



Research and Analysis of Behavioral Gesture Recognition Technology Based on Wireless Perception

Jiacheng Li

School of Artificial Intelligence, Hubei University, Wuhan, Hubei, 430062, China
202431121010006@stu.hubu.edu.cn

Abstract. With the rapid development and continuous progress of wireless communication technology, the role played by wireless perception technology in the field of behavior and gesture recognition has become increasingly important and prominent. Therefore, this paper summarizes several current behavior and attitude recognition technologies based on wireless perception and systematically combs and analyzes them. It is found that the recognition accuracy of CeHAR is higher than that of the best recognition method before. AFEE-MatNet can significantly shorten the training time and address the limitations of retrained models using Wi-Fi CSI data in new environments. CNN-ABLSTM achieves high recognition accuracy in a variety of environments. The average accuracy of DPWiT for the localization of the start and end time and category of the activity far exceeds that of the baseline model. WiPhase reduces the computational complexity and the number of model parameters. Therefore, the above methods can be flexibly selected according to the performance requirements of different application scenarios.

Keywords: WiFi-CSI, Wireless perception, Human activity recognition.

1 Introduction

Whether it is 5G or 6G technology in the future, it not only needs high-speed data transmission ability, but also needs wireless perception function. At present, WiFi-CSI-based wireless perception behavior recognition technology is widely used in many fields and has important value. For example, the behavior and posture recognition technology based on wireless perception can detect the fall of the elderly, so as to notify the family members of the rescue in time, effectively reducing the risk caused by the fall. The operation behavior of workers can be detected to avoid accidents caused by improper operation. In the field of smart homes, the automatic adjustment of household appliances can be realized by recognizing people's behavior and facial expressions. At present, the main methods of posture recognition are visual image analysis based on captured behavior and signal analysis based on wearable devices. However, there are some problems in visual image analysis, such as privacy, monitoring blind spots, and vulnerability to environmental interference. Wearable devices have problems such as forgetting to wear, high cost, and difficulty in applying to long-term health monitoring. The Wireless Device-Free Human Perception technology based on CSI and RSSI can

© The Author(s) 2026

S. Zhang (ed.), *Proceedings of the 2025 International Conference on Electronics, Electrical and Grid Technology (ICEEGT 2025)*, Advances in Engineering Research 292,

https://doi.org/10.2991/978-94-6463-986-5_31

realize non-intrusive human behavior and posture monitoring through wireless signals, which does not require users to wear devices, improves convenience, and also plays a role in protecting privacy, filling the technical gap of traditional device-dependent and visual perception.

This study systematically combed and compared five typical wireless perception behavior and posture recognition methods based on Wi-Fi CSI, such as CeHAR, AFEE-MatNet, CNN-ABLSTM, WiPhase and DPWiT, and carried out in-depth analysis from the aspects of technical principles, performance indicators, advantages and disadvantages and application scenarios. On this basis, the current challenges and coping strategies are sorted out. Finally, future research directions such as multi-modal fusion, cross-domain transfer learning and model lightweight are proposed, which provide a comprehensive and targeted reference for researchers and engineering practice in this field.

2 Analysis of Typical Techniques

2.1 CeHAR

CeHAR employs a scaling factor of batch Normalization as a criterion for every related channel and dynamically switches channels between subnetworks of magnitude and phase. To achieve the depth fusion of amplitude and phase, a feature channel's scaling factor is substituted with another feature's information at the same location when it approaches zero. Figure 1 depicts the structure of the CeHAR system, which primarily consists of four modules: The module for radio map construction, the feature extraction module, the depth fusion module and the classification module. The module for radio map construction is used to record the measurements of amplitude and phase, preprocess the measurement results and construct the radio map using the related position. The module for feature extraction uses ResNet50 as the backbone to train two subnetworks, amplitude and phase. The module for deep fusion uses the scaling factor of batch Normalization to determine the position of the exchange and fusion channel between the two subnetworks, and then exchanges the channels of the two subnetworks, and dynamically integrates the amplitude and phase. At the same time, the convolutional filters are shared; however, the private batch Normalization layers are kept with different features, so that the feature fusion network is almost as compact as the single feature network, and the efficiency and accuracy are balanced. The module for classification uses SoftMax to output activity categories. In the offline phase, the training dataset is used to train the model. In the offline phase, the model is trained, and in the online phase, the real-time CSI is input into the model to forecast the activity.

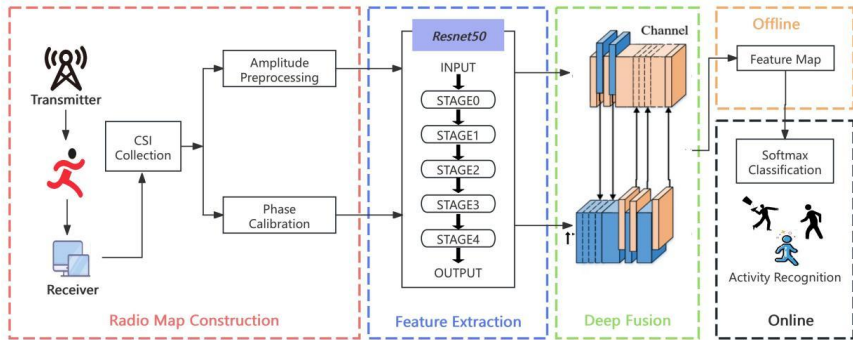


Fig. 1. CeHAR structure [1].

This method has high recognition accuracy, good generalization performance, high efficiency of online testing of the model, and low time cost. In the confusion matrix of CeHAR and the baseline methods in the office environment, the recognition accuracy of CeHAR for all actions is more than 90%. In the two test environments, CeHAR enhances the overall accuracy by 2.9% and 6.6%, respectively, compared with the previous best model ABLSTM, which indicates that its channel exchange and parameter sharing mechanism effectively enhance the generalization performance. Experiments based on office area data show that the training time of the deep learning model is much faster than that of the machine learning algorithm DWT-RF, but in the offline phase, the training is only performed once. The online inference time for CeHAR can fulfill real-time application needs, and its time overhead is lower than other models, reflecting higher efficiency [1]

2.2 AFEE-MatNet

The structure of AFEE-MatNet is shown in Figure 2, which consists of three main modules: CSI acquisition, CSI preprocessing, and activity recognition based on MatNet-PCC. In the CSI acquisition module, the Intel 5300 network interface Card is used to collect CSI data changed by human activities according to IEEE 802.11n. Each transceiver antenna pair has 30 subcarriers. In the CSI preprocessing module, the AFEE method is proposed. Firstly, conjugate multiplication and PCA are used to remove noise and irrelevant data to overcome phase offset, and then the FFT transformation is performed on the processed data to discard the high-frequency and zero-frequency parts to reduce the data dimension. In the activity recognition module based on MatNet-PCC, the MatNet architecture uses CNN and LSTM to automatically learn cross-environment transferable features and minimizes the loss function with multi-source environmental data during training. PCC corrects recognition errors based on the behavior state transition diagram and confusion matrix.

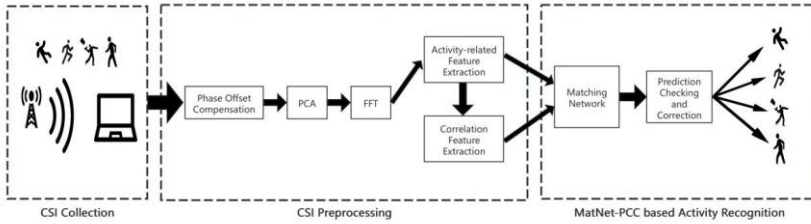


Fig. 2. Structure of AFEE-MatNet [2].

The innovation of our method is that the model can be directly used in new environments without retraining, while only requiring a limited quantity of source environments for training. Six types of activities are recognized in three new test environments, and AFEE-MatNet performs noticeably better than the other three sensing techniques in all environments, showing the advantage of directly adapting to new environments with no retraining. At the same time, compared with the scheme without AFEE, the training time is shortened by 75.2% after introducing AFEE, which significantly improves the recognition efficiency and reduces the training cost [2].

2.3 CNN-ABLSTM

The structure of CNN-ABLSTM is shown in Figure 3, which is divided into a physical layer and a network layer. The physical layer is responsible for CSI data acquisition and preprocessing. The network layer was composed of CNN and ABLSTM. The CNN three-layer convolution extracted spatial features, and the BLSTM in ABLSTM extracted temporal features, which were weighted and fused through the attention layer. Finally, the activity labels are classified and output by SoftMax.

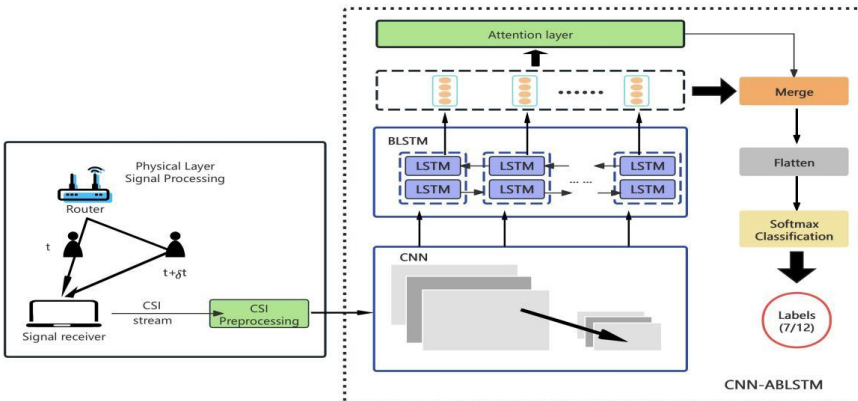


Fig. 3. CNN-ABLSTM structure [3].

CNN-ABLSTM introduces a model of an attention mechanism, which can independently determine each parameter's significance and assign greater weight to those

that are more crucial to improve the stability of the model when the environment or dataset changes. The proposed method reduces the computational complexity and enhances the robustness. In the comparison of the ultimate accuracy of each algorithm, CNN-ABLSTM is significantly higher than other methods. At the same time, a comparison is made regarding the recognition accuracy between CNN-ABLSTM and CNN-BLSTM. In the absence of the attention layer, CNN-BLSTM has no significant advantage, which further proves that perception performance can be effectively enhanced by adding the attention layer. CNN-ABLSTM has an accuracy of more than 96%-97% on the training set, which shows excellent performance. More importantly, when the label is extended from 7 classes to 12 classes and the scene is changed from single person to group interaction, the precision of each action recognition is still high, which reflects excellent robustness [3].

2.4 2.4 WiPhase

The structure of WiPhase is shown in Figure 4. Firstly, the CSI is filtered by a Butterworth low-pass filter, and then the signals related to the activity are separated by EEMD. Then the most relevant subcarriers are extracted by SSP, and the sparse CSI is reconstructed into CSI-PIR and phase map, respectively. The Gated Pseudo-Siamese network (GPSiam) and Graph Attention Network (DRGAT) were input, and finally, the decision fusion module (DD) integrated them to output the activity judgment.

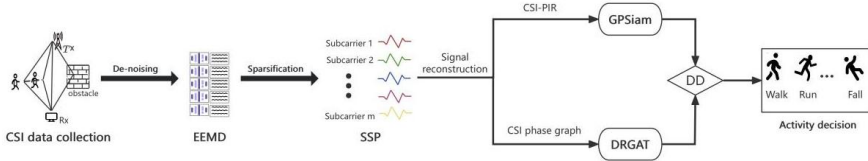


Fig. 4. WiPhase structure [4].

The dual-feature stream fusion architecture solves the problem of ignoring the complex correlation of subcarriers in traditional methods, and has low computational complexity, higher model efficiency, and strong cross-domain generalization ability. Wi-Phase performs well with a variety of commonly used methods on dataset1. Further tested under dataset2, the accuracy of other methods decreases, while WiPhase still has the highest recognition accuracy, which proves its excellent accuracy in HAR tasks. In addition, WiPhase saves the training time by at least 40.34%, reduces the computational complexity by 46.74%, and reduces the number of model parameters by 36.61%. Under CCD conditions of different test sets, the accuracy of other models significantly decreases, while WiPhase only slightly decreases, but still maintains at 90.571%, showing excellent cross-domain generalization ability [4].

2.5 DPWiT

The framework of DPWiT is shown in Figure 5. The input signal is passed through three CGR layers in turn to extract primary features, and then fed into the double pyramid temporal context modeling module, which includes Temporal Signal Semantic Encoder (TSSE) and Locality Sensitive Response Encoder (LSRE). The TSSE consists of a Transformer branch and a Conv-Pool branch, which are used to learn low frequency and high frequency respectively. These functions are integrated by ContraNorm normalization. LSRE uses channel Windows to slide on the time axis to extract regional information, and capture signal fluctuations in a learning-free manner. In this way, it achieves faster processing speed and lower memory usage. The aggregated features are mapped to a more robust discriminant space by MLP, and the functions of the latter two encoders are aligned by a cross-attention pyramid fusion mechanism. Finally, the detection results of the head are predicted for training and inference.

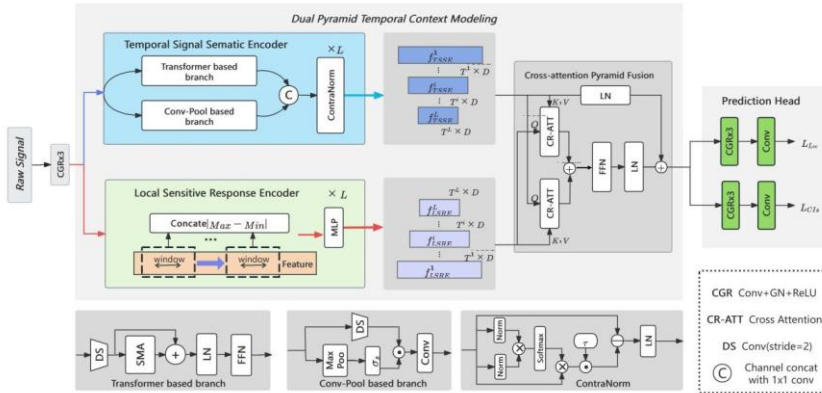


Fig. 5. DPWiT structure [5].

This method can process long-term untrimmed and continuous WiFi CSI signals, which is suitable for long-term monitoring. It can accurately locate the start and end time of activities and activity categories, and has low computational complexity and no learning mechanism. At a high tIoU threshold (0.7), the mAP of DPWiT reaches 54.5, far exceeding the baseline model's 40.1, and the overall average mAPavg is 74.5, which is significantly ahead of all baselines, indicating that the model can accurately classify and accurately locate the activity time series boundary at the same time, which is superior to all baselines [5].

3 Challenge Analysis

Although CeHAR fuses CSI amplitude and phase features through the channel exchange mechanism, it relies on BN scaling factor to judge channel importance, which may cause feature selection bias in complex environments. A possible solution is to increase feature complementarity evaluation based on mutual information to avoid the

dominance of a single feature. At the same time, although its shared parameter mechanism reduces the amount of calculation, the training cost of the dual-channel network structure is higher than that of the single feature model, so replacing the traditional convolution with the depthwise separable convolution to reduce the number of parameters may have certain improvement [6].

Although AFEE-MatNet does not need re-training and can be used directly in new environments, when the source environment is insufficient, the recognition accuracy will be significantly reduced. A possible solution is to combine meta-learning, such as MAML algorithm, so that the pre-trained model can learn fast adaptation ability on multiple source environments, so that it can be fine-tuned with only a small number of new environment samples [7].

Although CNN-ABLSTM introduces an attention mechanism to improve key feature focus, in small sample scenarios, the weight of attention may be biased towards noise or a few samples, and the introduction of contrastive learning may improve stability [8].

In this study, WiPhase does not consider the decoupling of multi-target signals. Specifically, the experiment is mainly for single-user scenarios, but multi-user activities will lead to signal superposition, which may reduce the feature discrimination of CSI phase difference and phase ratio, and it may be feasible to introduce multi-target signal separation algorithm to improve accuracy.

DPWiT only verifies a few daily activities in the experiment, but when extended to more fine-grained actions, the classification accuracy of the feature pyramid may decrease, and using 3D-CNN to improve the discrimination ability of similar actions may have a certain effect [9].

4 Application Scenarios

Human activity recognition technology based on Wi-Fi CSI is widely used in many fields. For example, in a smart home, the WiFall fall detection system uses commercial WiFi equipment to collect CSI data, and analyzes the signal multipath propagation changes caused by sudden changes in height when the human body falls. After data filtering and dimension reduction, the machine learning algorithm is used to realize non-contact fall detection [10]. In medical health, non-intrusive breath detection based on commercial WiFi can be realized. Wang et al. proposed to use the Fresnel zone model to convert chest respiratory motion into WiFi signal path changes, and extract respiratory rate through CSI data processing and frequency diversity [11]. In the aspect of intelligent transportation, the WiDriver system based on WiFi CSI can analyze the influence of the driver's hand and arm movements on the multipath propagation of WiFi signals, and use BP neural network to identify the static posture and DCFA model to track the rationality of continuous motion sequence. Combined with the time interval, amplitude stability and other characteristics of the action, whether the driver is tired driving can be monitored [12].

5 Conclusion

The five methods analyzed above in this study have different advantages in recognition accuracy, environmental adaptability, robustness, model efficiency, start and end time and category of localization activities. Therefore, according to the required high performance in the environment, the possibility of realization can be improved by selecting the appropriate method mentioned above. At the same time, the methods discussed in this paper provide a comprehensive and targeted reference for researchers and engineering practice in this field. In the current era of rapid development of wireless communication technology, HAR based on Wi-Fi signals is at the forefront of technological development, which indicates that the interaction between humans and the digital domain will be more seamless and intuitive. It also means that researchers will pay more efforts to explore in the future. The core direction of future research covers the following three aspects. Firstly, multimodal fusion technology aims to build a cross-modal complementary model to overcome the limitations of single-modal information. Secondly, cross-domain transfer learning is introduced to enhance the generalization performance of the algorithm in diverse application scenarios by introducing self-supervised learning and meta-learning mechanisms. Finally, the model is lightweight, which aims to adapt to the limitations of edge devices in memory and computing resources through the optimization of deep network structure.

References

1. X. Lu, Y. Li, W. Cui, H. Wang, CeHAR: CSI-based channel-exchanging human activity recognition. *IEEE Internet Things J.* 10, 5953-5961 (2022)
2. Z. Shi, Q. Cheng, J. A. Zhang, R. Y. Da Xu, Environment-robust WiFi-based human activity recognition using enhanced CSI and deep learning. *IEEE Internet Things J.* 9, 24643-24654 (2022)
3. Z. He, X. Zhang, Y. Wang, Y. Lin, G. Gui, H. Gacanin, A robust CSI-based Wi-Fi passive sensing method using attention mechanism deep learning. *IEEE Internet Things J.* 10, 17490-17499 (2023)
4. X. Chen, C. Li, C. Jiang, W. Meng, W. Xiao, WiPhase: A Human Activity Recognition Approach by Fusing of Reconstructed WiFi CSI Phase Features. *IEEE Trans. Mobile Comput.* (2024)
5. Z. Liu, L. Zhang, B. Li, Y. Zhou, Z. Chen, C. Zhu, WiFi CSI Based Temporal Activity Detection via Dual Pyramid Network. *Proc. AAAI Conf. Artif. Intell.* 39, 550-558 (2025)
6. M. G. D. Nascimento, R. Fawcett, V. A. Prisacariu, Dsconv: Efficient convolution operator. *Proc. IEEE/CVF Int. Conf. Comput. Vis.* 5148-5157 (2019)
7. C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks. *Int. Conf. Mach. Learn.* 1126-1135 (2017)
8. P. Khosla et al., Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* 33, 18661-18673 (2020)
9. S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 221-231 (2012)
10. Y. Wang, K. Wu, L. M. Ni, Wifall: Device-free fall detection by wireless networks. *IEEE Trans. Mobile Comput.* 16, 581-594 (2016)

11. H. Wang et al., Human respiration detection with commodity WiFi devices: Do user location and body orientation matter?. Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. 25-36 (2016)
12. S. Duan, T. Yu, J. He, WiDriver: Driver activity recognition system based on WiFi CSI. Int. J. Wireless Inf. Networks 25, 146-156 (2018)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

