IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 30, 2020, accepted July 12, 2020, date of publication July 27, 2020, date of current version August 6, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3012021

Temporal-Frequency Attention-Based Human Activity Recognition Using Commercial WiFi Devices

XIAOLONG YANG, (Member, IEEE), RUOYU CAO, MU ZHOU, (Senior Member, IEEE), AND LIANGBO XIE, (Member, IEEE)

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Corresponding author: Ruoyu Cao (ruoyucao@foxmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61771083 and Grant 61704015; in part by the Science and Technology Research Project of Chongqing Education Commission under Grant KJQN201800625; and in part by the Chongqing Natural Science Foundation Project under Grant cstc2019jcyj-msxmX0635.

ABSTRACT Human activity recognition has been growing for decades in a variety of technological disciplines. However, in the existing WiFi-based human activity recognition systems, there are the following problems: Firstly, in the processing of channel state information (CSI) data, mainly for the removal of noise in the superimposed signal, there is no effective removal of useless multipath signals; Secondly, the data segmentation algorithm based on the empirical threshold requires manual adjustment of the threshold in different environments, resulting in poor robustness and unstable segmentation; Thirdly, simple learning classification is applied without specific design for CSI data structure and sufficiently abstracting information features. In this paper, a device-free human activity recognition system with a temporal-frequency attention mechanism is proposed, which can be deployed on commercial WiFi devices to identify human's daily activities. Firstly, the multipath signal affected by the channel change is extracted by using the difference of the propagation delay of different multipath, thereby eliminating the delay and invalid multipath signals that have undergone multiple reflections and refractions. Secondly, a neural network model based on attention mechanism is proposed, which assigns different weights to different characteristics and sequences by imitating the human brain to dedicate more attention to important information. Then, the long short-term memory (LSTM) model is used to learn the correlation features of different dimensions to realize human activity recognition. Finally, the system performance is evaluated in different environments, and the experimental results show that our syetem holds a better performance in both line-of-sight (LOS) and non-line-of-sight (NLOS) than the existing human activity recognition systems.

INDEX TERMS Channel state information, multipath selection, human activity recognition, temporal-frequency attention.

I. INTRODUCTION

Human activity recognition is one of the most potential technologies at present, and plays an important role in human-computer interaction [1]–[7]. Such as smart home, security monitoring, medical assistance, etc. In previous studies, researchers have proposed a variety of human activity recognition systems that use different technologies, such as, methods based on wearable sensors [8], [9], methods based on computer vision [10], [11], methods based on environmental equipment [12]–[14], etc. However, these technologies have problems of inconvenience or environmental restrictions. In addition, special equipment needs to be deployed in the detection area to realize human activity recognition.

As WiFi devices have entered thousands of households, in recent years, WiFi-based human activity recognition technology has attracted the attention of researchers. The device-free human activity recognition system based on WiFi does not need any equipment carried by the detected target, and is not affected by the illumination environment,

The associate editor coordinating the review of this manuscript and approving it for publication was Qilian Liang.

and can even realize human activity recognition under the through-the-wall scenario, such as, Wi-Vi [15], CARM [16], TW-see [17], [18], etc. The use of wireless signals for behavior recognition has great application potential in complex environments. For example, in the fields of anti-terrorism rescue, fall detection and security monitoring, compared with other technologies, wireless signal-based behavior recognition has unique advantages.

In this paper, a device-free human activity recognition system based on temporal-frequency attention (WiTA) is proposed. To achieve WiTA, we mainly face three technical challenges. The first technical challenge is how to separate the effective multipath signals in the original channel state information (CSI). Since the CSI signal is a superposition of multipath signals, how to separate the multipath signals will be a challenge. The second technical challenge is how to cut effective motion fragments. Existing work can complete the segmentation of activities, but when the signal environment or the status of the transceiver changes, the traditional algorithm based on empirical thresholds needs to re-adjust the parameters. The third challenge is the effective extraction of correlation features. Although existing work is needed to obtain this correlation directly from the original CSI. However, in the process of pre-processing and feature extraction, the processing of effective information has destroyed its information integrity.

To overcome the first challenge, this paper proposes a relatively accurate and effective method for multipath signal extraction, which is used to isolate multipath signals related to human activities. WiTA implements the selection of multipath signals according to the different propagation delays of different multipath signals, eliminating the delay lag caused by equipment errors and multipath propagation delays, and the long-delay multipath signal that has been repeatedly reflected and refracted in the channel environment. In addition, in order to make full use of the CSI information, after the CSI error is removed by multipath selection, this paper uses linear transformations to reduce random noise on the phase, and removes the high-frequency channel noise contained in the amplitude by filtering at the same time. Through above processing, we extract the valid part of the effective multipath signal completely. In order to overcome the latter two challenges, unlike the existing work, a multi-layer Long Short-Term Memory (LSTM) network structure model with a temporal-frequency attention mechanism is designed. The change in CSI caused by human activity is a process that changes with the movement of the human body at different times. According to the changes with the time series, a time attention sub-network is designed, which can automatically assign different weights to the feature series according to the changes of the features, representing different contributions to the final classification result. The characteristics of the information contained in different sub-carriers are very different, but the change trends caused by the changes in human activities are generally similar. Finally, WiTA learns the forward and reverse laws through Bi-directional Long Short-Term Memory (Bi-LSTM), which has a higher feature learning efficiency.

Our work contributions are summarized as follows:

- In this paper, an adaptive multipath selection algorithm is proposed, which can extract effective multipath signals. The experimental results show that the multipath selection algorithm is effective and can meet the requirements of human activity recognition system.
- An end-to-end activity recognition method is designed to reduce the loss of effective information during data processing.
- A temporal-frequency attention network based on LSTM is proposed, which uses the attention mechanism instead of the traditional sequence segmentation algorithm, assigns different attention weights according to the value of the features of each dimension, and learn the relevance of special documents.

The rest of this paper is organized as following. A literature review of existing work in the second section is presented. The third section is the system overview. Section IV introduces the data pre-processing algorithms, including multipath selection and amplitude and phase feature processing. Section V introduces the framework of temporal-frequency attention network based on LSTM. Extensive experiments and evaluations are shown in Section VI. Finally, this paper is concluded in Section VII.

II. RELATED WORK

In this section, we summarize the related work in the field of human activity recognition in recent years.

A. WEARABLE SENSORS-BASED

Human activity recognition technology based on wearable sensors [2], [4], [8], [9] mainly obtains relevant data by using RFID [9], motion sensors [4], [8] and other devices [2]. Use motion sensors to acquire activity data for human activity recognition and so on. Motion sensor-based methods usually require users to wear sensors on some torsos of the body. Although these methods can capture activity information more accurately, they will bring inconvenience to users because users need to wear sensors.

B. COMPUTER VISION-BASED

Based on the computer vision method, visual features can be obtained through various visual equipment, and conduct activity recognition [3], [7], [10]–[14], [19]. It can recognize human activity by capturing relevant features of human activity from images. But computer vision has a large amount of data and complicated calculations. Bernardes Jr et al by using computer vision to identify human activity [19], it can be used to operate game interfaces, etc.

C. WIRELESS SIGNAL-BASED

One of the most universal methods of human activity recognition technology based on wireless signals [5], [6],

[15]–[18], [20]–[25] is to use ordinary household WiFi signals. The wireless signal based is low cost and easy to deploy.

The device-free human activity recognition system based on WiFi does not need any equipment carried by the detected target, and is not affected by the illumination environment, and can even realize human activity recognition under the through-the-wall scenario. Early activity recognition technology based on WiFi was mainly realized by using received signal strength indication (RSSI) [22]–[24]. However, RSSI is a coarse-grained signal belonging to WiFi, and information on channel changes reflected is seldom. With the development of WiFi sensing technology, CSI has entered the vision of researchers. CSI is a fine-grained physical layer information, including two data characteristics of amplitude and phase. Therefore, CSI has replaced RSSI as a more interesting research direction among researchers. For example, Wi-Vi [15] can realize object detection or human activity recognition by using Universal Software Radio Peripheral (USRP) to simulate WiFi signals and collect CSI. However, as a professional equipment, USRP is difficult to widely promote for its high cost. CARM [16] establishes the velocity and activity model based on the Doppler frequency shift, then performs human activity recognition. TW-see [17] proposes an or-PCA algorithm to sort out the low-rank matrix and sparse matrix of CSI data to complete the separation of dynamic data and static data. Zheng et al removes the data information without Doppler frequency shift for the purpose of motion detection [18], according to the principle that human dynamic will cause a Doppler frequency shift. Although these systems have the advantages of passive detection, but they are only coarse-grained separated of dynamic and static signals, and the processing of CSI data does not go deep into the multipath.

CSI data pre-processing of traditional human activity recognition systems mainly includes data pre-processing and feature extraction. Due to severe phase distortion of CSI, traditional human activity recognition systems generally discard phase information, and mainly analyze the change process of CSI amplitude [16], [17], [25]. According to changes of CSI amplitude characteristics, the CSI data are subjected to data preprocessing and then input to the classifier for learning classification. Phase information of CSI data is not effectively used. Data-preprocessing generally performs noise reduction on the CSI data by performing noise reduction processing on the multipath superimposed signal. However, this method only removes part of the noise contained in the multipath superimposed signal, and the invalid multipath part of the superimposed signal still exists. Therefore, in the traditional human activity recognition system, the invalid multipath channel for CSI has not been effectively dealt with. We know that the signal is mainly composed of valid information and invalid information, and the purpose of our signal processing is to remove invalid information and retain valid information. However, in the traditional human activity recognition system, due to insufficient removal of invalid information, feature extraction will also be performed based on the processed data to further reduce the impact of invalid information. Although the process of feature extraction reduces the content of invalid information in the extracted features, the integrity of valid information is also destroyed. Moreover, previous human activity recognition systems generally use simple machine learning classifiers, such as a k nearest neighbor (KNN) classification algorithms [26]–[28] and support vector machines (SVM) [25], [29], etc. There is no corresponding network model designed to learn data features and activity recognition according to the data structure of CSI. In real-world applications, this method of using a simple classifier cannot achieve satisfactory recognition accuracy.

In this paper, we will perform error elimination and noise reduction processing on CSI to remove the effects of invalid multipath signals and high-frequency noise in CSI. Then the amplitude and phase features contained in the CSI are completely extracted, and the feature information are directly used as the input features of the neural network to avoid the loss of effective information due to feature extraction and ensure the integrity of the effective information in the CSI features. Then a special deep learning network structure is designed according to the data characteristics of CSI, and an end-to-end high-precision behavior recognition system based on CSI is realized.

III. OVERVIEW OF WiTA

WiTA consists of two routing equipment, which act as transmitters and receivers, respectively. We use a transmitter with one antenna and a receiver with three antennas. In WiTA, access point (AP) and receivers are deployed in our conference room. If human activity occurs within the channel coverage that the transmitter can receive, the WiTA system can automatically detect and give the recognition result.

In WiTA, we mainly focus on the following issues: Firstly, extraction of multipath signals after human body reflection and refraction. After getting the original CSI measurement data, the multipath signal that does not pass through the human body should be removed. At the same time, the effective signal part is kept as complete and efficient as possible. Secondly, sequence segmentation. After extracting the effective signal, we must distinguish between the part of the signal sequence where the human body is active and the part where the human body is not active, and then focus on the sequence related to human activity. Thirdly, feature learning and activity recognition. According to different subcarriers in CSI, although they work on different frequencies and are slightly affected by channel changes, their overall trends over time are the same. By designing the corresponding model, according to their correlation and difference, the characteristics of different subcarriers are learned and a model of the human activity classifier is established.

In order to solve the above issues, the WiTA system is divided into two main modules, which contains a total of eight small modules, as shown in Fig. 1. The CSI data processing module includes a data acquisition module, a multipath

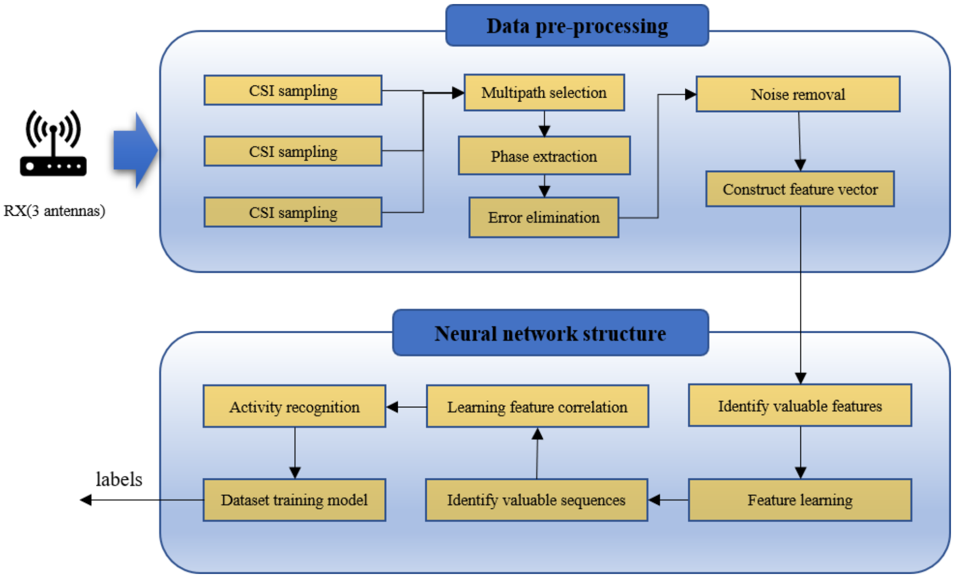

FIGURE 1. Framework of WiTA.

separation module, a feature extraction module and a noise removal module. The neural network module includes temporal attention module, frequency attention module, correlation feature learning and classification module. The data collection module mainly collects signals through the three antennas of the receiver. The receiver uses three antennas to sample the amplitude and phase of 30 sub-carrier signals of different frequencies. The multipath selection module mainly performs multipath separation according to the difference in the propagation delay in the channel caused by the different propagation paths of each multipath signal. We combine the principle of orthogonal frequency-division multiplexing (OFDM) modulation and the principle of the Fourier transform to calculate multipath signals with different delays, and then analyze the multipath signals to remove the multipath signals that are not related to human activity. The feature extraction module is mainly responsible for extracting amplitude and phase features in CSI. The noise cancellation module is mainly responsible for the removal of high-frequency noise in the effective multipath signal characteristics. The frequency attention module is mainly responsible for learning the difference between different sub-carrier features, and the time attention module is mainly responsible for solving the problem of poor robustness of traditional sequence segmentation methods. The correlation feature learning module is mainly responsible for learning the correlation features between feature sequences. a specially LSTM layer is designed to learn the nature of time-dependent features and the correlation features between different features. The classification module is responsible for distinguishing different human activities and outputting the recognition results.

IV. DATA PRE-PROCESSING

This section mainly describes the design details of the CSI data processing module, which consists of a data acquisition module, a multipath separation module, a feature extraction module and a noise removal module.

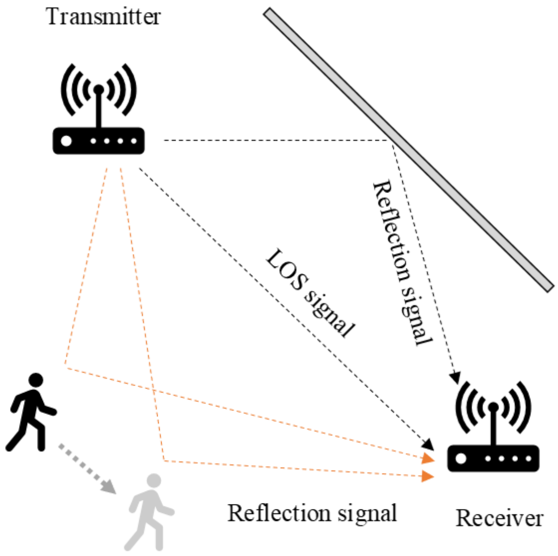

FIGURE 2. Schematic diagram of WiFi signal propagation path.

According to the signal processing theory, received signal can generally be expressed as the convolution of the transmitted signal and the channel impulse response, thus channel frequency response (CFR), also called CSI, can be obtained by Fourier transform of channel impulse response (CIR). As shown in Fig. 2, the line of-sight (LOS) signal has the shortest propagation delay through the LOS path. Other path signals have different path lengths due to the different refracted and reflected paths. Assume that $x(t)$ is the equivalent baseband signal, and there are N_p paths in the channel to reach the receiver. Equivalent baseband channel impulse response is used to model our wireless channel, the signal we receive can be equivalent to:

$$\begin{aligned} y(t) &= \int_{-\infty}^{+\infty} h(\tau, t)\, s(t-\tau)\, d\tau \\ &= \sum_{n=1}^{N_p} \kappa_n(t)\, e^{-j\phi_n(t)} s(t-\tau_n(t)). \end{aligned} \tag{1}$$

where $h(\tau, t)$ is the equivalent CIR at time t and $h(\tau, t) = \sum_{n=1}^{N_p} \kappa_n(t) e^{-j\phi_n(t)} \delta(\tau - \tau_n(t))$. $\kappa_n(t)$ represents the amplitude of the n^{th} path at time t, which is related to large-scale path loss and shadow fading. $\tau_n(t)$ represents the propagation delay of the n^{th} path, and $\phi_n(t)$ is the phase shift caused by the delay $\tau_n(t)$ and the Doppler frequency shift.$\delta(\tau - \tau_n(t))$ is impulse signal with delay $\tau_n(t)$. Especially, the path at $n = 1$ corresponds to LOS path. From equation (1), We can find that the response of different path signals in the propagation channel to the receiver through different delays can be expressed by CIR. Assume that the signal reaches the receiver through two paths, the delays of the signals in these two paths are $\tau_n(t)$ and $\tau_{n+1}(t)$. If the absolute difference between the propagation delays of these two multipath paths is greater than the discrete sampling interval of CIR, that is $|\tau_n(t) - \tau_{n+1}(t)| \geq 1/B_\omega$ (B_ω indicates channel bandwidth), multipath signals can be separated. For the WiFi, the signal transmission bandwidth is usually 20MHz or 40MHz. This means that the recognizable minimum delay difference of the multipath is 50ns or 25ns. the multipath signals are divided into $\overline{N_p}$ multipath components according to the minimum delay difference cluster. For multipath components with a delay difference less than the minimum delay resolution, we can ignore it because it exceeds the limit of WiFi resolution. Thus,

$$h(\tau, t) \approx \sum_{n=1}^{\overline{N_p}} \kappa_n(t) e^{-j\phi_n(t)} \delta(\tau - \tau_n(t)). \tag{2}$$

In the WiFi system, OFDM is used to divide the channel wideband selective fading into multiple overlapping, orthogonal narrowband flat fading channels, which not only weakens the impact of inter-symbol interference (ISI), but also greatly improves the wireless channel utilization. We can convert CIR to CFR through Fourier transform. $H(f, t)$ is defined as the Fourier transform of $h(\tau, t)$ at time t with respect to time delay τ:

$$\begin{aligned} H(f, t) &= F[h(\tau, t)] \\ &= \int_{-\infty}^{+\infty} h(\tau, t)\, e^{-j2\pi f \tau} d\tau \\ &= \int_{-\infty}^{+\infty} \sum_{n=1}^{N_p} \kappa_n(t) e^{-j\phi_n(t)} \delta(\tau - \tau_n(t))\, e^{-j2\pi f \tau} d\tau \\ &= \int_{-\infty}^{+\infty} \sum_{n=1}^{N_p} \varphi_n(f, t) \delta(\tau - \tau_n(t))\, e^{-j2\pi f \tau} d\tau \\ &= \sum_{n=1}^{N_p} \varphi_n(f, t) e^{-j2\pi f \tau_n(t)} \\ &\approx \sum_{n=1}^{\overline{N_p}} \varphi_n(f, t) e^{-j2\pi f \tau_n(t)}, \end{aligned} \tag{3}$$

where $F[\cdot]$ represents the Fourier transform operation and $\varphi_n(f, t)$ is a complex, which represents information such as the propagation attenuation and initial phase shift of a signal traveling through the n^{th} path. $-2\pi f \tau_n(t)$ represents the phase shift of the frequency component corresponding to the time delay $\tau_n(t)$. Practically, as shown in Fig. 2, commercially available WiFi equipment is used to collect CSI and recognize human activity by analyzing channel changes. The transmitter uses a fixed rate to continuously send WiFi signal data packets to the receiver. When the human body moves, the channel environment will change and the multipath signal passing through the human body will be affected as shown in Fig. 2. The receiver can distinguish small signal changes in the propagation path caused by human activities, and recognize human activities by monitoring changes in CSI. The CSI data we obtained is the sample data of $H(f, t)$.

Suppose that the wireless system is a narrow-band flat fading channel and we have

$$\mathbf{Y}_t = \mathbf{H}_t \mathbf{Z}_t + \mathbf{N}_t, \tag{4}$$

where $\mathbf{Y}_t$ and $\mathbf{Z}_t$ represent the data received by the receiver and the data sent by the transmitter at the t^{th} $(1 \leq t \leq T)$ packet,respectively. T represent the total number of received packets. $\mathbf{H}_t$ and $\mathbf{N}_t$ represent the CSI and noise, respectively. Let I and K represent the total number of antennas and subcarriers. The CSI is generally written as,

$$\mathbf{H}_t = \begin{bmatrix} csi_{t,1,1}, & \cdots, & csi_{t,1,k}, & \cdots, & csi_{t,1,K} \\ \vdots & & & & \\ csi_{t,i,1}, & \cdots, & csi_{t,i,k}, & \cdots, & csi_{t,i,K} \\ \vdots & & & & \\ csi_{t,I,1}, & \cdots, & csi_{t,I,k}, & \cdots, & csi_{t,I,K} \end{bmatrix}, \tag{5}$$

where $csi_{t,i,k}$ $(1 \leq i \leq I, 1 \leq k \leq K, 1 < t < T)$ represents the CSI of the k^{th} subcarrier on the i^{th} antenna in the t^{th} packet and $csi_{t,i,k} = |csi_{t,i,k}| e^{j\angle csi_{t,i,k}}$. $\angle csi_{t,i,k}$ denotes its phase. In order to facilitate the processing of CSI, we generally reconstruct the CSI data as:

$$\begin{aligned} \widetilde{\mathbf{H}}_t &= reshape(\mathbf{H}_t) \\ &= \left[csi_{t,1,1}, \cdots, csi_{t,1,K}, \cdots, csi_{t,I,K}\right]^T, \end{aligned} \tag{6}$$

where T is vector transpose operator. In order to analyze the change of the channel, we continuously detect the channel and collect the CSI.

The CSI contains not only the signal that is affected by the human body, but also the signals that have not passed through

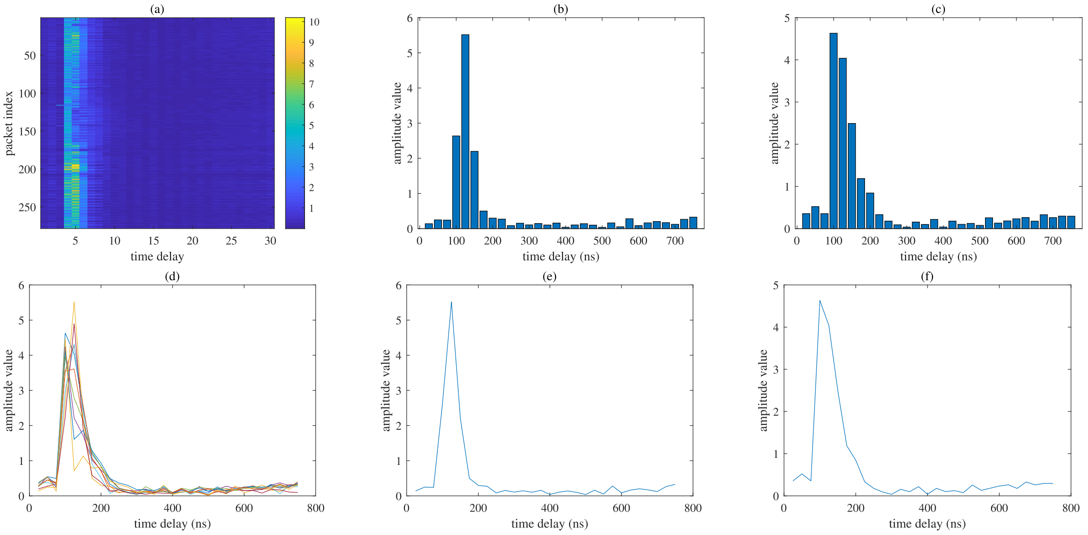

FIGURE 3. CIR of data packets: (a) and (d) show the changes in CIR of different data packets, (b) and (e) show the changes in CIR under the NLOS path, (c) and (f) show the changes in CIR under the LOS path.

the human body during the channel propagation process. Thus, the signal can be expressed as:

$$\widetilde{\mathbf{H}}_t = \sum_{i=1}^{n1} \mathbf{d}_{t,i} + \sum_{j=1}^{n2} \mathbf{d}'_{t,j}, \tag{7}$$

where $\mathbf{d}_{t,i}$ indicates the i^{th} multipath information that is not affected by human activity, and $\mathbf{d}'_{t,j}$ represents the j^{th} multipath information that is not affected by human activity. Note that $1 \leq n_1 \leq \overline{N_p}$, $1 \leq n_2 \leq \overline{N_p}$ and $1 \leq n_1 + n_2 \leq \overline{N_p}$. In order to get the data $\sum_{i=1}^{n1} \mathbf{d}_{t,i}$ affected by human activities in CSI, we propose a multipath selection algorithm to get the effective information related to human activity in CSI as efficiently and completely as possible.

Due to the influence of synchronization errors between devices and the propagation delay of the signal, there will be a certain delay lag in the CIR in the actual data obtained, and there are differences in the delay lag in the CIR obtained from different data packets [30]. According to previous related research [31], the maximum propagation delay of wireless signals indoors is less than 500ns and signals with a delay of more than 500ns are generally reflected signals from objects farther away from the transceiver, or signals that reach the receiver after multiple reflections and refractions in the channel. These multipath signals are meaningless for normal communication and human activity recognition, and they will cause random CSI jitter after superposition. We show the CIR in two different transmission and reception environments through Fig. 3.

As shown in Fig. 3, by analyzing the change trend of CIR through statistics, we find that when the signal of the LOS path arrives at the receiver, the amplitude will suddenly change, and the multipath amplitude after the mutation will gradually drop to a minimum value. This minimum path signal arrives about 500ns after the LOS signal arrives. Then the CIR gradually returns to a smooth fluctuation, which is very similar to the noise signal in the channel. From this we can deduce that in each CSI packet, the multipath of the CIR multipath amplitude between the abrupt value and the minimum value is the path we need. From (3), the CIR in our CSI is $\mathbf{h}_{t,i} = F^{-1}\left[\mathbf{H}_{t,i}\right] = \left(\mathbf{h}_{t,i,1}, \cdots, \mathbf{h}_{t,i,\Delta\tau}, \cdots, \mathbf{h}_{t,i,M}\right)$, $F^{-1}[\bullet]$ is the inverse Fourier transform. $\mathbf{h}_{t,i,\Delta\tau}$ represents the $\Delta\tau^{th}$ multipath information i^{th} antenna t^{th} packet. Therefore, the data within the required delay range is selected by the following calculation:

$$\Delta_{t,i,\Delta\tau} = \mathbf{h}_{t,i,\Delta\tau} - \mathbf{h}_{t,i,\Delta\tau-1}, \tag{8}$$

$$\Delta^{a}_{t,i,\Delta\tau} = \frac{\sum_{j=\Delta\tau-\frac{n}{2}}^{\Delta\tau+\frac{n}{2}} \mathbf{h}_{t,i,j}}{n+1}. \tag{9}$$

Fig. 4 is obtained by calculating (8) and (9), where n is the window length and set to be 6 based on experience. The maximum value of the first-order forward difference on the multipath signal is calculated to locate the starting point of the delay of multipath selection, which is $\max\left(\Delta_{t,i,\Delta\tau}, \forall(1 \leq \Delta\tau \leq M)\right)$.

Because of the influence of channel noise, the location of the minimum value cannot be directly determined, so the minimum point of the CIR moving average is used to determine the end of the delay for multipath selection, which is $\min\left(\Delta^{a}_{t,i,\Delta\tau}, \forall(1 \leq \Delta\tau \leq M)\right)$. Fig. 5 and Fig. 6 shows the effect comparison of the subcarrier amplitude of 5 consecutive CSI packets before and after multipath selection. It can be observed that after multipath selection, fluctuations in the original CSI amplitude affected by errors and noise become smoother.

Next, the amplitude and phase information of the CSI are calculated. The Butterworth filter is used to directly filter

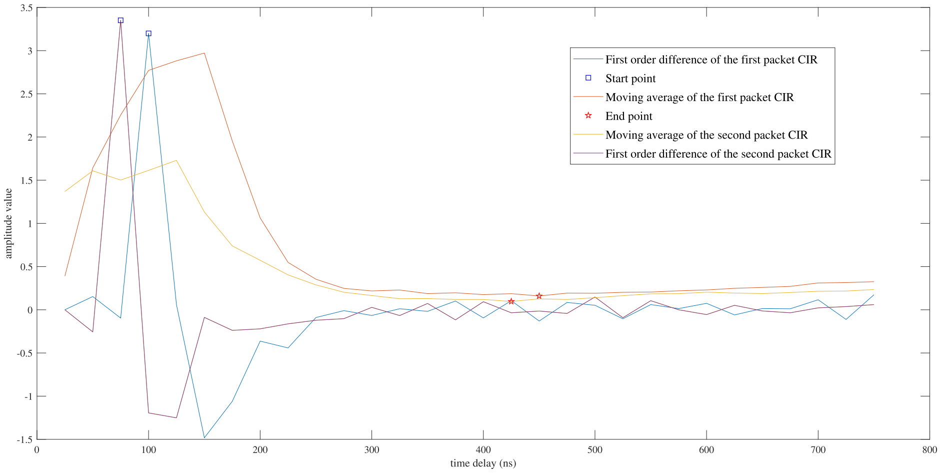

FIGURE 4. Delay range of multipath signals affected by humans.

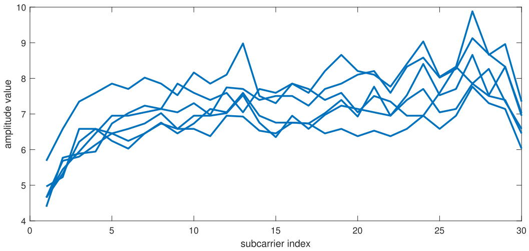

FIGURE 5. Subcarrier amplitude of original CSI.

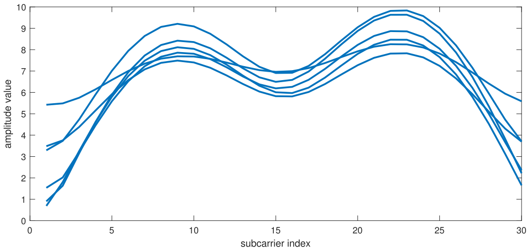

FIGURE 6. Subcarrier amplitude after multipath selection.

the amplitude information of CSI. Because the error and noise content in the phase information is much higher than the amplitude information, the phase information is obtained by improving PADS [32] and Wi-motion [33]. According to PADS, the phase $\hat{\phi}_k$ of the k^{th} subcarrier can be expressed as:

$$\hat{\phi}_k = \phi_k - 2\pi \frac{k}{K}\psi + \zeta + D, \tag{10}$$

where ϕ_k represents the true phase, and ψ is the timing offset of the receiver. It will cause a phase error as a middle term, and ζ is an unknown phase offset, and D is the measurement noise.

Although the index of the subcarriers is defined as -28 to 28 in the IEEE 802.11.n, we adjust the index to -15 to 15 for the reason that the WiFi driver has been modified to use the CSI information of 30 subcarriers. Compared with PADS, most δ and some D are removed through multipath selection, so more accurate CSI phase information can be gotten. Fig. 7 shows the results of our phase processing. After extracting the amplitude and phase information of the CSI, the Standardization algorithm is applied to uniformly normalize the amplitude and phase features to obtain the CSI normalized feature matrix $\mathbf{X}$ as:

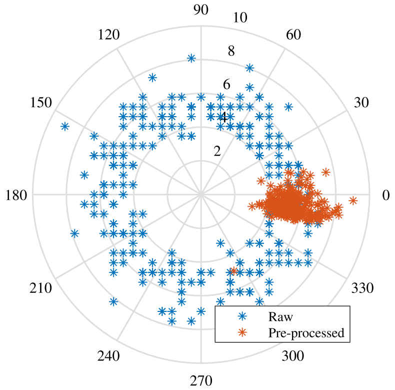

FIGURE 7. Phase before and after linear transformation.

$$\begin{aligned}\mathbf{X} &= \begin{bmatrix} \left[\left|\overline{\mathbf{H}}_1\right|, \cdots, \left|\overline{\mathbf{H}}_t\right|, \cdots, \left|\overline{\mathbf{H}}_T\right|\right] \\ \left[\angle\overline{\mathbf{H}}_1, \cdots, \angle\overline{\mathbf{H}}_t, \cdots, \angle\overline{\mathbf{H}}_T\right] \end{bmatrix} \\ &= \left(\mathbf{x}_1, \cdots, \mathbf{x}_t, \cdots, \mathbf{x}_T\right), \end{aligned} \tag{11}$$

among them:

$$\mathbf{x}_t = \begin{bmatrix} \left|\overline{\mathbf{H}}_t\right| \\ \angle\overline{\mathbf{H}}_t \end{bmatrix}, \tag{12}$$

$$\left|\overline{\mathbf{H}}_t\right| = \begin{bmatrix} \left|\overline{csi}_{t,1,1}\right| \\ \vdots \\ \left|\overline{csi}_{t,I,K}\right| \\ \vdots \\ \left|\overline{csi}_{t,I,1}\right| \\ \vdots \\ \left|\overline{csi}_{t,I,K}\right| \end{bmatrix}, \quad \angle\overline{\mathbf{H}}_t = \begin{bmatrix} \angle\overline{csi}_{t,1,1} \\ \vdots \\ \angle\overline{csi}_{t,I,K} \\ \vdots \\ \angle\overline{csi}_{t,I,1} \\ \vdots \\ \angle\overline{csi}_{t,I,K} \end{bmatrix}, \tag{13}$$

where $\overline{csi}_{t,1,1}$ and $\angle\overline{csi}_{t,1,1}$ respectively represent the amplitude and phase characteristics of the CSI of the t^{th} packet, i^{th} antenna, and k^{th} sub-carrier preprocessed by our proposed algorithm.

V. DESIGN OF ATTENTION NETWORK MODEL

In this section, we first introduce the basic structure of the LSTM neural unit of the network model we built, and then introduce each module in our neural network model in detail. When we humans observe something, we focus our attention on the things we want to observe, and we pay less attention to the surrounding environment. Based on this idea, the attention model is added to the neural network by imitating human attention. The attention mechanism can also be called a resource allocation method, so that more resources can be allocated to the features that need attention. The attention mechanism in the neural network can make the neural network have the ability to focus on the more favorable parts of the input features. Based on this idea, we design a neural network structure using the attention mechanism as the core, which can give more attention to the sequences and features required for human activity recognition. Then, in our main network, in order to better learn the sequence correlation feature in the sequence feature, the traditional structure is changed, and the learning dimension of the feature is transformed. Then use Bi-LSTM to learn the two-way law of sequence features to improve the recognition accuracy.

The recurrent neural network (RNN) andthe LSTM are firstly analyzed to make the paper stand alone. RNN is a popular model using sequential data modeling and feature extraction [34]. Fig. 8 shows the neurons of RNN. The output response at time sequence t is determined by the hidden layer weight generation $\mathbf{g}_{t-1}$ of the previous time sequence RNN neuron and the current time input:

$$\mathbf{g}_t = \mathbf{A}(\mathbf{w}_{xg}^T\mathbf{x}_t + \mathbf{w}_{gg}^T\mathbf{g}_{t-1}+\mathbf{b}_g), \tag{14}$$

where $\mathbf{A}$ represents a non-linear activation function, $\mathbf{w}_{xg}$ and $\mathbf{w}_{gg}$ represent learnable connection vectors, and $\mathbf{b}_g$ represents a bias value. This network structure of RNN is most suitable for solving the problem of continuous sequences, and is good at learning rules from samples with a certain correlation. Because the reverse error transmission of the general activation function can only maintain about 6 layers in the neural network. However, error transmission in RNN not only exists between layers, but also between the sample sequences of each layer, so the length of the sequence that RNN can learn is limited.

LSTM is a special type of RNN that can learn long-term dependent information [35]. LSTM avoids the problem of long-term dependence through deliberate design. Its structure is shown in Fig. 9. The core idea of this structure is to introduce a connection called cell state C_t to store what it wants to remember. At the same time, LSTM adds forget gate f_t, input gate i_t and output o_t gate, in order to make flexible choices in the status update and participation input in the LSTM.

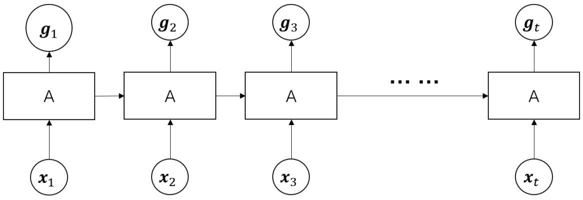

FIGURE 8. Structure of RNN.

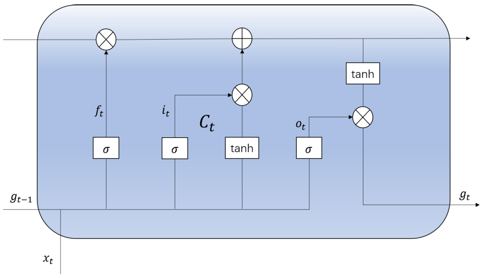

FIGURE 9. Structure of LSTM.

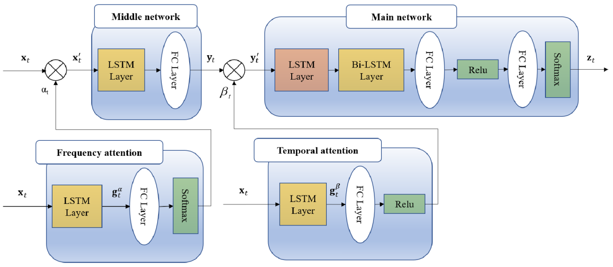

FIGURE 10. Structure of our proposed neural network model.

This paper proposes a multilayer LSTM network with temporal-frequency attention mechanism for human activity recognition. The model automatically selects the valuable subcarrier features in each sequence feature input through the frequency attention module, and assigns different attention to different time sequence through the time attention module. Fig. 10 shows the overall architecture, which is composed of an LSTM main network, middle network, a time attention subnetwork and a frequency attention subnetwork. α_t and β_t are the output of frequency attention sub-network and time domain attention sub-network, respectively, used to assign weights to sequences and features. y_t and z_t are the output of the middle network and the main network respectively. Since the four sub-networks interact with each other, training the network has certain challenges. In the following, we first discuss the structure of the attention subnetwork and the main network, and then introduce the learning and training process designed to overcome the difficulties of model training.

We propose a frequency attention model to automatically learn and utilize the different values of different subcarrier characteristics. By training the attention subnetwork, which can assign unique frequency attention weights to the subcarrier characteristics of each CSI sequence. This allows our model to better focus on features that are more sensitive to activity.

At each moment of step t, the set of subcarrier features is x_t, and a unique attention weight is assigned to each feature, and the output set of attention is α_t:

$$\begin{aligned}\alpha_t &= \mathit{soft}\max\left(\mathbf{W}_{g\alpha}\mathbf{g}_t^{\alpha} + \mathbf{b}_{\alpha}\right)\\ &= \left(\alpha_{t,1}, \alpha_{t,2}, \cdots, \alpha_{t,2K}\right),\end{aligned} \tag{15}$$

where $\mathbf{W}_{h\alpha}$ is a parameter matrix that can be learned, $\mathbf{b}_{\alpha}$ is a bias vector, and $\mathbf{g}_t^{\alpha}$ is a hidden variable of the LSTM layer. We use *softmax* as the activation function of the frequency attention subnet to get the frequency attention weight. This sub-network controls the amount of information that each feature flows to the main network. As shown in Fig. 10, the information entered into the next network module is $\mathbf{x}_t' = \alpha_t^{\mathrm{T}}\mathbf{x}_t$.

For time sequence, due to the difference in the value of information provided by different time sequence, only some sequences related to motion changes contain the most valuable information, while other sequences are used to provide channel environment information, etc. Based on this cognition, a time attention subnet is designed, using *Relu* with good convergence performance as the activation function, and assigning different attention β_t to different sequences:

$$\beta_t = \mathit{Relu}\left(\mathbf{W}_{g\beta}\mathbf{g}_t^{\beta} + \mathbf{b}_{\beta}\right), \tag{16}$$

which depends on the hidden output $\mathbf{g}_t^{\beta}$ of the LSTM layer at the current moment. For sequence information flow, similar to the frequency attention module, we calculate the information flowing to the main network from the output y_t of the LSTM middle network and the weight β_t of each time sequence. As illustrated in Fig. 10, the main network input has been adjusted to $\mathbf{y}_t' = \beta_t^T\mathbf{y}_t$.

Although different characteristics have different values for activity recognition, but they have a strong correlation. Therefore, we transform the LSTM network to use dimensions in the main network. Using the principle that the general LSTM network can extract the time correlation, the LSTM network is used to extract the correlation between different features of the sequence. As shown in Fig. 10, LSTM is used to learn the correlation between CSI features, and the correlation feature expression. Then, through the forward and reverse rules of the CSI sequence characteristics of the Bi-LSTM network, combining the forward and reverse will have a higher degree of fit than the LSTM network. For the classification of activity, we determine the score of each activity in all sequences based on the output $\mathbf{o} = \sum_{t=1}^{T}\mathbf{z}_t = (o_1, o_2, \cdots, o_C)$ of the main network, C is the number of actions we need to identify.

$$\mathbf{E} = \left(e^{o_1}, e^{o_2}, \cdots, e^{o_C}\right) = \left(E_1, E_2, \cdots, E_C\right), \tag{17}$$

$$\mathbf{p} = \frac{\mathbf{E}}{\sum_{j=1}^{C}E_j} = \left(p_1, p_2, \cdots, p_C\right). \tag{18}$$

where p_c represents the probability that the human body is performing the c^{th} action, and the activity with the highest final probability is the activity we recognize, namely $\max\left(p_1, p_2, \cdots, p_C\right)$.

We integrate temporal-frequency attention into the same network, as shown in Fig. 10, and learn and recognize through the main network. The sequence's regular cross-entropy loss function for the temporal-frequency attention network is applied:

$$\begin{aligned}L = -\sum_{i=1}^{c} l_i \log l_i' + \lambda_1\sum_{k=1}^{2K}\left(1 - \frac{\sum_{t=1}^{T}\alpha_{t,k}}{T}\right)^2\\ +\frac{\lambda_2}{2}\sum_{t=1}^{T}\|\beta_t\|^2 + \lambda_3\sum\|w\|.\end{aligned} \tag{19}$$

where l_i represents the label of the data, if it belongs to the i^{th} class, then l_i= 0, if not, then l_i= 0. l_i' means that the probability that the neural network predicts the i^{th} class is $l_i' = p_i$. $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters that control the degree of regularization. $\sum\|w\|$ represents the sum of all weight parameters in the network model.

The first regularization is to allow frequency-domain attention module to dynamically adjust the attention of different features. In tests, we find that the frequency attention model is easy to follow a sequence of training and fixedly focus on a certain feature. Some valuable features, while ignoring other features, are prone to overfitting, so regularization is introduced to avoid overfitting. The second regularization uses the L2 norm to regularize with the time attention of learning, reducing the disappearance of the back-propagation gradient. The third regularization uses the L1 norm to reduce the overfitting of the entire network.

Since temporal-frequency attention joint network is composed of four sub-networks, the four sub-networks interact during the training process, and some of the sub-networks may not be adequately trained due to the disappearance of the gradient, so a network step-by-step training method is proposed.

Step 1: Gauss initializes the network parameters of each sub-network.

Step 2: The frequency attention sub-network is combined with the middle network to train the frequency attention sub-network.

Step 3: Restore the parameters of the middle network.

Step 4: The middle network combines the time attention sub-network to train the time attention module.

Step 5: Restore the parameters of the middle network.

Step 6: Connect the four sub-networks, train the main network and fine-tune the remaining three sub-networks to obtain the entire model of convergence.

VI. RESULTS AND DISCUSSION

For network and parameter settings of our system, we set 180 LSTM neurons for the LSTM layer of the intermediate subnet and the temporal-frequency attention subnetwork, and

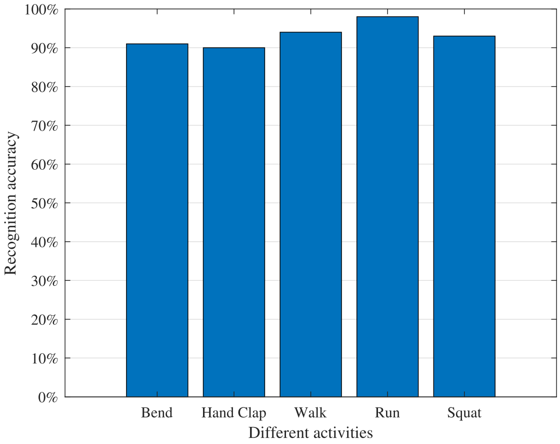

FIGURE 11. Test results using ***WiAR*** dataset.

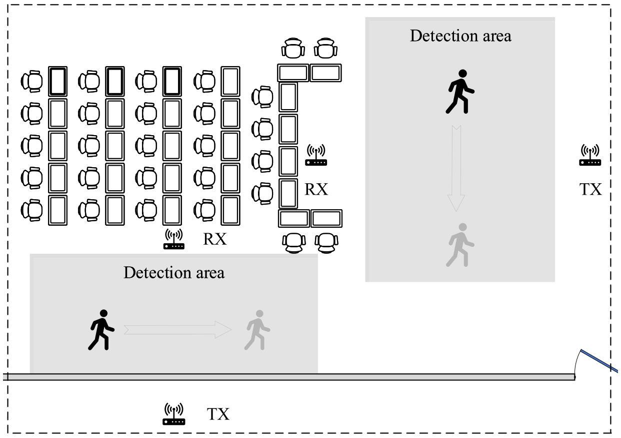

FIGURE 12. Test environment structure diagram.

500 LSTM neurons for the LSTM layer of the main network. Bi-LSTM has 50 LSTM neurons in the forward and backward layers, respectively. During the training and optimization process of collected data, through continuous testing and adjustment, the final adjustments of λ_1, λ_2, λ_3 are set to 0.01, 0.001, and 0.0004, respectively. Use dropout combined with regularization to alleviate the problem of overfitting, and use *Adam* [36] to automatically adjust the learning rate.

In order to verify the effectiveness of our proposed system, we first tested our system using the data set constructed by Guo et al in *WiAR* [37], which used a commercial TP-link wireless router as the transmitter and a ThinkPad notebook with three antennas as the receiver. The computer was equipped with an Intel 5300 network card and the network card firmware was modified. We extracted the six most common activity data in the dataset for testing. Fig. 11 shows our test results. It can be observed that the average recognition accuracy of our system is as high as 94%.

In order to test the performance of the system more comprehensively, we collected the data ourselves for testing. We collected activity data in two environments and demonstrated the experimental scenario in Fig. 12, which shows

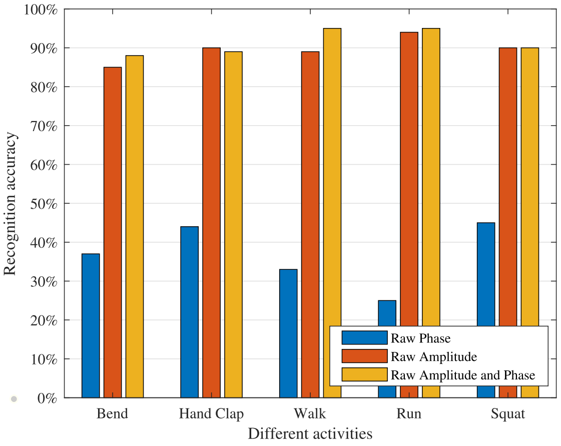

FIGURE 13. Row amplitude and phase as the feature accuracy.

the structure of the indoor LOS environment and NLOS environment. The former is in the conference room. There are many obstacles such as tables and chairs. It is a complex indoor environment with many multipath components. The latter is the channel environment that passes through the wall of the conference room and the receiver is in the conference room. The indoor real scene is shown in Fig. 12.

We use MSI PROBOX23 mini host with Intel 5300 network card, and use modified firmware to collect CSI on Ubuntu 14.04 system. As shown in Fig. 12, the transmitter uses a single antenna and the receiver uses three antennas.

In the two environments, we collected the data separately. Our database contains 5 activities, each of which is repeated in 100 groups, and 5 volunteers were invited to perform in two environments. At the same time, in order to distinguish whether the activity is performed and analyze the channel environment, we collected 100 sets of unmanned air environment data in each of the two environments.

In order to verify the effectiveness of our data preprocessing, that is, the validity of the model input features, we conducted separate tests using amplitude and phase features, respectively. Fig. 13 shows the average prediction accuracy of the three feature combinations without channel selection in our LOS environment under our network model. Fig. 14 shows the average prediction accuracy of the three feature combinations under our model in the LOS environment. We can find that our preprocessing of the data is effective and can improve the recognition performance of all activities tested.

Most human activity recognition systems today are implemented using simple classifiers (E.g. KNN, SVM and Random Forest). By using the same data, the proposed algorithm is compared with other common recognition algorithms to analyze the performance of the proposed system. From our experimental results, as shown in Fig. 15, the algorithm proposed in this paper has higher recognition accuracy than the existing conventional recognition method systems. The average recognition accuracy of the algorithm proposed in this paper is as high as 93%. At the same time, we find

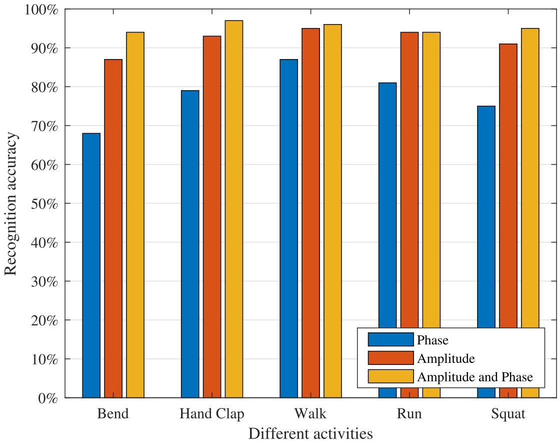

FIGURE 14. Pre-processed amplitude and phase as the feature accuracy.

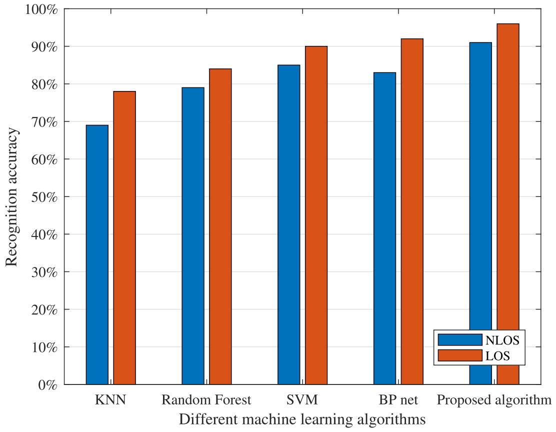

FIGURE 15. Performance comparison by five classification algorithms.

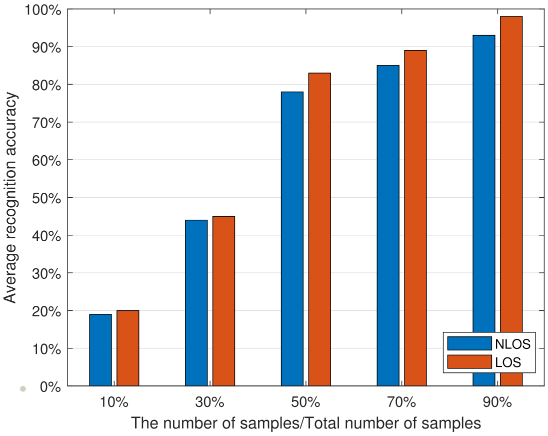

FIGURE 16. The effect of sample size on recognition accuracy.

that our proposed multipath selection algorithm has better performance than other algorithms in the NLOS environment due to the removal of delay and multipath.

In order to further test the performance of the system, we also tested the impact of the number of training samples on the performance. We test different performances of the system by changing the ratio of training data to our sample data set. After testing, we noticed that the number of samples had a positive relationship with the recognition accuracy, and Fig. 16 shows our experimental results. We can find that with the increase of training samples, the classification accuracy will also increase, mainly because the more training samples, the clearer and more accurate the model is for different types of activity. We can also speculate that if the sample data is further increased, the accuracy of the algorithm can be further improved.

VII. CONCLUSION

In this paper, a multipath selection algorithm is first proposed to extract the effective multipath information of CSI. It is adaptive in LOS and NLOS environments, especially in environments where there is more multipath information or where the LOS path propagation is blocked. Then, we propose a temporal-frequency attention network model based on the attention mechanism to perform feature learning and activity recognition on data, and use the attention mechanism to overcome the instability problem of existing data cutting algorithms. From our experimental results, our system has good performance in various experimental environments. The recognition accuracy under the LOS environment can reach 96.6%, and the recognition accuracy under the NLOS environment can reach 93%.

REFERENCES

[1] N. Twomey, T. Diethe, I. Craddock, and P. Flach, "Unsupervised learning of sensor topologies for improving activity recognition in smart environments," *Neurocomputing*, vol. 234, pp. 93–106, Apr. 2017.

[2] G.-M. Jeong, P. H. Truong, and S.-I. Choi, "Classification of three types of walking activities regarding stairs using plantar pressure sensors," *IEEE Sensors J.*, vol. 17, no. 9, pp. 2638–2639, May 2017.

[3] B. Boufama, P. Habashi, and I. S. Ahmad, "Trajectory-based human activity recognition from videos," in *Proc. Int. Conf. Adv. Technol. Signal Image Process. (ATSIP)*, May 2017, pp. 1–5.

[4] J. Wang, Z. Huang, W. Zhang, A. Patil, K. Patil, T. Zhu, E. J. Shiroma, M. A. Schepps, and T. B. Harris, "Wearable sensor based human posture recognition," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2016, pp. 3432–3438.

[5] J. Yang, H. Zou, H. Jiang, and L. Xie, "Device-free occupant activity sensing using WiFi-enabled IoT devices for smart homes," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3991–4002, Oct. 2018.

[6] H. Zou, Y. Zhou, J. Yang, H. Jiang, L. Xie, and C. J. Spanos, "DeepSense: Device-free human activity recognition via autoencoder long-term recurrent convolutional network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[7] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4263–4270.

[8] K. Yatani and K. N. Truong, "BodyScope: A wearable acoustic sensor for activity recognition," in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 341–350.

[9] D. Fortin-Simard, J.-S. Bilodeau, K. Bouchard, S. Gaboury, B. Bouchard, and A. Bouzouane, "Exploiting passive RFID technology for activity recognition in smart homes," *IEEE Intell. Syst.*, vol. 30, no. 4, pp. 7–15, Jul. 2015.

[10] G. Debard, P. Karsmakers, M. Deschodt, E. Vlaeyen, E. Dejaeger, K. Milisen, T. Goedemé, B. Vanrumste, and T. Tuytelaars, "Camera-based fall detection on real world data," in *Outdoor and Large-Scale Real-World Scene Analysis*. Springer, 2012, pp. 356–375.

[11] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim, "Robust human activity recognition from depth video using spatiotemporal multi-fused features," *Pattern Recognit.*, vol. 61, pp. 295–308, Jan. 2017.

[12] Q. Wu, W. Tao, Y. D. Zhang, and M. G. Amin, "Radar-based fall detection based on Doppler time–frequency signatures for assisted living," *IET Radar, Sonar Navigat.*, vol. 9, no. 2, pp. 164–172, Feb. 2015.

[13] S. Tao, M. Kudo, and H. Nonaka, "Privacy-preserved behavior analysis and fall detection by an infrared ceiling sensor network," *Sensors*, vol. 12, no. 12, pp. 16920–16936, Dec. 2012.

[14] Y. Li, K. C. Ho, and M. Popescu, "A microphone array system for automatic fall detection," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 5, pp. 1291–1301, May 2012.

[15] F. Adib and D. Katabi, "See through walls with WiFi!" in *Proc. ACM SIGCOMM Conf. SIGCOMM*, 2013, pp. 75–86.

[16] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of WiFi signal based human activity recognition," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 65–76.

[17] X. Wu, Z. Chu, P. Yang, C. Xiang, X. Zheng, and W. Huang, "TW-see: Human activity recognition through the wall with commodity Wi-Fi devices," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 306–319, Jan. 2019.

[18] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with Wi-Fi," in *Proc. 17th Annu. Int. Conf. Mobile Syst., Appl., Services*, Jun. 2019, pp. 313–325.

[19] J. L. Bernardes, Jr., R. Nakamura, and R. Tori, "Design and implementation of a flexible hand gesture command interface for games based on computer vision," in *Proc. 8th Brazilian Symp. Games Digit. Entertainment*, 2009, pp. 64–73.

[20] X. Liu, M. Jia, X. Zhang, and W. Lu, "A novel multichannel Internet of Things based on dynamic spectrum sharing in 5G communication," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 5962–5970, Aug. 2019.

[21] X. Liu and X. Zhang, "NOMA-based resource allocation for cluster-based cognitive industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5379–5388, Aug. 2020.

[22] H. Zhu, F. Xiao, L. Sun, R. Wang, and P. Yang, "R-TTWD: Robust device-free through-the-wall detection of moving human with WiFi," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 5, pp. 1090–1103, May 2017.

[23] H. Zhu, F. Xiao, L. Sun, X. Xie, P. Yang, and R. Wang, "R-PMD: Robust passive motion detection using PHY information with MIMO," in *Proc. IEEE 34th Int. Perform. Comput. Commun. Conf. (IPCCC)*, Dec. 2015, pp. 1–8.

[24] C. Li, H. J. Yang, F. Sun, J. M. Cioffi, and L. Yang, "Multiuser overhearing for cooperative two-way multiantenna relays," *IEEE Trans. Veh. Technol.*, vol. 65, no. 5, pp. 3796–3802, May 2016.

[25] H. Yan, Y. Zhang, Y. Wang, and K. Xu, "WiAct: A passive WiFi-based human activity recognition system," *IEEE Sensors J.*, vol. 20, no. 1, pp. 296–305, Jan. 2020.

[26] M. Zhou, Y. Wang, Y. Liu, and Z. Tian, "An information-theoretic view of WLAN localization error bound in GPS-denied environment," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 4089–4093, Apr. 2019.

[27] M. Zhou, Y. Wang, Z. Tian, Y. Lian, Y. Wang, and B. Wang, "Calibrated data simplification for energy-efficient location sensing in Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 6125–6133, Aug. 2019.

[28] C. Li, H. J. Yang, F. Sun, J. M. Cioffi, and L. Yang, "Adaptive overhearing in two-way multi-antenna relay channels," *IEEE Signal Process. Lett.*, vol. 23, no. 1, pp. 117–120, Jan. 2016.

[29] X. Qi, G. Zhou, Y. Li, and G. Peng, "RadioSense: Exploiting wireless communication patterns for body sensor network activity recognition," in *Proc. IEEE 33rd Real-Time Syst. Symp.*, Dec. 2012, pp. 95–104.

[30] S. Sen, B. Radunovic, R. R. Choudhury, and T. Minka, "You are facing the Mona Lisa: Spot localization using PHY layer information," in *Proc. 10th Int. Conf. Mobile Syst., Appl., Services*, 2012, pp. 183–196.

[31] I. V. Korogodin and V. V. Dneprov, "Impact of antenna mutual coupling on WiFi positioning and angle of arrival estimation," in *Proc. Moscow Workshop Electron. Netw. Technol. (MWENT)*, Mar. 2018, pp. 1–6.

[32] K. Qian, C. Wu, Z. Yang, Y. Liu, and Z. Zhou, "PADS: Passive detection of moving targets with dynamic speed using PHY layer information," in *Proc. 20th IEEE Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Dec. 2014, pp. 1–8.

[33] H. Li, X. He, X. Chen, Y. Fang, and Q. Fang, "Wi-motion: A robust human activity recognition using WiFi signals," *IEEE Access*, vol. 7, pp. 153287–153299, 2019.

[34] D. Britz. (2015). *Recurrent Neural Networks Tutorial, Part 1—Introduction to RNNS*. Accessed: May 2, 2019. [Online]. Available: http://www.wildml.com/2015/09/recurrent-neural-networkstutorial-part-1-introduction-to-rnns/

[35] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," in *Proc. 9th Int. Conf. Artif. Neural Netw.*, 1999, pp. 850–855.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[37] L. Guo, L. Wang, J. Liu, W. Zhou, and B. Lu, "HuAc: Human activity recognition using crowdsourced WiFi signals and skeleton data," *Wireless Commun. Mobile Comput.*, vol. 2018, pp. 1–15, Jan. 2018.

XIAOLONG YANG (Member, IEEE) received the M.Sc. and Ph.D. degrees in communication engineering from the Harbin Institute of Technology, in 2012 and 2017, respectively. From 2015 to 2016, he was a Visiting Scholar with Nanyang Technological University, Singapore. He is currently a Lecturer with the Chongqing University of Posts and Telecommunications. His current research interests include wireless sensing, indoor localization, and cognitive radio networks.

RUOYU CAO was born in Henan, China, in 1994. He received the B.S. degree in communication engineering from the Chongqing University of Posts and Telecommunications, in 2017, where he is currently pursuing the master's degree in information and communication engineering with the School of Communication and Information Engineering. His main research interests include wi-fi through-wall radar target detection, activity recognition, and deep learning.

MU ZHOU (Senior Member, IEEE) received the Ph.D. degree in communication and information systems from the Harbin Institute of Technology (HIT), China, in 2012. He was a Joint-Cultivated Ph.D. Student with the University of Pittsburgh (PITT), USA. He was also a Postdoctoral Research Fellow with The Hong Kong University of Science and Technology (HKUST), China. He joined the Chongqing University of Posts and Telecommunications (CQUPT), China, where he has been a Full Professor with the School of Communication and Information Engineering, since 2014. He was supported by the Chongqing Municipal Program of Top-Notch Young Professionals for Special Support of Eminent Professionals. Over the past five years, he was engaged in five national projects, nine provincial and ministerial projects (including two major projects), and seven enterprise projects. He has published more than 100 peer-reviewed research articles. His main research interests include wireless localization and navigation, signal reconnaissance and detection, convex optimization, and deep learning. He served on the technical program committees of the IEEE ICC, GLOBECOM, WCNC, IWCMC, VTC, IWCMC, and so on. He received the Outstanding Cooperation Project Award from Huawei Technologies Company Ltd.

LIANGBO XIE (Member, IEEE) was born in Sichuan, China, in 1986. He received the B.S. and M.S. degrees from Chongqing University, in 2007 and 2010, respectively, and the Ph.D. degree from the University of Electronic Science and Technology of China (UESTC), in 2016. He is currently an Associate Professor with the School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include ultra-low-power analog circuits, low-power SAR ADC, low-power digital circuits, anti-collision algorithm for RFID, and indoor-localization.