Human Activity Recognition Using MSHNet Based on Wi-Fi CSI

Fuchao Wang(✉), Pengsong Duan, Yangjie Cao, Jinsheng Kong, and Hao Li

School of Software, Zhengzhou University, Zhengzhou, China
wfc117@163.com, {duanps,caoyj,jskong}@zzu.edu.cn, lh442401597@163.com

Abstract. In recent years, with the prominent population aging problem, health conditions of aged solitaries are inherently gaining more and more attentions. Among the techniques allowing real-time health monitoring, activity perception has become an important and promising eld in both academia and industry. In this paper, a human activity perception recognition model, named MSHNet (Multi-Stream-Hybrid-Network) based on Deep Learning is proposed to solve the problems of difficulty in extracting perceptual features of Wi-Fi signals and low recognition accuracy in traditional Machine Learning methods. MSHNet adopts passive wireless sensing technology, it uses commercial off-the-shelf Wi-Fi devices to collect Channel State Information (CSI) based on underlying physical equipment and automatically extracts human activity features characterized by amplitude in CSI. Then MSHNet aggregates the data streams of the same receiving antenna using the wireless signal transceiving characteristics of Multiple Input Multiple Output (MIMO) and trains the aggregated data streams respectively. At last, the voting mechanism is adopted to select the best training result. The experimental results demonstrate that MSHNet's results on the public dataset have reached the state-of-the-art and on the datasets of four environments collected by ourselves the average recognition accuracy rate has reached 97.41%, satisfying the daily activity monitoring of the elderly, especially those living alone.

Keywords: Human activity recognition · Wi-Fi · CSI · MIMO · Voting mechanism

1 Introduction

Recently, with the aging of the population, it is a long-term obligation of society and families to protect the health of the elderly, especially those who living alone and independently [9]. The detection and recognition of falls [14] and certain diseases, such as Parkinson's disease [29], can be realized by monitoring the daily activities of people. In order to realize human activity monitoring and recognition, researchers mainly use APT, which is obtain information such as the

Supported by Zhengzhou University.

S. W. Loke et al. (Eds.): MONAMI 2020, LNICST 338, pp. 47–63, 2020.
https://doi.org/10.1007/978-3-030-64002-6_4

current position, activity and action trajectory of the target through hardware or software, to realize analysis and understanding of the current activity of the target. There are mainly three kinds of APTs: computer vision, special sensors and wireless sensing technology.

Computer vision technology [8,16,24] has high recognition accuracy and wide application range, but its shortcomings are obvious, which is easily affected by illumination and obstacles, invasion of user privacy, and existence of blind spot. Special sensors technology [18,22,27] uses special sensors or wearable devices to collect relevant human actions, thus realizing human activity perception. They can realize fine-grained activity perception with high accuracy, but their installation and maintenance usually require high cost, therefore preventing this technique to be widely used.

Wireless sensing technology overcomes the shortcomings of the aforementioned technologies and has gained increasing attention in recent years. It further consists of RF-based and Wi-Fi-based approaches. RF-based [11,35] requires special equipment to be customized, and the cost is usually relatively high, so it is not suitable for large-scale installation. In recent years, with the widespread deployment of Wi-Fi hotspots [1], using Wi-Fi signals to implement some wireless sensing applications [4,5,10] has attracted great attention of researchers. By utilizing the ubiquitous in-house Wi-Fi signals, freeing the personel from any on-body equipment, as well as avoiding the effect of illumination and unnecessary personal privacy invasion.

Initially, researchers realized indoor localization, gesture recognition and simple motion detection based on Received Signal Strength (RSS) in Wi-Fi signals. Due to RSS contains too single information to recognize fine-grained human activity, Channel State Information (CSI) has emerged to replace it. The raw CSI is not sufficient to represent human activity, in practical applications, researchers often use Machine Learning method to extract features from CSI manually, but this method is easy to cause feature loss and requires experts in signal field. With the development of Deep Learning, researchers use Deep Learning methods to automatically extract features from CSI and achieve better results.

In Deep Learning, for time series data, such as human activity data, RNN and its variants are often used for feature extraction. Recently, Convolution Neural Network (CNN) has shown excellent performance in processing time series data. Full convolution network (FCN) [31], residual network (ResNet) [31] and multi-scale convolution network (MCNN) [7] have full advantages in the processing of time series data. Inspired by these two network structures, this paper combines RNN's variant Bi-directional LSTM (BLSTM) with Temporal Convolution Network (TCN) to improve feature extraction and recognition performance, and proposed a Multi-Stream Hybrid Network (MSHNet) as shown in Fig. 1, which is used to realize automatic feature extraction and autonomous learning in human activity recognition perception based on CSI. MSHNet firstly extracts the amplitude information from CSI, and then according to the Multi Input Multi Output (MIMO) characteristics of Wi-Fi signals, it aggregates the data streams of different antennas and trains the aggregated data streams respectively. Finally,

it adopts the voting mechanism to select the classification results. Experiments demonstrate that MSHNet results are superior to the current several Deep Learning methods, and its accuracy rate in the experimental environment reaches 97.41%.

The contributions of this paper are summarized as follows:

- MSHNet is proposed to extract deep features of raw CSI obtained from Wi-Fi signals, and we conducted a lot of experiments to verify MSHNet is the best network structure.
- Based on the MIMO characteristics of Wi-Fi signals, this paper proposes a voting mechanism to improve the accuracy of activity recognition of MSHNet.
- On the public dataset, we compared the results of MSHNet and other models. The results demonstrate that MSHNet has best performance.
- To evaluate MSHNet's performance in other environments, we collect datasets from different environments and conduct training. The experimental results demonstrate that MSHNet has good environmental adaptability.

The remaining of this paper is organized as follows: Sect. 2 provides related work and progress of Wi-Fi sensing. We will introduce MSHNet architecture and analyze the function of each part in Sect. 3, followed by experimental results and evaluation in Sect. 4. We finally conclude the work in Sect. 5.

2 Related Work

Bahl et al. proposed—RADAR, an indoor localization system based on Wi-Fi RSS, which is the first research work based on wireless Wi-Fi sensing and provides a new idea for wireless Wi-Fi contactless sensing [4]. Nuzzer realized simple action detection based on RSS, but it can only identify whether there are actions in the environment, and cannot distinguish different actions [26]. WiGest used the influence of gestures on RSS to identify different gestures [3]. Due to multipath and fading effects in the process of signal propagation, RSS will be unstable and contain a large amount of noise during collection, resulting in limited performance, which is difficult to apply to fine-grained activity recognition.

However, in 2011, Halperin et al. released the CSI Tool [13], which greatly facilitates the extraction of CSI based on physical layer from commercial off-the-shelf Wi-Fi devices. CSI contains abundant and stable amplitude and phase information [13], which enables to be applied in finer-grained human activity monitoring and recognition systems, such as sleep monitoring [5], fall detection [14], gesture recognition [20], identity recognization [34], activity recognition [6,12,33], etc.

Affected by multipath effect, CSI contains environment and other noises, so the raw CSI is not sufficient to represent different human activities. In practice, the commonly used processing method is to manually extract relevant features from the raw CSI to distinguish different human activities [30]. However, manually extract features not only requires experts in relevant fields, but also easily ignores some features, resulting in unsatisfactory results. With the development

of Deep Learning, a large number of Deep Learning models have been applied in Wi-Fi sensing fields to automatically extract features and improve recognition performance. For Sleep monitoring, CBMR is based on CSI and uses Deep Learning model to realize single sleep monitoring, achieving better results [5]. In [36], authors combine CSI and Deep Learning model to realize number gesture recognition. In human activity recognition, Long short-term memory (LSTM) network was used to automatically extract features from CSI, which is superior to the traditional Machine Learning method of manually extract features [33]. ABLSTM model was used to extract features from the raw CSI data for action recognition in [6], achieving satisfactory performance.

To avoid hand-crafted feature extraction, we proposed a hybrid neural network model—BLSTM-TCN by combining BLSTM and TCN with deep learning knowledge. BLSTM simultaneously extracts features from the past and feature of human activity data to improve the classification performance of different human activities. TCN is a new type of time series feature extraction network and has good performance in some applications. In order to realize high-performance human activity recognition, we proposed MSHNet model, which first extract the amplitude from CSI, and then according to the MIMO characteristics of Wi-Fi signals, it aggregates and trains the data streams received by different antennas, and adopts a voting mechanism for the training results. The experimental results demonstrate that MSHNet we designed has very good performance not only on public dataset, but also on our own datasets.

3 MSHNet Architecture

This section introduces the overall process of MSHNet, as shown in Fig. 1. In Sects. 3.1 and 3.2, we focus on Wi-Fi signal characteristics and voting mechanism, In Sect. 3.3 network model construction and technologies are introduced.

3.1 Channel State Information

Channel State Information (CSI) is used to estimate channel attributes of communication links in orthogonal frequency division multiplexing (OFDM) technology [32]. Assuming physical space (including environmental objects and people) is described as a wireless channel, refraction, diffraction and scattering phenomena will occur when signals propagate in the wireless channel. Therefore, the received signals are the superposition of multipath signals. CSI data integrates the time delay, amplitude attenuation and phase shift effects of all signals propagating in wireless channels. In the frequency domain, a wireless channel having a plurality of transmit and receive antennas is described as $y = Hx + \theta$, where y, x, θ and H represent the reception vector, transmission vector, noise vector and channel matrix respectively. Channel matrix is the estimation of CSI.

In OFDM, CSI is presented in the form of subcarriers, and CSI of a single subcarrier can be described by the following mathematical expression $h = |h|e^{jsin\alpha}$, where $|h|$ represent amplitude and α represents phase. CSI provides a more fine-grained description of wireless channels.

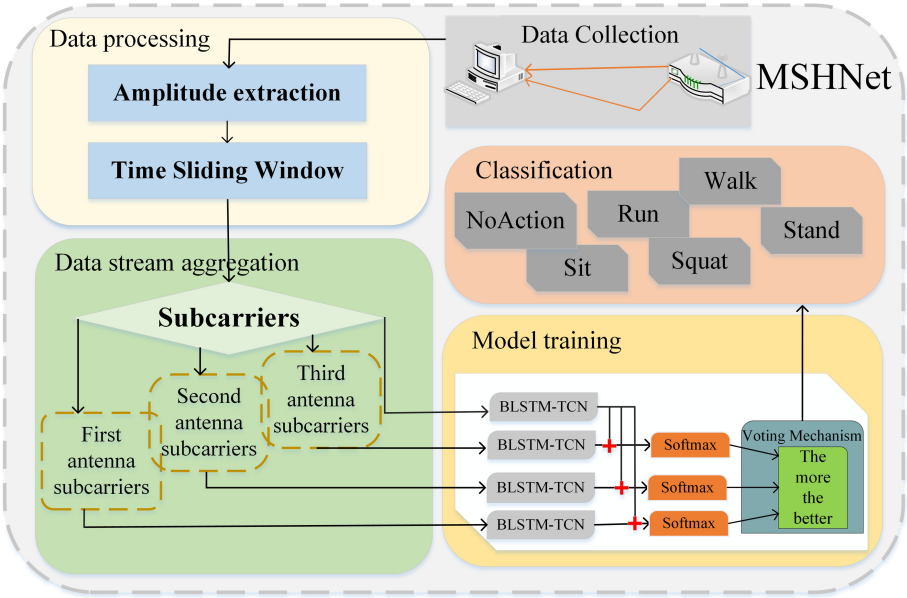


Fig. 1. The structure of MSHNet.

3.2 Multiple Input Multiple Output and Voting Mechanism

Multiple Input Multiple Output (MIMO) represents multiple transmit antennas and multiple receive antennas in a wireless signal transceiving system. MIMO uses multiple antennas to improve signal data throughput and transmission distance without increasing bandwidth and total transmission power. The CSI Tool used in this paper is a MIMO wireless signal under 802.11n standard. Figure 2 is a 2×3 MIMO system model. Assuming that there are N_t transmit antennas

Fig. 2. MIMO system model.

and N_r receive antennas in the MIMO system, all data streams (i.e., channel matrix H) in the wireless channel can be expressed as a matrix of $N_t \times N_r$:

$$H = \begin{pmatrix} H_{11} & \cdots & H_{1N_r} \\ \vdots & \ddots & \vdots \\ H_{N_t1} & \cdots & H_{N_tN_r} \end{pmatrix} \tag{1}$$

Under 802.11n standard, there are n CSI subcarriers in each data stream. All CSI subcarriers in each data stream can be expressed as

$$H_{N_tN_r} = \{h^{N_tN_r,1}, h^{N_tN_r,2}, \cdots, h^{N_tN_r,n}\} \tag{2}$$

refraction, diffraction and scattering phenomena will occur when signals propagate in wireless channel, resulting in multipath effect. Different data streams go through different paths, so that receive antennas receive different CSI [2,23].

The transmission and reception of Wi-Fi signals are carried out in MIMO mode, so multiple antennas receive multiple data streams. The amplitude information of the data streams corresponding to the three antennas is shown in Fig. 3. From the figure, it can be seen that the amplitude information of different data streams are all affected by human activity, moreover, owing to the subcarriers in the data streams tend to fade independently in OFDM, it provides an opportunity to study different data streams separately, and provides a basis for the voting mechanism to be used for the results of different antenna training. All the received data streams are treated as a whole in [6,33]. In this paper, the data received by different antennas are aggregated, and then the aggregated data streams are studied separately. MSHNet uses BLSTM-TCN model to train each data stream separately, and then uses voting mechanism to select the training results of all data streams. The experimental results reach the expectation, and the method provides a new idea for subsequent Wi-Fi signals research.

3.3 Network Model Construction

Bi-directional Long Short-Term Memory. Traditional Recurrent Neural Network (RNN) processing long sequence data will cause gradient vanishing and explosion. To solve those problems, Hochreiter and Schmidhuber proposed a variant of RNN called long short-term memory network (LSTM) in [15]. LSTM uses some gates with memory cells to solve the phenomena of gradient vanishing and explosion. LSTM introduces a connection of cell states, which are used to store information that needs to be memorized. Therefore, LSTM can handle the long-term information dependence in the sequence well. The internal structure of LSTM is shown in Fig. 4.

At time t, LSTM reads data C_{t-1}, h_{t-1}, and X_{t-1}, then updates and outputs information through forget gate f_t, input gate i_t, and output gate O_t in the cell. f_t decides which information to discard from the cell state. i_t has two functions, one is to find the cell state that needs updating, and the other is to update the

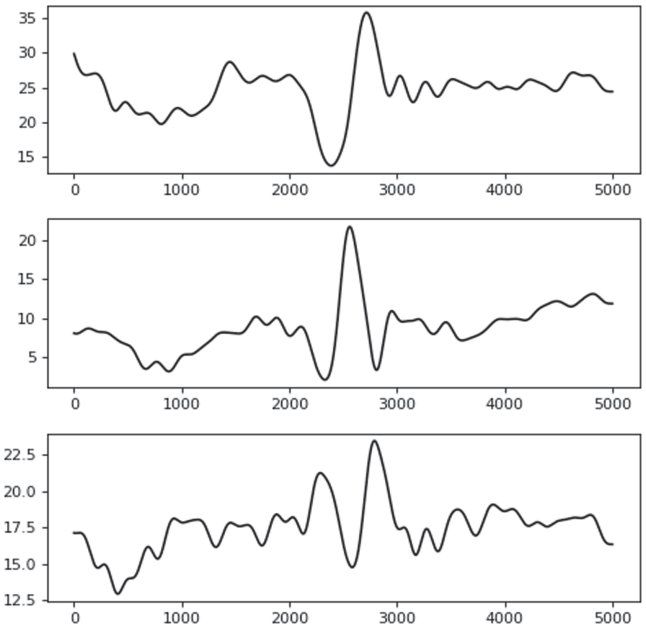


Fig. 3. Amplitude information of different antenna subcarriers. These three diagrams show that the amplitude information contained in the data stream received by different antennas is different, and each antenna has its own characteristics, which provides a basis for the voting mechanism to be used for the data stream training results of different antennas.

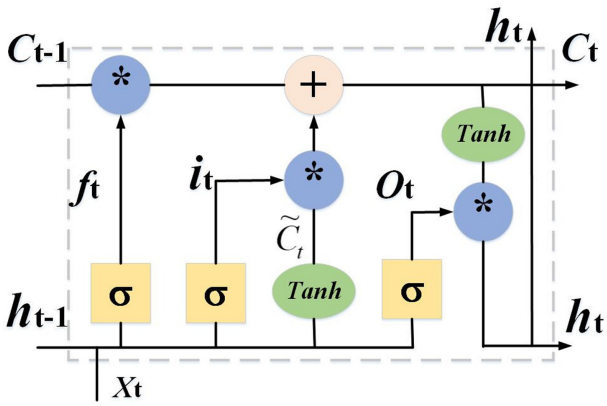


Fig. 4. The structure of LSTM.

information into the cell state. The internal mathematical calculation of LSTM can be described as

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, X_{t-1}] + b_f), \\ i_t &= \sigma(W_i \cdot [h_{t-1}, X_{t-1}] + b_i), \\ \widetilde{C}_t &= \tanh(W_c \cdot [h_{t-1}, X_{t-1}] + b_c), \\ C_t &= \sigma(f_t * C_{t-1} + i_t * \widetilde{C}_t), \\ O_t &= \sigma(W_o * [h_{t-1}, X_{t-1}] + b_o), \\ h_t &= O_t * \tanh(C_t). \end{aligned} \tag{3}$$

Where W_f, W_i, W_c, W_o and b_f, b_i, b_c, b_o respectively represent corresponding weight and bias. $\sigma(\cdot)$ and $\tanh(\cdot)$ represent sigmoid and tanh activation functions, respectively. h_t indicates the hidden state of cells. C_t represents the updated cell state.

Traditional RNN and its variants can only remember the past information when extracting data features. However, for sequence data, the feature information is also of great significance to the current moment. Therefore, Bi-directional recurrent neural network (Bi-RNN) is proposed [25], which can extract not only the past information but also the feature information, Bi-RNN can be defined as follows:

$$h_t = \overrightarrow{h_t} \bigoplus \overleftarrow{h_t} \tag{4}$$

Temporal Convolution Network. Temporal Convolution Network (TCN) is used to solve the classification of time series data. The validity of TCN in time series data classification is fully proved in [19]. In [19], let $X_t \in \Re^{F_0}$ be the input feature vector of length F_0 for time step t $(0 < t \leq T)$. Note that the length of time T is fixed in this paper. The true label for each time series is given by $y_t \in \{1, \cdots, C\}$, where C is the number of classes.

Consider L convolutional layers. We apply a set of 1D filters on each of these layers that capture how the input signals evolve over the course of an action. According to [19], the filters for each layer are parameterized by tensor $W^l \in \Re^{F_l \times d \times F_{l-1}}$ and biases $b^l \in \Re^{F_l}$, where $l \in \{1, \cdots, L\}$ is the layer index and d is the filter duration. For the l-th layer, the i-th component of the (unnormalized) activation $\hat{E}_t^l \in \Re^{F_l}$ is a function of the incoming (normalized) activation matrix $\hat{E}_{l-1}^l \in \Re^{F_{l-1} \times T}$ from the previous layer

$$\hat{E}_{i,t}^l = Relu({b_i}^l + \sum_{t'=1}^{d} \langle W_{i,t',.}^l, E_{.,t+d-t'}^{l-1} \rangle) \tag{5}$$

for each time t where $Relu(\cdot)$ is a Rectified Linear Unit.

When extracting data features, LSTM extracts potential information in chronological order, while TCN extracts potential information layer by layer. Considering the difference between the two, we designed a hybrid network—BLSTM-TCN as shown in Fig. 5. On the left side of Fig. 5, we use TCN to

extract data features, each 1D convolution layer is followed by Batch Normalization (BN) [17], Activation and Dropout layer [28]. BN layer prevents data distribution from changing during model training and accelerates the convergence and training speed [17], and Dropout layer prevents model from overfitting [28]. To make full use of the low-dimensional information and improve the model performance, BLSTM-TCN refers to the residual mechanism in [31] and concats the features extracted from the first TCN block with the features extracted from the second and third TCN blocks respectively. At the end of the left is the global average pooling, which regularizes the structure of the entire network to prevent overfitting, and reduces the data dimension and parameter amount [21]. On the right side of Fig. 5 is a Bi-directional Recurrent Neural Network based on LSTM, the purpose of which is to extract the past and future information of data simultaneously. The BLSTM-TCN model finally concats the features extracted from the left and right sides.

In the model training part of Fig. 1, the aggregated data streams are respectively input into BLSTM-TCN for training. In order to obtain best performance, all data streams are respectively concat with training results of different data streams after being trained by BLSTM-TCN, and then input into Softmax layer for classification. Finally, voting mechanism is used to select the results, if the results are different, the classification results of the data streams with the highest training accuracy are selected as the result.

4 Experiment and Evaluation

In this section, we experimentally confirm the effectiveness of the proposed MSHNet using real human activity data. Firstly, the dataset we used is described in Sect. 4.1. The evaluation metrics in Sect. 4.2. Finally, the experimental results are presented and discussed in detail in Sect. 4.3.

4.1 Dataset Description

Four datasets are used in this paper, one is public dataset, and the others are the datasets of different environments that we collected.

The public dataset [33] was collected in an indoor office area where the transmitter (Tx) and receiver (Rx) are 3 m apart in line-of-sight (LOS) condition. The Rx is equipped with a commercial Intel 5300 NIC. During data collection, each subject starts moving and doing an activity within a period of 20 s in LOS condition, while in the beginning and at the end of the time period the subject remains stationary. The whole data collection process is recorded by the camera to label the data. The dataset includes 6 persons, 6 activities, denoted as "Lie down, Fall, Walk, Run, Sit down, Stand up" and 20 trials for each one.

To comprehensively evaluate the impact of environment on MSHNet performance, we collected our own datasets. We considered four environments, i.e., Data Collection Room (DCRoom), Conference Room (CRoom), Exhibition Room (ERoom) and Psychology Consultation Room (PCRoom). The layout of

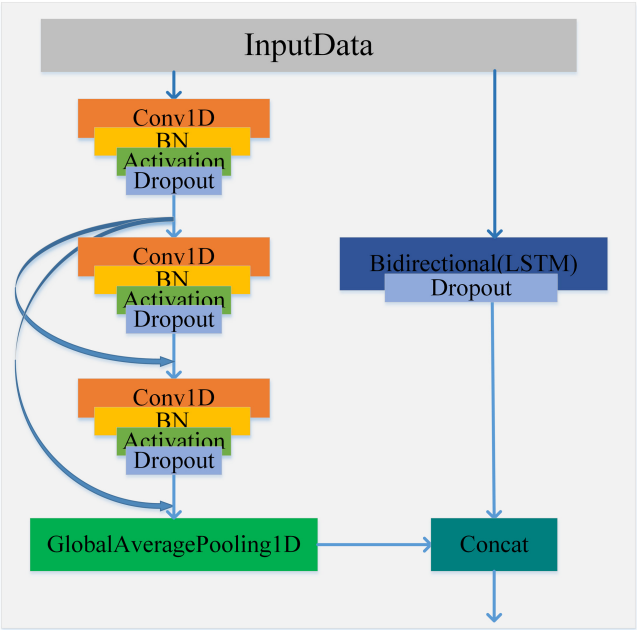
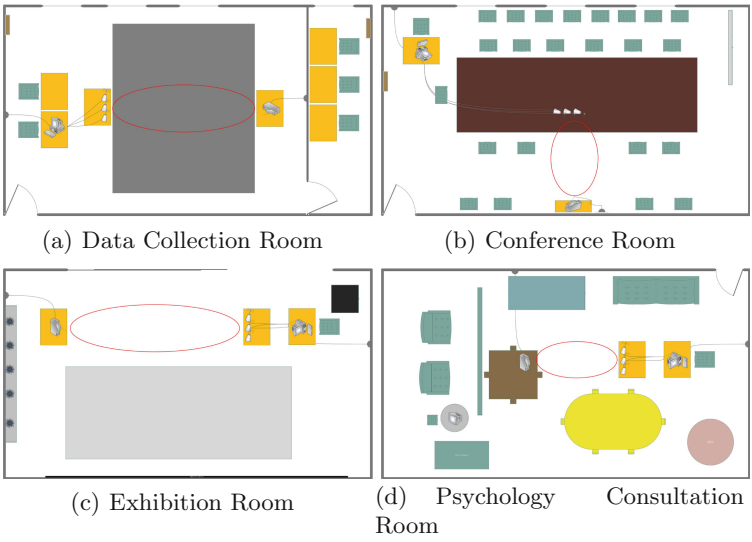


Fig. 5. The structure of BLSTM-TCN.

the four environments is shown in Fig. 6. We use the same equipment in the four environments. The TP_LINK AC1750 wireless router as the Tx, and a desktop computer equipped with an Intel 5300 802.11n Wi-Fi NIC as the Rx. In the process of data collection, each subject has been doing an action repeatedly. We take 5s as a segment to collect data. To make the data more diversified, we collected 10 subjects, each with six common activities, including "noAction, Walk, Run, Sit down, Stand up and Squat". Each action was repeated 10 times. For the case of noAction, we divide it into two kinds of situations, one is that there is someone in the environment and the other is no one. To improve the robustness of MSHNet, we arranged non-subjects to move freely on the non-LOS (outside the red circle in Fig. 6) during data collection.

The sampling rate of all datasets is 1 kHz and Rx has three antennas, each receiving 30 subcarriers, so the amplitude extracted from the raw CSI is 90 dimensions. Note that an Intel 5300 NIC can only be connected with three antennas, if multiple 5300 NICs are used to expand more antennas, it is difficult to achieve synchronous communication between different NICs, therefore, in this paper, we only use one 5300 NIC (i.e. only three receiving antennas). In addition, we use a time sliding window size of $T = 800$ ms to segment data in this paper.



(a) Data Collection Room (b) Conference Room

(c) Exhibition Room (d) Psychology Consultation Room

Fig. 6. Layout of the four environments.

4.2 Evaluation Metrics

To evaluate MSHNet performance, we use the following four metrics. The first is the most commonly used Accuracy, and others are Precision, Recall and F1. Take current category as a True or Positive (TP) example and non-current category as a False or Negative (FN) example. TP refers to the number of samples that current category is correctly judged as current category; FN refers to the number of samples that current category is judged as non-current category; FP refers to the number of sample that Non-current category is incorrectly judged as current category; and TN refers to the number of samples that non-current category is correctly judged as non-current category. The confusion matrix of TP, FN, FP and TN is shown in Table 1.

Table 1. The confusion matrix of TP, FN, FP and TN.

		Predict label	
		Positive	Negative
True label	True	TP	FN
	False	FP	TN

Accuracy indicates the proportion of both current category and non-current category that are correctly predicted, it can defined as

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}. \tag{6}$$

Precision is defined as the ratio of the number of correctly predicted as current category to the number of predicted as both current category and non-current category, which is computed as

$$Precision = \frac{TP}{TP + FP}. \tag{7}$$

Recall refers to the proportion that current category is truly predicted when the true label is current category, which is computed as

$$Recall = \frac{TP}{TP + FN}. \tag{8}$$

To avoid extreme situations in which the precision or recall is 1 and the other one is 0, the harmonic average of precision and recall, F1, is used to evaluate the performance of MSHNet, which is computed as

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{9}$$

The above formulas can solve a two-class problem, but a multi-class classification is required in this paper. Since the number of each huamn activity in our dataset is relatively balanced, the average F1 value can be defined as

$$F1 = \frac{2}{k} \frac{N_k}{N_{total}} \sum_{k} \frac{Precision_k \times Recall_k}{Precision_k + Recall_k}, \tag{10}$$

where k is class index of human activity, N_k is the number of samples of k-th class, and N_{total} is the total number of dataset. $Precision_k$ and $Recall_k$ are the Precision and Recall of human activity of k-th class, respectively.

4.3 Experimental Results

Comparison of Results of Different Structural Models. To choose the best model structure, this paper has done a series of comparative experiments. Comprises the following five experiments:

- Training all data streams by using BLSTM-TCN model (All streams);
- Without voting mechanism, i.e. the training results of different data streams are concated and then classified;
- The training results of all data streams are not concated on the training results of different data streams (Without all streams);
- Without Dropout;
- The low-dimensional features extracted from the first TCN block are not concated (Without residual).

It can be seen from the Table 2 that MSHNet has the best results on the four metrics of Accuracy, Precision, Recall and F1, so the paper finally determines MSHNet as the best model structure, and it also shows that adding Dropout, adding low-dimensional features and voting mechanism can improve the generalization performance and classification accuracy of the model.

Table 2. Comparison of results of different structural models.

	Precision	Accuracy	Recall	F1
All streams	0.9581	0.956	0.9606	0.9593
Without voting mechanism	0.9567	0.9508	0.9549	0.9558
Without all streams	0.9402	0.9275	0.9391	0.9396
Without Dropout	0.9694	0.9663	0.9706	0.97
Without residual	0.9627	0.9585	0.9641	0.9634
MSHNet	**0.9762**	**0.9741**	**0.9773**	**0.9767**

Comparison of Different Model Results on Public Dataset. The confusion matrix can represent the classification performance of the model for each category. This paper takes Long short-term memory [33] and ABLSTM [6] as benchmark experiments. Table 3, Table 4, and Table 5 are respectively the confusion matrices of Long short-term memory, ABLSTM and MSHNet on the public dataset.

From the confusion matrices, MSHNet not only performs well in the overall classification performance, but also achieves the best performance in each category, which mainly due to the fact that BLSTM and TCN show more powerful advantages in feature extraction of sequence data after mixing, and the voting mechanism we designed further improves the performance of MSHNet. The accuracy of Sit down in MSHNet's confusion matrix is slightly lower than other actions, because Sit down and Stand up are very similar in action characteristics.

Random Forest and Hidden Markov Model are mentioned in [33], both of which extract features manually and then use Machine Learning for training and classification. Table 3, Table 4, and Table 5 all use Deep Learning method for automatic feature extraction. The results are obvious, Automatic feature extraction by Deep Learning method can not only reduce labor cost but also improve classification performance.

Table 3. Confusion matrix of long short-term memory's classification results [33].

		Predict					
		Lie down	Fall	Walk	Run	Sit down	Stand up
Actual	Lie down	**0.95**	0.01	0.01	0.01	0.00	0.02
	Fall	0.01	**0.94**	0.05	0.00	0.00	0.00
	Walk	0.00	0.01	**0.93**	0.04	0.01	0.01
	Run	0.00	0.00	0.02	**0.97**	0.01	0.00
	Sit down	0.03	0.01	0.05	0.02	**0.81**	0.07
	Stand up	0.01	0.00	0.03	0.05	0.07	**0.83**

Table 4. Confusion matrix of ABLSTM's classification results [6].

		Predict					
		Lie down	Fall	Walk	Run	Sit down	Stand up
Actual	Lie down	**0.96**	0.00	0.01	0.00	0.02	0.01
	Fall	0.00	**0.99**	0.00	0.01	0.00	0.00
	Walk	0.00	0.00	**0.98**	0.02	0.00	0.00
	Run	0.00	0.00	0.02	**0.98**	0.00	0.00
	Sit down	0.01	0.01	0.01	0.00	**0.95**	0.02
	Stand up	0.01	0.00	0.00	0.00	0.01	**0.98**

Table 5. Confusion matrix of MSHNet's classification results.

		Predict					
		Lie down	Fall	Walk	Run	Sit down	Stand up
Actual	Lie down	**1.00**	0.00	0.00	0.00	0.00	0.00
	Fall	0.00	**1.00**	0.00	0.00	0.00	0.00
	Walk	0.00	0.00	**1.00**	0.00	0.00	0.00
	Run	0.00	0.00	0.00	**1.00**	0.00	0.00
	Sit down	0.00	0.00	0.00	0.00	**0.97**	0.03
	Stand up	0.00	0.00	0.00	0.00	0.00	**1.00**

Comparison of MSHNet Results in Different Environments. To verify the adaptability of MSHNet in different environments, this paper collects datasets in four environments: DCRoom, CRoom, ERoom and PCRoom, and trains with MSHNet. The four evaluation metrics (Accuracy, Precision, Recall, F1) of different environments are shown in Fig. 7 and Table 6 shows the accuracy and overall classification performance of each action in the four environments. As can be seen from Fig. 7 and Table 6, DCRoom has the best results, because DCRoom is an ideal experimental environment with the least obstacles and thus has the least influence on the action characteristics. There are relatively few non-experimental related items in CRoom and ERoom. The PCRoom experimental environment is the most complex, and there are many non-experimental related objects, so the results are relatively slightly worse. However, on the whole, static objects in different environments have little influence on the results, and the activities of non-subjects on non-LOS also have little influence.

The experimental results in different environments show that MSHNet has good adaptability in different environments, and has good performance for individual activity monitoring.

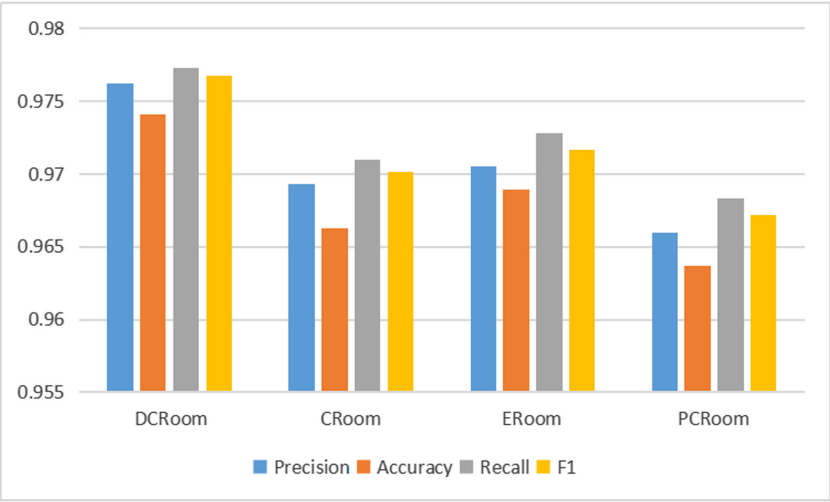


Fig. 7. Comparison of MSHNet results in different environments.

Table 6. Accuracy of different activities in different environments.

	noAction	Run	Sit down	Squat	Stand up	Walk	Overall
DCRoom	1.00	0.9722	1.00	1.00	0.9375	0.9538	0.9772
CRoom	1.00	0.9444	1.00	1.00	0.9125	0.9538	0.9684
ERoom	1.00	0.9583	1.00	1.00	0.9125	0.9538	0.9702
PCRoom	1.00	0.9305	0.9848	1.00	0.925	0.9583	0.9664

5 Conclusion

In this paper, an MSHNet deep neural network model is proposed for the daily health monitoring of elderly people who living alone and independently. The model not only achieves the best performance on public dataset, but also achieves good results on datasets collected in different environments by ourselves. However, MSHNet is only for single person, so in future work, we will mainly study human activity monitoring in multi-person situation. In addition, the paper does not use the phase information of CSI, so we will study the relationship between phase and human activity in the future. For multi-environment situations, we hope that the model trained in one environment can adapt to other environments quickly, which is also the focus of our future research.

References

1. Cisco Mobile, VNI, Cisco visual networking index global mobile data traffic forecast update 2016–2021, pp. 1–17. Cisco Visual Networking Index, San Jose, USA (2017)

2. Abdelnasser, H., Samir, R., Sabek, I., Youssef, M.: MonoPHY: mono-stream-based device-free WLAN localization via physical layer information. In: IEEE Wireless Communications and Networking Conference (WCNC 2013) (2013)
3. Abdelnasser, H., Youssef, M., Harras, K.A.: WiGest: a ubiquitous WiFi-based gesture recognition system. In: IEEE Conference on Computer Communications, pp. 1472–1480 (2015)
4. Bahl, P., Padmanabhan, V.N.: RADAR: an in-building RF-based user location and tracking system. In: Proceedings of the Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM 2000. IEEE (2000)
5. Cao, Y., et al.: Contactless body movement recognition during sleep via WiFi signals. IEEE Internet Things J. **7**(3), 2028–2037 (2019)
6. Chen, Z., Zhang, L., Jiang, C., Cao, Z., Cui, W.: WiFi CSI based passive human activity recognition using attention based BLSTM. IEEE Trans. Mob. Comput. **18**(11), 2714–2724 (2019)
7. Cui, Z., Chen, W., Chen, Y.: Multi-scale convolutional neural networks for time series classification. CoRR abs/1603.06995 (2016)
8. Du, J., Zheng, Z., Li, G., Ying, S., Ju, Z.: Gesture recognition based on binocular vision. J. Yangtze Univ. (3), 1–11 (2018)
9. Ghasemzadeh, H., Jafari, R.: Physical movement monitoring using body sensor networks: a phonological approach to construct spatial decision trees. IEEE Trans. Ind. Inform. **7**(1), 66–77 (2011)
10. Gu, Y., et al.: EmoSense: computational intelligence driven emotion sensing via wireless channel data. IEEE Trans. Emerg. Top. Comput. Intell. https://doi.org/10.1109/TETCI.2019.2902438
11. Gu, Y., Wang, Y., Liu, Z., Liu, J., Li, J.: SleepGuardian: an RF-based healthcare system guarding your sleep from afar. IEEE Netw. (2019). https://doi.org/10.1109/MNET.001.1900235
12. Gu, Y., Zhang, X., Liu, Z., Ren, F.: BeSense: leveraging WiFi channel data and computational intelligence for behavior analysis. IEEE Comput. Intell. Mag. **14**(4), 31–41 (2019)
13. Halperin, D., Hu, W., Sheth, A., Wetherall, D.: Tool release: gathering 802.11n traces with channel state information. ACM SIGCOMM Comput. Commun. Rev. **41**(1), 53–53 (2011)
14. Han, C., Wu, K., Wang, Y., Ni, L.M.: WiFall: device-free fall detection by wireless networks. IEEE Trans. Mob. Comput. **16**(2), 581–594 (2017)
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
16. Hong, T.K.N., Fahama, H., Belleudy, C., Pham, T.V.: Low power architecture exploration for standalone fall detection system based on computer vision (2015)
17. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. CoRR abs/1502.03167 (2015)
18. Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. IEEE Commun. Surv. Tutor. **15**(3), 1192–1209 (2013)
19. Lea, C., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks: a unified approach to action segmentation. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 47–54. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_7
20. Li, H., Yang, W., Wang, J., Xu, Y., Huang, L.: WiFinger: talk to your smart devices with finger-grained gesture. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (2016)

21. Lin, M., Chen, Q., Yan, S.: Network in network. Computer Science (2013)
22. Ma, C., Li, W., Gravina, R., Cao, J., Li, Q., Fortino, G.: Activity level assessment using a smart cushion for people with a sedentary lifestyle. Sensors **17**(10), 2269 (2017)
23. Perahia, E., Stacey, R.: Next Generation Wireless LANs: 802.11 n and 802.11 ac. Cambridge University Press, Cambridge (2013)
24. Ramakic, A., Bundalo, Z., Bundalo, D.: A method for human gait recognition from video streams using silhouette, height and step length. J. Circuits Syst. Comput. **29**(7), 2050101:1–2050101:18 (2020)
25. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Trans. Sig. Process. **45**(11), 2673–2681 (1997)
26. Seifeldin, M., Saeed, A., Kosba, A.E., El-Keyi, A., Youssef, M.: Nuzzer: a large-scale device-free passive localization system for wireless environments. IEEE Trans. Mob. Comput. **12**(7), 1321–1334 (2013)
27. Selvabala, V.S.N., Ganesh, A.B.: Implementation of wireless sensor network based human fall detection system **30**, 767–773 (2012)
28. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
29. Wang, T., et al.: Recognizing parkinsonian gait pattern by exploiting fine-grained movement function features. ACM Trans. Intell. Syst. Technol. **8**, 1–22 (2016)
30. Wang, W., Liu, A.X., Shahzad, M., Ling, K., Lu, S.: Device-free human activity recognition using commercial WiFi devices. IEEE J. Sel. Areas Commun. **35**(5), 1118–1131 (2017)
31. Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: a strong baseline, pp. 1578–1585 (2017)
32. Wu, K., Jiang, X., Yi, Y., Min, G., Ni, L.M.: FILA: fine-grained indoor localization. In: Proceedings of the IEEE INFOCOM, pp. 2210–2218 (2012)
33. Yousefi, S., Narui, H., Dayal, S., Ermon, S., Valaee, S.: A survey of human activity recognition using WiFi CSI. CoRR abs/1708.07129 (2017). http://arxiv.org/abs/1708.07129
34. Zhang, J., Wei, B., Hu, W., Kanhere, S.S.: WiFi-ID: human identification using WiFi signal. In: International Conference on Distributed Computing in Sensor Systems, pp. 75–82 (2016)
35. Zhao, M., et al.: Through-wall human pose estimation using radio signals. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7356–7365 (2018)
36. Zhou, Q., Xing, J., Li, J., Yang, Q.: A device-free number gesture recognition approach based on deep learning. In: 2016 12th International Conference on Computational Intelligence and Security (CIS) (2016)