Human Activity Recognition based on Collaboration of Vision and WiFi Signals

Shibo Li[1,2], Yao Ge[2], Minjian Shentu[1], Shuyuan Zhu[1], Muhammad Imran[2]
Qammer Abbasi[2], Jonathan Cooper[2]
[1]University of Electronic Science and Technology of China, Chengdu, China
[2]James Watt School of Engineering, University of Glasgow, Glasgow, UK

***Abstract*—In the WiFi protocol, channel state information (CSI) is the modulated as the fine-grained data to assess the channel efficiency. Meanwhile, it contains the information about the environment change, including the movement of human in a specific environment. Therefore, the CSI data can be used to recognize the human activity. In this paper, we design a vision and WiFi collaboration-based human activity recognition scheme to classify the human activities. More specifically, we collect the CSI data from the WiFi signals and the human skeleton points from the video signals. Then, we construct a long-short-term Transformer network to build up the collaboration of the CSI data and the skeleton points. Based on this collaboration, we can use the CSI data to well recognize the human activities.**

***Index Terms*—WiFi, channel state information, human activity recognition, long-short-term, Transformer**

I. Introduction

WiFi is one of the significant communication protocols in Internet of Things system. It can be used to link the internet and detect the environment change with channel state information (CSI). Therefore, WiFi signal can be utilized to recognize the movement of human in a specific environment. With the development of WiFi protocol and commercial off-the-shelf (COTS) devices, it brings more and more potential applications of the WiFi-based human activity recognition (HAR) [1], [2]. HAR aims to identify the actions of human, which can be adopted in the design of human computer interaction and remote healthcare monitoring systems.

In the propagation of WiFi signal, the transmitted CSI data is able to record the characteristics of physical space with multi-path effect. When a person is in the physical space, the additional sub-carrier paths will be introduced due to the signal reflection and diffraction occurred to the human body. The CSI data contains the information about the change of sub-carrier paths. By establishing the information mapping of CSI and human movement, the WiFi-based HAR can be implemented.

The WiFi-based HAR has a number of advantages. Firstly, the propagation of WiFi signals is not limited by the line-of-sight area. The radio frequency of WiFi, normally at 2.4GHz and 5GHz, makes the signals can be easily penetrated through the blocking object for transmission. Secondly, it is not affected by the lighting condition, which allows it to monitor the human activities independent to the light. Thirdly, this solution protects the private information of users. For instance, the user's appearance is not visualized and cannot be recognized or restored by the CSI data either. Finally, the adopted hardware to implement it is simple and it does not need any other equipments.

Recently, many researches focus on the human activity recognition by using the WiFi signals, including the detection of falling [3], gesture [4], [5], smoking [6], and so on. However, the current WiFi-based solutions can only coarsely classify the human movement. Besides of the WiFi-based solution, the vision-based methods were also proposed to recognize the human activity [7], [8]. These methods were developed based on the video signals and required lots of memory spaces to save data as well as high-efficiency algorithms to achieve the recognition task. Collaborating WiFi signal and video signal offers a potential solution to construct a more effective HAR scheme. The recently-developed AlphaPose [9] and OpenPose [10] proposed to estimate the skeleton points of human, which offer a potential solution to achieve this goal. Typically, the OpenPose [10] has been used in the WiFi-based human skeleton points detection [11], and it can also be adopted in the design of WiFi-based HAR.

Inspired by [11], we propose a new method in this paper to construct an HAR scheme which is developed based on the collaboration of both vision and WiFi signals. More specifically, we utilize the Transformer neural network [12] to comppse the long-short-term Transformer (LSTT) network so that we can obtain the skeleton points of human from the CSI data. Then, with the generated skeleton points, we accomplish the HAR task by using the support vector machines (SVM).

II. Background

The CSI data is the channel response that contains the information of channel condition. It is transmitted in the physical layer and obtained from the decoded sub-carriers of the orthogonal frequency division multiplexing system. The CSI of single sub-carrier can be modeled with a given frequency f at time t [13]

$$H(f,t) = e^{-j2\pi\Delta ft}\left(H_s(f) + \sum_{i=1}^{I_d} a_i(f,t)e^{-j2\pi d_i(t)\lambda}\right) \quad (1)$$

where $e^{-j2\pi\Delta ft}$ is the phase offset caused by carrier frequency offset, packet detection delay, sampling frequency offset [14], H_s represents the CSI reflected from stationary objects and stably transmitted in line-of-sight, I_d is the dynamic path

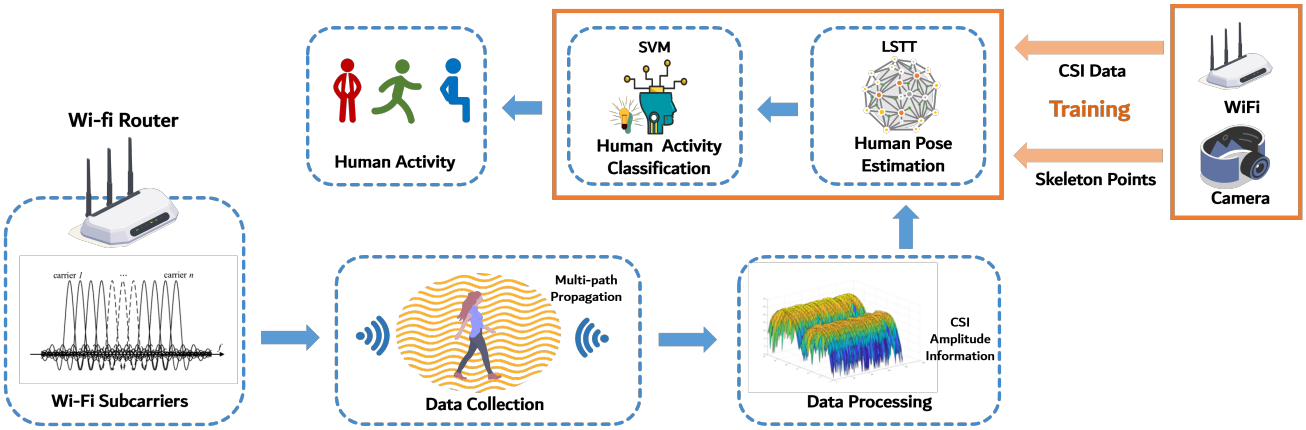

Fig. 1. Our proposed scheme.

index, $a_i(f,t)$ stands for the complex attenuation factor and initial phase on each path, $e^{-j2\pi d_i(t)\lambda}$ and $d_i(t)$ stand for the phase change and its length of i^{th} path, and λ is wavelength of wireless signal. The CSI date is able to capture the change of environment [13]. Therefore, it can be used to recognize the movement and action of human.

III. Our Proposed Method

A. Processing of CSI data

As mentioned above, the CSI data can be used to recognize the movement and activity of person. However, the phase offset of network interface cards (NICs) cannot be exactly obtained, which often produces phase noise in the received data. These data cannot be used for the recognition of human activity. To tackle this problem, we only employ the amplitude of CSI to implement the HAR task.

In this work, we aim at obtaining the pose of human by constructing the mapping between the CSI data and the skeleton points which are generated from the video data. Our proposed pose estimation highly depends on the data of both CSI and skeleton points. However, the packet loss happens frequently in the transmission of WiFi signal, which accordingly induces the losing of CSI. To tackle this problem, when the CSI data is missed, we will utilize the neighboring data to linearly interpolate the lost one. Meanwhile, if the the amplitude of the data is infinite, we will replace it by using the amplitude of the previous datap.

Our proposed scheme is illustrated in Fig. 1. To construct the vision and WiFi collaborated scheme for HAR, we use a webcam to collect videos of person and produce the skeleton points with OpenPose [10]. These skeleton points are related to the corresponding CSI data. To build up the collaboration between CSI and video frames, we get the timestamp of standard time from the internet for both devices. An example of the potential correlation between CSI and human pose is illustrated in Fig. 2.

B. LSTT for Human Pose Estimation

In this work, we compose the LSTT neural network to build up the mapping between the CSI data and the human pose. The Transformer model [12] has demonstrated impressive performance in the wireless signal-based HAR [15]. Our LSTT network consists of four layers, including input layer, self-attention layer, aggregation layer and prediction layer, and its architecture is illustrated in Fig. 3.

In LSTT, the input layer receives the processed CSI data and splits it into the long-term and short-term data streams, respectively, which are fed into the self-attention layer to extracts discriminative features. The short-term stream aims to acquire the movement information of different body parts in a short time period, while the long-term stream works to obtain the static information in a long period to provide a temporal constraint for the prediction of skeleton points. To save the memory, the average pooling is adopted is this layer to reduce the dimensions of the streams.

The self-attention layer receives the long-term stream and short-term stream from the input layer, where each of stream is divided into a temporal stream and a channel stream. The size of the divided streams are $T_l \times C$ and $T_s \times C$, respectively. Then, these streams are fed into the multi-scale convolution augmented transformer (MCAT) to extract discriminative features, which can directly integrate the information within the whole sequence. The structure of MCAT is illustrated in Fig. 4 and it is composed by two sub-layers, including a multi-head self-attention and a multi-scale CNN with self attention. Moreover, these sub-layers are connected by residual connection [16] and normalization layer [17].

The aggregation layer receives the temporal streams and channel streams, and then aggregates them into vectors via four separate convolutional blocks. After that, the resulting temporal and channel vectors are concatenated to be fed as the input of the prediction layer.

The prediction layer uses a fully-connected layer to generate the coordinate vectors for the human pose which is composed by 17 skeleton points in a frame, where the first vector is composed by the horizonal coordinates of the points and

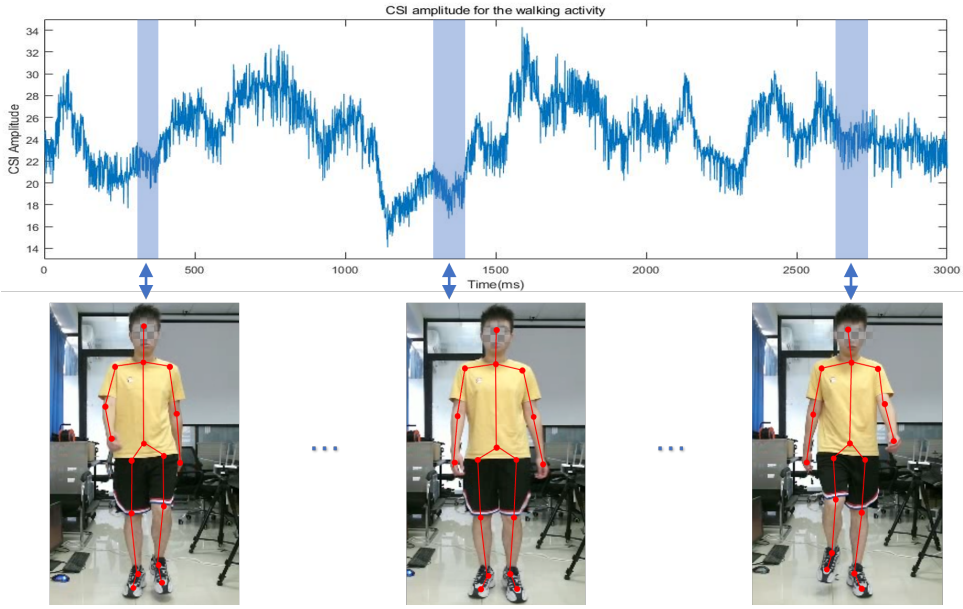

Fig. 2. Correlation between CSI and human pose.

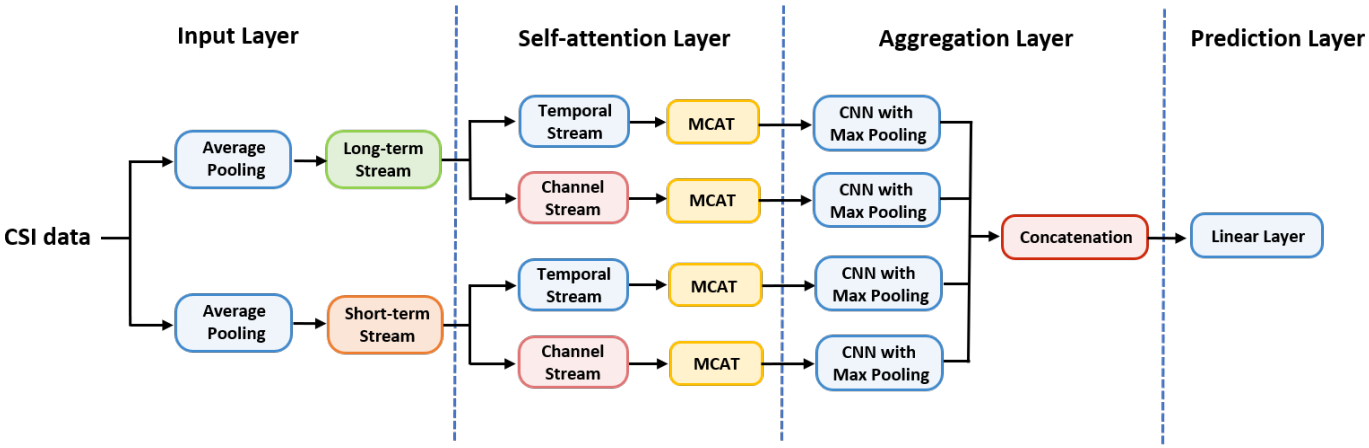

Fig. 3. Architecture of LSTT model.

the second one is composed by the vertical coordinates. The total loss of pose estimation is defined based on the sum of Euclidean distances between labels and predicted points

$$Loss = \sum_{i=1}^{17} \sqrt{(x_i - \bar{x}_i)^2 + (y_i - \bar{y}_i)^2}. \tag{2}$$

where (x_i, y_i) is the predicted location of ith joint and $(\bar{x}_i, \bar{y}_i)$ is the label of location.

C. Human Activity Classification

With the LSTT network, we can estimate the human pose from a sequence of CSI data. As human activity changes, the moving tracks of points, including along both the horizonal and vertical directions, are apparent differences, which can be used to distinguish and classify the human activities. To obtain efficient features from the 17 skeleton points, we adopt the variance of these points as descriptors for our HAR task. The variance descriptor is calculated as

$$\sigma^2 = \frac{1}{K}\sum_{i=1}^{K}(n_i - \frac{1}{K}\sum_{j=1}^{K} n_j)^2 \tag{3}$$

where σ^2 is the variance of the same point along the horizonal or vertical direction in K frames. Then, we can acquire a feature vector which contains 34 variance descriptors in horizonal and vertical directions. These descriptors are used for the detection of various activities. After that, we employ a non-linear SVM model [18], [19] to process the resulting descriptors for activity classification, where the descriptors are normalized by the unit L_2 norm. Finally, the HAR task can be implemented.

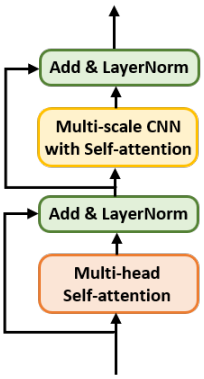

Fig. 4. Structure of MCAT.

TABLE I
PREDICTION ERRORS OF POSE ESTIMATION FOR DIFFERENT NETWORK SCHEMES

Model	Short-term	Long-term	Short-long-term
Error	3.67	2.24	**1.96**

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

Our WiFi devices are equipped with Linux 802.11n CSI Tool with 5300 NICs [20], on Ubuntu 14.04 operating system of portable computer. We apply one COTS WiFi transmitter and three WiFi receivers to get the human activity profile from different directions. The employment of multiple devices in our work can well collect the spatial information of human activity. Figure 5 illustrates the construction of the scheme, where the transmitter is placed the same location with camera and three receivers are placed around the sensing area. A person is invited to do different types of activities in the center of sensing area.

In the data collection, we set the sample frequency to 100Hz for CSI and the frame rate to 25Hz for camera. The length of the long-term and short-term CSI inputs are 8 and 42, respectively. In this experiment, the data of 3 persons are collected to compose the dataset for both training and testing. More specifically, each person does 6 actions, including walking, waving hands, picking up, jumping, raising hands and squatting, where each action lasts for three seconds, recorded by 300 CSI packets as well as 75 corresponding video frames. Moreover, each action is repeated 100 times. As a result, we obtain over all 1,800 actions and use them as 1,800 groups of data. Then, we separate the whole dataset into two parts, where 900 groups of data are selected to train our LSTT network and the other 900 groups are employed to train the SVM model and also used as the test dataset.

B. Results and Analysis

1) Human Pose Estimation: To verify the effectiveness of our proposed LSTT network, an ablation experiment is firstly conducted on human pose estimation. The performances of three models with the short-term, long-term and short-long-term structures are verified, respectively. The corresponding results, including the average mean square error between the prediction and the label are all given in Table I. It can be seen from Table I that coupling the short-term and long-term streams together is able to achieve better prediction for human pose estimation.

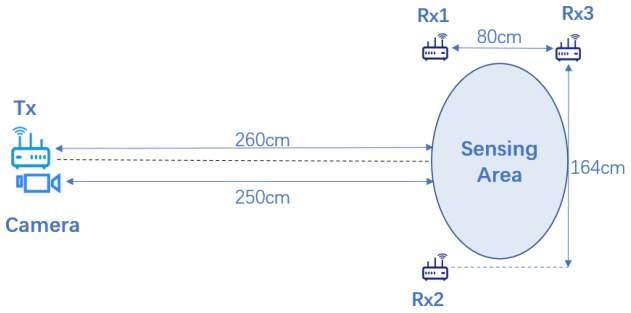

Fig. 5. Data collection for our proposed HAR scheme.

TABLE II
ACTION RECOGNITION ACCURACY (%)

	Accuracy			Average
	User 1	User 2	User 3	
Walking	100	100	98	99
Waving Hands	100	76	88	88
Picking Up	100	96	100	98
Jumping	96	98	96	97
Raising Hands	100	100	100	100
Squatting	100	98	100	99
Average	99	94	96	96

2) Human Activity Classification: We apply 3-fold cross validation to quantitatively evaluate the performance of our activity classification method. In this experiment, the test dataset which contains 900 groups of data is divided into three sub-sets according to the three persons, where 600 groups of data collected from two persons are used for training and the other 300 groups of data obtained from the third person are utilized for testing. We offer the recognition accuracy for six activities in Table II and it is found that our method achieves 96% accuracy (on average).

V. CONCLUSION

In this paper, we propose a WiFi and vision-based HAR scheme to classify the human activities. More specifically, we compose a long-short-term Transformer network to construct the collaboration between the CSI data obtained from the WiFi signal and the human skeleton points extracted from the video signal. With this collaboration, we can use the WiFi signal to effectively recognize the human actions. Our proposed scheme is constructed based on the COTS WiFi NICs, which can be easily transferred to any indoor environment, such as care-home, hospital and office.

REFERENCES

[1] Y. Ma, G. Zhou, and S. Wang, "WiFi sensing with channel state information: A survey," *ACM Computing Surveys*, vol. 52, no. 3, 2019.

[2] S. A. Shah and F. Fioranelli, "RF Sensing Technologies for Assisted Daily Living in Healthcare: A Comprehensive Review," *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 11, pp. 26–44, 2019.

[3] F. Wang, J. Han, S. Zhang, X. He, and D. Huang, "Csi-net: Unified human body characterization and pose recognition," *arXiv preprint arXiv:1810.03064*, 2018.

[4] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Gesture Recognition Using Wireless Signals," *GetMobile: Mobile Computing and Communications*, vol. 18, no. 4, pp. 15–18, 2015.

[5] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-Effort Cross-Domain Gesture Recognition with Wi-Fi," pp. 313–325, 2019.

[6] X. Zheng, J. Wang, L. Shangguan, Z. Zhou, and Y. Liu, "Design and implementation of a csi-based ubiquitous smoking detection system," *IEEE/ACM Transactions on Networking*, vol. 25, no. 6, pp. 3781–3793, 2017.

[7] B. Xu, H. Ye, Y. Zheng, H. Wang, T. Luwang, and Y.-G. Jiang, "Dense dilated network for video action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4941–4953, 2019.

[8] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[9] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2334–2343.

[10] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[11] F. Wang, S. Zhou, S. Panev, J. Han, and D. Huang, "Person-in-wifi: Fine-grained person perception using wifi," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5452–5461.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[13] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015, p. 65–76.

[14] X. Wang, C. Yang, and S. Mao, "Tensorbeat: Tensor decomposition for monitoring multiperson breathing beats with commodity wifi," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 1, pp. 1–27, 2017.

[15] B. Li, W. Cui, W. Wang, L. Zhang, Z. Chen, and M. Wu, "Two-stream convolution augmented transformer for human activity recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 1, 2021, pp. 286–293.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," pp. 770–778, 2016.

[17] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[18] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "Liblinear: A library for large linear classification," *the Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[19] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *23th International Joint Conference on Artificial Intelligence*, 2013.

[20] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *ACM SIGCOMM CCR*, vol. 41, no. 1, p. 53, Jan. 2011.