Research and Development of Gesture Recognition Technology Based on WiFi

Xuhui Deng

{3440945297@qq.com }

Sino-British International College of University of Shanghai for Science and Technology, Shanghai, 200000, China

Abstract. As human-computer interaction continues to evolve towards a natural and non-contact mode, gesture recognition technology based on wifi has become a hot topic in the field of intelligent interaction due to its advantages such as relatively low cost, weak correlation with devices, and strong anti-interference ability. This article provides a systematic review of the core modules, typical solutions, key challenges and future applications of this technology. This study first compared traditional sensing technologies, highlighting the advantages of WiFi-based channel state information, namely CSI. It also provided an overview of key technologies such as CSI feature extraction and model design. Then, it investigated the principles, performance, and limitations of five representative methods. These five methods are HandFi, CSI-DeepNet, TransferSense, GESFI and C + SVM respectively. The solutions to challenges such as the cross-scenario generalization gap were summarized, and potential applications like hearing impairment assistance were explored. The research results show that the existing technologies have achieved high-precision and low-resource recognition, but there are some limitations in subtle gesture detection and cross-scenario adaptability. Future research should focus on refining fine-grained CSI feature extraction and exploring parametric-free domain adaptive methods.

Keywords: WiFi gesture recognition, Channel State Information (CSI), lightweight models, domain generalisation, contactless interaction.

1 Introduction

Gestures, as a natural and intuitive medium for human-machine interaction, transcend the limitations of traditional interfaces such as physical buttons and touchscreens. They hold irreplaceable value in scenarios including smart homes, assistive technologies for special needs populations, and industrial control systems. For instance, in smart homes, gestures can adjust lighting brightness and regulate air conditioning temperatures [1]; individuals with hearing impairments can utilise gesture recognition for real-time conversion between sign language and spoken text [2]; divers employ visual gestures to communicate with autonomous underwater vehicles [3]; and non-contact gesture-based page-turning is possible during meetings [4].

ICMEEA 2026, April 24-26, Singapore, Singapore

DOI 10.4108/eai.24-4-2026.2364872

Current gesture recognition technologies typically rely on visual, ultrasonic, or millimetre wave sensing, but suffer from shortcomings: visual sensing experiences light variation and privacy threats [2]; ultrasonic sensing is limited effective and vulnerable to blocking [5], while millimetres wave sensing has high hardware cost and deployment [6]. Gesture sensing finds amplitude and phase variations of the subcarrier of WiFi signals, using current WiFi networks, which can achieve low-cost, device independence, no light and obstruction, and wide coverage, making it preferred for large scale interactive applications.

Researchers both domestic and international have explored WiFi gesture recognition techniques in recent years, for example, CSI feature design, lightweight model design, and cross-scenario robustness. Sruthi P proposed the feature extraction method based on CSI ratio and interquartile distance (IQD) to correctly recognize left and right hand gestures [2], Kabir M H proposed the lightweight deep learning model CSI-DeepNet (e.g., for edge devices) [7], Bu Q and Zhang X explored transfer learning and domain generalization techniques to solve small-sample cross- scenario recognition problems [8,9].

The essay will first include technical aspects, typical solutions, challenges and applications. First, we will outline key technologies (e.g., CSI feature extraction and model design) and analyze the performance, benefits, disadvantages and case studies of five typical solutions. Then we will address key challenges and solutions for solving problems and some emerging applications. Finally we will summarize its limitations and future directions.

2 Typical Technical Analysis

2.1 HandFi

The team led by Sruthi P at the University of Hyderabad, India, has proposed a gesture recognition technology named HandFi to address the issue of feature distortion caused by hardware noise interference in WiFi-based gesture recognition. This technology establishes a three-tier processing framework comprising CSI ratio denoising, IQD feature selection, and KNN classification. Its core objective is to achieve high-precision recognition of both left- and right-handed gestures through lightweight processing.

If CSI data for the Intel 5300 network interface card has been collected by averaging CSI values across antennas of the same receiver, than can filter out hardware-inherent noise like carrier frequency offset (CFO), sampling frequency offset (SFO) and other hardware-infherent noise. To avoid unwanted features we select the top 30 interquartile distances (IQDs) of the extracted CSI ratios as core features, keeping CSI dynamic variations caused by hand gestures as unaffected, and filtering for static noise. A KNN classifier with K=3 is also used to reduce model complexity to make it more suitable for low-computational-power devices.

Among a total of 12 classification tasks including 6 left-handed gestures and 6 right-handed gestures, the accuracy rate obtained from the test reached 98.68%. Among the two types of tasks distinguishing left and right hands, the accuracy rate reached 99.7%. This result represents the highest level that single-antenna WiFi gesture recognition can achieve during the same period, and it does not require GPU acceleration. The CPU inference latency is less than 1ms, and it can respond in real time to human gesture operations. This system was successfully deployed on the hardware platform of TP-Link 1750 AC router plus Intel 5300 network card, which verified that it is feasible to reuse the existing WiFi equipment without adding additional sensor

hardware. It has many main advantages, such as a very prominent hardware noise suppression effect. Compared with the pure CSI amplitude characteristics, the noise is reduced by 80%, and only a basic Python environment is required, with a very low deployment cost. However, it also has its drawbacks. It can only be used in one scenario. When the environment is changed from a laboratory to a home bedroom, its accuracy rate drops to 62%. Moreover, it cannot recognize continuous gestures. For instance, when performing circular motion, double-clicking a gesture cannot be recognized. It is mainly suitable for simple gesture control in a single scenario, such as turning on and off smart lights Or adjust the speed of the air conditioner fan [2].

2.2 CSI-DeepNet

The team led by Kabir M H from Asia University in South Korea has developed a lightweight deep learning technology called CSI-DeepNet to address the challenges brought by traditional deep learning models, which are affected by excessive parameters. This makes deployment on edge devices like ESP32 very challenging. The CSI-DeepNet technology achieves edge adaptation of the gesture recognition model by optimizing the structure, and also keeps the model with relatively high recognition accuracy.

The architecture it adopts is a Conv2D initialization layer plus four DS-Conv blocks, and finally a GAP classification layer. Among them, two DS-Conv blocks contain the Feature Attention block, that is, the FA block. And Residual blocks, that is, RB blocks, are used to balance lightweight design and feature extraction capabilities. To address the computational limitations of edge devices like ESP32, it adopts a design with 119k trainable parameters. Compared with the traditional ResNet50, which has 256M parameters, this indicates that its parameter count has been reduced by 99.95%. It combines Butterworth low-pass filtering and Gaussian smoothing preprocessing pipelines, which enhance the noise recovery capability of CSI data and make it particularly suitable for indoor multipath environments in the real world.

In a test with 21 classification tasks, which included 20 alphanumeric gestures and one steady-state category, the average accuracy rate of the test reached 96.31%, an increase of 4.9% compared to the 91.4% of the concurrent lightweight model WiGR. On ESP32, a low-power SoC with memory consumption less than 1MB, It has achieved a CPU inference latency of 18μs, and it can also support continuous operation for 72 hours without any lag during operation .

This means that for the first time, a WiFi gesture recognition model has been deployed on ESP32, eliminating the need to rely on external computing devices like Raspberry PI. It also provides a solution for the interaction of edge iot devices. This model has many advantages. For instance, it has a strong ability to resist multipath interference, even in non-line-to-sight conditions with three walls as obstacles Its accuracy has only dropped by 5%. It is fully compatible with low-power edge devices and has a relatively low deployment cost. However, this model also has some drawbacks. It requires a large amount of training data. If the number of samples available for each gesture category is less than 10, its accuracy rate will drop to 72%. For micro-gestures, such as finger tapping, its recognition accuracy can be maintained at 65%. This model is mainly suitable for complex gesture interaction scenarios driven by edge devices, such as the control of Internet of Things gateways and multi-gesture operations inside smart refrigerators [7].

2.3 TransferSense

The research team led by Bu Q from the Chinese Academy of Sciences has successfully addressed the challenges, solving the high cost of gesture data annotation and the frequent problem of small sample cross-scene recognition in practical scenarios. They proposed a transfer learning technology called TransferSense, which adopts a strategy that combines pre-training transfer and multi-feature fusion. This method reduces the workload of annotating gesture data and also enhances the generalization ability of the model in different scenarios.

The first instance of the CSI amplitude-phase fusion feature was sent into the pre-trained VGG16 network, and the 11th to 13th layers of the network were fine-tuned. This enables the model to reuse its general feature extraction function from the image domain and reduce its reliance on those gesture data. Here, a triple loss function was designed. That is, the loss function composed of anchor point samples, positive samples and negative samples, which reduces the distance between similar gesture features and increases the distance between different features, thereby solving the problem of inconsistent feature distribution among different users or different scenarios. Finally, the support vector machine was chosen as the final classifier to balance the granularity of the transferred features and the lightweight nature of the classifier, adapting to scenarios with a relatively small number of samples.

When conducting cross-domain recognition, only 3 to 5 labeled samples are needed to identify each gesture category. Compared with traditional models that require over 50 samples, this cross-domain recognition reduces the labeling workload by 90%. For cross-user gait recognition, its accuracy rate exceeds 77%. In the case of cross-domain, that is, from the laboratory to the home, The accuracy rate of sign language recognition has reached 88%. Compared with the 75% accuracy of non-transfer learning methods, it has increased by 13%. In three different indoor layouts, the fluctuation of accuracy can be kept below 5%, thus overcoming the limitations of specific scene models. The advantage of this recognition method is its strong adaptability to small samples, and it is particularly suitable for those challenging annotation scenarios. For instance, scenes involving sign language and rehabilitation gestures. In terms of cross-scenario generalization, it performs better than GESFI of the same period.

However, since this thing relies on the pre-trained VGG16 model, it is difficult for devices with limited resources like ESP32 to load it. It should be noted that the model initialization requires 500MB of memory, and the accuracy rate of micro-action recognition is only 58%. Nevertheless, it also has applicable scenarios. It is suitable for cross-scenario recognition with sparse samples, such as monitoring gestures during medical rehabilitation and providing assistance to the hearing-impaired when interacting with sign language [9].

2.4 GESFI

At Tsinghua University in China, there is a research team led by Zhang Xin. They have proposed the GESFI domain generalization technology, which is designed to address the problems existing in traditional domain generalization methods. Traditional domain generalization methods rely on subjective scene labels, such as position and direction labels, and this dependence can lead to inherent accuracy loss issues. The research team led by Zhang Xin proposed a method that adopts an objective latent domain mining and adversarial alignment framework, which can achieve cross-scenario gesture recognition without the need for target domain data.

This method disposes of subjective domain labeling. It relies on studying the performance deficiencies of gesture classifiers, that is, clustering the samples with the lowest recognition accuracy. It proposes a worst-case criterion, that is, mining potential domains, which can avoid deviations in manual labeling. It also builds a three-module adversarial network. This network is composed of several parts: feature extractors, domain classifiers and gesture classifiers. The feature extractor learns features that have no relation to the domain, while the domain classifier distinguishes the associations between samples and potential domains. Both of these components are optimized through adversarial interaction. It adopts a three-stage training process, namely pre-learning, mining, and calibration. In this way, only the data from the source scene can be used to adapt to new scenes. And there is no need to use data from the target domain.

In terms of location recognition, the recognition accuracy rate reached 95.01% at the intersection from the laboratory to the office and then to the bedroom. In terms of environmental recognition, the recognition accuracy rate reached 93.2% in the intersection environment from quiet scenes to those with many pedestrians interfering. This recognition accuracy rate is 6% higher than the 89% of the subjective marking method. It doesn't need to collect data in the target field. Trained with just 1,000 source scene samples, it can adapt to three new scenarios, thus reducing the cost of on-site data collection by 80%. In the case of multiple device interferences, such as microwave ovens and Bluetooth devices, its accuracy only drops by 7%, which is better than the concurrent WiGRUNT model. It has many advantages. One of them is the objective domain division, which can be consistent with the CSI signal distribution pattern, eliminate the need for target domain data, and reduce the cost of large-scale deployment. However, it also has disadvantages. One of the disadvantages is that the training process is relatively complex, and the training for a single scenario requires more than 12 hours. The training time is 30% longer than that of the traditional model, and it is also necessary to manually adjust the parameters of the latent domain count. This recognition method is suitable for dynamic cross-scene recognition applications, such as gesture interaction in vehicles and gesture control in public Spaces of shopping malls [8].

2.5 C+SVM

Kim Ji Soo from Seoul National University in South Korea proposed a lightweight C+SVM recognition technology. This technology is used to address the challenges faced when running deep learning models on resource-constrained devices. Devices like WiFi access points, which have only 4MB of memory, belong to resource-constrained devices. This method combines gesture classification with a lightweight SVM solution, which simplifies the task and enables low-cost recognition of complex gestures.

At the beginning, based on the number of action segments, ten commonly seen gestures such as pushing and pulling were divided into four categories, with each category containing two to three gestures. In this paper, the multi-classification task was split into two steps: classification and intra-class recognition, which reduced the decision-making complexity of the support vector machine (SVM). Next, The CSI data was subjected to SVD denoising and DWT smoothing processing, extracting a total of 8 time-domain features such as variance, mean, and correlation coefficient. This avoids the need to calculate complex frequency-domain features. By optimizing the kernel function, that is, selecting the RBF kernel, and adjusting the penalty

parameters, a dedicated SVM classifier was trained for each of the four gesture categories to balance accuracy and computational load.

The four SVM classifiers finally implemented have a total memory usage of only 0.28MB, which is better than CSI-DeepNet. The memory usage of CSI-DeepNet is less than 1MB. Moreover, these four SVM classifiers can be directly deployed on the TP-Link WiFi access point. It should be noted that this TP-Link WiFi access point only has 4MB of memory. In terms of CPU inference latency, it is 0.21ms under the condition of Intel Xeon 2.20GHz. It can support 4,761 gesture recognitions per second. Such a situation can meet the requirements of real-time interaction. For those untrained new users, the average recognition accuracy can reach 89%. Compared with 58% of independent support vector machines, it has increased by 31%, which has solved the problem of poor multi-class classification performance of traditional support vector machines. Its main advantage lies in the extremely low resource demand. It is precisely because of this advantage that it is suitable for use in devices that are severely restricted, such as WiFi access points and low-power iot gateways. Additionally, its classification errors can be controlled, with only 2-3 gestures distinguished in each category, ensuring that the decision-making boundary is clear.

However, gesture classification errors propagate; for instance, misclassifying a two-segment action as a single segment reduces subsequent accuracy by 40%. Furthermore, it only supports simple gestures, with accuracy below 65% for actions exceeding three segments. Consequently, it is particularly suited to resource-constrained edge scenarios, such as WiFi AP gesture control or battery-powered wireless gesture switches [10].

3 Core Challenges and Solutions

3.1 Poor cross-scenario generalization

Static clutter and dynamic interference across diverse settings cause misalignment in CSI feature distributions. Conventional methods, reliant on subjective domain labels such as position and orientation, experience accuracy declines of 40% to 50% in unseen scenarios. To address this, an objective latent domain mining approach can be employed, abandoning subjective labels. By mining latent domains through the model's worst-case criterion (the sample cluster yielding the poorest classifier performance) and combining adversarial learning to minimise inter-domain differences, cross-location accuracy is elevated to 95.01%. Through transfer learning, leveraging a source-scene pre-trained model with a small number of target-domain samples, cross-domain generalisation is achieved, yielding cross-user accuracy exceeding 77%.

3.2 Small sample scarcity

The high cost of annotating gesture data (e.g., collecting sign language vocabulary takes over 15 minutes per term) and the scarcity of samples in practical scenarios often fewer than five per category, lead to model overfitting. To address this, a combined strategy of GAN data augmentation and transfer learning was employed. GANs generated 20 times more virtual CSI spectrograms consistent with the real sample distribution, alleviating sample scarcity. Integrated with the TransferSense transfer learning framework, this leveraged the general feature

extraction capabilities of pre-trained models, reducing reliance on annotated samples. Sign language recognition accuracy improved from 75% to 88%.

3.3 Hardware resources are constrained

Deep learning models (such as ResNet50) require over 200MB of memory, rendering them unsuitable for edge devices like the ESP32 or Raspberry Pi. To address this, Deep Separable Convolution (DS-Conv) replaces traditional convolutions, reducing parameters to 119k and memory requirements to under 1MB. Through gesture classification combined with a lightweight SVM, total memory usage is only 0.28MB with CPU latency of 0.21ms, enabling deployment on WiFi access points or IoT gateways. Building upon this, an integrated optimisation scheme combining pruning, quantisation, and knowledge distillation further pushes hardware boundaries. Phase 1 further compressed parameters on the CSI-DeepNet foundation through iterative amplitude pruning (removing 30% of static noise-sensitive channels) and uniform symmetric quantisation-aware training (4/8-bit). Phase 2 utilises redundant weights identified during pruning to construct a teacher network, which distils knowledge onto the pruned lightweight model (student network). This achieves a 90% reduction in total parameters (retaining only 10% of the original model's parameters) Memory footprint is reduced to under 0.25MB, with accuracy loss below 3% (CSI-DeepNet maintains over 93% accuracy after PQK optimisation). This framework is perfectly suited to edge device characteristics: inference latency on Raspberry Pi 4B drops to 5μs, while memory usage on ESP32, after INT8 quantisation, is merely 0.25MB. It compensates for accuracy losses caused by single pruning/quantisation, and proves more adaptable than standalone CSI-DeepNet or C+SVM for resource-constrained edge scenarios. This forms a complete solution: a foundational lightweight model combined with integrated PQK optimisation. quantisation, while proving more suitable than standalone CSI-DeepNet or C+SVM for resource-constrained edge scenarios. This forms a comprehensive hardware adaptation solution combining a lightweight base model with PQK integrated optimisation [11].

3.4 CSI Interference and Noise

Multipath effects and hardware noise, such as carrier frequency offset and sampling frequency offset, can cause distortion of CSI signals, reducing recognition accuracy by 20% to 30%. To address this issue, one can reduce hardware noise by calculating the CSI ratio or rely on IQD feature selection to retain key information. And the noise reduction work is carried out by combining the use of Hampel filters and Butterworth low-pass filters. By integrating the amplitude and phase features of CSI, all the fine-grained information of the phase can make up for the deficiencies of the amplitude characteristics, thus significantly enhancing the anti-interference capability.

3.5 Modal deficiency and interference

In practical applications, CSI modalities may become unavailable due to signal interruptions or obstructions. This causes traditional models to experience a sharp decline in accuracy when modalities are missing. To address this issue, a multimodal fusion approach combined with an attention mechanism can be employed. TransferSense uses CSI amplitude-phase fusion as its

core modality. When a modality is missing, asymmetric multi-head attention maps features from the core modality to the missing modality. GESFI achieves 93.2% cross-environment accuracy, even with partial modality loss, through latent domain feature alignment. This represents an 18% improvement on traditional models.

4 Future Application Scenarios Outlook

4.1 Sign language interaction assistance for the hearing impaired

Sign language interaction is still urgently needed by the hearing impaired, yet existing solutions are largely confined to laboratory settings. However, by integrating HandFi's left-right hand recognition and GAN-based data augmentation for sample generation with CSI-DeepNet's edge deployment capabilities, it is possible to design a portable ESP32 device combined with a real-time sign language translation system. The device incorporates a Wi-Fi module to collect CSI data and train models using GAN-enhanced sign language datasets. It can achieve real-time speech-to-text conversion for over 60 items of Chinese sign language vocabulary. This addresses the deaf community's immediate outdoor communication needs.

4.2 Edge-aware smart homes

The proliferation of smart home devices requires a unified interaction gateway. Leveraging the lightweight nature of CSI-DeepNet, the model can be deployed on an IoT gateway (ESP32). Using WiFi CSI, the model recognises gestures that can be used to adjust lighting brightness, switch television channels, and perform other actions. This approach is compatible with existing Wi-Fi infrastructure, eliminating the need for additional sensors. The response latency of gesture control remains below 100 milliseconds, meeting the requirements for real-time interaction.

4.3 Medical rehabilitation gesture monitoring

Traditional rehabilitation monitoring relies on wearable sensors, which can interfere with therapeutic movements. Taking advantage of the non-contact nature of Wi-Fi, a rehabilitation gesture monitoring system can be designed. Wi-Fi access points collect CSI data from patient movements, such as arm elevation and finger extension. Using TransferSense, cross-ward generalisation can be achieved with just five annotated samples, enabling real-time tracking of rehabilitation progress while avoiding movement constraints imposed by sensors.

4.4 Gesture Interaction in Vehicle Networking

In automotive environments, drivers require the ability to operate in-vehicle systems without touching them (e.g., adjusting the air conditioning or answering calls). However, dynamic conditions such as vehicle vibrations and pedestrian interference demand high recognition robustness. Deploying the GESFI framework on in-vehicle Wi-Fi modules, therefore, enables adaptation to such environments through objective latent domain mining. Recognition accuracy reaches 92% without compromising driving safety.

5 Conclusion

This essay provides a systematic review of Wi-Fi-based gesture recognition technology. Research indicates that Wi-Fi gesture recognition can achieve high accuracy and low resource consumption through techniques such as channel state information (CSI) feature extraction, lightweight model design, and cross-domain learning. This demonstrates significant advantages in scenarios such as intelligent interaction and assistance for special populations. Compared to visual, ultrasonic and millimetre-wave sensing technologies, Wi-Fi sensing's low cost and device-independent nature make it more suitable for large-scale deployment. However, several areas require further advancement. Recognition accuracy for micro-gestures such as finger taps needs to be improved, and CSI fine-grained feature extraction requires optimisation. Furthermore, multi-scenario adaptability necessitates the exploration of domain mining methods that do not require manual parameter tuning. Additionally, multimodal fusion (e.g. WiFi and visual) could enhance robustness in complex environments. Future research focusing on these areas will advance WiFi gesture recognition from laboratory development to practical implementation.

References

[1] Kim, J.S., Jeon, W.S., Jeong, D.G.: WiFi-based low-complexity gesture recognition using categorization. In: 2022 IEEE 95th Vehicular Technology Conference: (VTC2022-Spring), pp. 1–7. IEEE (2022)

[2] Sruthi, P., Satapathy, S., Udgata, S.K.: HandFi: WiFi sensing based hand gesture recognition using channel state information. Procedia Computer Science 235, 426–435 (2024)

[3] Hożyń, S.: Advancements in visual gesture recognition for underwater human–robot interaction: A comprehensive review. IEEE Access 12, 163131–163142 (2024)

[4] Tang, L., et al.: Research on recognition algorithm for gesture page turning based on wireless sensing. Intelligent and Converged Networks 4(1), 15–27 (2023)

[5] Wang, Y., Hao, Z., Dang, X., Zhang, Z., Li, M.: UltrasonicGS: A highly robust gesture and sign language recognition method based on ultrasonic signals. Sensors 23(4), 1790 (2023)

[6] Hao, Z., Sun, Z., Li, F., et al.: Millimeter wave gesture recognition using multi-feature fusion models in complex scenes. Scientific Reports 14, 13758 (2024)

[7] Kabir, M.H., Hasan, M.A., Shin, W.: CSI-DeepNet: A lightweight deep convolutional neural network based hand gesture recognition system using Wi-Fi CSI signal. IEEE Access 10, 114787–114801 (2022)

[8] Zhang, X., Feng, Y., Huang, J., et al.: Objective gesture recognition based on WiFi. TechRxiv (2024).

[9] Bu, Q., Ming, X., Hu, J., et al.: TransferSense: towards environment independent and one-shot wifi sensing. Personal and Ubiquitous Computing 26, 555–573 (2022)

[10] Cui, J., Cui, L., Huang, Z., Li, X., Han, F.: IoT wheelchair control system based on multi-mode sensing and human-machine interaction. Micromachines 13(7), 1108 (2022)

[11] Kim, J., Chang, S., Kwak, N.: PQK: model compression via pruning, quantization, and knowledge distillation. arXiv preprint arXiv:2106.14681 (2021).