A Novel WiFi Gesture Recognition Method Based on CNN-LSTM and Channel Attention

Yu Gu*
Jiang'an Li
yugu.bruce@ieee.org
2019170966@mail.hfut.edu.cn
School of Computer and Information, Hefei University of Technology
Hefei, Anhui, China

ABSTRACT

With the rapid development of wireless sensing, intelligent human-computer interaction, and other fields, gesture recognition based on WiFi has become an important research field. Gesture recognition based on WiFi has the advantages of non-contact and privacy protection. In addition, the use of home WiFi makes the technology have a broad application scenario. At present, most gesture recognition models based on WiFi can only achieve good results in a specific domain. When changing the environment or the orientation of gesture action, the performance of the model becomes very poor. This paper proposes a gesture recognition system based on the channel attention mechanism and CNN-LSTM fusion model. On the one hand, the channel attention mechanism can consider the importance of different channel characteristics; On the other hand, the CNN-LSTM fusion model can extract richer features in the time domain and space domain. The system has achieved good classification results in multiple domains of the public data set widar3.0.

CCS CONCEPTS

• **Human-centered computing** → **Gestural input**.

KEYWORDS

gesture recognition, CSI, attention, neural network

ACM Reference Format:
Yu Gu and Jiang'an Li. 2021. A Novel WiFi Gesture Recognition Method Based on CNN-LSTM and Channel Attention. In *2021 3rd International Conference on Advanced Information Science and System (AISS 2021) (AISS 2021), November 26–28, 2021, Sanya, China.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3503047.3503148

1 INTRODUCTION

The development of human-computer interaction technology has greatly facilitated people's modern life. With the development of technology, various interactive methods have emerged. Gestures are the most powerful means of human nonverbal expression and play the most important role in body language. Therefore, gesture recognition has become an important way of human-computer interaction, which has attracted wide attention from researchers. The current gesture recognition methods mainly include sensor-based gesture recognition, vision-based gesture recognition, and radio frequency-based gesture recognition. At the beginning of the research on gesture recognition, people's research on gesture recognition mainly focused on the sensor, especially the research on the data glove [5]. Researchers wear sensor equipment and connect the equipment to a computer. The computer obtains information such as the position of the hand and finger extension obtained by the sensor for gesture recognition. Later, with the development of computer vision and image processing technology, people began to recognize gestures through image processing technology and computer vision methods, and some representative products were born, such as Microsoft's Kinect [4]. Although these two methods can recognize gestures to a certain extent, sensor-based gesture recognition requires additional equipment, which imposes certain constraints on the human body, while vision-based gesture recognition needs to be performed under light conditions, and may Bring troubles to personal privacy. With the development of wireless sensing technology, people use radio frequencies for gesture recognition. This technology solves the aforementioned problems well. Among various radio frequency perceptions, WiFi perception has unique advantages, because WiFi devices have been popularized in every family. When WiFi perception technology matures, ordinary home routers can be used for gesture recognition.

WiGest [1] is the first to propose the use of WiFi perception for gesture recognition, which has achieved 87.5% accuracy of gesture classification and recognition. Because WiGest uses coarse-grained RSS information for classification, it does not achieve high accuracy and is sensitive to environmental changes. The use of CSI information can provide more fine-grained features related to gestures, which improves the recognition accuracy but still cannot solve the interference caused by environmental changes to the model. Because the WiFi signal is easily affected by the environment, the CSI information in different environments has various noises and environment-related interference. WiDar3 [10] designed a physical feature BVP for gesture recognition in response to this problem and achieved high-accuracy cross-domain gesture recognition.

However, BVP features have certain limitations. The characteristics of manual design have certain limitations. Manual design is difficult to restore common features on different domains. For this reason, we use deep learning to automatically perform feature extraction. Using deep learning to automatically extract features

AISS 2021, November 26–28, 2021, Sanya, China

ACM ISBN 978-1-4503-8586-2/21/11…$15.00
https://doi.org/10.1145/3503047.3503148

faces two difficulties. On the one hand, it may not be able to extract sufficiently rich features, on the other hand, it may extract some interference features, such as the influence of environment and noise. To solve these two problems, this paper comprehensively considers the time and space characteristics of CSI information, uses the LSTM-CNN fusion model to extract features, and uses the channel attention mechanism to reduce the interference of environmental changes and noise on effective features.

We evaluated our gesture recognition method on the public gesture recognition data set WiDar3. The experimental results show that our method can achieve 97.33%, 87.17%, 92.78%, 89.01% accuracy across locations, directions, environments, and users, respectively.

The main contributions of this paper are as follows:

- Using the CNN-LSTM parallel network architecture, it can extract richer features in the time domain and spatial domain for gesture recognition.
- The channel attention mechanism is used for the spatial features extracted by CNN, which effectively suppresses the interference caused by environment and noise during cross-domain gesture recognition.
- Our gesture recognition method achieved 97.33%, 87.17%, 92.78%, and 89.01% accuracy across locations, directions, environments, and users, respectively.

We organize the remainder of this paper as follows: we introduced the related work in Section 2 and the system design in Section 3. Then, we show our experimental results in Section 4. Finally, we conclude our work in Section 5.

2 RELATED WORK

Gesture recognition based on WIFI can be roughly divided into three categories: pattern-based methods, model-based methods, and deep learning-based methods.

The purpose of pattern-based methods is to find unique patterns that recognize human behavior. This depends on constructing effective features, which can be used to identify different human behaviors [8]. If the signal change pattern has a unique and consistent relationship with certain human activities, then pattern-based methods may be able to accurately identify human behavior from the signal pattern. Therefore, pattern-based methods usually require samples to construct patterns to realize CSI-based behavior recognition.

The design of the model-based method relies on the physical model, which associates the signal space with the physical space, and establishes the laws of physics through the relationship between the received signal and the sensing target [6]. Model-based recognition methods associate the user's movement with the received signal changes. Therefore, it recognizes human behavior by exploring the laws of physics and establishing models based on physics or mathematics. Compared with the mode-based method, the model-based method uses the model, so it can achieve good performance with a small amount of measurement data.

Compared with the first two methods, the method based on deep learning has many obvious advantages because it can automatically identify and extract effective features and input them into the classifier. The process of feature extraction and classification is realized through the middle layer of the deep neural network model. By designing a reasonable deep learning model, it is possible to find common features in different domains and achieve better cross-domain gesture recognition.

3 SYSTEM DESIGN

The entire process of using WiFi for gesture recognition is shown in Fig. 1. The CSI information affected by the gesture is collected by the data collection module, and then the phase information of the CSI information is extracted by the data preprocessing module and converted into a format acceptable by the feature extraction module for feature extraction. The feature extraction module extracts Spatio-Temporal features and finally uses them for gesture classification. In the whole process, the two most important modules are the data preprocessing module and the feature extraction module.

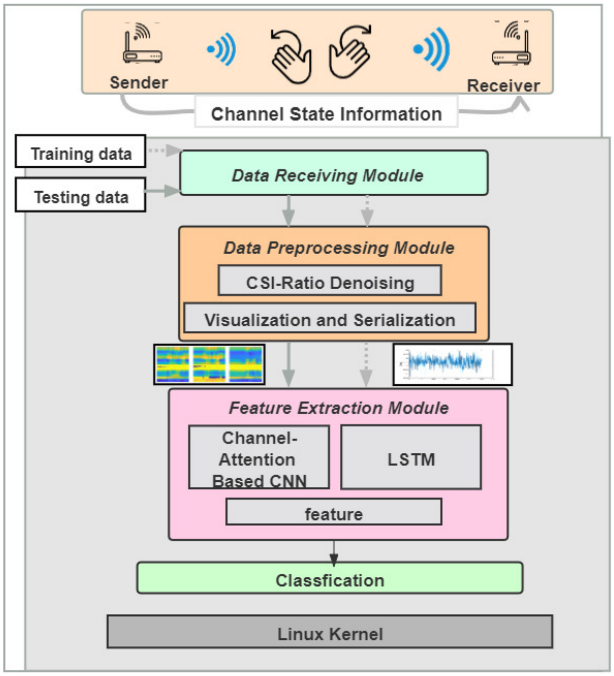
Sender
Receiver
Channel State Information
Training data
Testing data
Data Receiving Module
Data Preprocessing Module
CSI-Ratio Denoising
Visualization and Serialization
Feature Extraction Module
Channel-Attention Based CNN
LSTM
feature
Classfication
Linux Kernel

Figure 1: System Design

3.1 Data preprocessing module

CSI represents how a wireless signal travels from the transmitter to the receiver along multiple paths at a specific carrier frequency.

Let $\vec{Y}$ and $\vec{X}$ be the signal vectors received and transmitted by WiFi, and $\vec{N}$ is Gaussian noise, then the relationship between them can be characterized as:

$$\vec{Y} = \vec{H} \cdot \vec{X} + \vec{N} \tag{1}$$

.

Where $\vec{H}$ is the channel state matrix, which is a four-dimensional matrix. When a person makes different gestures within the coverage of the WiFi signal, the receiver's $\vec{X}$ changes differently, and the $\vec{H}$,

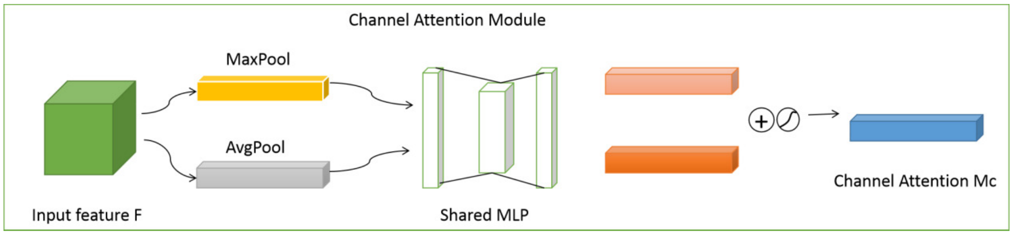

Figure 2: Channel attention module

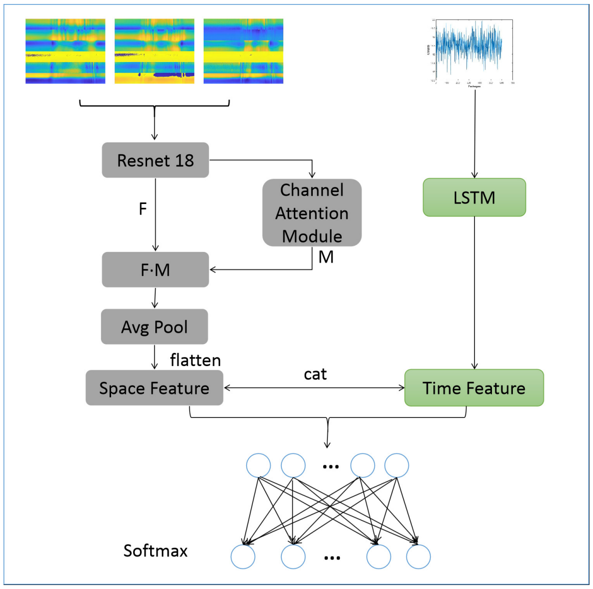

Figure 3: Feature Extraction module

which characterizes the relationship between the two, also changes accordingly. In other words, $\vec{H}$ carries information about the gesture, and through $\vec{H}$, we can recognize the gesture.

We denoise according to the CSI ratio model theory [9], and use the CSI information of different antennas to eliminate random phase shifts, which is convenient for identifying changes caused by tiny fingers such as gestures. Two conversion operations are performed on the denoised CSI matrix. On the one hand, we convert the CSI matrix into image format data according to the method of [2]. On the other hand, we serialize the matrix into time-series data.

3.2 Feature extraction module

The feature extraction module mainly relies on three parts, convolutional network, channel attention, an LSTM network.

CNN can extract local spatial features of image data, and finally, extract spatially-related high-dimensional features through modules such as convolution and pooling. Resnet [3] is a CNN network architecture. We first put the CSI data into the pre-trained Resnet network in the form of images for preliminary feature extraction. According to the extracted features, we use the attention mechanism to obtain the channel weights.

The channel attention model [7] is shown in Fig. 3, where F is the feature initially extracted by Resnet, and Mc is the attention weight map. The calculation formula of Mc is:

$$\mathbf{M_c}(\mathbf{F}) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\mathbf{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\mathbf{F}))) \tag{2}$$

After maximum pooling, average pooling, and multi-layer perceptrons, the channel weight is obtained. By setting this weight, it is possible to focus on features that have a greater impact on classification, thereby improving the accuracy of classification.

The function of the LSTM network is to extract the features on the time series. We flatten the CSI information into time series data and put it into the LSTM network to extract the time-series features.

The architecture of the entire feature extraction module is shown in Fig. 2. The image data is extracted by Resnet18 to obtain the feature F. The feature F is passed through the channel attention module to generate the channel weight M, then F and M are matrices multiplied, then average pooling is performed, and finally flattened Obtain the spatial feature, and the time series data is extracted by LSTM to obtain the temporal feature. The two features are classified using Softmax after a fully connected neural network.

4 EXPERIMENTAL EVALUATION

We use the Widar3 dataset to evaluate our gesture recognition model. Widar3 contains gesture data in three environments. The data in each environment is divided according to the user, location, and direction. We divide the data set according to the environment, user, location, and direction. The comparison with the experimental results of Widar3 using BVP for classification is shown in Fig. 4.

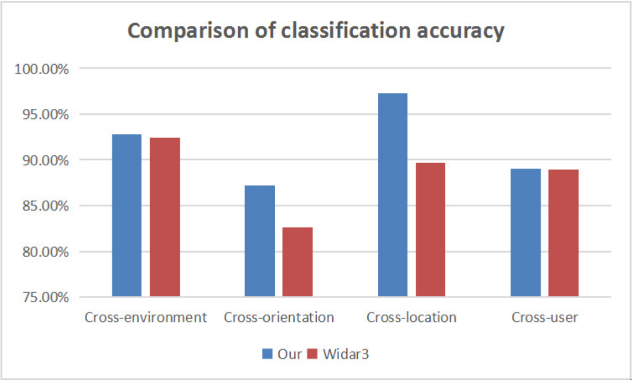

Figure 4: Comparison of classification accuracy

The classification accuracy of our model across environments, locations, and users is slightly higher than that of the BVP model. In particular, the classification accuracy of our model across locations is significantly higher than that of the BVP model.

5 CONCLUSION

We designed a gesture recognition method based on the LSTM-CNN fusion model and channel attention mechanism. This method first performs two different pre-processing on CSI data, and converts the data into image form and sequence form respectively. Then use the CNN-LSTM fusion model for automatic feature extraction, extracting temporal and spatial features respectively. The channel attention mechanism is used for spatial features, and the weights of different channel features are redistributed to reduce the impact of environmental interference. Finally, the two features are combined for gesture classification.

We use the CSI ratio model to preprocess the CSI data and convert the data into two data formats: image and sequence. Then extract the features of the two formats of data, extract the features through the CNN model and use the channel attention mechanism to assign weights to the features of different channels to obtain the spatial features, and then merge them with the temporal features extracted by LSTM, and perform gestures on this basis Classification. On the Widar3 data set, this method has achieved better classification results than the BVP model. We use the Widar3 data set to test the recognition effect of this method in cross-domain. The experimental results show that the method achieves higher accuracy than the BVP classification of the Widar3 data set in four different scenarios across locations, directions, environments, and users.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation of China (61772169) and the Key Research and Development Plan of Anhui Province (202004b11020018).

REFERENCES

[1] Heba Abdelnasser, Moustafa Youssef, and Khaled A. Harras. 2015. WiGest: A ubiquitous WiFi-based gesture recognition system. In *2015 IEEE Conference on Computer Communications, INFOCOM 2015, Kowloon, Hong Kong, April 26 - May 1, 2015*. IEEE, 1472–1480. https://doi.org/10.1109/INFOCOM.2015.7218525

[2] Y. Gu, X. Zhang, Z. Liu, and F. Ren. 2020. WiFE: WiFi and Vision based Intelligent Facial-Gesture Emotion Recognition. (2020).

[3] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition. *IEEE* (2016).

[4] Zhou Ren, Junsong Yuan, Jingjing Meng, and Zhengyou Zhang. 2013. Robust Part-Based Hand Gesture Recognition Using Kinect Sensor. *IEEE Transactions on Multimedia* 15, 5 (2013), 1110–1120. https://doi.org/10.1109/TMM.2013.2246148

[5] Shohel Sayeed, Rosli Besar, and Nidal S. Kamel. 2006. Dynamic Signature Verification Using Sensor Based Data Glove. In *2006 8th international Conference on Signal Processing*, Vol. 3. https://doi.org/10.1109/ICOSP.2006.345880

[6] Zhengjie Wang, Kangkang Jiang, Yushan Hou, Wenwen Dou, Chengming Zhang, Zehua Huang, and Yinjing Guo. 2019. A Survey on Human Behavior Recognition Using Channel State Information. *IEEE Access* 7 (2019), 155986–156024. https://doi.org/10.1109/ACCESS.2019.2949123

[7] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. CBAM: Convolutional Block Attention Module. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 3–19.

[8] Dan Wu, Daqing Zhang, Chenren Xu, Hao Wang, and Xiang Li. 2017. Device-Free WiFi Human Sensing: From Pattern-Based to Model-Based Approaches. *IEEE Commun. Mag.* 55, 10 (2017), 91–97. https://doi.org/10.1109/MCOM.2017.1700143

[9] Y. Zeng, D. Wu, J. Xiong, E. Yi, R. Gao, and D. Zhang. 2019. FarSense: Pushing the Range Limit of WiFi-based Respiration Sensing with CSI Ratio of Two Antennas. (2019).

[10] Yi Zhang, Yue Zheng, Kun Qian, Guidong Zhang, Yunhao Liu, Chenshu Wu, and Zheng Yang. 2021. Widar3.0: Zero-Effort Cross-Domain Gesture Recognition with Wi-Fi. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), 1–1. https://doi.org/10.1109/TPAMI.2021.3105387