WIFI ACTION RECOGNITION VIA VISION-BASED METHODS

Jen-Yin Chang[⋆] *Kuan-Ying Lee*[⋆] *Kate Ching-Ju Lin*[†] *Winston Hsu*

National Taiwan University, Taipei, Taiwan, [†]Academia Sinica, Taipei, Taiwan
[⋆] Co-primary authors

ABSTRACT

Action recognition via WiFi has caught intense attention recently because of its ubiquity, low cost, and privacy-preserving. Observing Channel State Information (CSI, a fine-grained information computed from the received WiFi signal) resemblance to texture, we transform the received CSI into images, extract features with vision-based methods and train SVM classifiers for action recognition. Our experiments show that regarding CSI as images achieves an accuracy above 85%. Our contributions include:

- To our best knowledge, we are the first to investigate the feasibility of processing CSI by vision-based methods with extendable learning-based framework.
- We regard CSI of each *Tx-Rx* pair as a channel and investigate early and late fusion of multi-channels.
- We could know where and what action user performs with location-awareness classification.

Index Terms— WiFi, vision-based, texture, action recognition

1. INTRODUCTION

Researches have been devoted to action recognition, which is fundamental and essential to human computer interaction, video content-based retrieval, elders monitoring, shoppers behavior analysis and so on. As the most intuitive way, vision-based methods have been widely investigated. [1] harnessed deep convolutional neural network to learn temporal-spatial relationship and achieved an average 90% accuracy on KTH dataset and [2] achieved an 95% accuracy on KTH by learning the action trajectory.

However, not in every place and scenario are cameras applicable. For example, in restroom where privacy is of first priority or in places where lighting is scarce, cameras are of little use. Nevertheless, action recognition could not be spared in these places. For instance, timely detection of falling in bathroom limits the damage to minimum. Hence, previous works proposed using wearable devices such as accelerometer to obtain the speed profile and detect the action [3].

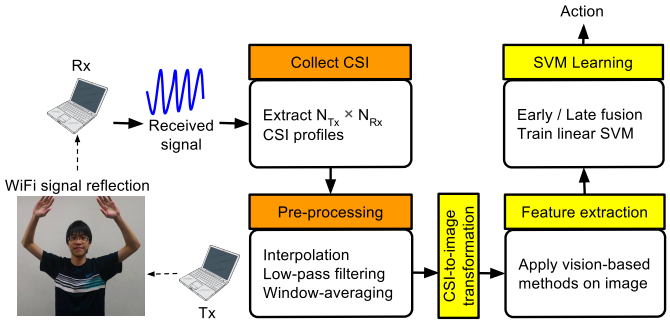

Fig. 1: Framework Overview (Blocks with yellow mark are the main differences between previous works and our work.)

The problem of this approach is that it requires users to wear devices, which is unrealistic in some scenarios like taking a shower, suggesting non-contact method is preferable. Since cameras typically have privacy constraints, wearable devices are not so extendable and extra apparatus should be as little as possible, we chose WiFi as our media.

In the beginning, researchers analyzed the frequency fluctuation of raw signals received by USRP, a software defined radio, to pinpoint timestamp of motion [4]. However, this method requires transforming a large amount of time domain signals to fine-grained frequency domain signals by performing Fourier transform, which costs too much time, making it impossible to be applied on real-time system. Hence, following works suggested directly analyzing the troughs and peaks on time domain signals [5]. However, commercial APs do not provide raw signals, urging later works [6, 7, 8] to perform analysis on CSI with modified driver [9]. We also realize our work based on this CSI toolkit.

Though extensive efforts have been put into action recognition harnessing CSI, to the best of our knowledge, most existing works extract ad-hoc features which might encounter accuracy debasement as scenarios change. Observing different textural appearances on transformed images, we investigate whether a general solution is possible via vision-based methods. The promising accuracy on predicting action and its location proves the feasibility of applying vision and learning based methods on transformed images for action recognition.

The rest of the paper is organized as follows. We briefly

describe how action recognition via WiFi is realized and the relationship between CSI and textural appearance in Section 2. Our framework is introduced in Section 3. Then, we present experiment results in Section 4. Finally, we conclude about image processing applied on WiFi action recognition in Section 5.

2. BACKGROUND

As WiFi wave propagates from transmitter (Tx) through the air to receiver (Rx), it bumps into objects and goes through several reflections. When an action takes place, paths reflecting from human body differ. Action recognition could thus be realized.

2.1. Raw Signal

From the frequency aspect, if we view human body as a source of the reflected signal, when the user pushes toward receiver, the relatively approaching speed causes a positive Doppler shift at the receiver. On the contrary, a negative Doppler shift occurs as the user's hand departs from the receiver. Harnessing Doppler effect, [4] achieves a 94% accuracy differentiating between nine gestures.

As from the amplitude angle, since the total path from Tx to human body and to Rx is shorten as pushing happens, the power dissipation decreases, rendering a rising amplitude on Rx side. Utilizing this phenomenon, [5] successfully lessens computational cost by performing analysis directly on time domain with an 91% accuracy classifying four gestures .

However, commercial APs do not provide raw signals, nudging researches toward CSI-based approaches.

2.2. Channel State Information

Modern WiFi protocols such as 802.11n implement OFDM (Orthogonal Frequency-Division Multiplexing) for reducing interference and fading. It segments the bandwidth into several closely-spaced sub-carriers, each carrying a data stream. Due to space constraint, for more details about OFDM please refer to [10].

Channel Frequency Response (CFR) describes the combined effect of fading, scattering and decay of a specific sub-carrier, usually a complex number detailing the phase shift and power decay. CSI is the union of these CFRs. When an action happens, the number of reflecting paths and their distance change accordingly and thereby, by extracting information from the received CSI, one could classify the action performed and even locate where it happened.

[7] presents a user-feedback system, separating actions into walking and in-place activity, which is capable of identifying several trajectories and activities. [8] proposes a PCA-based denoising method followed by discrete wavelet transform and introduces CARM, a system capable of human activity recognition independent of environment variances. Nonetheless, parts of their features are related to time duration of the action, which to our knowledge, might render classifiers highly dependent on duration.

2.3. CFR power and texture

From [8], we know CFR of a sub-carrier with frequency f at time t could be expressed as a sum of static CFR and dynamic CFR, expressed as $H(t, f)$ in equation (1), where $H_s(f)$ is the sum of CFRs for static paths, $H_d(f, t)$ is the sum of CFRs for all dynamic paths and Δf is the frequency offset between Tx and Rx.

$$H(f,t) = e^{-j2\pi\Delta f t}\left(H_s(f) + H_{\mathrm{d}}(f,t)\right) \tag{1}$$

$H_{\mathrm{d}}(f) = \sum_{k\in P_d} a_k(f,t)e^{-j\frac{2\pi d_k(t)}{\lambda}}$, where $a_k(f, t)$ is attenuation of the k_{th} path at time t and frequency f, $d_k(t)$ is the distance of k_{th} path and P_d is the set of all dynamic paths.

If an object moves at a constant speed, the distance of the k_{th} path, $d_k(t)$, could be expressed as $d_k(t) = d_k(0) + v_k$. Thus CFR power $|H(f,t)|^2$ at time t and frequency f is then derived as (Details omitted due to space constraint):

$$|H(f,t)|^2 = \tag{2}$$
$$\sum_{k\in P_d} 2|H_s(f)a_k(f,t)|\cos\left(\frac{2\pi v_k t}{\lambda} + \frac{2\pi d_k(0)}{\lambda} + \phi_{sk}\right) + C(f,t) +$$
$$\sum_{\substack{k,l\in P_d \\ k\neq l}} 2|a_k(f,t)a_l(f,t)|\cos\left(\frac{2\pi(v_k - v_l)t}{\lambda} + \frac{2\pi(d_k(0) - d_l(0))}{\lambda} + \phi_{kl}\right)$$

where $C(f,t)$ is a constant given sub-carrier frequency and time, ϕ_{sk} and ϕ_{kl} represent initial phase offsets.

We observe that in equation (2), frequency of cosine waves are determined by the action speed v_k. A faster speed leads to a larger phase change and renders denser stripes on transformed images, as shown in Figure 2. Since actions of different speeds present different textures on transformed images, we propose applying vision-based methods on transformed images.

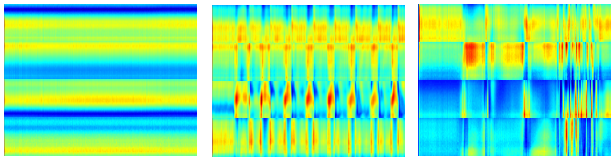

Fig. 2: Transformed images of standing still, clapping, boxing (X-axis: timestamp, Y-axis: 30 sub-carriers × 4 channels). We could observe a faster punching speed leads to denser stripes, as in the rear part of boxing.

3. FRAMEWORK

In this section, we describe the overall flow of the proposed framework, from collecting CSI, pre-processing, extracting features to training classifier, as shown in Figure 1.

3.1. Collecting CSI

We use MacBook Pro 2014 as Tx and Fujitsu SH560 with Intel 5300 NIC as Rx, each having two antennas and communicating on 2.4GHz. With $N_{Tx} = 2$, $N_{Rx} = 2$, we have $2 \times 2 = 4$ *Tx-Rx* pairs, each generating a set of CSI with dimension 30 (sub-carriers) $\times$ t (samples). We then process these four sets of CSI separately and investigate whether early or late fusion yields better performances.

3.2. Pre-processing

Due to interference caused by other devices in the same WiFi channel, packets received are not evenly distributed in time. Thus, we linearly interpolate raw CSI to 1000 samples/second. We then apply 5th-order Butterworth filter with cutoff frequency 50Hz to remove high-frequency noises. And since power distributions of different sub-carriers vary, we normalize each sub-carrier by subtracting an average of a moving-window, width set as 300ms, from each sample.

3.3. Feature Extraction

After transforming a set of CSI into an image of specific size, we experiment with Gabor and BoF-SIFT on it. Though deep features are potentially more powerful, due to the scarcity of current data we will not address it in our work.

3.3.1. Gabor Filter

A Gabor filter is defined by a plane wave multiplied by a Gaussian function. By setting different scales and orientations, a set of filters are obtained (details could be found in [11]). These filters are convoluted with a transformed image. When a local patch resembles the filter, a high response will be obtained. Finally, a response map is produced, of which we then take two statistics, mean and standard deviation.

We set *#scale*, *#orientation* and size of the Gabor filters to 8, 6 and 15 respectively, which usually produce better accuracy from our measurements. Hence, the dimension of our final Gabor feature is $8 \times 6 \times 2 = 96$.

3.3.2. Bag of Feature-SIFT

SIFT (Scale Invariant Feature Transform) seeks to transform an image into a collection of keypoints, each described by a feature vector invariant to illumination, translation, rotation and scaling [12]. We take all feature vectors of the training images from a *Tx-Rx* pair and perform K-means clustering to find 48 centroids. BoF-SIFT feature is then generated by quantizing vectors of an image to the nearest centroid, producing a histogram of dimension 48.

Thus, in testing phase, we quantize the feature vectors of the input image into centroids found during training and feed the produced 48-dimension feature into the trained classifier.

3.4. Training Classifier

For each of the four *Tx-Rx* pairs, we obtain a feature vector. We investigate fusing them before or after training linear SVM classifiers.

3.4.1. Early Fusion

We concatenate four features from four *Tx-Rx* pairs into a new feature. Then, we train a single classifier and take action with the highest probability as the predicted result.

3.4.2. Late Fusion

Instead of concatenating four feature vectors and training a single classifier, we train one for each pair, so there would be four classifiers. Given a testing instance, probability of each action is obtained, rendering four probability vectors of length *#action* (seven in our case). Summing these four vectors, we take action with the highest probability as the predicted result.

4. EXPERIMENTS

4.1. Datasets

Dataset A is collected in a seminar room, as shown in Figure 3, for verifying if our method could recognize actions as well as locations. We define seven actions: *Box, Clap, Wave hand, Kick, Quick squat, Jump, Stand still* and six locations *a, b, c, d, e, f*. A single subject performs each action 10 times on each location, so in total we have $7 \times 6 \times 10 = 420$ data. Dataset B is collected to compare with vision-based methods on video action recognition, and thus we define actions the same as the benchmark dataset, KTH [13]. These actions include: *Box, Clap, Wave hand, Walk, Jog, Run, Stand still*. Two subjects are asked to perform each action 10 times and in total we have $2 \times 7 \times 10 = 140$ data.

All actions are performed in a 5-second period, each generating four sets of CSI with dimension 30×5000 (interpolated to 1000 samples/second). We then transform them into four images of size 576×432.

4.2. Experiment Results

Our experiments mainly focus on examining the feasibility of the proposed method. We leave comparing the strength and limitation of different approaches as our future study.

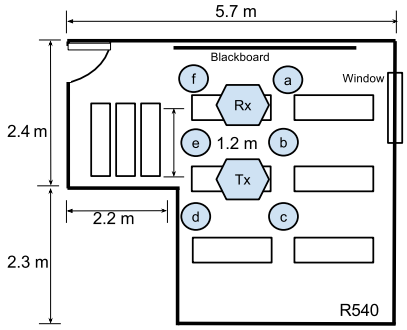

Fig. 3: Subject performs actions in each circle.

We evaluate the performance using 10-fold cross validation (Due to the superiority of late fusion to early fusion in Table 4, we only list results of late-fusion). First, we conduct experiments on dataset A, with cross-validation accuracy shown in Table 1. We could verify that viewing CSI as texture is feasible and Gabor filters, particularly suitable for texture recognition, perform better. Hence, following experiments are primarily based on Gabor.

Location	BoF-SIFT	Gabor
a	0.6429	0.8436
b	0.8143	0.9657
c	0.7714	0.8979
d	0.8000	0.8671
e	0.5143	0.8121
f	0.7429	0.8150
all	0.5048	0.7745

Table 1: Accuracy of Gabor and BoF-SIFT on dataset A.

Wondering if location affects the accuracy, we split the classification process into two stages, namely, location identification followed by action recognition. Location is predicted using classifier trained on all data first. Then a classifier trained with data of the suggested location is employed to obtain the action. The results are shown in Table 2.

Target	Accuracy
location	0.98
location + action	0.83

Table 2: Accuracy of location-awareness classification.

The accuracy boost from 77% to 83% of action classification reveals that our features still embed location information, which causes a slight accuracy decrease when classifying action using only a single classifier trained with data from all locations. To further testify, we test the trained classifiers on data of unseen location and find the accuracy drops, showing that our methods are still location-dependent, which will be discussed later in Section 5.

Table 3 shows the results of cross-validation on both datasets using early fusion and late fusion. As the statistics show, late fusion performs better since it exploits four different channels with four classifiers. Though each classifier is weaker compared to that of early fusion, more channels provide more information for recognition. Also from results on dataset B, we believe WiFi could actually supports cameras in differentiating actions that are visually similar but of different CSI patterns.

	Dataset A	Dataset B
Early	0.7024	0.8023
Late	0.7745	0.8696

Table 3: Accuracy of early and late fusion applying Gabor.

Finally, we conduct an experiment exploring whether size of the transformed images affects accuracy, as shown in Table 4. The result demonstrates that as the size of images becomes smaller, performance remains excellent as long as the size of filters alters accordingly, implying the proposed framework is computationally efficient.

Size	Filter size	Accuracy
5000×30	15	0.6654
576×432	15	**0.8696**
72×54	15	0.7564
72×54	9	0.8446

Table 4: Accuracy on dataset B between different sizes (in pixels) of image.

5. CONCLUSION

We observe the resemblance of CSI to texture and apply vision-based methods on images transformed from CSI. With this brand new angle, we propose a method which achieves accuracy above 85% identifying the predefined seven actions. Though environment dependency mentioned in Section 4.2 is still a challenging issue, which lowers the performance when the user deviates from the training locations too much, we believe, as the amount of data increases, techniques such as deep neural network could be capable of finding the hidden factor more clearly and thus mitigate such degradation.

For future work, we would collect more training examples and more actions to further verify that the proposed method is promising. Also, since collecting CSI is time-consuming, we would be working on utilizing surveillance camera to automate data collection and also explore data augmentation which is commonly utilized in training neural network [14].

6. REFERENCES

[1] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3d convolutional neural networks for human action recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 1, pp. 221–231, Jan 2013.

[2] Yemin Shi, Wei Zeng, Tiejun Huang, and Yaowei Wang, "Learning deep trajectory descriptor for action recognition in videos using deep neural networks," in *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, June 2015, pp. 1–6.

[3] Pierluigi Casale, Oriol Pujol, and Petia Radeva, "Human activity recognition from accelerometer data using a wearable device," in *Proceedings of the 5th Iberian Conference on Pattern Recognition and Image Analysis*, Berlin, Heidelberg, 2011, IbPRIA'11, pp. 289–296, Springer-Verlag.

[4] Qifan Pu, Sidhant Gupta, Shyamnath Gollakota, and Shwetak Patel, "Whole-home gesture recognition using wireless signals," in *Proceedings of the 19th Annual International Conference on Mobile Computing & Networking*, New York, NY, USA, 2013, MobiCom '13, pp. 27–38, ACM.

[5] R. Nandakumar, B. Kellogg, and S. Gollakota, "Wi-Fi Gesture Recognition on Existing Devices," *ArXiv e-prints*, Nov. 2014.

[6] Yunze Zeng, Parth H. Pathak, and Prasant Mohapatra, "Analyzing shopper's behavior through wifi signals," in *Proceedings of the 2Nd Workshop on Workshop on Physical Analytics*, New York, NY, USA, 2015, WPA '15, pp. 13–18, ACM.

[7] Yan Wang, Jian Liu, Yingying Chen, Marco Gruteser, Jie Yang, and Hongbo Liu, "E-eyes: Device-free location-oriented activity identification using fine-grained wifi signatures," in *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, 2014, MobiCom '14, pp. 617–628, ACM.

[8] Wei Wang, Alex X. Liu, Muhammad Shahzad, Kang Ling, and Sanglu Lu, "Understanding and modeling of wifi signal based human activity recognition," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, New York, NY, USA, 2015, MobiCom '15, pp. 65–76, ACM.

[9] Daniel Halperin, Wenjun Hu, Anmol Sheth, and David Wetherall, "Tool release: Gathering 802.11n traces with channel state information," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 1, pp. 53–53, Jan. 2011.

[10] Taewon Hwang, Chenyang Yang, Gang Wu, Shaoqian Li, and G.Y. Li, "OFDM and its wireless applications: A survey," *Vehicular Technology, IEEE Transactions on*, vol. 58, no. 4, pp. 1673–1694, May 2009.

[11] J. R. Movellan, "Tutorial on Gabor Filters," *Tutorial paper http://mplab.ucsd.edu/tutorials/pdfs/gabor.pdf*, 2008.

[12] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[13] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: A local svm approach," in *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, Washington, DC, USA, 2004, ICPR '04, pp. 32–36, IEEE Computer Society.

[14] R. Wu, S. Yan, Y. Shan, Q. Dang, and G. Sun, "Deep Image: Scaling up Image Recognition," *ArXiv e-prints*, Jan. 2015.