Towards Environment-independent Human Activity Recognition using Deep Learning and Enhanced CSI

Zhenguo Shi*, J. Andrew Zhang*, Richard Xu*, Qingqing Cheng†, Andre Pearce*
*Global Big Data Technologies Center, University of Technology Sydney, Sydney, Australia
† University of New South Wales, Sydney, Australia
Emails:{zhenguo.shi, andre.pearce}@student.uts.edu.au; {andrew.zheng, yida.xu}@uts.edu.au
{qingqingcheng}@unsw.edu.au

***Abstract*—Deep learning has shown a strong potential in device-free human activity recognition (HAR). However, a fundamental challenge is ensuring accuracy, without re-training, when exposing a previously trained architecture to a new or unseen environment. To overcome the aforementioned challenge, this paper proposes an environment-robust channel state information (CSI) based HAR by leveraging the properties of a matching network (MatNet) and enhanced features (HAR-MN-EF). To improve the CSI quality, we propose a CSI cleaning and enhancement method (CSI-CE) that includes two key stages: activity-related information extraction (ARIE) and correlation feature extraction based on principal component analysis (CFE-PCA). The ARIE stage is able to effectively enhance the activity-dependent features whilst mitigating behavior-unrelated information. The CFE-PCA stage further improves the extracted features by filtering out the residual activity-unrelated data and the residual noise contained in signals from the former stage. The extracted features are then sequenced into the MatNet to create an environment-robust HAR. Experimental results confirm that an architecture trained by the proposed HAR-MN-EF can be directly adapted to a new environment, achieving reliable sensing accuracies without requiring additional effort.**

***Index Terms*—WiFi, Device free sensing, Channel state information, Human activity recognition, One-shot learning.**

I. Introduction

Over the past decades, WiFi signal has had an influential role in device-free human activity recognition (HAR), attracting significant interest from the research community [1]. The key insight of WiFi-based HAR is to investigate the different influences on WiFi signal propagations caused by various human activities [2]. This relationship between human activities and WiFi signal propagations enables WiFi-based HAR to function without the need of a wearable device, thereby bringing several advantages, such as convenience, low-cost and privacy protection [3]. In this paper, we study WiFi-based HAR by ultimately exploring the properties of the channel state information (CSI).

Recent innovations in CSI-based HAR have employed deep learning (DL) networks to extract inherent features from CSI for classification. Various DL networks have been explored in pioneering HAR approaches. For instance, the authors in [4] investigated HAR by leveraging a sparse auto-encoder network. Furthermore, convolutional neural networking (CNN) [5] and long-short term memory recurrent neural networking (LSTM-RNN) [6] have also been widely explored for recognizing different activities. Although these methods are able to achieve desirable sensing results, their performance is strongly environment-dependent. In other words, a DL network trained under one environment cannot be applied to another, severely prohibiting applicabilities of DL-HAR in practice. To address this problem, various pioneering efforts have been made [7], [8]. To name a few, a recent solution, [7], proposed a cross-environment recognition method drawing support from an adversarial learning network. The authors in [8] developed an environment-independent HAR by leveraging the property of transfer learning. A transfer neural network is explored in [9], to ultimately remove the environment-specific information included in human behaviors. Despite the effectiveness in environment-robustness, these methods usually fail to achieve reliable sensing accuracies. A large number of different source environments are required for training purposes, which restricts the potential in practical applications. In addition to this limitation, some studies, such as that conducted by [8], cannot effectively classify light human activities, such as laying and standing.

In order to address the aforementioned problems and limitations, we investigate an environment-robust CSI-based HAR, drawing support from a matching network (MatNet) [10] and enhanced features (HAR-MN-EF). Under the proposed HAR-MN-EF scheme, an architecture trained with a limited number of source environments can be used to directly identify different activities in a new (testing) environment, without the requirement of re-training. In comparison with the existing HAR methods, our proposed scheme is able to achieve more reliable sensing accuracy under dynamic parameters and conditions. To improve the quality of the CSI, we developed a CSI cleaning and enhancement method (CSI-CE), which ultimately improves activity-dependent features whilst removing environment-specific information from the raw CSI. Furthermore, the dimension of the signals provided to the MatNet can be reduced by CSI-CE, thereby significantly decreasing the training complexity. Two key

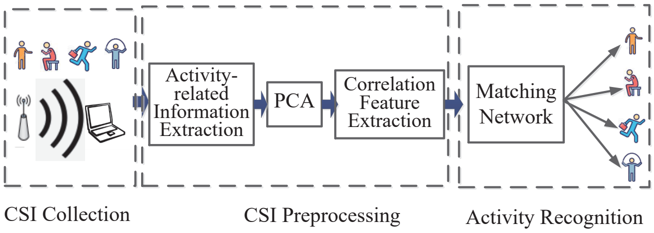


Fig. 1. Main processing modules of the HAR-MN-EF Scheme.

stages are included in CSI-CE: activity-related information extraction (ARIE) and correlation feature extraction based on principle component analysis (CFE-PCA). For the ARIE stage, we propose a method to mitigate activity-independent features, whilst retaining the activity-related information from the raw CSI. For the CFE-PCA stage, we further filter out the residual activity-unrelated information and residual noise contained in signals from the former stage. As a result, the extracted information will be more activity-related and less environment-dependent. The properties of a MatNet, an advanced one-shot learning technology [10], is then leveraged to learn and extract inherent and transferable features from the enhanced features. To evaluate the sensing performance, we conducted extensive experiments under various conditions and parameters. Experimental results demonstrate that the proposed HAR-MN-EF is superior to the existing methods, facilitating a successful environment-robust HAR with notable higher sensing accuracy.

II. Overview of HAR-MN-EF Scheme

The architecture of the proposed HAR-MN-EF scheme, illustrated in Fig. 1, includes three main modules: CSI collection, CSI preprocessing and activity recognition. The detailed introduction of the last two modules are presented in Section III and Section IV, respectively.

CSI Collection: The intention of this module is to collect the CSI, which reflects differences of human behaviors on wireless channels. Suppose there is a person performing different activities in an indoor environment equipped with a WiFi network, as illustrated in Fig. 1. This would induce various distortions to wireless signal propagation, such as phase shift, the number of multi-path and amplitude attenuation. These variations are essential in achieving a successful HAR as they contain discriminative characteristics of different activities. To acquire CSI, we adopt a widely accepted commercial off-the-shelf WiFi, the Intel 5300 network interface card (NIC). The CSI tools [11] are employed in this work to extract CSI from 30 subcarriers for each pair of transmitter-receiver antennas, which is in accordance with the IEEE 802.11n protocol. The detailed experimental setup is provided for reference in Section V-A.

CSI Preprocessing: The purpose of this module is to improve the quality of the CSI matrix, by mitigating the activity-unrelated information and condensing activity-dependent features. To that end, we propose CSI-CE that included two main stages: *ARIE* and *CFE-PCA*. In the first stage, we start by computing the static objects contained in the received signals by conducting a linear recursive operation, then removing it from the raw CSI matrix via subtraction. In the second stage, we perform principle component analysis (PCA) and correlation operations on the channel matrix from stage 1, obtaining the correlation feature matrix (CFM). Through these stages, the CFM is expected to contain considerably reduced activity-unrelated information. Furthermore, the dimension of the CFM is largely smaller in comparison with the raw CSI matrix, significantly decreasing the processing overhead.

Activity Recognition This module is designed to classify different human activities with the help of the extracted CFM and a MatNet. This is achieved by utilizing the MatNet to automatically learn and extract inherent and transferable features from the CFM. Using these extracted features, the relationship between source environments and the testing environment can be built, which is essential for facilitating a successful HAR in the testing environment. Note that, the trained architecture can be directly used in a new environment, without requiring any re-training process. This makes the proposed scheme suitable for practical deployments, which largely credits the unique property of a MatNet. To ultimately achieve HAR in this module, we first train the MatNet offline using the extracted information from Module 2; Then, we use the well-trained network online to distinguish different human behaviors in the testing environment.

III. CSI-CE based CSI Preprocessing

In this section, we describe in detail the proposed CSI-CE for CSI processing, including ARIE and CFE-PCA.

A. Activity-related information extraction

Let $\mathbf{h}(m)$ be the magnitude of CSI vector at the m-th received packet, which can be given by

$$\mathbf{h}(m) = [H_{1,1}(m),\dots,H_{1,k}(m),\dots,H_{l,k}(m),\dots,H_{L,K}(m)]^T, \quad (1)$$

where $H_{l,k}(m)$ is the CSI information in the lth wireless link for the kth subcarrier; K stands for the total number of subcarriers in each wireless link; L denotes the number of wireless links in total, and $L = N_t \times N_r$, N_t and N_r represents the number of transmitter and receiver antennas, respectively; T represents the transpose operation. The CSI matrix $\mathbf{H}$ can be obtained by acquiring CSI vectors from M packets, by

$$\mathbf{H} = [\mathbf{h}(1), \dots, \mathbf{h}(m) \dots, \mathbf{h}(M)]. \quad (2)$$

The key task of ARIE is to extract feature signals that are more activity-dependent and environment-independent. For that, it is necessary to mitigate the activity-unrelated data and retain the activity-related information, making the extracted features more robust to various experimental environments. To do that, we partition $\mathbf{H}$ into two parts: dynamic CSI and static CSI, which can be written as

$$\mathbf{H}(m) = \mathbf{H}_s(m) + \mathbf{H}_d(m), \quad (3)$$

where $\mathbf{H}_d(m)$ represents the dynamic CSI vector induced by human behaviors; $\mathbf{H}_s(m)$ stands for the static CSI vector that is unrelated to human activities. Note that, although $\mathbf{H}_s(m)$ does not present characters of human activities, it has greater impact on the sensing performance than $\mathbf{H}_d(m)$. The reason is that a person's activities generally induce limited impact on the whole environment, especially for light activities, such as sitting and standing. Under this situation, the recognition performance would be severely dropped if directly using $\mathbf{H}(m)$ for HAR. Therefore, it is necessary to remove the static CSI vector $\mathbf{H}_s(m)$ from $\mathbf{H}(m)$, in order to significantly improve the quality of extracted feature signals and simplify the signal structure.

To filter out $\mathbf{H}_s(m)$ from $\mathbf{H}(m)$, we propose a recursive algorithm by leveraging the exponentially weighted moving average approach [12]. In such a case, $\mathbf{H}_s(m)$ from the m-th recursion can be estimated as

$$\widehat{\mathbf{H}}_s(m) = \gamma\mathbf{H}(m) + (1-\gamma)\widehat{\mathbf{H}}_s(m-1), \tag{4}$$

where $\widehat{\mathbf{H}}_s(m)$ stands for the estimated value of $\mathbf{H}_s(m)$, and γ is the forgetting factor. The initial value of $\widehat{\mathbf{H}}_s(1)$ is selected as $\mathbf{0}$. Under this situation, the estimated dynamic CSI, $\widehat{\mathbf{H}}_d(m)$, can be expressed as

$$\widehat{\mathbf{H}}_d(m) = \mathbf{H}(m) - \widehat{\mathbf{H}}_s(m). \tag{5}$$

In this regard, the estimation of the dynamic CSI matrix for M packets can be given by

$$\widehat{\mathbf{H}}_d = [\widehat{\mathbf{H}}_d(1), \ldots, \widehat{\mathbf{H}}_d(m), \ldots, \widehat{\mathbf{H}}_d(M)]. \tag{6}$$

Through the above steps, $\widehat{\mathbf{H}}_d$ contains mostly activity-dependent information, which is significant to extract distinctive features for classifying different activities.

B. Correlation Feature Extraction based on PCA

It should be noted that $\widehat{\mathbf{H}}_d$ can reflect discriminative features of different human behaviors, while it still contains some residual noise and residual activity-independent information. To deal with this problem, we propose a CFE-PCA operation, to acquire more reliable features for HAR.

We conduct a PCA operation on $\widehat{\mathbf{H}}_d$ to eliminate the residual activity-unrelated information and noise, while retaining activity-related data. For the maximum preservation of activity-dependent information, all the available principal components are selected, which is

$$\mathbf{C}_p = \widehat{\mathbf{H}}_d \times \mathrm{PCA}(\widehat{\mathbf{H}}_d), \tag{7}$$

where $\mathbf{C}_p$ denotes the extracted features via the PCA operation, with size of $L \times M$; PCA(.) represents the operation of principle component analysis. It is important to note that different principal components are correlated, which can be used to offer extra information for HAR. As a result, we perform a correlation operation on the output of the PCA, obtaining the PCA based CFM, by

$$\mathbf{D}_C = \mathbf{C}_p \times \mathbf{C}_p^T. \tag{8}$$

where $\mathbf{D}_C$ denotes the CFM that is treated as the input signal for training the MatNet.

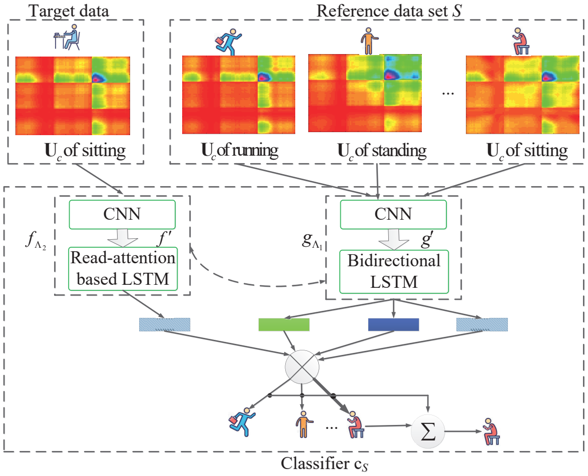

Target data
Reference data set S
$\mathbf{U}_c$ of sitting
$\mathbf{U}_c$ of running
$\mathbf{U}_c$ of standing
...
$\mathbf{U}_c$ of sitting
f_{Λ_2}
CNN
f'
Read-attention based LSTM
g_{Λ_1}
CNN
g'
Bidirectional LSTM
...
Σ
Classifier c_S

Fig. 2. Structure of a MatNet based HAR using CFM $\mathbf{U}_C$ as input.

IV. MatNet-based Human Activity Recognition

As aforementioned in the previous section, we proposed a CSI preprocessing method to eliminate the impact of the environment on the extracted information, ultimately making it environment agnostic. In this section, we employ a MatNet, a one-shot learning network augmented with external memory, to learn transferable features from the enhanced CSI. This can effectively bridge the relationship between source environments and the test environment, resulting in a successful HAR and improving its environment-robustness.

A. Architecture of matching network

The architecture of HAR using a MatNet is illustrated in Fig 2. The core function of a MatNet is to build a classifier c_s for a given source data set S, which maps S to c_S, $S \rightarrow c_S(.)$. In this regard, the expression of S with N samples can be written as

$$S = \{(x_i, y_i)\}_{i=1}^{N}, \tag{9}$$

where (x, y) stands for the CFM-label pairs; $x = \{\mathbf{D}_c\}$ with a size of $L \times L$ denotes the input data for the CFM; y represents the label for the corresponding behavior.

Given a target sample $\hat{x}$, we define the probability distribution of the output $\hat{y}$ as

$$P(\hat{y}|\hat{x}, S) \triangleq S \rightarrow c_S(\hat{x}), \tag{10}$$

where P denotes the probability distribution that is parameterized by the CNN and LSTM, as demonstrated in Fig. 2. Following this, we can obtain the estimated output label $\hat{y}$ given a source data set S and input $\hat{x}$, by

$$\hat{y} = \arg\max_{y} P(y|\hat{x}, S). \tag{11}$$

To achieve the estimated $\hat{y}$, we calculate the linear combination of y with the source data set S. Suppose x_i represents the CFM, and y_i denotes the corresponding label from the source data set $S = \{(x_i, y_i)\}_{i=1}^{N}$, then equation (11) is equal to

$$\hat{y} = \sum_{i=1}^{N} a(\hat{x}, x_i) y_i \tag{12}$$

where $a(.)$ stands for an attention mechanism in the form of softmax over the *cosine similarity*, given by

$$a(\hat{x}, x_i) = \frac{e^{\cos(f(\hat{x}), g(x_i))}}{\sum_{j=1}^{N} e^{\cos(f(\hat{x}), g(x_j))}}, \tag{13}$$

where $\cos(\alpha, \beta)$ denotes the cosine similarity function, which can be written as

$$\cos(\alpha, \beta) = \frac{\alpha \cdot \beta}{\parallel \alpha \parallel \parallel \beta \parallel}. \tag{14}$$

In equation (13), f and g denote the embedding functions of $\hat{x}$ and x_i, respectively, which can be used to extract features from input signals. As shown in Fig. 2, both f and g act as a bridge to input data for obtaining the maximum performance with the classifier as discussed in equation (12).

To learn the discriminative and transferable features for one-short learning, we design f and g to embed $\hat{x}$ and x_i fully based on the whole source data set S. Under this situation, f and g can be expressed as $f(\hat{x}, S)$ and $g(x_i, S)$, respectively.

According to Fig. 2, the embedding function g consists of a CNN with a bidirectional LSTM [13]. The CNN includes several stacked modules such as the convolution layer, ReLU non-linearity and max-pooling layer. The output of the CNN, $g'(x_i)$, which can be treated as the distinguishable features of x_i, is put into the bidirectional LSTM for processing. We can obtain $g(x_i, S)$ by

$$g(x_i, S) = \vec{h}_i + \overleftarrow{h}_i + g'(x_i), \tag{15}$$

$$\vec{h}_i, \vec{c}_i = \text{LSTM}(g'(x_i), \vec{h}_{i-1}, \vec{c}_{i-1}), \tag{16}$$

$$\overleftarrow{h}_i, \overleftarrow{c}_i = \text{LSTM}(g'(x_i), \overleftarrow{h}_{i+1}, \overleftarrow{c}_{i+1}), \tag{17}$$

where $\vec{h}_i$ and $\overleftarrow{h}_i$ stand for the output of the forward and backward LSTM, respectively; $\vec{c}_i$ and $\overleftarrow{c}_i$ denote the cell of the forward and backward LSTM, respectively; and $\text{LSTM}(g', h, c)$ is in accordance with the definition in [14]. Note that g plays a significant role in embedding x_i, especially when an element x_j is very close to x_i. Suppose x_i and x_j are input signals for two similar activities (such as sitting and sit down), respectively, g is able to map x_i and x_j to two discriminative domains conditioned on the whole source data set. This significantly improves the recognition accuracy when classifying these two behaviors.

For the embedding function f, it is also composed by a CNN with LSTM. The CNN adopted here is the same as the one in g, but the architecture of the LSTM is different, that being the read-attention based LSTM [15]. Suppose attLSTM(.) stands for the read-attention based LSTM, given a target sample $\hat{x}$, the output of attLSTM(.) based on the whole source data set S can be obtained by

$$f(\hat{x}, S) = \text{attLSTM}(f'(\hat{x}), g(S), N_p), \tag{18}$$

where $f'(\hat{x})$ denotes the input data for the read-attention based LSTM, which is extracted from the CNN (similar to the case in g); $g(S)$ stands for the output achieved by embedding the signal x_i from the source data set S; and N_p is the number of unrolling steps in the LSTM. In such a case, the state of the LSTM for the n_pth step can be written as

$$h_{n_p} = \hat{h}_{n_p} + f'(\hat{x}), \tag{19}$$

$$\hat{h}_{n_p}, c_{n_p} = \text{LSTM}(f'(\hat{x}), [h_{n_p-1}, r_{n_p-1}], c_{n_p-1}), \tag{20}$$

where $\text{LSTM}(f'(\hat{x}), [h_{n_p-1}, r_{n_p-1}], c_{n_p-1})$ follows the implementation described in [14]; r_{n_p-1} denotes the read-out from $g(S)$, which is concatenated to h_{k-1}, and it can be written as

$$r_{n_p-1} = \sum_{i=1}^{N_s} a(h_{n_p-1}, g(x_i)) g(x_i), \tag{21}$$

where N_s stands for the length of $g(S)$; $a(\cdot, \cdot)$ represents the attention function of softmax, which is

$$a(h_{n_p-1}, g(x_i)) = \text{softmax}(h_{n_p-1}^T g(x_i)). \tag{22}$$

From the above, N_p steps of "reads" are performed, so we obtain $\text{attLSTM}(f'(\hat{x}), g(S), N_p) = h_{N_p}$, where h_{n_p} is provided in (19).

B. Training Strategy

We let $\mathcal{T}$ be a task that can be treated as a distribution over possible label sets of human behaviors. In each episode, we sample a set of human activities (L) from $\mathcal{T}$, $L \sim \mathcal{T}$, including several behaviors: $\{empty, lying, standup, standing, walk, fall\}$. Next we use L to sample a batch of target set B and the source data set S, getting $\mathcal{B} = B \sim L$ and $\mathcal{S} = S \sim L$. The task of training the MatNet is to minimize the error by estimating the labels in the batch $\mathcal{B}$ conditional on $\mathcal{S}$. In such a case, the loss function of MatNet based HAR, $\mathcal{L}$, can be obtained by

$$\mathcal{L} = -\text{E}_{L \sim \mathcal{T}} \left[\text{E}_{\mathcal{S}, \mathcal{B}} \left[\sum_{(x,y) \in B} \log P_\Lambda \left(y | x, \mathcal{S} \right) \right] \right], \tag{23}$$

where $\Lambda = \{\Lambda_1, \Lambda_2\}$, Λ_1 and Λ_2 denote the parameter sets of embedding functions g and f, respectively. The key objective of the training process is to minimize the loss function given a batch for a source data set $\mathcal{S}$, which is

$$\Lambda = \arg\min_{\Lambda} \mathcal{L}(\Lambda). \tag{24}$$

V. Implementation and Evaluation

In this section, we design and perform a wide range of experiments to verify the performance of the proposed HAR-MN-EF.

A. Experimental Setup

For implementing the developed HAR-MN-EF, two computers, each equipped with an Intel WiFi NIC5300 network card, are employed as a transmitter and receiver. The number of antennas at transmitter and receiver are $N_t = 1$ and $N_r = 3$, respectively. For each pair of transmitter-receiver antennas, the total number of subcarriers is 30 ($S = 30$). The operating frequency band is 5.32 GHz. A sliding window with

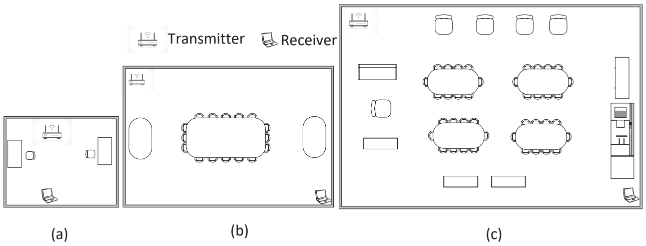


Fig. 3. Layout of three indoor experimental areas: (a) $3m \times 4m$ office. (b) $4m \times 6m$ meeting room. (c) $6m \times 7m$ laboratory.

TABLE I
AVERAGE RECOGNITION ACCURACY OF DIFFERENT METHODS IN THREE INDOOR CONFIGURATIONS

Method	1st Exp.	2nd Exp.	3rd Exp.
Proposed HAR-MN-EF	0.701	0.652	0.608
EI	0.523	0.471	0.413
RNN	0.289	0.254	0.201
TNNAR	0.475	0.421	0.362

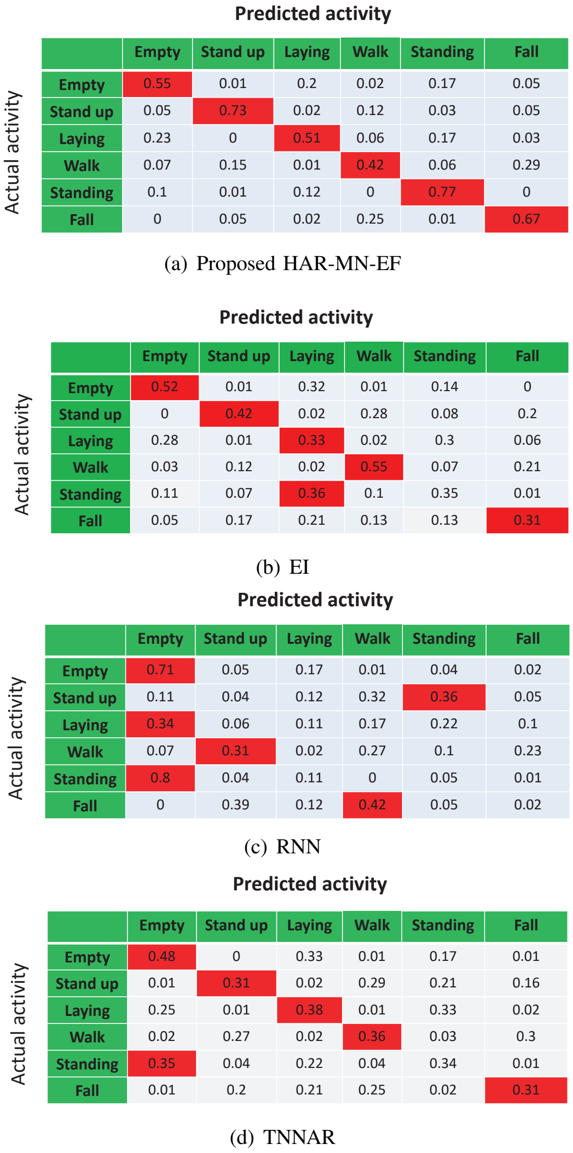
a time length of 2s is used to collect the CSI samples of each activity, and the rate of samples is 1 KHz. Under this situation, the dimension of the CSI matrix ($\mathbf{H}$) is 90×2000. Each embedding function of HAR-MN-EF consists of a CNN with 8 convolutional layers, with each layer including a 3×3 convolution, a 2×2 max-pooling, and a ReLU non-linearity operation. We use a 3.4 GHz PC with an Nvidia P5000 graphic card (16GB memory) to train the proposed HAR-MN-EF scheme. We set the number of training iterations, batch size, and learning rate as 1000, 64 and 0.001, respectively.

The proposed HAR-MN-EF scheme is deployed in three indoor environments, with different configurational complexities, as shown in Fig. 3. The size of first, second, and third configurations are $3m \times 4m$, $4m \times 6m$, and $6m \times 7m$, respectively. In each environment, there are serval objects placed between the transmitter and receiver. Five people are performing six different activities in each configuration. The total number of rounds for each activity is 200. These six behaviors are collected as the dataset that is divided into the training dataset and testing dataset.

B. Performance Evaluation

To validate the recognition performance, we first compare the sensing accuracy of the proposed HAR-MN-EF and other three HAR methods (i.e., RNN [6], EI [7], and TNNAR [9]), under various configurations and parameters. Then, we explore the influence of the developed CSI-CE on the performance of the proposed HAR-MN-EF. Note that, each HAR scheme is trained using source environments, and the well-trained architecture is used for HAR in the testing environment, without any re-training process.

In Table I, we show the average sensing accuracy of four HAR methods for six activities. In this Table, when one of three configurations in Fig. 3 is selected as the testing environment, the other two are chosen as source environments. From this table, it is clear that the proposed HAR-MN-EF is notably superior in all configurations in comparison to the other three methods. This is because we proposed a CSI-CE method to enhance and condense the activity-related features whilst mitigating the activity-independent information from input signals. As a result, the behavior-dependent information can be effectively learned and extracted, contributing to a reliable recognition.


Predicted activity
Actual activity

	Empty	Stand up	Laying	Walk	Standing	Fall
Empty	0.55	0.01	0.2	0.02	0.17	0.05
Stand up	0.05	0.73	0.02	0.12	0.03	0.05
Laying	0.23	0	0.51	0.06	0.17	0.03
Walk	0.07	0.15	0.01	0.42	0.06	0.29
Standing	0.1	0.01	0.12	0	0.77	0
Fall	0	0.05	0.02	0.25	0.01	0.67

(a) Proposed HAR-MN-EF

	Empty	Stand up	Laying	Walk	Standing	Fall
Empty	0.52	0.01	0.32	0.01	0.14	0
Stand up	0	0.42	0.02	0.28	0.08	0.2
Laying	0.28	0.01	0.33	0.02	0.3	0.06
Walk	0.03	0.12	0.02	0.55	0.07	0.21
Standing	0.11	0.07	0.36	0.1	0.35	0.01
Fall	0.05	0.17	0.21	0.13	0.13	0.31

(b) EI

	Empty	Stand up	Laying	Walk	Standing	Fall
Empty	0.71	0.05	0.17	0.01	0.04	0.02
Stand up	0.11	0.04	0.12	0.32	0.36	0.05
Laying	0.34	0.06	0.11	0.17	0.22	0.1
Walk	0.07	0.31	0.02	0.27	0.1	0.23
Standing	0.8	0.04	0.11	0	0.05	0.01
Fall	0	0.39	0.12	0.42	0.05	0.02

(c) RNN

	Empty	Stand up	Laying	Walk	Standing	Fall
Empty	0.48	0	0.33	0.01	0.17	0.01
Stand up	0.01	0.31	0.02	0.29	0.21	0.16
Laying	0.25	0.01	0.38	0.01	0.33	0.02
Walk	0.02	0.27	0.02	0.36	0.03	0.3
Standing	0.35	0.04	0.22	0.04	0.34	0.01
Fall	0.01	0.2	0.21	0.25	0.02	0.31

(d) TNNAR

Fig. 4. Confusion matrix for different human activity recognition methods.

To further investigate the sensing performance, we demonstrate the confusion matrix of each scheme, as illustrated in Fig. 4. In this figure, the third configuration is selected as the testing environment. As the figure highlights, the proposed HAR-MN-EF significantly outperforms the other three methods, in terms of identifying different behaviors. To be specific, for the proposed HAR-MN-EF, the predicted activity is in accordance with the corresponding actual activity, which implies that the proposed scheme is able to accomplish a successful HAR. In contrast, the predicted activities of the other three methods do not match their corresponding actual behaviors, alluding to the fact that those methods have difficulties predicting activities correctly.

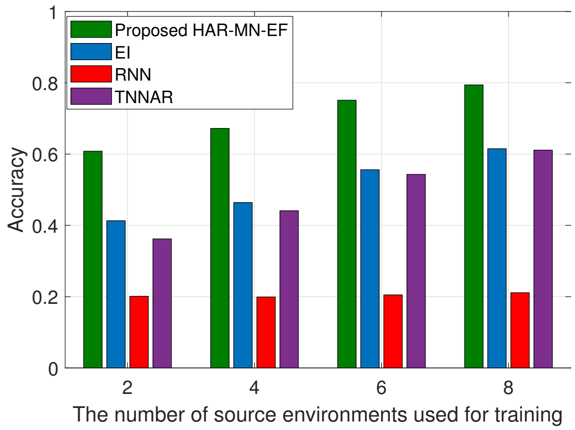


Fig. 5. Recognition accuracy with increased number of source environment

TABLE II
IMPACT OF CSI-CE ON PROPOSED HAR-MN-EF

Method	Accuracy	Training Time
With CSI-CE	0.751	34.5 min
Without CSI-CE	0.582	89.7 min

Fig. 5 illustrates how the accuracy of each method varies with the increased number of source environments. From this figure, the sensing accuracies of the proposed HAR-MN-EF, EI and TANNR can be improved with more source environments. This is due to the fact that increasing the number of source environments allows these three methods to extract more transferable features shared by source environments, which is beneficial in distinguishing different activities. In contrast, more source environments do not help the RNN in [6] to achieve a better sensing accuracy. This is because, the RNN in [6] does not have the capability to extract common features shared by source environments and the testing environment.

The impact of the proposed CSI-CE method on the performance of HAR-MN-EF is shown in Table II, in the context of sensing accuracy and training time. In this table, we train the HAR scheme using six source environments. It is clear that the proposed CSI-CE method significantly improves the sensing accuracy of HAR-MN-EF, in addition to reducing the training time. This is largely due to the property of the proposed CSI-CE, which ultimately enhances the activity-related information and reduces the size of input signals.

VI. CONCLUSION

In this paper, we propose the HAR-MN-EF scheme to accomplish environment-independent HAR by leveraging the property of MatNet and the proposed CSI-CE method. Using CSI-CE, the activity-dependent features can be enhanced, whilst mitigating the behavior-unrelated information from input signals. Furthermore, the CSI-CE method is also able to reduce the training complexity of the proposed scheme via decreasing the size of input data. To achieve successful cross-environment HAR, the MatNet is adopted to process features extracted by CSI-CE. Through extensive experiments, we validate that an architecture trained by the proposed HAR-MN-EF with source environments can be directly used in new environments for HAR without requiring additional effort. In addition, the experimental results illustrated in this paper demonstrate that the proposed HAR-MN-EF scheme significantly outperforms the state-of-the-art methods in terms of sensing accuracy.

REFERENCES

[1] X. Guo, B. Liu, C. Shi, H. Liu, Y. Chen, and M. C. Chuah, "Wifi-enabled smart human dynamics monitoring," in *Proceedings of the 15th ACM Conference on Embedded Network Sensor Systems*, ser. SenSys '17. New York, NY, USA: ACM, 2017, pp. 16:1–16:13.

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

[3] Y. Wang, K. Wu, and L. M. Ni, "Wifall: Device-free fall detection by wireless networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 2, pp. 581–594, Feb 2017.

[4] Q. Gao, J. Wang, X. Ma, X. Feng, and H. Wang, "Csi-based device-free wireless localization and activity recognition using radio image features," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10 346–10 356, Nov 2017.

[5] F. Wang, W. Gong, and J. Liu, "On spatial diversity in wifi-based human activity recognition: A deep learning-based approach," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 2035–2047, April 2019.

[6] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A survey on behavior recognition using wifi channel state information," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 98–104, Oct 2017.

[7] W. Jiang, C. Miao, F. Ma, S. Yao, Y. Wang, Y. Yuan, H. Xue, C. Song, X. Ma, D. Koutsonikolas, W. Xu, and L. Su, "Towards environment independent device free human activity recognition," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18. New York, NY, USA: ACM, 2018, pp. 289–304.

[8] J. Zhang, Z. Tang, M. Li, D. Fang, P. Nurmi, and Z. Wang, "Crosssense: Towards cross-site and large-scale wifi sensing," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '18. New York, NY, USA: ACM, 2018, pp. 305–320.

[9] J. Wang, V. W. Zheng, Y. Chen, and M. Huang, "Deep transfer learning for cross-domain activity recognition," in *Proceedings of the 3rd International Conference on Crowd Science and Engineering*, ser. ICCSE?8. New York, NY, USA: Association for Computing Machinery, 2018.

[10] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 3630–3638.

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

[12] S. W. Roberts, "Control chart tests based on geometric moving averages," *Technometrics*, vol. 1, no. 3, pp. 239–250, 1959.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[15] O. Vinyals, S. Bengio, and M. Kudlur, "Order matters: Sequence to sequence for sets," *arXiv preprint arXiv:1511.06391*, 2015.