LSTM-CNN network for human activity recognition using WiFi CSI data

Shunuo Shang¹,a, Qingyao Luo²,b, Jinjin Zhao³,c, Ran Xue⁴,d, Weihao Sun⁵,e and Nan Bao⁶,f*

¹²³⁴⁵⁶ College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, Liaoning, 110169, China
email: ²a20185504@stu.neu.edu.cn,
b²20185451@stu.neu.edu.cn,
c²20185467@stu.neu.edu.cn,
d²20185434@stu.neu.edu.cn,
e²20185452@stu.neu.edu.cn,
f³baonan@bmie.neu.edu.cn

Abstract. Human Activity Recognition (HAR) has had a diverse range of applications in various fields such as health, security and smart homes. Among different approaches of HAR, WiFi-based solutions are getting popular since it solves the problem of deployment cost, privacy concerns and restriction of the applicable environment. In this paper, we propose a WiFi-based human activity recognition system that can identify different activities via the channel state information from WiFi devices. A special deep learning framework, Long Short-Term Memory-Convolutional Neural Network (LSTM-CNN), is designed for accurate recognition. LSTM-CNN is going to be compared with the LSTM network and the experimental results demonstrate that LSTM-CNN outperforms existing models and has an average accuracy of 94.14% in multi-activity classification.

1. Introduction
In recent years, with the rise of smart homes [1,2], game entertainment [3] and intelligent medical treatment, Human Activity Recognition (HAR) technology has been applied in numerous fields and has attracted intensive attention. Some obvious benefits have been seen especially in the area of health care. The global aging problem is becoming increasingly serious, and the fall of the elderly has become a serious public health problem [4]. Studies have confirmed that the earlier falls are reported, the lower the morbidity and mortality are [5,6]. However, when the human activity monitoring system is used, the fall of the elderly can be detected and warned real-early and forthwith [7]. In addition to fall detection, this monitoring technology can also detect routine movements of the elderly to obtain their physical status indicators and evaluate the health level, reduce the risk of death, and become the key to intelligent nursing in the future [8].

Nevertheless, there are hinde rs of traditional human body sensing technologies since they usually require some special devices (e.g., wearable sensors [9,10]). As a result, it is generally inconvenient for users to wear or carry sensors all the time, and it also limits the fields that can be applied. Besides, some
video-based behavior recognition methods [11,12] may also trigger privacy concerns or some of them may be affected by environmental factors such as the indoor light condition. Compared with these approaches, WiFi-based human behavior recognition technologies have become a cutting-edge research topic because of their convivence and low-cost infrastructure. With the wide deployment of the commodity WiFi devices installed at home, the device-free behavior recognition technology based on WiFi has received a wide range of attention. It can also solve the problems of deployment cost, privacy infringement and limitation of the application environment.

WiFi-based technology usually includes two kinds of wireless signals, namely RSS (Radio Signal Strength) and CSI (Channel State Information) [13]. RSS describes coarse-grained information about the communication link, while CSI provides fine-grained information about the state of the communication channels. Currently, the most commonly used signal is RSS, which has been widely used in indoor localization [14], tracking [15], as well as Radio Tomographic Imaging (RTM) [16]. However, due to multipath fading and time dynamic characteristics, RSS cannot perform well in complex situations. The other kind of wireless signal, CSI, can be represented as a complex matrix, with each entry recording the amplitude and phase response of the signal transmission channel for each subcarrier. We can calculate the amplitude and phase of the received signal for each channel from a complex number and evaluate the quality of the channel. Thus, the amplitude of a CSI signal demonstrates the signal power attenuation caused by the multipath effect. This has a wider field of application, for example, respiration detection [17], gesture recognition [18] and human behavior recognition [19] and has achieved great performance. Thus, this article concentrates on HAR with CSI-based WiFi signals.

At present, most activity recognition applications combine deep learning algorithm to classify different activities. Deep learning algorithm has several advantages. To be specific, it enables a large amount of data to be processed at the same time. Besides, it can achieve good performance by automatically extracting effective features then feeding them into a neural network and classifying samples. Traditional classification methods include the Support Vector Machine (SVM) [20, 21], the Long Short-Term Memory (LSTM) [22, 23] and the convolutional neural network (CNN) [24, 25]. However, they are not powerful or intelligent enough to extract informative features. To address this issue, we proposed a novel CSI-based system with a deep learning framework, Long Short-Term Memory-Convolutional Neural Network (LSTM-CNN) which integrates both temporal features and spatial features to distinguish activities accurately.

Specifically, our main contributions are summarized as follows:
- We leverage the temporal stability and frequency diversity of CSI and design a CSI-based human activity recognition system.
- We propose a novel deep learning framework Long Short-Term Memory-Convolutional Neural Network (LSTM-CNN) for recognition. It can automatically extract features from both temporal and spatial domains and effectively classify different activities.
- We have conducted extensive experiments with commodity WiFi devices in the indoor environment under various situations. Our system outperforms most current human activity recognition systems.

2. Proposed approach

2.1. Data collection and pre-processing
As for the underlying operating theory, CSI collection consists of a WiFi signal transmitter and a CSI receiver $N_{TX}$ transmitting antennas, $N_{RX}$ receiving antennas, and $N_{SC}$ CSI subcarriers. The transmitter adopts a commercial WiFi device with 5GHz frequency band signals. The receiver adopts the csitool [26] to collect CSI. Finally, we can extract a total of $N_{TX} \times N_{RX} \times N_{SC}$ CSI streams from each data pack. The detection range of the WiFi signal between the transmitter and the receiver constitutes the space for human body motion detection so as to recognize the human activity.
The experiment carries out HAR including two static movements of standing and sitting, and three dynamic movements of falling, standing up and stepping. The five types of activities are observed from five volunteers in a classroom. The existing pieces of literature [10] have proposed that the amplitude of CSI has a positive effect on distinguishing between dynamic and static movements. The phase difference of CSI can more effectively judge the variety of specific human movements in static and dynamic, because the dynamic movements can cause the phase difference fluctuation over time. For example, standing up and falling cause the phase difference to display an obvious conversion from the fluctuation state to the stable one. Therefore, extracting the amplitude and the phase difference from the raw CSI can improve the model classification performance better. We contrast them by inputting CSI with 2 channels and 4 channels, which produce various classification results. The input with 2 channels is the product of CSI amplitudes and the phase differences detected by two antennas and the input with 4 channels is added CSI amplitudes detected by two antennas.

However, the raw CSI collected is quite noisy so that cannot be identified well. It is mainly high-frequency components in the signal. Hence, we use the Butterworth low pass filter to denoise. The valuable features can be extracted from the raw CSI and then converted into corresponding multi-dimensional data with multiple channels to become the form to meet the neural network model.

![Figure 1. The CSI amplitude of the stepping before and after the denoising scheme.](image)

2.2. LSTM-CNN neural network

As a variant of Recurrent Neural Network (RNN), LSTM is suitable for processing time series and it can handle the continuous temporal relationship well. CNN can extract the spatial features and reduce frequency domain changes.[28] The LSTM-CNN model combines the advantages of the two models. It is important that for CSI data input, different variants of LSTM have different effects. We compare LSTM and LSTM-CNN with concrete result to human activity recognition. The structure of LSTM-CNN is shown in figure 2.

As it can be seen, the input passes through the LSTM layer. The information of input is captured from the signal fluctuation changes, and the time-domain features of the original signal are then extracted. The output of LSTM passes through a 2D-CNN convolutional layer. High-dimensional implicit feature extraction is performed by achieving convolution. Next, the results of the convolution pass through a maximum pooling layer to acquire the best feature sequence, which is used as a direct basis for the final classification. Finally, the predicted probability of each type of movement is obtained by using the MLP classifier.

The following section explains data collection, pre-processing and each layer of the network model.
2.2.1. LSTM.
LSTM can tackle the exploding and vanishing phenomenon in gradients as a variant of RNN [29]. Specially, LSTM can extract the hidden relationships of time series, and process the potential time information of the data. It has a great advantage of processing the time features of signals [30]. The basic LSTM unit is shown in figure 3. Each LSTM unit has an input gate layer, a forget gate layer, and an output gate layer to achieve the update and output of the memory state. The formulas of the LSTM unit are as follows:

\[
f_t = \sigma(W_f \times [h_{t-1}, x_t] + b_f) \\
i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \\
\tilde{c}_t = \tanh(W_c \times [h_{t-1}, x_t] + b_c) \\
c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t \\
o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_o) \\
\hat{h}_t = o_t \times \tanh(c_t)
\]

Where: \(\sigma\) is the logistic sigmoid function; \(\tanh\) is the hyperbolic tangent function; \(h_t\) is the output of the unit; \(b\) is the bias vector; \(W\) is the weight matrix.

The input CSI consists of a 400-length sequence of the 10-batch size. According to the classification result of LSTM on CSI, the input passes through a hidden layer with the number of hidden units setting to 28 to abstract temporal features.

Proper regularization is necessary for network training. Due to the complex training structure of LSTM, the parameters of the previous layer easily affect the changes in the parameters of the latter layer. Any increase in the number of network layers will make the model more complex, so that creates effects on the distribution of results. To prevent overfitting, we add a dropout layer at the end of each layer. The dropout rate is set to 0.2.

The output of the LSTM layer is a type of list data with the size of \(b \times s \times r\). The output of LSTM needs dimensional promotion to form matrix \(Q\) as the input of the CNN layer, because CNN can only accept input of 4-dimensional data. The size of \(Q\) is \(b \times s \times r \times 1\). \(b\) is batch size. \(s\) is the time step. \(r\) is the number of hidden units.
2.2.2. CNN
Convolutional neural network (CNN) is currently one of the most popular architectures for deep learning, which can automatically extract deep high-dimensional features, instead of being limited to some shallow ones. Inspired by [31], we believe that CNN and LSTM are complementary operations in the process of deep learning. LSTM is suitable for time-domain analysis and has a better reaction of short-duration movement, while CNN works mainly on changes in frequency-domain and has a better reaction of long-duration movement. In addition, convolution on a consecutive time series can extract the dominant information [32]. In the convolutional layer, 5 convolution kernels are used for convolution and the size of each convolution kernel is 30×28. The convolution kernel slides on the sequence with a step size of 1 to achieve feature recognition of different position vectors. A filter \( M \) convolves with the vectors validly, through a maxpooling layer, to generate the feature map \( S \) as:

\[
\begin{align*}
    c_j &= ReLU(q_{j,j+k-1} \times m + b) \\
    s_j &= Max(c_j)
\end{align*}
\]

Where \( q_{j,j+k-1} \) is the continuous \( k \) vectors in \( Q \), \( m \) and \( b \) are the weights and bias in the convolution process respectively, the size of \( c_j \) is 371×1, \( s_j \) is the element in \( S \). \( ReLU \) function is used as nonlinear activation function in the convolution layer. The results of the 5 filters are connected to get \( S \), which is used as the direct basis for classification. At this time, we can follow the general design idea of CNN and make the final classification prediction in the output layer. In this part, multi-layer perceptron (MLP) can be used to conduct activity classification.

3. Result

3.1. Data and evaluation metrics
In this section, we try to use several indicators such as accuracy, loss, confusion Matrices to evaluate the two-classification and multi-classification performance of the model. A total of 2800 groups of data for training and testing were collected in the laboratory, including 1020 groups of static data (540 groups for standing and 480 groups for sitting) and 1780 groups of dynamic data (640 groups for falling, 540 groups for standing up and 600 groups for stepping).

3.2. LSTM-CNN model parameter optimization
For the collected WiFi signal data, we select two channels (two parameters of signal amplitude product and phase difference of two antennas) and four channels (signal amplitude of two antennas and their amplitude product and phase difference) input parameters into LSTM-CNN network for binary
classification. When setting the step size to 400 and the number of iterations to 50, the accuracy and loss of the training and verification process are shown in figure 4. It can be seen that the training result of the four-channel input model is better than that of the two-channel input model. The loss value lowers at 0.25 after 42 iterations, and the accuracy rate is stable at 89% after 40 iterations. The highest accuracy rate of the verification set is 90.85%. There is no big difference in accuracy and loss between the training data set and the verification data set. Therefore, over-fitting is limited within the acceptable range.

![Two-channel progress over iterations](image1)

![Four-channel progress over iterations](image2)

Figure 4. Accuracy and loss curves of two-channel and four-channel inputs.

In machine learning, two of the important parameters are the batch size and the learning rate. It can be seen from figure 5, selecting different vectors will result in different classification performances. We finally shift the vector learning rate to 0.001, batch size to 10. As a result, the classification of network at this time shows the best effect. When selecting the parameter iterative for 70 times, the rate tends to be stable, without a significant fluctuation.

![Histogram of classification effect of different parameter models](image3)

(a) (b)

Figure 5. Histogram of classification effect of different parameter models.

3.3. Comparison with other models

In order to verify the classification performance of LSTM-CNN network, we compare it with LSTM network. Both of them have the same number of LSTM hidden layers for input data, and both use TensorFlow package on Python platform. As shown in table 1, LSTM network has better classification effect. The average accuracy rate of binary classification is 94.97%, the loss value is 0.1380, and all indexes are better than LSTM network. The average accuracy of multi-classification is 94.14%, which is slightly lower than the accuracy of two-classification of this network, but better than that of LSTM network by 11.42%. To sum up, the proposed LSTM-CNN network is a simple, efficient and high classification accuracy method.

The confusion matrix of the final results of LSTM-CNN model two classifications and multi-classifications is shown in figure 6. Obviously, the classification effect of falling action is the best, with an accuracy rate of 95.63%, followed by standing action, with an accuracy rate of 94.82%. Among them,
however, the classification effect of "sitting" is relatively poor. It can be seen that some "sitting" actions are divided into "standing", probably because the signals of the two have no obvious classification characteristics and the amplitude is more similar. However, on the whole, the classification accuracy rate is stable at over 93%, which means the classification effect is good.

Table 1. Comparison of LSTM model and other model binary classification results.

| Model     | Average accuracy | Average Loss | Precision | Recall  | F1-score |
|-----------|------------------|--------------|-----------|---------|----------|
| LSTM      | 89.84%           | 0.2354       | 88.98%    | 87.64%  | 88.13%   |
| LSTM-CNN  | 94.97%           | 0.1380       | 91.49%    | 91.55%  | 91.52%   |
| KNN       | 84.29%           |              | 86.93%    | 90.54%  | 84.34%   |
| SVM       | 77.01%           |              | 80.59%    | 75.86%  | 78.05%   |

Table 2. Comparison of LSTM model and other model multi-classification results.

| Model     | Average accuracy | Average Loss | Precision | Recall  | F1-score |
|-----------|------------------|--------------|-----------|---------|----------|
| LSTM      | 82.72%           | 0.5081       | 80.70%    | 79.82%  | 79.95%   |
| LSTM-CNN  | 94.14%           | 0.1269       | 91.03%    | 83.81%  | 87.10%   |
| KNN       | 74.13%           |              | 76.65%    | 74.12%  | 72.28%   |
| SVM       | 71.33%           |              | 78.96%    | 71.02%  | 72.97%   |

Figure 6. The normalized confusion matrix of the LSTM-CNN model for binary classification and multi-classification.

4. Conclusions
In this paper, we have compared architectures of the network models of LSTM and LSTM-CNN. We have compared and contrasted their specific results of HAR for CSI data inputs with two channels and four channels. In addition, we have set up a complete CSI data collection system and perform a series of data pre-processing. The LSTM-CNN can extract the detailed hidden details of time and spatial in the CSI data comprehensively. It shows the ability and reliability of the LSTM and CNN models to process time series CSI data. The classification result indicators of the LSTM-CNN model can prove that it is better than the other models. Therefore, it has a significant effect on the HAR of CSI data.

Acknowledgements
Gratitude and appreciation to fellow lecturers and students who have provided guidance and support in this research.
References

[1] Abdelnasser, H., Youssef, M., & Harras, K. A. (2015) Wigest: A ubiquitous wifi-based gesture recognition system. In: 2015 IEEE conference on computer communications (INFOCOM). Hong Kong. pp. 1472-1480.

[2] Yang, Z., Wu, C., Zhou, Z., Zhang, X., Wang, X., & Liu, Y. (2015) Mobility increases localizability: A survey on wireless indoor localization using inertial sensors. ACM Computing Surveys (Csur), 47(3): 1-34.

[3] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., ... & Blake, A. (2011) Real-time human pose recognition in parts from single depth images. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs. pp. 1297-1304.

[4] Cai, L., Lin, C., Zhou, L., & Fang, Y. (2018) Research progress on elderly care using generative and discriminative models. Personal and ubiquitous computing, 14(6): 489-498.

[5] Noury, N., Fleury, A., Rumeau, P., Bourke, A. K., Laighin, G. O., Rialle, V., & Lundy, J. E. (2007) Fall detection-principles and methods. In: 2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Lyon. pp. 1663-1666.

[6] Van Kasteren, T. L. M., Englebienne, G., & Kröse, B. J. (2010) An activity monitoring system for elderly care using generative and discriminative models. Personal and ubiquitous computing, 14(6): 489-498.

[7] Maurer, U., Smailagic, A., Siewiorek, D., & Deisher, M. (2006) Activity recognition and monitoring using multiple sensors on different body positions. In: Proceedings of the International Workshop on Wearable and Implantable Body Sensor Networks. Cambridge, Massachusetts. pp. 99-102.

[8] Parkka, J., Ermes, M., Korpiopaa, P., Mantyjarvi, J., Peltola, J., & Korhonen, I. (2006) Activity classification using realistic data from wearable sensors. IEEE Transactions on information technology in biomedicine, 10(1): 119-128.

[9] Bodor, R., Jackson, B., & Papanikolopoulos, N. (2003) Vision-based human tracking and activity recognition. In: Proceedings of the 11th Mediterranean Conference on Control and Automation. Greece. pp. 18-20.

[10] Wilson, J., & Patwari, N. (2010) See-through walls: Motion tracking using variance-based radio tomography networks. IEEE Transactions on Mobile Computing, 10(5): 612-621.

[11] Liu, X., Cao, J., Tang, S., Wen, J., & Guo, P. (2015) Contactless respiration monitoring via off-the-shelfWiFi devices. IEEE Transactions on Mobile Computing, 15(10): 2466-2479.

[12] Raja, M., Ghaderi, V., & Sigg, S. (2018) WiBot! In-vehicle behaviour and gesture recognition using wireless network edge. In: 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS). Vienna. pp. 376-387.

[13] Ohara, K., Maekawa, T., & Matsushita, Y. (2017) Detecting state changes of indoor everyday objects using Wi-Fi channel state information. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, 1(3): 1-28.
[20] Wang, Y., Wu, K., & Ni, L. M. (2016) Wifall: Device-free fall detection by wireless networks. IEEE Transactions on Mobile Computing, 16(2): 581-594.

[21] Wang, H., Zhang, D., Wang, Y., Ma, J., Wang, Y., & Li, S. (2016) RT-Fall: A real-time and contactless fall detection system with commodity WiFi devices. IEEE Transactions on Mobile Computing, 16(2): 511-526.

[22] Yousefi, S., Narui, H., Dayal, S., Ermon, S., & Valaee, S. (2017) A survey on behavior recognition using WiFi channel state information. IEEE Communications Magazine, 55(10): 98-104.

[23] Feng, C., Arshad, S., Yu, R., & Liu, Y. (2018) Evaluation and improvement of activity detection systems with recurrent neural network. In: 2018 IEEE International Conference on Communications (ICC). Kansas. pp. 1-6.

[24] Zhou, Q., Xing, J., Chen, W., Zhang, X., & Yang, Q. (2018) From signal to image: enabling fine-grained gesture recognition with commercial Wi-Fi devices. Sensors, 18(9): 3142.

[25] Jiang, W., Miao, C., Ma, F., Yao, S., Wang, Y., Yuan, Y.,... & Su, L. (2018) Towards environment independent device free human activity recognition. In: Proceedings of the 24th Annual International Conference on Mobile Computing and Networking. New Delhi. pp. 289-304.

[26] Halperirr, D., Hu, W., Sheth, A., & Wetherall, D. (2011) Tool release: gathering 802.11n traces with channel state information. Acm Sigcomm Computer Communication Review, 41(1): 53-53.

[27] Wang, H., Zhang, D., Wang, Y., Ma, J., Wang, Y., & Li, S. (2017) Rt-fall: a real-time and contactless fall detection system with commodity wifi devices. IEEE Transactions on Mobile Computing, 16(2): 511-526.

[28] Hammerla, N. Y., Halloran, S., & Ploetz, T. (2016) Deep, convolutional, and recurrent models for human activity recognition using wearables. Journal of Entific Computing, 61(2): 454-476.

[29] Hochreiter, S., Schmidhuber, J. (1997) Long short-term memory. Neural Computation, 9(8): 1735-1780.

[30] Zhang, J., Wu, F., Wei, B., Zhang, Q., Huang, H., Shah, S. W., & Cheng, J. (2020) Data augmentation and dense-lstm for human activity recognition using wifi signal. IEEE Internet of Things Journal.

[31] Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., & Schmidhuber, J. (2016) LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10): 2222-2232.

[32] Kalchbrenner, N., Grefenstette, E., Blunsom, P., Kartsaklis, D., Kalchbrenner, N., Sadrzadeh, M.,... & Blunsom, P. (2014) A convolutional neural network for modelling sentences. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore. pp. 212-217.