Wi-Multi: A Three-Phase System for Multiple Human Activity Recognition With Commercial WiFi Devices

Chunhai Feng, Sheheryar Arshad, Siwang Zhou, *Member, IEEE*, Dun Cao, and Yonghe Liu

***Abstract*—Channel state information-based activity recognition has gathered immense attention over recent years. Many existing works achieved desirable performance in various applications, including healthcare, security, and Internet of Things, with different machine learning algorithms. However, they usually fail to consider the availability of enough samples to be trained. Besides, many applications only focus on the scenario where only single subject presents. To address these challenges, in this paper, we propose a three-phase system Wi-multi that targets at recognizing multiple human activities in a wireless environment. Different system phases are applied according to the size of available collected samples. Specifically, distance-based classification using dynamic time warping is applied when there are few samples in the profile. Then, support vector machine is employed when representative features can be extracted from training samples. Lastly, recurrent neural networks is exploited when a large number of samples are available. Extensive experiments results show that Wi-multi achieves an accuracy of 96.1% on average. It is also able to achieve a desirable tradeoff between accuracy and efficiency in different phases.**

***Index Terms*—Activity recognition, channel state information (CSI), recurrent neural network (RNN).**

I. Introduction

CHANNEL state information (CSI)-based application development, including healthcare, security and Internet of Things (IoT), has received immense attention over the past years. In terms of activity recognition, many traditional approaches with carry-on sensors [1]–[4] or cameras [5] have achieved desirable performance. However, it is now considered as inconvenience to ask users to wear sensors all the time because of the battery charge issues. Besides, it also raises many privacy security concerns on camera-based systems. On the contrary, CSI can be extracted from commercial wireless devices instead of requiring extra costs of equipments as in traditional approaches. Therefore, it has attracted broad interests in recent years.

Because of the multipath propagation phenomenon, human presence or body movement around the wireless devices can affect the strength quality of Wi-Fi signal [6]–[10]. CSI can record this detailed physical layer information from different subcarriers of the channel. By modifying the driver of Intel 5300 network interface card (NIC) [11], many existing papers proposed various systems to detect human activities, such as keystroke [12], gestures [13], and breathing [14]. However, many of them only focus on scenarios where only a single subject presents. In this case, the potentials of these systems are limited since there are usually multiple subjects per household [15].

In this paper, we propose a novel human activity recognition system termed as Wi-multi. Different from existing works, Wi-multi targets at *recognizing multiple activities of different subjects in the same environment*, with only commercially off the shelf WiFi devices. Potential applications of multiple human activity recognition include home security and emergency care. For instance, the system can raise an alert if additional unexpected subjects are detected in the house. Besides, an emergency alarm can be triggered when dangerous activity such as fall by particular person is not noticed by other family members. In designing Wi-multi, we have to overcome several challenges. First, the size of available samples in the profile can affect the performance of activity recognition [16]. Each sample is extracted by detecting the start and end of performed activity. Many existing papers assume that enough samples can be provided for model training [17], [18]. However, it is usually unlikely to have hundreds or even more samples in reality, especially at the early stage of system deployment. Since machine learning or deep neural networks usually require adequate samples, it becomes infeasible to apply them in activity detection if there are few samples available. On the other hand, signal processing methodologies [12] take longer time if size of dataset increases. Therefore, it is crucial to design a system that is able to accommodate the size of available samples in the profile. In addition, reflections of wireless signals by multiple subjects owing to multipath effects are more challenging to identify than those by single subject, further complicating the extraction of activities.

In order to address these challenges, we propose a three-phase system according to the size of available samples in the profile. In the first phase, principal component analysis (PCA)

Manuscript received January 30, 2019; revised April 3, 2019 and May 1, 2019; accepted May 4, 2019. Date of publication May 9, 2019; date of current version July 31, 2019. *(Corresponding author: Chunhai Feng.)*

C. Feng, S. Arshad, and Y. Liu are with the Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019 USA (e-mail: chunhai.feng@mavs.uta.edu; sheheryar.arshad@mavs.uta.edu; yonghe@cse.uta.edu).

S. Zhou is with the College of Information Science and Technology, Hunan University, Changsha 410082, China (e-mail: swzhou@hnu.edu.cn).

D. Cao is with the School of Computer and Communications Engineering, Changsha University of Science and Technology, Changsha 410114, China (e-mail: caodun@csust.edu.cn).

Digital Object Identifier 10.1109/JIOT.2019.2915989

that reduces dimensions of features and dynamic time warping (DTW) that calculates the distance between various length signals so as to measure similarities are employed at the beginning stage of building the system when limited training samples are available. In the second phase, when more samples are available, support vector machine (SVM) is exploited for model training and testing where a number of representative features can be extracted from both time and frequency domain. In this case, a model can be pretrained and hence it is not necessary to compare the similarities among all samples as in the first phase of the system, which may largely reduce time cost as sample size increases. In the third phase, we propose a deep learning system structure based on long short term memory (LSTM) unit if a large number of samples are available in the profile. LSTM, as one type of recurrent neural network (RNN) units, is capable of remembering and filtering the past information in the input sequence during training process. The proposed deep learning network is able to automatically select high level features without any preprocessing modules. The evaluation results demonstrate that our proposed system achieves a desirable tradeoff performance between accuracy and efficiency in different phases. In general, Wi-multi can achieve 96.1% accuracy on average.

Before activity classification, we also have developed an effective activity sample extraction algorithm to identify the start and end points of multiple subject activities. First, we apply outlier filtering and differential algorithms to the variance of CSI values among different subcarriers. Second, we calculate the largest eigenvalues from both amplitude and calibrated phase correlation matrices so as to eliminate potential false detection. The extracted samples are then presented to different system phases for further analysis. It is shown that our algorithm can extract activity samples in both noisy and non-noisy environments with multiple subjects.

The key contributions of this paper can be summarized as follows.

1) We propose a three-phase system that can recognize multiple human activities, where each phase is designed according to the size of available dataset in the profile during different stages of a system deployment.
2) We evaluate the system in terms of various aspects. Extensive results show that Wi-multi is able to achieve desirable tradeoff between accuracy and efficiency in different phases.
3) We propose a novel activity extraction algorithm that is able to identify the start and end point of an activity even in noisy environment with multiple subjects.

In the remaining sections, we briefly introduce the related works in Section II, and presents CSI in the Section III. Section IV introduces the activity extraction algorithm. Design of different system phases is given in Section V. Section VI presents the experiments evaluation results. We finally conclude in Section VII.

II. Related Works

An extensive set of works, employing CSI for human activity recognition, have been done recently. They can be generally divided into the following two groups according to different activity levels.

A. Fine-Grained Activity Detection

Fine-grained activity refers to activities performed at micro levels. For instance, Kaltiokallio *et al.* [19] employed CSI to detect the breath rate and monitor sleep quality with predeployed antennas and transceivers. A gesture identification system is proposed by Abdelnasser *et al.* [13] in order to promote human–computer interface (HCI). In addition, keystroke systems implemented on either laptop [12] or smartphone [20] yield desirable accuracy in keystroke recognition. Furthermore, WiHear in [21] can analyze and detect people speech based on the disturbance of CSI caused by lip movement. In general, these systems achieved desirable accuracy in a relatively controlled environment.

B. Coarse-Grained Activity Detection

Coarse-grained activity usually refers to activities performed at macro levels. For example, many previous works [22]–[24] focus on daily activity recognition (including walking and running) and achieve desirable performance in both line-of-sight and nonline-of-sight circumstances by the support of two mathematical model in [25] and [26]. Both WiFall [25] and RT-Fall [27] build an alarm system to detect human falls in realtime. Zheng *et al.* [28] developed a Smokey system that can detect the smokers behavior without deploying special devices. Moreover, Wang *et al.* [29] are able to detect different humans based on the features of their behaviors.

However, many of these works only target at single subject environment. This limits the potentials of applications according to a national survey [15], which indicates that there are usually around 2 to 3 people in each household in 2015.

III. CSI

WiFi signals arriving at a receiver usually come from different paths [30]. This multipath effect often causes interference, phase shifting and fading of the signal. Compared with the environment with only one person, multiple humans can cause even more disturbances in wireless channels.

There are usually two ways to evaluate the channel condition without chip level access. On one hand, radio signal strength indicator (RSSI) is the most common way because of its accessibility. However, it only provides limited amplitude information due to its low resolution. On the other hand, CSI, which represents the channel frequency response (CFR), can capture both the amplitude and phase variance for each OFDM subcarrier.

By modifying the driver of Intel 5300 NIC in 802.11n network [11], we are able to get CSI values of 30 subcarriers between one pair of transmit-receive antennas. Let T and R represent the WiFi signals in the frequency domain from the transmitter and receiver, respectively. The wireless channel model can be modeled as

$$R = H \times T + \mathcal{N}_o \tag{1}$$

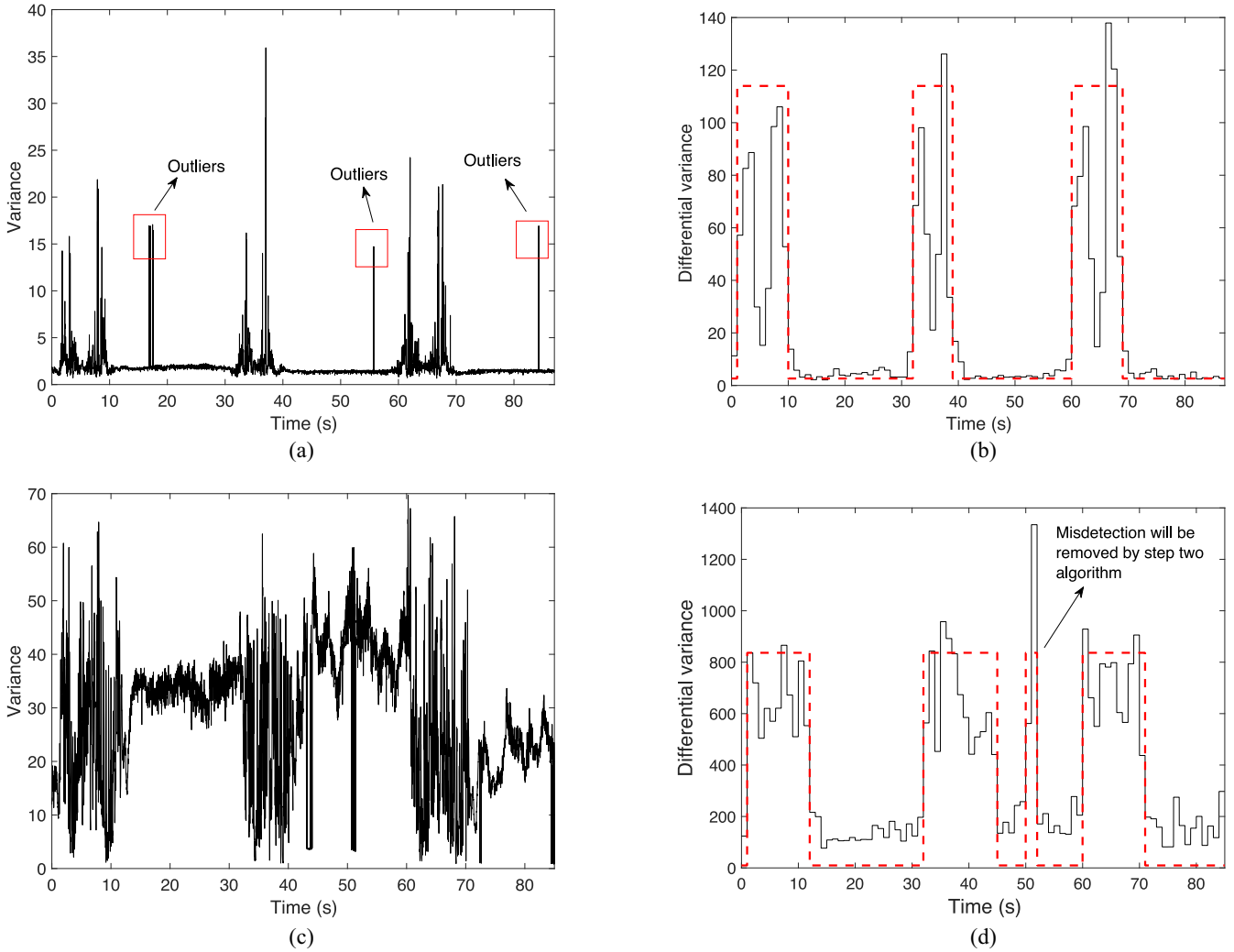


Fig. 1. Performance of activity extraction algorithm in non-noisy and noisy environment. Variance under (a) non-noisy and (c) noisy environment. (b) and (d) Differential variance and extraction result.

where H represents the CSI estimation of CFR. Note that H can then be approximated as

$$\widehat{H} = \frac{R}{T} \tag{2}$$

assuming noise $\mathcal{N}_o$ follows zero mean complex normal distribution of circularly symmetric, i.e., $\mathcal{N}_o \sim N_c(0, \Gamma)$.

$\widehat{H}$ is a matrix consisted of CSI from all subcarriers. $\widehat{h}$, CSI of one subcarrier, can also be represented as

$$\widehat{h} = ||\widehat{h}||e^{j\angle\widehat{h}} \tag{3}$$

where $||\widehat{h}||$ and $\angle\widehat{h}$ represent the amplitude and phase information, respectively.

IV. Activity Extraction

In order to extract the activity sample, it is necessary to detect the period of time where activity occurs. We first observe that the variances of CSI amplitudes among different subcarriers can be used as an indicator for activity presence. An example of CSI amplitude variance among 30 subcarriers can be found in Fig. 1(a). It can be easily observed that variance when no activity presents is much more stable than that when activities occur. In this section, we design a two-step algorithm to extract the start and end points of each activity. Note that only one CSI stream of 30 subcarriers is required for activity extraction.

A. Step 1: Differential Threshold Estimation

As shown in Fig. 1(a), the outliers caused by the internal hardware errors improve the difficulty for activity extraction. In order to address this challenge, we compare the CSI value at the mth time point with a threshold defined as $\delta_v = \lambda|V(m+1) - V(m-1)|$, where λ is an empirical coefficient. Assume the amplitude variance of 30 subcarriers is V, then we remove the outlier by setting it as the average of the values at prior and posterior packets

$$V(m) = \begin{cases} (V(m-1) + V(m+1))/2, & \text{if } V(m) > \delta_v \\ V(m), & \text{otherwise.} \end{cases} \tag{4}$$

Next, we split the signal into even slots in time domain and determine if there is activity presence in each slot. Here

we assume that the length of the activity is longer than one time slot. In our case, we set the length of each time slot as 1 s. It can be changed to larger or smaller values according to different scenarios. We then consider continuous time slots with activity presence as one activity sample.

Denote the sampling rate of CSI as R_s, the variances of CSI values in the ith slot can then be described as $V(j)$, where $j = 1, 2, \ldots, R_s$. In ascending order, they can be represented as $V'(1) < V'(2) < \cdots < V'(R_s)$. We then compute the difference between the sum of largest half values and the sum of smallest half values as

$$D_i = \sum_{j=1}^{R_s/2} \big(V'(j + R_s/2) - V'(j)\big). \tag{5}$$

Fig. 1(b) depicts the corresponding results of Fig. 1(a). By comparing D_i with a self adaptive threshold δ_D, we can obtain an initial result that determines whether there is an activity in this slot

$$I_i = \begin{cases} 1, & \text{if } D_i > \delta_D \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Here I is the indicator of the presence of activity, $\delta_D = (\sum_{i=1}^{L/R_s} D_i)/(L/R_s)$ and L here represents the total length of the signal. Afterwards, we consider continuous time slots where I_i is 1 as one activity sample. In non-noisy environment, it can usually detect the boundary of activity correctly as shown in Fig. 1(b). However, it may also result in detection errors in some cases, especially in extremely noisy environments or/and multiple subjects activities. Fig. 1(c) and (d) shows an example that the first step algorithm misdetects an activity. In order to address this challenge, our algorithm uses a second step to double check the results and remove potential, although rare, misdetections.

B. Step 2: Eigenvalues Comparison

As discussed above, the CSI values in stationary environment tends to be more stable than that with human presence. Therefore, the correlation between consecutive CSI values can be much higher in stationary environment. In this case, we build the correlation matrices of both amplitude and calibrated phase [31] within a time window. Assume the size of window is W, the CSI measurements in this window can be described as

$$H(i) = [H_1(i), H_2(i), \ldots, H_K(i)]$$

where $i = 1, 2, \ldots, W$ and K is the total number of subcarriers. Thus the covariance matrices of amplitude and phase can be computed as

$$\mathbf{A} = \begin{bmatrix} \text{cov}(|H(1)|, |H(1)|) & \cdots & \text{cov}(|H(1)|, |H(W)|) \\ \vdots & \ddots & \vdots \\ \text{cov}(|H(W)|, |H(1)|) & \cdots & \text{cov}(|H(W)|, |H(W)|) \end{bmatrix}$$

and

$$\mathbf{P} = \begin{bmatrix} \text{cov}\big(\widetilde{\angle H(1)}, \widetilde{\angle H(1)}\big) & \cdots & \text{cov}\big(\widetilde{\angle H(1)}, \widetilde{\angle H(W)}\big) \\ \vdots & \ddots & \vdots \\ \text{cov}\big(\widetilde{\angle H(W)}, \widetilde{\angle H(1)}\big) & \cdots & \text{cov}\big(\widetilde{\angle H(W)}, \widetilde{\angle H(W)}\big) \end{bmatrix}.$$

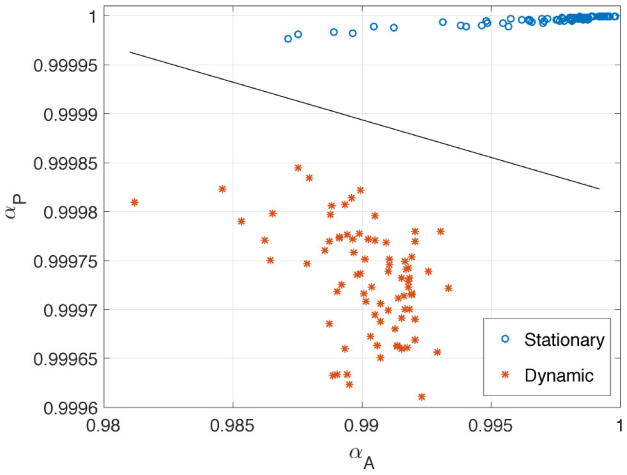


Fig. 2. Eigenvalues comparison between stationary and dynamic cases.

Afterwards, the largest normalized eigenvalues of **A** and **P** can be computed as

$$\alpha_A = \max(\text{norm}(\text{eigen}(\mathbf{A})))$$

and

$$\alpha_P = \max(\text{norm}(\text{eigen}(\mathbf{P})))$$

respectively. After conducting several experiments, we observe that α_A and α_P tend to be larger in stationary environment. As depicted in Fig. 2, scenarios with activities can be easily separated from stationary scenarios. Moreover, this threshold is independent from different environmental background since eigenvalues are power independent. As shown in Fig. 1(d), the misdetection shown in the result of step one can then be removed. Therefore, step two can further improve the accuracy-based step one result.

V. Three-Phase System Design

In this section, we describe the system structure of Wimulti based on a three phase design, corresponding to different phases of system deployment. Phase 1 is applied when only few samples are available in the profile. Phase 2 is employed only when the effective features can be extracted and trained with SVM. Lastly, phase 3 based on LSTM is employed when there are a large number of samples for deep learning networks. An overview of the system structure is shown in Fig. 3.

Before applying phases 1 and 2, we first apply interpolation to the raw data as data points can be missing because of errors in the system or the collecting tool [14]. Afterwards, we also apply a low pass Butterworth filter and phase calibration on amplitude and phase in order to remove outliers and random noises [31]. Note that phase 3 does not need any preprocessing as it can automatically extract representative features.

A. Phase 1

In this phase, we assume very limited availability of samples in the profile, i.e., at the early stage of system deployment. First, PCA is exploited to remove correlated information and

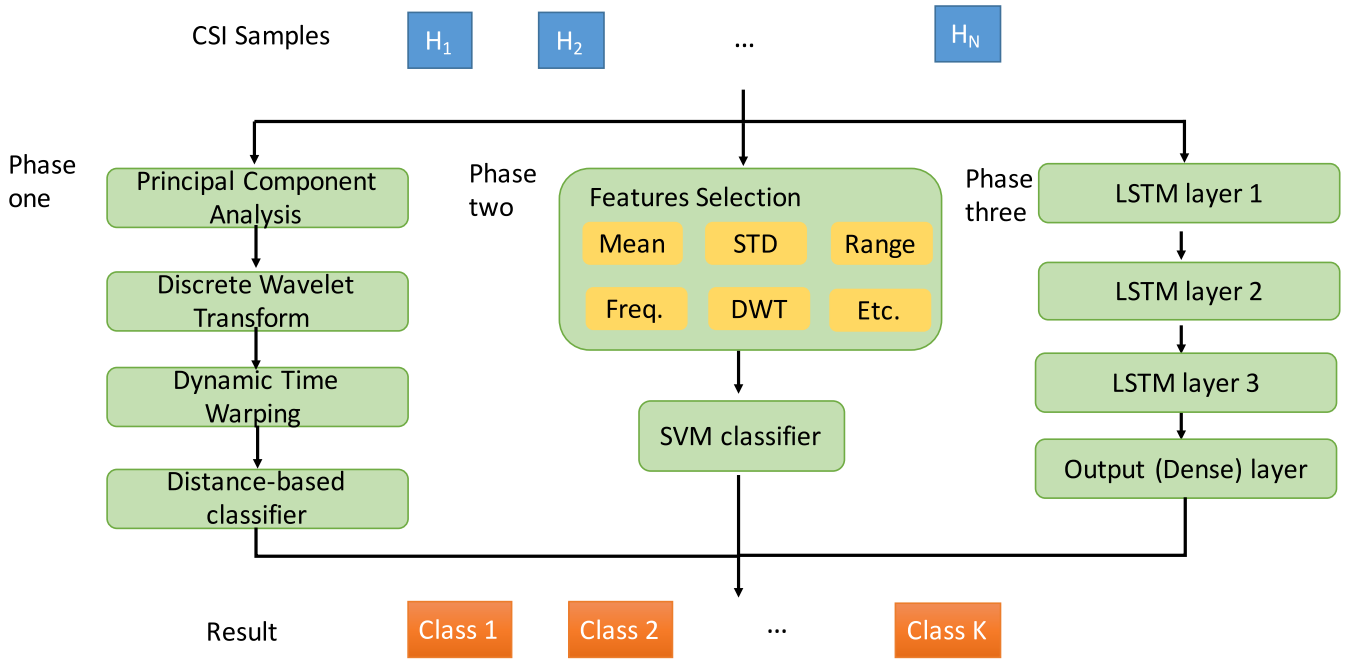


Fig. 3. Three-phase system overview.

reduce feature dimensions among the 30 subcarriers. Second, DWT is employed in order to compress the length of signal data without losing much representative information. Lastly, a distance-based method is applied to identify the label of activity.

1) Principal Component Analysis: As discussed earlier, PCA is used to remove correlated information among 30 subcarriers. The detailed PCA implementation is as follows.

a) Training samples combination: Denote CSI matrix of each training sample as H_i, and the size of each matrix is $L_i \times 30$, where $i = 1, 2, \ldots, N$. Thus, the combined matrix of training data can be represented as H, whose size is $\sum_i^N L_i \times 30$.

b) Static component removal: In this step, the static component is calculated by the average of the signal in each subcarrier. Denote it as avg_j, where $j = 1, 2, \ldots, 30$ represents subcarrier index. By subtracting avg_j from each column of H, we can get a centered matrix H_D.

c) Covariance matrix computation: The covariance matrix is then computed as $H_D^T \times H_D$.

d) Eigenvectors calculation: The eigenvectors corresponding to the covariance matrix can be calculated as $q_1, q_2, \ldots, q_n$, respectively.

e) Training samples projection: In order to project training samples to the eigenspace, the first k components can then be computed as $[c_1, c_2, \ldots, c_k] = H_D \times [q_1, q_2, \ldots, q_k]$, where c_i represents the ith component.

f) Testing samples projection: Similarly, denote the centered CSI matrix of the testing samples as T_D, then the first k components of the testing data can be calculated as $Z = T_D \times [q_1, q_2, \ldots, q_k]$ by projecting testing samples to eigenspace toward the same direction as training data.

g) Matrix separation: Let $C = [c_1, c_2, \ldots, c_k]$, whose size is $\sum_i^N L_i \times k$. Therefore, the first k components of each training sample can be obtained by splitting it into N separate matrices. Similar separation process is applied to matrix Z computed in the last step.

Note that this approach requires the computation of the eigenvectors only once, meaning that both the training data and testing data are projected to the eigenspace in the same direction. A flow chart of PCA implementations is presented in Fig. 4.

2) Discrete Wavelet Transform: Because of the high computation cost often associated with longer sample activity data, we further employ discrete wavelet transform (DWT) to compress the signal. It is able to reduce the length of the signal without losing much representative information. Denote the measured discrete signal as

$$s[n] = \frac{1}{\sqrt{M}} \sum_i \alpha[j_0, i] \phi_{j_0,i}[n] + \frac{1}{\sqrt{M}} \sum_{j=j_0}^{\infty} \sum_i \beta[j, i] \psi_{j,i}[n] \tag{7}$$

where M is the length of the signal. $\phi_{j_0,i}[n]$ and $\psi_{j,i}[n]$ are defined as orthogonal to each other and they represent scaling functions and wavelet functions, respectively. Similarly, $\alpha[j_0, i]$ and $\beta[j, i]$ are termed as approximation coefficients and detail coefficients, respectively. They can be modeled as

$$\alpha[j_0, i] = < s[n], \phi_{j_0+1,i}[n] >= \frac{1}{\sqrt{M}} \sum_n s[n] \phi_{j_0+1,i}[n] \tag{8}$$

$$\beta[j, i] = < s[n], \psi_{j+1,i}[n] >= \frac{1}{\sqrt{M}} \sum_n s[n] \psi_{j+1,i}[n] \tag{9}$$

in the jth level.

In this paper, we adopt approximation coefficients in order to reduce computation cost. Note that different levels of computation of DWT will lead to various compression lengths of

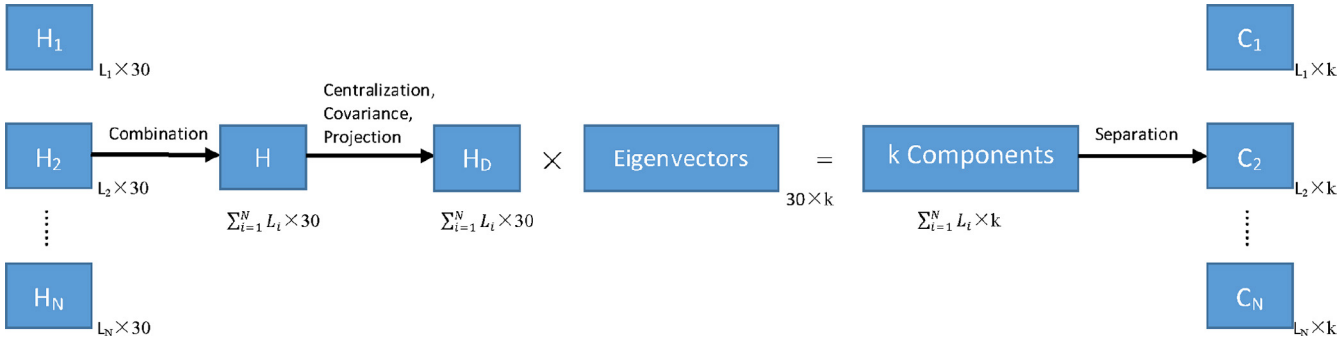


Fig. 4. PCA implementation.

the signal. The higher level of DWT, the shorter length the signal can be compressed. We will discuss the impact of different DWT levels in the next section.

3) Distance-Based Classification: In order to compare the similarities of different waveforms, DTW is employed to calculate the Euclidean distances between signals. By aligning the waveforms, DTW yields the addition of Euclidean distances between their corresponding points. Smaller distance usually represents higher similarity of waveforms. After the construction of different profiles, the label of the new test sample is predicted by comparing its distances from different activities.

Denote the Euclidean distances between a test sample and each sample in the profile as D_i, where $i = 1, 2, \ldots, N$ represents N different activity samples in the profile. Denote the increasingly ordered K distances for test sample as D'_i, where $D'_1 < D'_2 < \cdots < D'_K$ and K is chosen empirically. Assume that there are n kinds of different activities in the profile and the corresponding label of sample that is associated with D'_i is B_i. The final predicted label of the test sample can be presented as

$$F = \begin{cases} 1, & \text{if } \sum_{i=1}^{K}(B_i == 1) \geq \sum_{i=1}^{K}(B_i \neq 1) \\ 2, & \text{if } \sum_{i=1}^{K}(B_i == 2) \geq \sum_{i=1}^{K}(B_i \neq 2) \\ \vdots & \\ n, & \text{if } \sum_{i=1}^{K}(B_i == n) \geq \sum_{i=1}^{K}(B_i \neq n). \end{cases} \tag{10}$$

Note that our scheme predicts labels based on the similarity between test signal and sample signals in the profile. Therefore, it still works well even if only a few number samples are available in the profile.

B. Phase 2

When more samples become available as time progresses, it may become difficult to still use the system of phase 1 for activity recognition. Distance-based classification requires calculation with every activity sample in the dataset for each test data. It leads to inefficiency especially with longer signals and larger sample sizes. In this case, we first extract representative features from time and frequency domain, then utilize SVM to train a model for classification. Since the model can be pretrained, it can achieve higher efficiency with larger sample size as compared with phase 1.

1) Feature Extraction: We manually select representative features from both time and frequency domains. As stated in [31], CFR power can be considered as an indicator of the speed of paths length change caused by multiple human activities. Therefore, we employ six different features, including the standard deviation, median absolute deviation, max, mean, and first and third quartile of the filtered CFR power, respectively. Besides, same features from calibrated phase [18] are also extracted. In addition, we also select six frequency domain features from different energy levels of DWT [26] which indicate the intensity of movement in each speed range. In general, we can get a group of 18 features for each sample in total.

2) Support Vector Machine: Let $x_i = \{f_1, f_2, \ldots, f_{N_f}\}$ be the ith sample and y_i be the corresponding label in feature space, where N_f is the number of extracted features. In other words, the training dataset can be expressed as $T = (x_i, y_i)$ with uncertain distribution. By applying Gaussian kernel, it converts features to a higher-dimensional feature space where classifier hyperplane can be computed by solving quadratic function as the following:

$$\underset{\text{maximum}}{q(c_1, c_2, \ldots, c_n)} = \sum_{i=1}^{n} c_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i c_i k(x_i, x_j) y_j c_j \tag{11}$$

$$\text{subject to} \quad \sum_{i=1}^{n} c_i y_i = 0, \text{ and } \quad 0 \leq c_i \leq \frac{1}{2n\lambda}. \tag{12}$$

Here, $k(x_i, x_j)$ represents kernel function that satisfies $k(x_i, x_j) = \varphi(x_i)\varphi(x_j)$. The wights ω and bias b can then be calculated as $\omega = \sum_{i=1}^{n} c_i y_i \varphi(x_i)$ and $b = \omega\dot{\varphi}(x_i) - y_i$, respectively.

In this phase, all subcarriers are used independently for classification since they show similar fluctuations for the same activity [31]. In other words, we can get predicted labels from each subcarrier CSI series. Assume the predicted result of one activity sample is $g = [g(1), g(2), \ldots, g(30 \times N_s)]$, where N_s is the number of CSI streams used for recognition. The final predicted label can be computed with majority voting as the following:

$$L = \max_{j \in [1,2,\ldots,n]} \left(\frac{\sum_{i=1}^{30 \times N_s} (g(i) == j)}{30} \right). \tag{13}$$

C. Phase 3

As abundant samples are available in the profile, we propose a deep learning network structure based on LSTM for activity recognition. It is able to automatically extract effective

features from raw signals rather than manually selecting as in phase 2, which can potentially be subjective in choosing different features [16].

1) Long Short Term Model: LSTM, as one type of RNN units, is able to remember and filter the past information in the input sequence during training process. It is proposed to solve the problem of exploding and vanishing gradient during learning long-term dependencies with back propagation in traditional RNNs. Therefore, it becomes one of the most popular system structure in many areas, including speech identification and sequence classification.

An LSTM block consists of three gates that can be configured to control information through the cell state. The first one is termed forget gate, which decides to how much previous information is removed from the memory. The forget gate vector can be denoted as

$$f_t = \sigma_g\big(W_f x_t + U_f h_{t-1} + b_f\big) \tag{14}$$

where σ_g is a sigmoid function, W and U are input and forget weight matrixes, x_t is input vector, h_{t-1} is output vector, and b is bias vector. The black block in the figure represents time delay of the self-loop. Besides, input gate decides the amount of new information allowed to flow into the memory. Input vector can be presented as

$$i_t = \sigma_g(W_i x_t + U_i h_{t-1} + b_i). \tag{15}$$

In addition, output gate decides how much information is filtered to produce the output. Similarly, the output gate vector is

$$o_t = \sigma_g(W_o x_t + U_o h_{t-1} + b_o). \tag{16}$$

Then, the cell state vector is computed by

$$c_t = f_t c_{t-1} + i_t \sigma_c(W_c x_t + U_c h_{t-1} + b_c). \tag{17}$$

Therefore, output vector can be derived as

$$h_t = o_t \sigma_h(c_t) \tag{18}$$

where both σ_c and σ_h are hyperbolic tangents.

2) Deep Learning Structure: Based on the LSTM unit as discussed above, we propose to obtain representative features from recorded CSI samples by a multilayers neural network. As shown in Fig. 3(right), it is composed of three LSTM hidden layers and one fully connected dense layer. During training process, each LSTM layer learns to filter the information from CSI input samples or the output from last layer. Denote the input CSI sample as X, the output of three LSTM layers can be represented as

$$Y_1 = \text{LSTM}(X, \Omega_1) \tag{19}$$

$$Y_2 = \text{LSTM}(Y_1, \Omega_2) \tag{20}$$

$$Y_3 = \text{LSTM}(Y_2, \Omega_3) \tag{21}$$

where Ω_1–Ω_3 are set of LSTM parameters in different layers. In this paper, the number of LSTM cells are configured as 128, 64, and 32 in three layers. In this case, the shape of Y_1 and Y_2 will be (*timesteps*) $\times$ 128 and (*timesteps*) $\times$ 64, respectively, where timesteps indicate the length of an activity sample. Differently, the output of the third LSTM layer is the hidden states of the last timestep, which length will be 32. By concatenating the LSTM layers one by one together, the network is capable of learning representatively high level features from collected CSI samples. In order to project the extracted features from the third LSTM layer into activity label probability distribution, a fully connected layer with softmax activation function is added as the output layer as follows:

$$\widehat{Y} = \text{Softmax}(W_l Y_3 + b_l) \tag{22}$$

where W_l and b_l are trainable parameters. The predicted label is computed as the corresponding index with the largest probability.

Similar as in phase 2, each CSI series from different subcarriers are considered as independent samples in order to enlarge the dataset for training purpose. It is reasonable as we observe that CSI on different subcarriers show similar fluctuation patterns after normalization [23], [31]. Besides, majority voting is applied afterwards the same as depicted in (13).

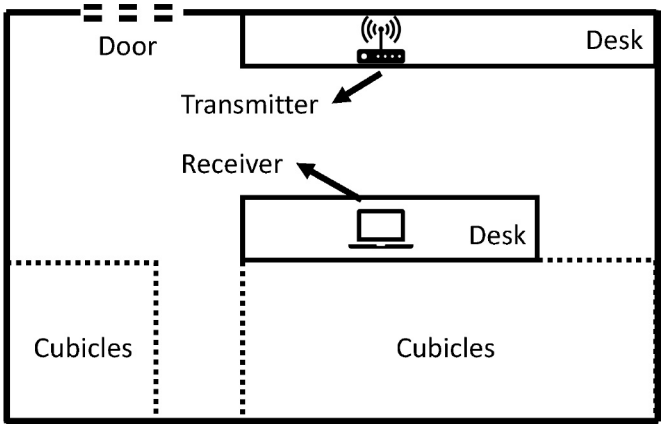


Fig. 5. Layout of laboratory where CSI samples are collected.

VI. Experiments and Evaluation

In this section, we evaluate and present the performance of the proposed three phase design, namely Wi-multi, from a variety of aspects.

A. Experimental Setup

In order to collect CSI of different activities, we deployed two off-the-shelf wireless devices in our lab, which size is around 6 $\times$ 8 m. The layout of the laboratory is shown in Fig. 5. Besides, we use a Linksys EA4500 router equipped with three antennas as transmitter and a Sony laptop equipped with two antennas as receiver. As discussed in Section III, we can obtain CSI of 30 subcarriers (one stream) from each pair of transmit–receive antennas. Therefore, we are able to record 2 $\times$ 3 CSI streams between different pair of transmit–receive antennas by installing the tool on the Intel 5300 NIC on the laptop. Considering that the frequency of most human activities is below 10 Hz [32], we configure the CSI sampling rate as 80 pkts/s. In this case, CSI is capable of capturing the movement of human bodies. Afterwards, we ask ten volunteers, with different body shape, age, and sex, to perform

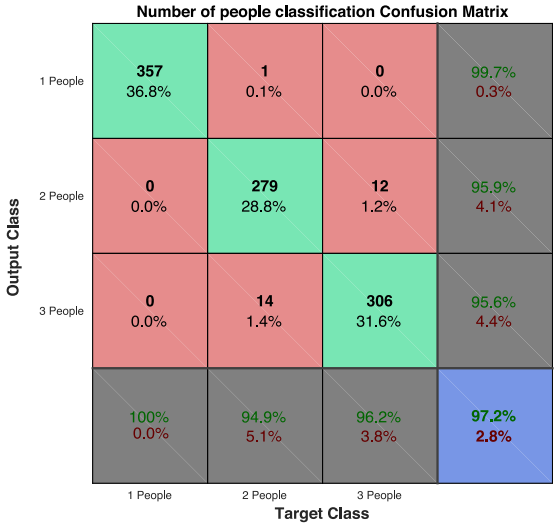


Fig. 6. Accuracy of predicting the number of subjects.

different activities in the lab. Different number of people, from 1 to 3, may be asked to perform activities at the same time. Each subject performs one of the three activities including walk (W), run (R), and hand movement (H). Different subjects may perform activities at different speeds. There are nine different combinations, including W, R , H, W&R, W&W, W&H, W&W&W, W&W&H, and W&R&H. Unless otherwise specifically mentioned, half of the samples are used for training and the other half for testing. In summary, we collect 936 samples in total. The detailed experimental results are shown as below.

B. Evaluation of Subjects Number Classification

Based on the collected CSI samples, we classify the number of subjects present in the scene. Because of space limitation in this paper, hereby we only present the results using the second phase of the system. Fig. 6 illustrates the confusion matrix of subjects number classification. Specifically, it achieves an accuracy of 100%, 94.9%, and 96.2% in predicting 1–3 subjects, respectively. Similar results can be observed with the other two phases. In summary, our system is able to detect the number of subjects present in the scene.

C. Evaluation of Phase 1

First, we compare the accuracies of activity recognition using different numbers of streams. As shown in Fig. 7(a), 1–6 streams CSI are exploited to evaluate the impact on accuracies. It is observed that all activities achieve an accuracy above 90% even with only one stream employed. Besides, it is also found that the accuracy slightly increases with more number of streams. For example, the average accuracy rises from 97.43% at one stream to 99.23% at six streams. This is reasonable since the space diversity of different antennas can provide more diverse information.

Second, we evaluate the impact of different DWT levels on recognition accuracy and efficiency. As shown in Fig. 7(b), the overall accuracy has a slight fall with larger DWT levels. We can also observe from Fig. 7(c) that the time cost of predicting new samples dramatically decreases with larger DWT levels. This is reasonable since DWT can reduce the length of the signal while keeping most of the features and thus reducing time needed for DTW computation. Indeed, the time for predicting new sample is more than 20 s without DWT compression, making it nearly impossible for real time applications. On the contrary, the time is dramatically reduced to as low as 0.14 s with a four-level DWT. To achieve a tradeoff between accuracy and efficiency, DWT level 3 seems to be the ideal choice as it achieves an average accuracy of 97.5% while the time needed for predicting a new sample is only around 0.38 s with one stream CSI.

Third, we evaluate the impact of training sample size (each activity) on recognition accuracy. As shown in Fig. 7(d), our scheme achieves desirable accuracy even with only a few number of samples available in the profile. Here the training size refers to the number of samples of each activity in the profile and the result is achieved with three levels of DWT. As shown in this figure, it achieves an average accuracy of around 85% with only ten training samples and one CSI stream. Moreover, the corresponding prediction time drops to as low 0.078 s with the decreasing DWT computation cost. Therefore, we conclude that phase 1 is able to achieve desirable performance without requiring a large number of training samples. It is suitable at the beginning stage of establishing a system when only few samples can be provided in the profile.

D. Evaluation of Phase 2

As presented earlier, it is impractical to apply phase 1 system when there are more samples (say 200 in total) available in the profile, since the computation time cost is over 1 s for predicting test data even with one CSI stream. In this case, phase 2 design can be employed. Since the SVM model can be pretrained with representative features, the time cost for predicting test samples is usually constant. We evaluate the system of phase 2 from the following aspects.

First, we evaluate the impact of features numbers. In this paper, we compare the results of 6, 12, and 18 features, respectively. These amplitude features include standard deviation, median absolute deviation, max, mean, and first and third quartile of the filtered CFR power with different cut-off frequencies. As shown in Fig. 8(a), it demonstrates the accuracy results under different number of features. For instance, the accuracy of recognizing W&H rises from 65% to 80.42% when the number of features increases from 6 to 18. In general, the average accuracy increases from 79.19% with six amplitude features to 82.87% with 18 amplitude features. Because of the limited space of this paper, we only present the result using one CSI stream here. Similar results can be observed with more CSI streams.

Second, we evaluate the impact of CSI stream numbers. As shown in Fig. 8(b), we compare the results of activity recognition accuracy with different number of CSI streams. Similar as shown in Section VI-C, it is observed that more CSI streams can be utilized in order to achieve higher accuracy. For example, the accuracy of recognizing W&R&H climbs from

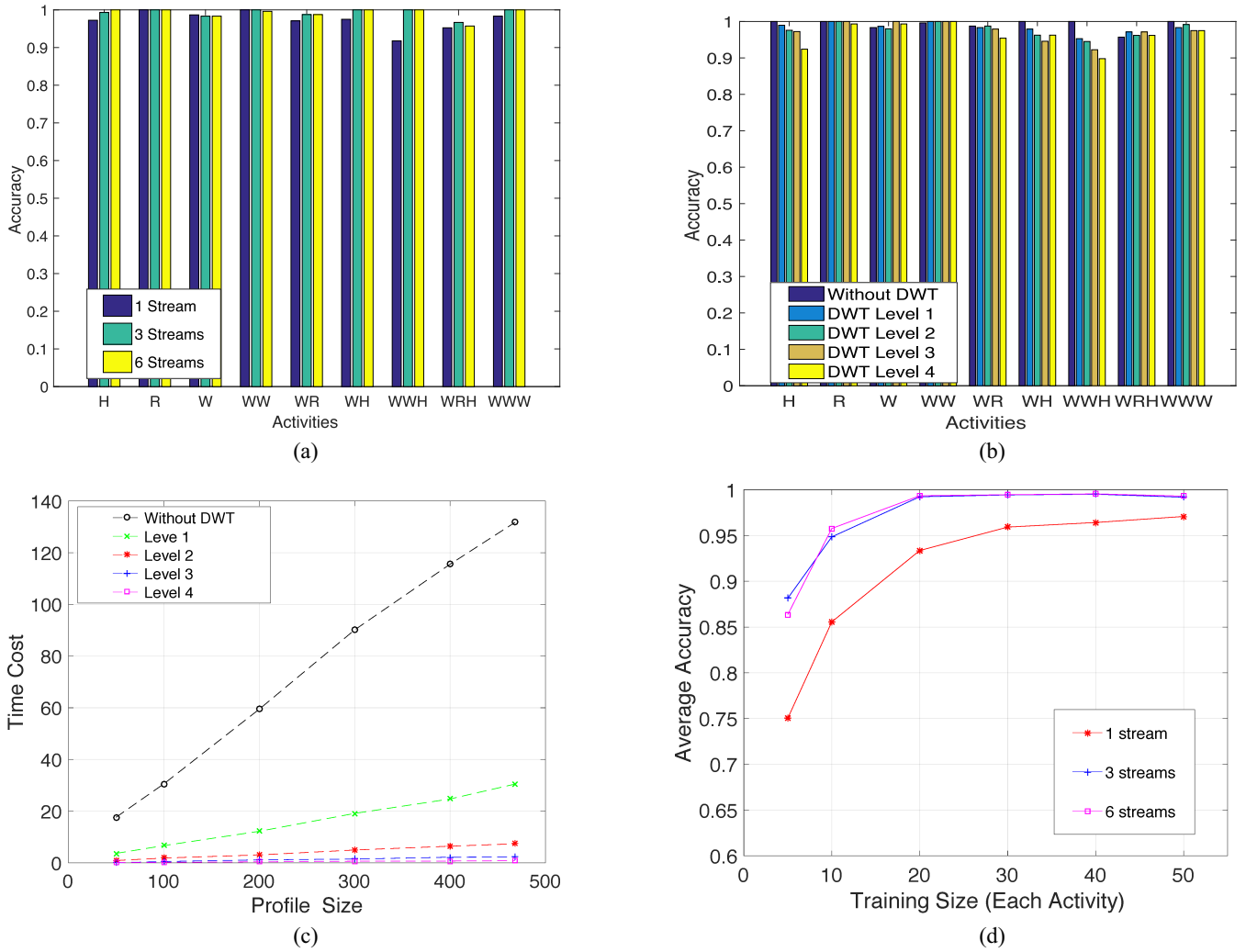


Fig. 7. Evaluation of phase 1. (a) Accuracy comparison with one stream CSI. (b) Accuracy of various DWT levels. (c) Time cost of various DWT levels. (d) Accuracy with different number of profile samples.

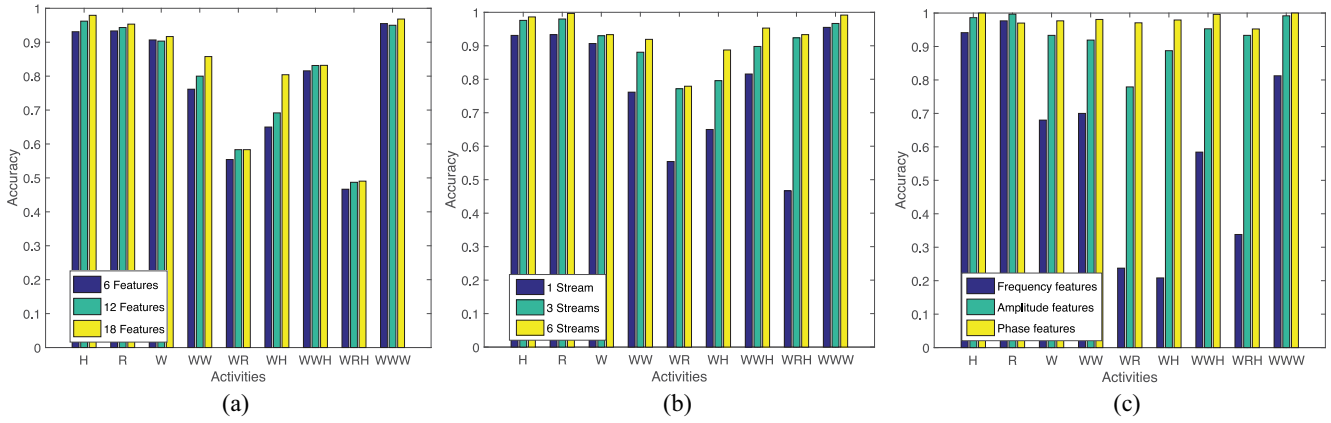


Fig. 8. Evaluation of phase 2. Impact of (a) features numbers on accuracy, (b) stream numbers on accuracy, and (c) feature metrics on accuracy.

46.67% to 93.33% when the number of CSI streams increases from 1 to 6. Besides, the average accuracy of all activities is 79.19% with one CSI stream, which is 14.31% less than the average accuracy with six CSI streams. As we discussed in Section III, each CSI stream is collected from 30 subcarriers between one pair of transmit–receive antennas. Since different pairs of antennas provide spatial diversities, it is reasonable that higher accuracy can be achieved with more CSI streams.

Third, we evaluate the impact of different feature metrics. As discussed in Section V-B, we select six different features

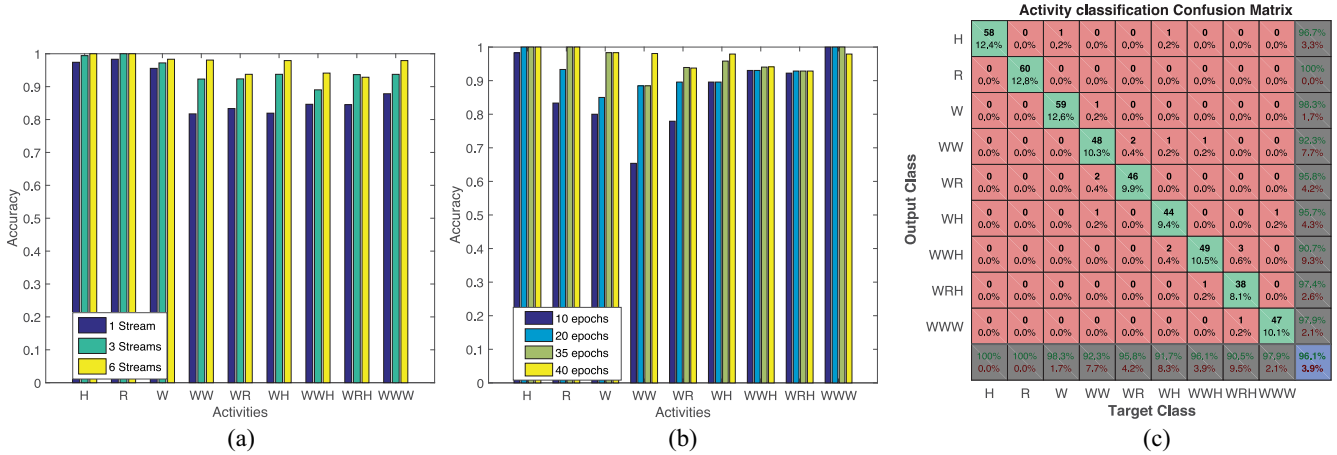


Fig. 9. Evaluation of deep learning system. Impact of (a) stream numbers on accuracy and (b) iteration numbers. (c) Performance of deep learning structure.

from amplitude, phase, and frequency domains, respectively. As shown in Fig. 8(c), it compares the accuracy results achieved by different feature metrics. Besides, we observe that results are similar using different number of CSI streams. Because of space limitation, we only present the results using 6 CSI streams in this paper. It is shown from Fig. 8(c) that the accuracies achieved by frequency domain features are much lower than the other two. For instance, the accuracy of recognizing W&R using frequency domain features is 23.75%, which is much lower compared with 77.92% of amplitude features and 97.08% of phase features. In general, the average accuracy with frequency domain features is 63.13%, which is 30.27% and 34.99% less than accuracy with amplitude and phase features, respectively. This is reasonable considering that some activities (such as W&R and W&R&H) may have similar intensity of movement, which results in alike patterns of DWT energy level.

E. Evaluation of Phase 3

As depicted in Section VI-D, it can be subjective to select different features. With increasing numbers of samples in the profile, phase 3 based on deep learning network is exploited by automatically extracting representative features. The experimental results are shown as below.

First, we evaluate the impact of different stream numbers. Similar as shown in Sections VI-C and VI-D, we observe that the accuracies with more number of streams are much higher. As shown in Fig. 9(a), the accuracy of recognizing W&W with six CSI streams is 98.08%, which is 16.35% and 5.98% higher than accuracy with one CSI stream and three CSI streams, respectively. Overall, the average accuracy increases from 88.36% to 97.22% when the number of CSI streams increases from 1 to 6.

Second, we evaluate the impact of epoch numbers. Because of the limitation of hardware, the input training samples are usually divided into small batches, which go through the network one by one. One epoch is the process of passing the entire training dataset (all batches) once through the neural network. It is known that different number of epochs may cause overfitting and underfitting of the trained model. As shown in Fig. 9(b), we compare the classification accuracies of models trained with different number of epochs. It is observed that the accuracy increases with more epochs and becomes stable after 35 epochs. This is reasonable since it requires enough epochs to update the network parameters and train a robust model.

Lastly, we present the detailed evaluation result of phase 3 in Fig. 9(c). We configure the number of CSI streams as 6, the number of epochs as 35 and the batch size as 64. From this confusion matrix, it is observed that the overall accuracy of phase 3 is 96.1%, which is higher than 95.18% of phase 1 and 93.4% of phase 2. Compared with phase 1, it requires no additional time cost when more training samples are available in the profile. This is because the model can be pretrained and does not require to compute similarities between test sample and all training samples. Moreover, compared with phase 2, it is able to automatically extract effective features. In conclusion, phase 3 of deep learning networks achieves better performance when abundant samples are collected.

F. Discussion

Based on the results shown above, it is suggested that different phases are employed separately according to the size of available CSI samples in the profile. Phase 1 is recommended at the early stage of the system deployment in new environments. This is because only few CSI samples may be available and thus machine learning or deep learning algorithms are not applicable. In this phase, our system is able to achieve desirable performance even with fewer samples provided in the profile. However, as shown in Fig. 7(c), the evaluation time to predict activity also increases with larger sample sizes. When the sample size reaches around 200, the time cost for evaluating each test sample is over 1 s. In this case, phase 2 can be employed as it can train a model beforehand. Therefore, the evaluation time is stable despite of the size of CSI samples. As we collect more and more samples, phase 3 is desired since deep learning structure can automatically extract representative features instead of manual extraction as in phase 2, where features selection can be subjective as shown in Fig. 8. Besides, similar as phase 2, the evaluation time for predicting

new samples becomes stable (around 0.40 s in our case) in phase 3 as the model can be trained beforehand.

As discussed in Section I, many existing papers [12], [25], [26] achieved desirable performance in detecting single people activities. Compared with these works, our proposed system outperforms in several aspects. First, our system is able to recognize activities with fewer CSI samples while many other systems such as WiFall [25] employ machine learning algorithms and thus can not be applicable if only few CSI samples are provided. Second, our phase 3 system can extract effective features automatically rather than manually as in CRAM [26], where different feature selections can be subjective as shown in Fig. 8. Third, our system achieves desirable tradeoff between accuracy and efficiency while systems like Keystroke [12] lack such flexibility in their presented work.

In this paper, we achieved desirable performance in multiple human activity recognition. However, there are still few challenges remaining. For example, how to extract individual activity from group activities remains unsolved. We consider this as our future work.

VII. Conclusion

Targeting at multiple human activity recognition, in this paper we propose Wi-multi, a three-phase system using CSI. At the initial stage of system deployment, it is infeasible to apply machine learning algorithms as there are usually few samples available in the profile. In this case, our designed phase 1 of the system that utilizes distance-based classification is exploited. As more samples become available for training, phase 2 of our design that employs SVM with representative features from both time and frequency domain is applied. It dramatically reduces computation costs as compared with phase 1 which requires computing similarities between the test sample and all samples in the profile. Finally, when we have enough samples for deep learning networks, phase 3 based on LSTM is proposed. It can achieve higher accuracy and efficiency since it can automatically choose representative features and pretrain the model. Given the availability of samples, each phase of our design achieves a desirable tradeoff between accuracy and efficiency.

References

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

[2] G. Fortino, S. Galzarano, R. Gravina, and W. Li, "A framework for collaborative computing and multi-sensor data fusion in body sensor networks," *Inf. Fusion*, vol. 22, pp. 50–70, Mar. 2015.

[3] G. Fortino, R. Giannantonio, R. Gravina, P. Kuryloski, and R. Jafari, "Enabling effective programming and flexible management of efficient body sensor network applications," *IEEE Trans. Human–Mach. Syst.*, vol. 43, no. 1, pp. 115–133, Jan. 2013.

[4] H. Ghasemzadeh, P. Panuccio, S. Trovato, G. Fortino, and R. Jafari, "Power-aware activity monitoring using distributed wearable sensors," *IEEE Trans. Human–Mach. Syst.*, vol. 44, no. 4, pp. 537–544, Aug. 2014.

[5] R. Bodor, B. Jackson, and N. Papanikolopoulos, "Vision-based human tracking and activity recognition," in *Proc. 11th Mediterr. Conf. Control Autom.*, vol. 1, 2003, pp. 1–6.

[6] S. Arshad, C. Feng, I. Elujide, Z. Siwang, and Y. Liu, "SafeDrive-Fi: A multimodal and device free dangerous driving recognition system using WiFi," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2018, pp. 1–6.

[7] W. Jiang *et al.*, "Towards environment independent device free human activity recognition," in *Proc. 24th ACM Annu. Int. Conf. Mobile Comput. Netw.*, 2018, pp. 289–304.

[8] A. Virmani and M. Shahzad, "Position and orientation agnostic gesture recognition using WiFi," in *Proc. 15th ACM Annu. Int. Conf. Mobile Syst. Appl. Services*, 2017, pp. 252–264.

[9] J. Yang, H. Zou, H. Jiang, and L. Xie, "Fine-grained adaptive location-independent activity recognition using commodity WiFi," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2018, pp. 1–6.

[10] R. H. Venkatnarayan, G. Page, and M. Shahzad, "Multi-user gesture recognition using WiFi," in *Proc. 16th ACM Annu. Int. Conf. Mobile Syst. Appl. Services*, 2018, pp. 401–413.

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

[12] K. Ali, A. X. Liu, W. Wang, and M. Shahzad, "Keystroke recognition using WiFi signals," in *Proc. 21st ACM Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 90–102.

[13] H. Abdelnasser, M. Youssef, and K. A. Harras, "WIGest: A ubiquitous WiFi-based gesture recognition system," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, 2015, pp. 1472–1480.

[14] X. Liu, J. Cao, S. Tang, and J. Wen, "Wi-Sleep: Contactless sleep monitoring via WiFi signals," in *Proc. IEEE Real Time Syst. Symp. (RTSS)*, 2014, pp. 346–355.

[15] *Average Size of Households in the U.S. 1960–2015*. Accessed: Dec. 2018. [Online]. Available: https://www.statista.com/statistics/183648/average-size-of-households-in-the-us/

[16] C. Feng, S. Arshad, Y. Ruiyun, and Y. Liu, "Evaluation and improvement of activity detection systems with recurrent neural network," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2018, pp. 1–6.

[17] J. Wang, L. Zhang, Q. Gao, M. Pan, and H. Wang, "Device-free wireless sensing in complex scenarios using spatial structural information," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2432–2442, Apr. 2018.

[18] X. Huang, S. Guo, Y. Wu, and Y. Yang, "A fine-grained indoor fingerprinting localization based on magnetic field strength and channel state information," *Pervasive Mobile Comput.*, vol. 41, pp. 150–165, Oct. 2017.

[19] O. Kaltiokallio, H. Yiğitler, R. Jäntti, and N. Patwari, "Non-invasive respiration rate monitoring using a single cots Tx-Rx pair," in *Proc. IEEE 13th Int. Symp. Inf. Process. Sensor Netw. (IPSN)*, 2014, pp. 59–69.

[20] M. Li *et al.*, "When CSI meets public WiFi: Inferring your mobile phone password via WiFi signals," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2016, pp. 1068–1079.

[21] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with Wi-Fi!" *IEEE Trans. Mobile Comput.*, vol. 15, no. 11, pp. 2907–2920, Nov. 2016.

[22] Y. Wang, J. Liu, Y. Chen, M. Gruteser, J. Yang, and H. Liu, "E-eyes: Device-free location-oriented activity identification using fine-grained WiFi signatures," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, 2014, pp. 617–628.

[23] S. Arshad *et al.*, "Wi-chase: A WiFi based human activity recognition system for sensorless environments," in *Proc. IEEE 18th Int. Symp. World Wireless Mobile Multimedia Netw. (WoWMoM)*, 2017, pp. 1–6.

[24] S. Arshad, C. Feng, R. Yu, and Y. Liu, "Leveraging transfer learning in multiple human activity recognition using WiFi signal," in *Proc. IEEE 20th Int. Symp. World Wireless Mobile Multimedia Netw. (WoWMoM)*, 2019, pp. 1–6.

[25] C. Han, K. Wu, Y. Wang, and L. M. Ni, "WiFall: Device-free fall detection by wireless networks," in *Proc. IEEE Conf. Comput. Commun. INFOCOM*, 2014, pp. 271–279.

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

[27] H. Wang, D. Zhang, Y. Wang, J. Ma, Y. Wang, and S. Li, "RT-fall: A real-time and contactless fall detection system with commodity WiFi devices," *IEEE Trans. Mobile Comput.*, vol. 16, no. 2, pp. 511–526, Feb. 2017.

[28] X. Zheng, J. Wang, L. Shangguan, Z. Zhou, and Y. Liu, "Smokey: Ubiquitous smoking detection with commercial WiFi infrastructures," in *Proc. IEEE 35th Annu. Int. Conf. Comput. Commun. (IEEE INFOCOM)*, 2016, pp. 1–9.

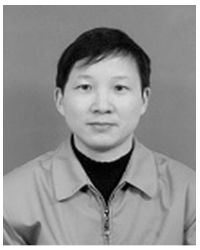
[29] W. Wang, A. X. Liu, and M. Shahzad, "Gait recognition using WiFi signals," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2016, pp. 363–373.
[30] Z. Zhou, Z. Yang, C. Wu, W. Sun, and Y. Liu, "LiFi: Line-of-sight identification with WiFi," in *Proc. IEEE Conf. Comput. Commun. INFOCOM*, 2014, pp. 2688–2696.
[31] C. Feng, S. Arshad, and Y. Liu, "MAIS: Multiple activity identification system using channel state information of WiFi signals," in *Proc. Int. Conf. Wireless Alg. Syst. Appl.*, 2017, pp. 419–432.
[32] Y. Zeng, P. H. Pathak, and P. Mohapatra, "WiWho: WiFi-based person identification in smart spaces," in *Proc. 15th Int. Conf. Inf. Process. Sensor Netw.*, 2016, p. 4.

Chunhai Feng received the B.S. and M.S. degrees from Soochow University, Suzhou, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree in computer engineering at the University of Texas at Arlington, Arlington, TX, USA.

His current research interests include wireless communications, computer networks, and machine learning.

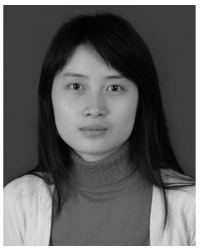
Sheheryar Arshad received the B.Sc. and M.Sc. degrees from the University of Engineering and Technology, Lahore, Lahore, Pakistan, in 2008 and 2013, respectively. He is currently pursuing the Ph.D. degree at the Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA.

His current research interests include wireless communications and networks, sensor networks, wireless sensing, and system design using machine and deep learning.



Siwang Zhou (M'18) received the B.S. degree from Fudan University, Shanghai, China, the M.S. degree from Xiangtan University, Xiangtan, China, and the Ph.D. degree from Hunan University, Changsha, China.

He has been an Associate Professor with the College of Computer Science and Electronic Engineering, Hunan University. His current research interests include sensor network, compressive sensing, and image processing.



Dun Cao received the B.S. degree in communication engineering from Central South University, Changsha, China, in 2001, the M.S. degree in information systems and communications from Hunan University, Changsha, in 2006, and the Ph.D. degree in vehicle engineering from the Changsha University of Science and Technology, Changsha, in 2017.

She is a faculty member with the School of Computer and Communication Engineering, Changsha University of Science and Technology. She was a Visiting Scholar with the National Mobile Communications Research Laboratory, Southeast University, Nanjing, China, and the University of Texas at Arlington, Arlington, TX, USA.

Yonghe Liu received the B.S. and M.S. degrees from Tsinghua University, Beijing, China, in 1998 and 1999, respectively, and the Ph.D. degree from Rice University, Houston, TX, USA, in 2004.

He is an Associate Professor with the Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA. His current research interests include wireless networking, sensor networks, security, and system integration.