts the mean absolute error. Table 1 shows the test results of the two algorithms on the 300 W-LP data set. It can be concluded that the error of yaw angle differs by 1 degree, while there is a big difference between pitch and roll. The reason is that the classical method assumes that all faces are suitable for the general 3D face model, but this is not the case, which leads to a large error in the pitch and roll.

Table 1. Error comparison between FSA-Net and classical method, unit (degree)

	Yaw mean error	Pitch mean error	Roll mean error	Population mean error
Classic methods	5.92	11.86	8.27	8.68
FSA-Net	4.96	6.34	4.78	5.36

The instance-centered method in human interaction recognition uses Faster R-CNN network for target detection. The V-COCO data set includes 26 human actions. The HICO-DET data set includes 600 human actions, and it has modeled human actions as <human, verb, object>. The result of the target detection is fed into the iCAN network model to calculate the score for each type of action, and then the gradient is updated according to the loss function. Weights are backed up every 20 times in the network iteration to prevent all training progress from being lost due to server crash, memory overflow and other problems. The training process is shown in Fig. 6. Figure 7 is the recognition result of the test picture. The improved network based on YOLO adopts DarkNet network framework and adds a branch on the basis of YOLO network structure. The branch outputs the boundary frame of people and objects as well as the confidence of actions. The training process is shown in Fig. 8. The data set USES V-Coco data set. In order to speed up the training, we used the first 23 feature extraction layers of YOLO's pre-training model as the initial weights, and randomly generated the following network weights. The model also USES a weight backup every 20 iterations. Figure 9 shows the result of image recognition based on YOLO improved network.

On the V-COCO data set, we use the average precision (AP) [21] commonly used in target detection to evaluate the accuracy of the two algorithms. The resulting pair is shown in Table 2. It can be seen that the iCAN network has a high average accuracy, but because the iCAN network is a two-level recognition method that object detection and interactive recognition are carried out separately, the computing speed is relatively slow. In the process of interactive recognition, the three factors of people, objects and the spatial relationship between people and objects are considered to make the accuracy higher. YOLO improved network interaction and target detection at the same time, which is fast, but because the network is relatively simple, the prediction accuracy is worse than that of iCAN.

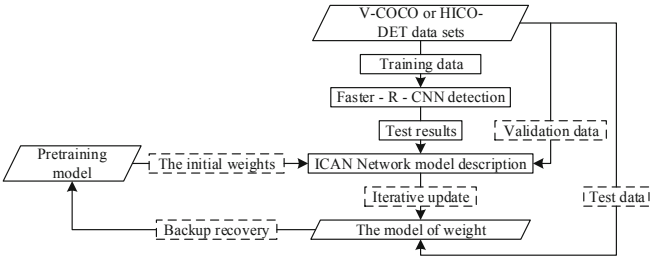

Fig. 6. ICAN action recognition network training process

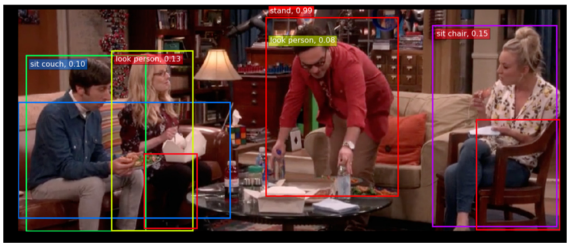

Fig. 7. Test image recognition results for the iCAN network model

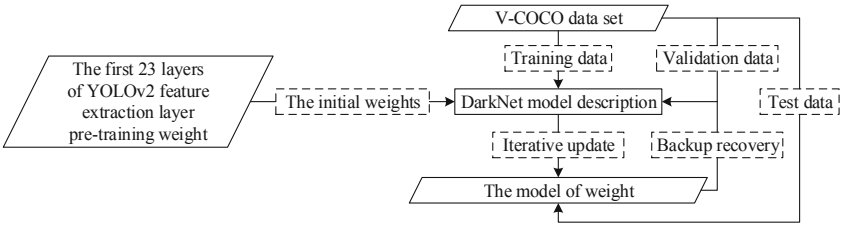

Fig. 8. Based on YOLO improved action recognition network training process

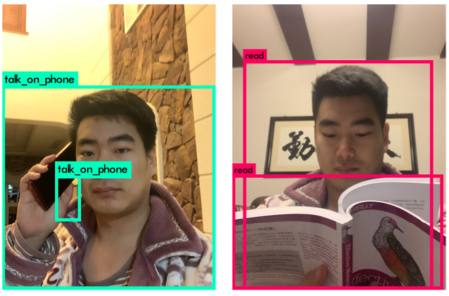

Fig. 9. Results of image recognition based on YOLO improved network

Table 2. Comparison of test results of human-object interaction detection algorithm

	Average precision(AP)	Mean operating time
iCAN	45.3	1.82 s
Improved network based on YOLO	35.6	0.028 s

4 Conclusion

The intelligent monitoring system realizes a variety of action recognition functions, including hand action recognition and head pose estimation for local human body, as well as global human-object interaction detection. The system is tested and compared using video from the camera in the laboratory environment. Hand action recognition using SSD and OpenPose hand positioning method, combined with SqeezeNet classifier, to recognize seven types of hand-related actions, including making phone call, using keyboard, using mouse, etc. It is suitable for simple indoor scenes such as secret rooms. Head pose estimation uses SSD to detect head position, and then uses classical method and FSA-Net network for deep learning to identify head angle respectively. It can output the head yaw, pitch and roll, so as to determine the direction of people's attention. It is suitable for monitoring and warning of dangerous areas such as high voltage cable. The method based on deep learning has higher accuracy, but the classical method has simple network structure and low requirements for equipment performance. Human-object interaction detection uses Faster R-CNN to detect objects and people, and then uses iCAN network and improved network based on YOLO to conduct interaction recognition respectively, so as to identify multiple action types of multiple people. It is suitable for multi-action recognition in complex scene. Experiments show that the accuracy of iCAN network is higher, but the improved network based on YOLO has a simple structure and short time-consuming which can meet real-time performance. According to different scenarios and application purposes, the intelligent monitoring system can use different methods to identify actions.

However, after many experiments, we found that the intelligent monitoring system has some problems. First of all, because action recognition requires object detection and classification of people and objects, there is a problem that relies heavily on objects and environment. Secondly, when people and objects are largely shielded, the detection effect is not good. And because the duration of different actions is different, the analysis with the same action recognition model has errors. In the next step, we plan to unify the advantages of each network, improve the practicability of the system, and try to use radar sensors to assist in identification to reduce dependence on the environment.

Acknowledgment. This work is supported by Beijing Natural Science Foundation (Grant no. L191004).

References

1. Ji, S., Xu, W., Yang, M., et al.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)
2. Wei, S., Ramakrishna, V., Kanade, T., et al.: Convolutional pose machines. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4732 (2016)
3. Sun, K., Xiao, B., Liu, D., et al.: Deep high-resolution representation learning for human pose estimation. In: CVPR (2019)
4. Cao, Z., Simon, T., Wei, S.E., et al.: Realtime Multi-person 2D pose estimation using part affinity fields. In: CVPR (2017)
5. Cheng, B., Xiao, B., Wang, J., et al.: HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. arXiv: 1908.10357 [cs.CV] (2019)
6. Gkioxari, G., Girshick, R., Dollár, P., et al.: Detecting and recognizing human-object interactions. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8359–8367 (2018)
7. Ulutan, O., Iftekhar, A., Manjunath, B.: VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions. ArXiv preprint arXiv:2003.05541 (2020)
8. Liu, W., Anguelov, D., Erhan, D., et al.: SSD: single shot multibox detector. Lecture Notes in Computer Science, pp. 21–37 (2016)
9. Wu, Z., Lu, M., Ji, C.: The design of an intelligent monitoring system for human hand behaviors. In: ACM International Conference Proceeding Series. ICMIP 2020-Proceedings of 2020 5th International Conference on Multimedia and Image Processing. 125–129 (2020)
10. Yang, T.Y., Chen, Y.T., Lin, Y.Y., et al.: FSA-net: learning fine-grained structure aggregation for head pose estimation from a single image. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
11. Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. arXiv: 1506.01497 [cs.CV] (2015)
12. Gao, C., Zou, Y., Huang, J.B.: ICAN: Instance-centric attention network for human-object interaction detection. In: British Machine Vision Conference (2018)
13. Lepetit, V., Moreno-Noguer, F., Fua, P.: EPnP: an accurate o(n) solution to the PnP problem. Int. J. Comput. Vis. **81**, 155–166 (2009)
14. Zhang, Z.: Iterative point matching for registration of freeform curves and surfaces. Int. J. Comput. Vis. **13**, 119–152 (1994)
15. Xu, K., Ba, J., Kiros, R., et al.: Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. arXiv: 1502.03044 [cs.LG] (2015)
16. Lin, T.Y., Maire, M., Belongie, S., et al.: Microsoft COCO: Common Objects in Context. arXiv: 1405.0312 [cs.CV] (2014)
17. Redmon, J., Farhadi, A.: YOLO9000: Better, Faster, Stronger. ArXiv preprint arXiv:1612.08242 (2016)
18. Zhu, X., Liu, X., Lei, Z., et al.: Face alignment in full pose range: a 3D total solution. IEEE Trans. Pattern Analy. Mach. Intell. **41**(1), 78–92 (2019)
19. Gupta, S., Malik, J.: Visual Semantic Role Labeling. ArXiv preprint arXiv:1505.04474 (2015)
20. Chao, Y.W., Liu, Y., Liu, X., et al.: Learning to Detect Human-Object Interactions. arXiv: 1702.05448 [cs.CV] (2017)
21. Everingham, M., Eslami, S.M.A., Van Gool, L., et al.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. **111**(1), 98–136 (2015)