Environment-Robust Device-free Human Activity Recognition with Channel-State-Information Enhancement and One-Shot Learning

Zhenguo Shi, *Student Member, IEEE,* J. Andrew Zhang, *Senior Member, IEEE,*
Richard Yida Xu, *Member, IEEE,* Qingqing Cheng, *Student Member, IEEE,*

Abstract—Deep Learning plays an increasingly important role in device-free WiFi Sensing for human activity recognition (HAR). Despite its strong potential, significant challenges exist and are associated with the fact that one may require a large amount of samples for training, and the trained network cannot be easily adapted to a new environment. To address these challenges, we develop a novel scheme using Matching Network with enhanced channel state information (MatNet-eCSI) to facilitate one-shot learning HAR. We propose a CSI Correlation Feature Extraction (CCFE) method to improve and condense the activity-related information in input signals. It can also significantly reduce the computational complexity by decreasing the dimensions of input signals. We also propose novel training strategy which effectively utilizes the data set from the previously seen environments (PSE). In the least, the strategy can effectively realize human activity recognition using only one sample for each activity from the testing environment and the data set from one PSE. Numerous experiments are conducted and the results demonstrate that our proposed scheme significantly outperforms state-of-the-art HAR methods, achieving higher recognition accuracy and less training time.

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

1 Introduction

Device-free human activity recognition (HAR) using WiFi signals has drawn considerable interest from the research community. In contrast to traditional device-based sensing techniques, WiFi-based HAR removes the requirement of equipping the target with any devices, and accomplishes the classification task by analyzing the differences in WiFi signal propagation induced by different human activities [1], [2]. WiFi-based HAR possesses several advantages, including convenience, wide availability, and privacy protection, making it an attractive sensing solution for a wide range of applications in smart home, health care, and intelligent monitoring [3].

Channel state information (CSI) based HAR receives particular interest recently as CSI provides fine-grained channel information such as amplitude, phase, and frequency diversity [4], [5], [6]. Various pioneering approaches for CSI-HAR have been proposed by exploring the properties of machine learning (ML) networks and signal processing techniques [7], [8]. While they have achieved some promising results, the performance of these methods nonetheless heavily depends on the precursor step of careful selection and fusion of features. Therefore, should the precursor steps fail to achieve its goal, the recognition accuracy may degrade significantly [9]. A variety of Deep Learning Network (DLN) methodologies have been proposed in an effort to overcome this problem. For instance, the authors in [10] proposed an activity recognition scheme, in which the sparse autoencoder (SAE) is used to extract discriminative features from CSI signals. Then, the learned features are fed into a softmax regression algorithm to recognize different activities. Using the same SAE architecture, the authors of [11] developed a recognition method by transferring the CSI measurements into radio images before using SAE network for feature extraction. The extracted information from radio images are then processed by the softmax regression method for activity recognition. The authors in [12] applied convolutional neural network (CNN) and long-short term memory (LSTM) for behavior recognition, by exploiting the characteristics of spatial information collected from multiple antenna pairs. In order to improve the sensing accuracy, the authors in [13] presented a recognition scheme by combining three DLNs for feature extraction, which is able to achieve a reliable accuracy at the cost of high complexity. Apart from the above DLNs, the long-short term memory recurrent neural networking (LSTM-RNN) has also been adopted in [14] for feature extraction, through which the representative features in CSI signals can be effectively learned and extracted. Our preliminary work [15] also investigated CSI-HAR by leveraging properties of feature extraction techniques and LSTM-RNN. As a result, the inherent features from the input data can be automatically extracted and transformed, thereby improving the accuracy and robustness of human behavior recognition [16].

Despite its effectiveness in improving recognition performance, DLN-based methods suffer from several inherent drawbacks. First, they require a large number of training examples from the testing/targeted environment to train the corresponding DLN [17], [18]. Consequently, the performance is dependant upon the number of training samples, which becomes particularly problematic when large

- *Z. Shi, A. Zhang, R. Y. D. Xu and Q. Cheng are with the School of Electrical and Data Engineering, University of Technology Sydney, Australia.*
E-mail: {zhenguo.shi, qingqing.cheng}@student.uts.edu.au {andrew.zhang, yida.xu}@uts.edu.au

amounts of training samples from testing environments are not accessible. The problem is further exacerbated by the fact that a network trained under one setting may not be applied to another. This may severely restrict the applicabilities of DLN-based methods in practice [19]. To address these problems, recent CSI-HAR schemes have attempted various advanced DLNs and the corresponding learning methodologies to reduce the number of training samples from the testing environment in order to improve the recognition performance in an environment-invariant fashion [20], [21], [22]. To name a few, the recent solutions in [20] and [21] exploited transfer learning to realize environment-robust recognition. Although the model in [20] can facilitate reliable sensing results, it requires many *previously seen environment*s (PSEs) for training. PSE is referred to the environment where a large number of training samples are collected. These samples from PSE are used only for training deep learning networks, but not for testing. The model in [21] does not need multiple PSEs for training, while it still requires several hundreds samples from the testing environment to perform network refinement. The authors in [22] exploited the property of adversarial learning to enable environment-independent recognition. In this work, a recognition model can be built and applied to a new environment without requiring samples from the testing environment. However, its sensing performance heavily depends on the number of PSEs used in training. When both the number of PSEs and the amount of samples from the testing environment are quite limited (e.g., one PSE and at the minimum one sample for each activity from the testing environment), the above methods fail to accomplish successful recognitions. In [23] the authors developed a cross-environment recognition model by extracting environment-independent features. Although this work does not require multiple PSEs or many samples from the testing environment, it is difficult to effectively identify the light activities (e.g., standing and laying).

One-shot learning (OSL) can be considered as a promising candidate to help address the above challenges. OSL has been successfully applied in many vision-based activity recognition and object classification problems [24], [25], [26], and therefore, making it a plausible technique to solve CSI-HAR issues. Its key insight is that, instead of learning the information about the testing/unseen environment with thousands of training samples, one can accomplish the task using just one sample by drawing support from the knowledge of PSEs [27], [28]. In other words, only one sample is enough to learn/extract discriminate features about the environment by bridging the gap between this environment and PSEs, no matter how different they can be. To the best of our knowledge, most of the OSL approaches have been focusing on vision-based scenarios in which video signals are analyzed for recognition. For CSI-HAR, very little has been investigated so far. In particular, a shortcoming of OSL is that although it only needs one sample from the testing environment, it still requires a large amount of samples from a wide variety of PSEs. This may not be accessible under many CSI-HAR settings, as obtaining samples from diverse environments is usually expensive or impractical.

Given the above, in this paper, we employ the state-of-the-art one-shot learning methodology, i.e., matching network (MatNet) [29], to recognize the type of human activities. As aforementioned, it is very difficult to achieve a satisfactory recognition performance if the diversity of PSEs or the samples from testing environment is insufficient. To overcome this challenge, we propose a novel training strategy to better sense and distinguish human behaviors. This training strategy does not require a large amount of samples from the testing environment. Instead, it only needs as few as one sample from the new/testing environment, which can be easily realized in practice. Moreover, our work is able to accomplish a successful recognition with, at the minimum, one PSE, which can hardly be achieved by the conventional learning-based HAR methods.

The main contributions of this paper are summarized below:

- We propose a HAR scheme using Matching Network with enhanced CSI (MatNet-eCSI) to successfully perform one-short learning to recognize human activities in a new environment. Our proposed scheme can largely improve the recognition accuracy in the new environment with much less training complexity, i.e., it requires only one training sample from the new environment.
- We propose a CSI correlation feature enhancement (CCFE) method to enhance the activity-dependent information and eliminate the activity-unrelated information. CCFE consists of two steps: activity-related information extraction (ARIE) and correlation feature extraction (CFE). The proposed CCFE can reduce the dimension of the signals input to the MatNet, significantly decreasing the training complexity.
- We propose a novel training strategy to leverage the properties of MatNet for the successful HAR. The proposed strategy can facilitate a reliable recognition performance even for the situation in which only one PSE is available. For completing the training task, only one sample from the testing environment and the data set from the PSE are required.
- To evaluate the performance of our proposed scheme, we conduct numerous experiments. The extensive results show that the proposed MatNet-eCSI achieves significantly higher recognition performance than state-of-the-art sensing methods, with much less training complexity.

The rest of this paper is structured as follows. The proposed MatNet-eCSI scheme is briefly described in Section 2. Section 3 presents the details for the CCFE method. The details of MatNet based human activity recognition are described in Section 4. The experimental settings and results are provided in Section 5. Conclusions are provided in Section 6.

2 The MatNet-eCSI Scheme

The diagram of the proposed MatNet-eCSI scheme is shown in Fig. 1, consisting of three main modules/stages: CSI Collection, CSI Preprocessing and Activity Recognition. In the first stage, the CSI that represents the variation of wireless channels induced by human activities is collected at the WiFi receiver. In the second stage, the collected CSI is then

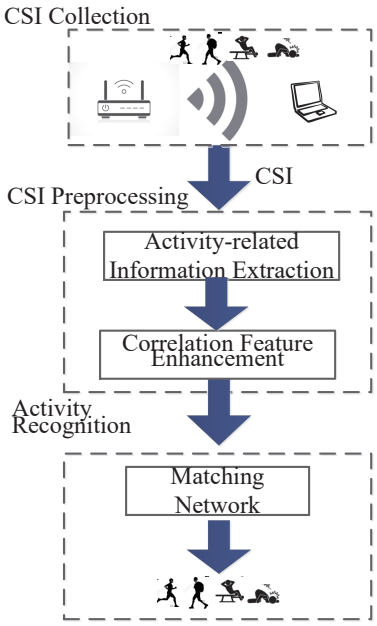


Fig. 1. Main processing modules of the MatNet-eCSI Scheme.

processed, including reducing activity-unrelated information such as scattering signals from the background objects, compressing and reducing the signal input to Stage 3 and enhancing the feature signals. Ideally, only the activity-related signals are transferred to the next stage. In the third and last stage, the MatNet is utilized to automatically learn and extract the hidden features from the enhanced CSI for human behavior classification. Next, we provide a brief overview for each stage, and then detail the last two stages in Section 3 and 4 respectively.

2.1 CSI Collection

Supposing in an indoor environment covered by a WiFi network, a person is moving around, which unavoidably changes wireless signal propagations. Some of the WiFi signals are absorbed, diffracted, reflected or scattered, leading to variations of amplitude, phase shift and the number of multiple paths. Since these variations contain characteristics of different human activities, it is possible to realize HAR by utilizing the CSI measurements from the WiFi signals. To that end, we adopt the Intel 5300 network interface card (NIC), a popular commercial off-the-shelf (COTS) WiFi device, to acquire and collect CSI. According to the protocol of IEEE 802.11n, the CSI tools [30] are used to effectively extract the CSI from 30 subcarriers for each pair of transmitter-receiver antennas. More details of experimental setup are referred to Section 5.1.

The CSI vector $\mathbf{h}(i)$, acquired from the i-th packet, is written as

$$\mathbf{h}(i)=[H_{1,1}(i),\ldots,H_{1,m}(i),\ldots,H_{n,m}(i),\ldots,H_{N,M}(i)]^T, \quad (1)$$

where $H_{n,m}(i)$ stands for the CSI measurement at the mth subcarrier in the nth wireless link; M denotes the total number of available subcarriers in each wireless link; N represents the total number of wireless links, and $N = N_t \times N_r$ where N_t and N_r are the number of antennas at the transmitter and receiver, respectively; and T stands for the transpose operation. The CSI matrix $\mathbf{H}$, made up of CSI vectors obtained from I packets, is

$$\mathbf{H} = [\mathbf{h}(1), \ldots, \mathbf{h}(i) \ldots, \mathbf{h}(I)]. \quad (2)$$

2.2 CSI Preprocessing

The CSI Preprocessing stage intends to reduce CSI for static background objects and condense the CSI matrix. On the one hand, the CSI matrix $\mathbf{H}$ represents the raw CSI measurements and contains multiple channel paths from static background objects and hence a lot of activity-unrelated information. Such information is generally environment-dependent and can largely reduce the robustness of the sensing system. It will also affect the quality of extracted features in the following processing. On the other hand, the size of $\mathbf{H}$ is quite large, and it is computationally intensive and time-consuming to utilize $\mathbf{H}$ directly for training and classification using neural networks. To address these problems, we use the CCFE method that consists of two main steps: *activity-related information extraction* and *correlation feature extraction*.

In the first step, we use a linear recursive operation to construct the CSI for static objects and then subtract it from the received signal. The output is expected to have significantly reduced activity-unrelated information. In the next step, we compute the correlation of the output channel matrix from Step 1, and obtain the correlation feature matrix (CFM). CFM contains condensed activity-related information, with largely reduced dimension compared to original CSI matrix $\mathbf{H}$.

2.3 MatNet based Activity Recognition

This module aims to recognize human activities using the MatNet technology, by automatically learning and extracting the hidden information and features from CFM.

To realize feature extraction, we utilize MatNet that can automatically learn and extract deeper features from CFM. Note that, the proposed training strategy is able to bridge a gap between the testing environment and the PSE. The training process requires the data set from the PSE and at the minimum one sample from the testing environment, facilitating one-shot learning in the testing environment. For realizing human activity recognition, the deep learning network is firstly trained offline using the training data; Then the well-trained network is used online to recognize different human activities.

3 CCFE for CSI Preprocessing

In this section, we present detailed design of CCFE for CSI preprocessing. We will first describe the linear recursive operation based activity-related information extraction method, and then discuss the correlation feature extraction method.

3.1 Activity-related Information Extraction

The core task of this step is to mitigate activity-unrelated information whilst retaining activity-related information. Consequently, we can extract feature signals more correlated with activities and less dependent on the environment. To

that end, we partition $\mathbf{h}(i)$ in (2) into two parts: dynamic CSI and static CSI, given by

$$\mathbf{h}(i) = \mathbf{h}^{st}(i) + \mathbf{h}^{dy}(i), \tag{3}$$

where $\mathbf{h}^{st}(i)$ represents the static CSI vector that is unrelated to human activities. $\mathbf{h}^{dy}(i)$ denotes the dynamic CSI vector that is caused by human activities. In such a case, most of the information contained in dynamic CSI vector in CCFE is activity-related, such as multipath signals which contain the movement information of the object. Note that $\mathbf{h}^{st}(i)$ is generally the dominating component in $\mathbf{h}(i)$ and much larger than $\mathbf{h}^{dy}(i)$. The reason is that the impact induced by human activities on the whole environment is generally limited. This is especially true when a person is performing minor actions, e.g., raising hands, sitting, standing, etc. Under this situation, the accuracy of human activity recognition may drop severely if directly utilizing $\mathbf{h}(i)$ (refer to Fig. 12). Therefore, we want to filter out the static information $\mathbf{h}^{st}(i)$ from $\mathbf{h}(i)$ by exploiting its stability over time. To that end, we propose a recursive algorithm leveraging the exponentially weighted moving average (EWMA) approach [31].

There is one major problem here: the timing offset between the WiFi transmitter and receiver, which are not clockwise synchronized, varies over packets. Such timing offset causes linear phase rotation over subcarriers. It must be estimated and compensated before applying the recursive operation.

Let $\widehat{\mathbf{h}}^{st}(i)$ denote the recursive output at the i-th packet, which is supposed to be the estimate for the static CSI. The recursive operation from continuous packets is respresented as follows:

$$\widehat{\mathbf{h}}^{st}(i) = \theta(\widehat{\mathbf{\Phi}}^{*}(i) \otimes \mathbf{I}_N)\mathbf{h}(i) + (1-\theta)\widehat{\mathbf{h}}^{st}(i-1), \tag{4}$$

where θ stands for the forgetting factor, the superscript $*$ denotes conjugate, $\mathbf{I}_N$ represents an $N \times N$ identity matrix, $\otimes$ represents the Kronecker product, $\widehat{\mathbf{\Phi}}(i) = \text{diag}\{\exp(j\widehat{\phi}_{m,i})\}$ is a diagonal matrix with the m-th element $\exp(j\widehat{\phi}_{m,i})$, and $\widehat{\phi}_{m,i}$ is an estimate of the actual $\phi_{m,i}$ associated with the timing offset. Since signals for all the antennas are typically tied to the same clock, the timing offset, as well as the phase shifts $\phi_{m,i}$ are the same for all antennas at subcarrier m in packet i.

The phase shift $\phi_{m,i}$ can be represented by

$$\phi_{m,i} = m\psi_i + \theta_i, \tag{5}$$

where ψ_i and θ_i are phase shifts related to the timing offset.

In order to estimate ψ and θ_i, we first compute the dot product $\odot$ between $\mathbf{h}(i)$ and $(\widehat{\mathbf{h}}^{st}(i-1))^*$, generating

$$\begin{aligned}\mathbf{r}(i) \triangleq& \mathbf{h}(i) \odot (\widehat{\mathbf{h}}^{st}(i-1))^* \\ =& (\mathbf{h}^{st}(i) + \mathbf{h}^{dy}(i)) \odot (\widehat{\mathbf{h}}^{st}(i-1))^* \\ \approx& \mathbf{h}^{st}(i) \odot (\widehat{\mathbf{h}}^{st}(i-1))^* \\ \approx& (\mathbf{\Phi}(i) \otimes \mathbf{I}_N)|\widehat{\mathbf{h}}^{st}(i-1)|^2 \end{aligned} \tag{6}$$

where $|\widehat{\mathbf{h}}^{st}(i-1)|^2$ denotes element-wise square of the absolute value. In (6), the first approximation is based on the fact that static paths typically have much larger power than dynamic ones, and the second approximation is based on the assumption that the estimate $\widehat{\mathbf{h}}^{st}(i-1)$ is close to the actual static CSI.

We can then stack $\mathbf{r}(i)$ into an $M \times N$ array, with each column containing CSI for one antenna, and compute the mean over each row to get a new $M \times 1$ vector $\bar{\mathbf{r}}(i)$. Computing the cross-correlation for neighbouring elements with equal spaced subcarrier indices in $\bar{\mathbf{r}}(i)$ and then computing the mean of the output, we can obtain a sample denoted by γ_i. Then we can obtain the estimate for ψ_i as

$$\widehat{\psi}_i = \angle(\gamma_i)/K_s, \tag{7}$$

where K_s is the index intervals between the used subcarriers that are equally spaced. For the Intel NIC5300 card used in the experiments in this paper, $K_s = 2$.

Let $\bar{r}_{m,i}$ be the m-th element in $\bar{\mathbf{r}}(i)$. The parameter θ_i in (5) can then be estimated as

$$\widehat{\theta}_i = \angle\left(\sum_m \bar{r}_{m,i} e^{-jm\hat{\psi}_i}\right), \tag{8}$$

where the sum can be over a selected number of samples with larger energy to mitigate the noise.

We then obtain the estimate $\widehat{\mathbf{\Phi}}(i)$ and can obtain the recursive output $\widehat{\mathbf{h}}^{st}(i)$. Note that, the initial value of $\widehat{\mathbf{h}}^{st}(1)$ can be obtained using (4) in a quiet environment.

At packet i, the estimated value of dynamic CSI, $\widehat{\mathbf{h}}^{dy}(i)$, is then given by

$$\widehat{\mathbf{h}}^{dy}(i) = (\widehat{\mathbf{\Phi}}^{*}(i) \otimes \boldsymbol{I}_N)\mathbf{h}(i) - \widehat{\mathbf{h}}^{st}(i). \tag{9}$$

Over I packets , the whole estimated dynamic CSI matrix, $\widehat{\mathbf{H}}^{dy}$, is written as

$$\widehat{\mathbf{H}}^{dy} = [\widehat{\mathbf{h}}^{dy}(1), \ldots, \widehat{\mathbf{h}}^{dy}(i), \ldots, \widehat{\mathbf{h}}^{dy}(I)]. \tag{10}$$

Let $\widehat{\mathbf{A}}^{dy}(i)$ and $\widehat{\mathbf{\Psi}}^{dy}(i)$ stand for the amplitude and phase parts of $\widehat{\mathbf{h}}^{dy}(i)$, respectively. Thus we can decompose the dynamic CSI matrix $\widehat{\mathbf{H}}^{dy}$ into *dynamic amplitude matrix* $\widehat{\mathbf{A}}^{dy}$ and *dynamic phase matrix* $\widehat{\mathbf{\Psi}}^{dy}$ as

$$\begin{aligned}\widehat{\mathbf{A}}^{dy} &= [\widehat{\mathbf{a}}^{dy}(1), \ldots, \widehat{\mathbf{a}}^{dy}(i), \ldots, \widehat{\mathbf{a}}^{dy}(I)], \\ \widehat{\mathbf{\Psi}}^{dy} &= [\widehat{\boldsymbol{\psi}}^{dy}(1), \ldots, \widehat{\boldsymbol{\psi}}^{dy}(i), \ldots, \widehat{\boldsymbol{\psi}}^{dy}(I)]. \end{aligned} \tag{11}$$

where $\widehat{\boldsymbol{A}}^{dy}(i)$ and $\widehat{\boldsymbol{\psi}}^{dy}(i)$ are amplitude and phase vector of $\widehat{\mathbf{h}}^{dy}(i)$. Note that $\widehat{\mathbf{A}}^{dy}$ and $\widehat{\mathbf{\Psi}}^{dy}$ contains mostly activity-related information. Thus they can be used to extract more distinctive features that are less dependent on environment for recognizing human activities.

3.2 Correlation Feature Extraction

It is noteworthy that we can divide a person's activity into different stages. Each stage can be represented by a feature signal, and different stages are dependent and correlated. For example, the activity "sit down" may involve a series of stages, from static, sitting down with accelerating, and sitting down with decelerating to sitting still. The features of different stages, e.g., speed and spatial positions of that person, are different but mutually correlated. While for the activity "sitting", its features among different stages, e.g., speed and spatial positions of human beings, remain similar, but not identical due to, e.g., the breathing activity. Hence "sit down" and "sitting" can be largely distinguished

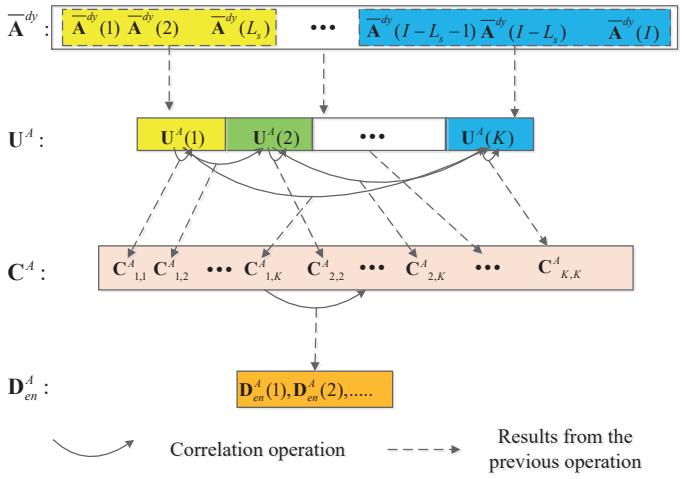


Fig. 2. Correlation feature extraction in the proposed CCFE.

via these differences, while relative static activities such as "sitting", "standing" and "empty" are differentiated via the different impacts of these activities on signal propagation associated with both body positions and minor body dynamics caused by, e.g., breathing.

Note that all the feature signals for each activity are contained in $\widehat{\mathbf{H}}^{dy}$. Such connections and dependency can typically be captured by a Markov chain, or a Markov chain combined with Recurrent Neural Networks, which are typically applied for natural language processing. In this paper, we investigate a correlation based method, which can not only capture such dependency, but also significantly reduce the complexity in at least the training stage.

We partition $\widehat{\mathbf{A}}^{dy}$ and $\widehat{\mathbf{\Psi}}^{dy}$ into several segments and calculate the correlation features between different segments, respectively. Besides, $\widehat{\mathbf{A}}^{dy}$ and $\widehat{\mathbf{\Psi}}^{dy}$ in different subcarriers are also correlated, providing additional information for recognizing human activities. Thus, our proposed CCFE conducts correlation operation over both packets and subcarriers, compressing correlated features between different segments and subcarriers, as shown in Fig. 2.

Next we refer to $\widehat{\mathbf{A}}^{dy}$ to present the process of correlation operation. For $\widehat{\mathbf{\Psi}}^{dy}$, the process is similar.

Assume that I is divisible by K. Then we evenly divide $\widehat{\mathbf{A}}^{dy}$ into K non-overlapped segments, with a length of I/K for each segment. The resulted signal matrix $\mathbf{U}^A$ is represented by

$$\mathbf{U}^A = [\mathbf{U}^A(1), \mathbf{U}^A(2), \ldots, \mathbf{U}^A(k), \ldots, \mathbf{U}^A(K)], \quad (12)$$

where $\mathbf{U}^A(k)$ stands for the $NM \times I/K$ dynamic amplitude matrix of the kth segment. Next, we calculate the covariance matrix between different segments

$$\mathbf{C}^A_{i,j} = \mathrm{E}\big[(\mathbf{U}^A(i) - \mathrm{E}[\mathbf{U}^A(i)])(\mathbf{U}^A(j) - \mathrm{E}[\mathbf{U}^A(j)])^T\big], \quad (13)$$

where $\mathrm{E}[.]$ represents the operation of taking the mean, $\mathbf{C}^A_{i,j}$ is the covariance matrix of $\mathbf{U}^A(i)$ and $\mathbf{U}^A(j)$, $i = 1, 2, \ldots, K$, $j = i, i+1, \ldots, K$. The whole covariance matrix, $\mathbf{C}^A$, is written as

$$\mathbf{C}^A = [\mathbf{C}^A_{1,1}, \mathbf{C}^A_{1,2}, \ldots, \mathbf{C}^A_{1,K}, \mathbf{C}^A_{2,2}, \mathbf{C}^A_{2,3}, \ldots, \mathbf{C}^A_{K,K}]. \quad (14)$$

Algorithm 1: Correlation feature extraction.

begin
  Initialize: the length of input data I,
  the number of non-overlapped segments K,
  the length of each segments L_s;
  $L_s = I/K$;
  For $1 \le k \le K$
    $\mathbf{U}^A(k) = [\widehat{\mathbf{A}}^{dy}((k-1)L_s + 1), \ldots, \widehat{\mathbf{A}}^{dy}(kL_s)]$;
    $\mathbf{U}^\Psi(k) = [\widehat{\mathbf{\Psi}}^{dy}((k-1)L_s + 1), \ldots, \widehat{\mathbf{\Psi}}^{dy}(kL_s)]$;
  end
  Construct $\mathbf{U}^A$ and $\mathbf{U}^\Psi$ based on Eq. (12);
  For $1 \le i \le K$
    For $i \le j \le K$
      Compute $\mathbf{C}^A_{i,j}$ and $\mathbf{C}^\Psi_{i,j}$ based on Eq. (13);
    end
  end
  Construct $\mathbf{C}^A$ and $\mathbf{C}^\Psi$ based on Eq. (14);
  Compute CFM $\mathbf{D}^A_{en}$ and $\mathbf{D}^\Psi_{en}$:
    $\mathbf{D}^A_{en} = \mathbf{C}^A \times (\mathbf{C}^A)^T$, $\mathbf{D}^\Psi_{en} = \mathbf{C}^\Psi \times (\mathbf{C}^\Psi)^T$;
end

Note that $\mathbf{C}^A$ can only reveal the correlation between different segments. The correlation between signals across subcarriers can be obtained by

$$\mathbf{D}^A_{en} = \mathbf{C}^A \times (\mathbf{C}^A)^T, \quad (15)$$

where $\mathbf{D}^A_{en}$ is the correlation feature matrix (CFM) of amplitude, which will be used to train MatNet.

Following the above steps, we can similarly obtain the segmented signal matrix $\mathbf{U}^\Psi$, covariance matrix $\mathbf{C}^\Psi$ and the CFM $\mathbf{D}^\Psi_{en}$ for $\widehat{\mathbf{\Psi}}^{dy}$.

For clarity, the procedure of correlation feature extraction is summarized in Algorithm 1. Note that the size of both $\mathbf{C}^A$ and $\mathbf{C}^\Psi$ are $NM \times \frac{(K+1)KNM}{2}$. For both $\mathbf{D}^A_{en}$ and $\mathbf{D}^\Psi_{en}$, their sizes are $NM \times NM$ and much smaller. Since the training complexity is highly influenced by the size of input data, reducing the size of input signal can result in a notable reduction in training complexity, much higher than the complexity associated with the correlation computation. Therefore the computational complexity can be significantly reduced when using $\mathbf{D}^A_{en}$ and $\mathbf{D}^\Psi_{en}$, instead of $\mathbf{C}^A$ and $\mathbf{C}^\Psi$, as the input for training MatNet (refer to Fig. 12).

4 MatNet based Human Activity Recognition

CSI based HAR is very sensitive to environment. In the previous section, we have introduced CSI preprocessing to reduce the impact of environment on the feature signals and improve its robustness to the environment. However, it cannot fully remove the impact as dynamic CSI can also be environment-related via, for example, signals sequentially scattered by human body and environmental objects, as well as the residual errors in preprocessing. One approach to further improving the robustness is to train DLN with data from massive different environment, which is however very costly. Although some data processing techniques, e.g., data augmentation and regularization [32], [33], can help to alleviate the problem of overfitting caused by insufficient training data, the improvement is limited due to the high correlation between the generated data and the original data.

In this section, we propose to use MatNet, a neural network augmented with external memory, to improve the environmental robustness via one-shot learning. The input to MatNet is the enhanced CSI (i.e., $\mathbf{D}_{en}^{A}$ and $\mathbf{D}_{en}^{\Psi}$). In particular, we propose a tailored training strategy for better utilizing the property of MatNet, which is capable of realizing the sensing task using at the minimum, one set of training data from the new environment.

4.1 Architecture of MatNet

The architecture of MatNet based HAR is illustrated in Fig. 3. For a given reference data set R, the function of MatNet is to build a classier c_R for each R, mapping R to c_R, $R \rightarrow c_R(.)$.

Let (x, y) stand for the CFM-label pairs, $x = \{\mathbf{D}_{en}^{A}, \mathbf{D}_{en}^{\Psi}\}$ is the input CFM with a size of $NM \times NM \times 2$, y is the output label for the corresponding human activity. Then the reference data set R with N_k samples can be written as

$$R = \{(x_i, y_i)\}_{i=1}^{N_k}. \tag{16}$$

For a given target sample $\hat{x}$, the probability distribution of the output $\hat{y}$ can be defined as

$$P(\hat{y}|\hat{x}, R) \triangleq R \rightarrow c_R(\hat{x}), \tag{17}$$

where P stands for the probability distribution, which is parameterised by the CNN and LSTM (shown in Fig. 3). As a result, the estimated output label $\hat{y}$ for a reference data set R and a given input $\hat{x}$ can be obtained by

$$\hat{y} = \arg\max_{y} P(y|\hat{x}, R). \tag{18}$$

One simple way to estimate $\hat{y}$ is calculating the linear combination of y in the reference data set R, so (18) is equal to

$$\hat{y} = \sum_{i=1}^{N_k} a(\hat{x}, x_i) y_i \tag{19}$$

where x_i, y_i are the CFM and the corresponding label from the reference data set $R = \{(x_i, y_i)\}_{i=1}^{N_k}$, and a is an attention mechanism in the form of softmax over the *cosine similarity*, which is defined as

$$a(\hat{x}, x_i) = \frac{e^{\cos(f(\hat{x}), g(x_i))}}{\sum_{j=1}^{N_k} e^{\cos(f(\hat{x}), g(x_j))}}, \tag{20}$$

where $\cos(\alpha, \beta)$ is the cosine similarity function [34], defined as

$$\cos(\alpha, \beta) = \frac{\alpha \cdot \beta}{\| \alpha \| \| \beta \|}. \tag{21}$$

In (20), f and g stand for the embedding functions to embed $\hat{x}$ and x_i, which can be seen as extracting features from the input data. As is illustrated in Fig. 3, both f and g are CNN with LSTM, acting as a lift to input features for achieving the maximum accuracy via the classifier as defined in (19).

In order to extract distinguishable and generalised features from input data for one-shot learning, g and f are designed to embed x_i and $\hat{x}$ fully conditioned on the whole reference data set R. Thus, g and f can be represented as $g(x_i, R)$ and $f(\hat{x}, R)$, respectively.

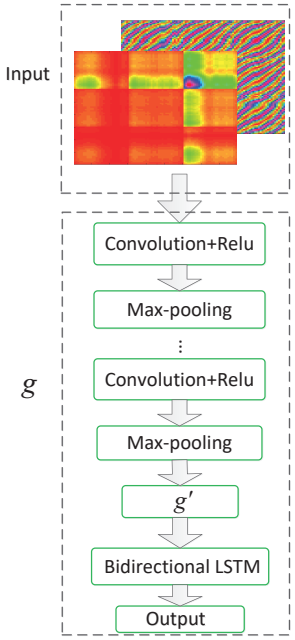


Fig. 4. Structure of embedding function g: CNN with bidirectional LSTM.

The structure of g is shown in Fig. 4, which consists of a CNN with a bidirectional LSTM [35]. The CNN adopted here is a classical structure including several stacked modules, e.g., convolution layer, Relu non-linearity and max-pooling layer. The output of CNN, $g'(x_i)$, which can be seen as discriminative features of x_i, is the input of the bidirectional LSTM. The value of $g(x_i, R)$ can be obtained by

$$g(x_i, R) = \vec{h}_i + \overleftarrow{h}_i + g'(x_i), \tag{22}$$

$$\vec{h}_i, \vec{c}_i = \text{LSTM}(g'(x_i), \vec{h}_{i-1}, \vec{c}_{i-1}), \tag{23}$$

$$\overleftarrow{h}_i, \overleftarrow{c}_i = \text{LSTM}(g'(x_i), \overleftarrow{h}_{i+1}, \overleftarrow{c}_{i+1}), \tag{24}$$

where $\vec{h}_i$ and $\vec{c}_i$ represent the output and cell of the forward LSTM, respectively; $\overleftarrow{h}_i$ and $\overleftarrow{c}_i$ stand for the output and cell of the backward LSTM, respectively; and $\text{LSTM}(g', h, c)$ follows the same definition in [36]. Note that g, a function of the whole reference set R, can play a key role in embedding x_i, which is especially useful when an element x_j is very close to x_i. In other words, if x_i and x_j are input features of two similar activities (e.g., sitting and sitdown), respectively, g can be trained to map x_i and x_j to two distinguishable spaces considering the whole reference data set.

The embedding function f is also composed by CNN and LSTM. The architecture of CNN is the same as the one in g, while the structure of LSTM is different which is the read-attention based LSTM [37]. Let attLSTM(.) denote the read-attention based LSTM, then for a given target sample $\hat{x}$, the output of attLSTM(.) over the whole reference data set R can be written as

$$f(\hat{x}, R) = \text{attLSTM}(f'(\hat{x}), g(R), N_p), \tag{25}$$

where $f'(\hat{x})$, the extracted feature via CNN (similar to g above), is the input of read-attention based LSTM; $g(R)$ denotes the data set obtained by embedding each sample x_i from the reference data set R via g; and N_p represents

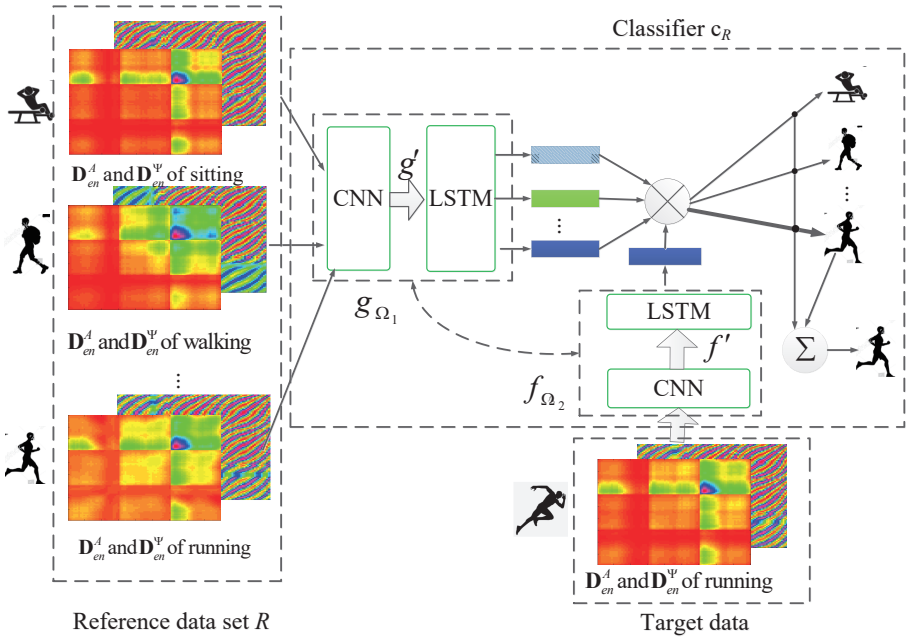


Fig. 3. Structure of MatNet based activity recognition using CFM $\mathbf{D}_{en}^{A}$ and $\mathbf{D}_{en}^{\Psi}$ as the input.

the number of unrolling steps in LSTM. Thus, for the n_pth processing step, the state of the read-attention based LSTM can be expressed as follows:

$$h_{n_p} = \hat{h}_{n_p} + f'(\hat{x}), \tag{26}$$

$$\hat{h}_{n_p}, c_{n_p} = \text{LSTM}(f'(\hat{x}), [h_{n_p-1}, r_{n_p-1}], c_{n_p-1}), \tag{27}$$

where $\text{LSTM}(f'(\hat{x}), [h_{n_p-1}, r_{n_p-1}], c_{n_p-1})$ follows the implementation described in [36]; r_{n_p-1} stands for the read-out from $g(R)$ and is concatenated to h_{k-1}. We can represent r_{n_p-1} as

$$r_{n_p-1} = \sum_{i=1}^{N_s} a(h_{n_p-1}, g(x_i))g(x_i), \tag{28}$$

where N_s is the length of $g(R)$; $a(\cdot,\cdot)$ denotes the attention function in the form of softmax, and is given by

$$a(h_{n_p-1}, g(x_i)) = \text{softmax}(h_{n_p-1}^{T} g(x_i)). \tag{29}$$

Since N_p steps of "reads" are conducted, we have $\text{attLSTM}(f'(\hat{x}), g(S), N_p) = h_{N_p}$, where h_{n_p} is given in (26).

4.2 Training Strategy and Testing procedure

In this subsection, we propose a tailored training procedure to realize HAR in a new (testing) environment using the training data set from one PSE and at the minimum, one sample, from the new testing environment. Our training procedure borrows the idea from episode-based training [29]. However, the training process in [29] requires many PSEs for feature extraction, hence, it cannot be directly applied to our problem. To overcome this issue, we develop a two-step training process to bridge the PSE and the new testing environment, so as to extract desired signal features using the training data from one PSE only.

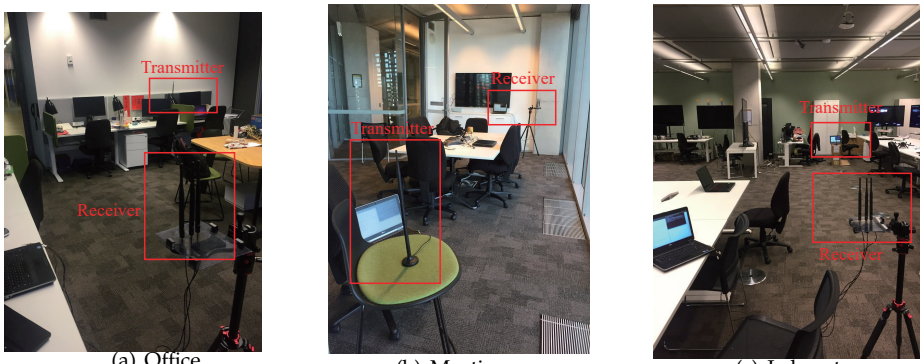
Let $\mathcal{T}$ denote a task which can be seen as a distribution over possible label sets of human activities. In each episode, L, a set of human activities, is sampled from $\mathcal{T}$, $L \sim \mathcal{T}$. L can be a label set $\{sitting, running, walk, running, standup, sitdown, empty\}$. Then L is used to sample the reference data set R and a batch of target set B, obtaining $\mathcal{R} = R \sim L$ and $\mathcal{B} = B \sim L$. The basic idea of training MatNet is to minimize the error from estimating the labels in the batch $\mathcal{B}$ conditional on $\mathcal{R}$. Thus, the loss function of MatNet based human activity recognition, $\mathcal{L}$, is expressed as

$$\mathcal{L} = -\text{E}_{L\sim\mathcal{T}}\left[\text{E}_{\mathcal{R},\mathcal{B}}\left[\sum_{(x,y)\in B} \log P_{\Omega}\left(y|x,\mathcal{R}\right)\right]\right], \tag{30}$$

where $\Omega = \{\Omega_1, \Omega_2\}$, Ω_1 and Ω_2 are the parameter sets of embedding functions g and f, respectively. The training objective is to minimize the loss function over a batch for a given reference data set $\mathcal{R}$, which can be represented as

$$\Omega = \arg\min_{\Omega} \mathcal{L}(\Omega). \tag{31}$$

It is important to note that, for each episode, our proposed training strategy includes two key steps with different data in R and B. Specifically, in the first step, the samples in R are only from the PSE, while the samples in B are from both the testing environment and the PSE. Notably, there is no overlap between R and B. The aim of this step is to build a relationship between the testing environment and the PSE. The essential features for recognizing different activities are also extracted. Then, the trained network coefficients are frozen for the next training step. In the second step, the samples in both R and B are from the testing environment. The network is trained based on R and B using the parameters obtained from the first step. This training step can be seen as a fine tuning process which can help the MatNet to better learn and extract the distinguishable features of human behaviors in the testing environment.



(a) Office (b) Meeting room (c) Laboratory

Fig. 5. Layout of three indoor experimental areas: (a) $3m \times 4m$ office. (b) $4m \times 6m$ meeting room. (c) $6m \times 7m$ laboratory.

5 Implementation and Evaluation

In this section, we perform extensive experiments to validate the performance of the proposed MatNet-eCSI scheme.

5.1 Experimental Setup

To implement the proposed MatNet-eCSI, we use two computers with Intel WiFi NIC5300 network card, serving as the transmitter and receiver. The WiFi cards operate in the 802.11n mode. The transmitter, using one antenna ($N_t = 1$), operates on the 5.32 GHz frequency band and continuously sends packets. The receiver, equipped with three antennas ($N_r = 3$), keeps collecting and storing CSI using the CSI tools in [30]. The number of subcarriers for each pair of the transmitter-receiver antennas is 30 ($S = 30$). We use a sliding window with time length 2s to get samples for each activity from raw CSI streams. During training, if the time window spans over multiple activities, it is labeled as the activity with the maximum proportion. This enables the training and the applications of the trained model to actual recognition. It may be better if a windowing method with window length adapting to activities can be developed and applied. However, this is a non-trivial task and we note it as an open research problem here. The rate of samples is 1 KHz, so the size of CSI matrix ($\mathbf{H}$) is 90×2000. The number of segments K in the proposed CCFE method is 5. For each embedding function of MatNet-eCSI, it contains a CNN with 8 convolutional layers. Each layer contains a 3×3 convolution, a ReLU non-linearity operation, and a 2×2 max-pooling. The proposed MatNet-eCSI is trained using a 3.4 GHz PC with Nvidia P5000 graphic card (16GB memory). The number of training iterations is 1000. The batch size and learning rate are set as 64 and 0.001, respectively.

We deploy our proposed MatNet-eCSI in three indoor configurations with different environmental complexities. The layout of each indoor configurations is illustrated in Fig. 5. Specifically, the first configuration is a $3m \times 4m$ square area. The second one is a $4m \times 6m$ meeting room, and third one is a $6m \times 7m$ laboratory room. Note that the wireless environments of different configurations are determined by not only the size of room but also several other factors, such as the locations of the transmitter and receiver, and the objects placed between transmitter and receiver. The latter can significantly influence the sensing performance. Moreover, the similarity between wireless environments in different configurations also has noticeable impact on recognition performance of the proposed scheme, because more common features can be learned and extracted if wireless environments are similar. We then compare the difference of different environments, via calculating the similarity of wireless environments involved in different configurations. To do this, we compute the cosine similarity function [34] for the received CSI. The similarity of wireless environment between the first and second configurations, between the second and third, and between the first and third configurations are 0.679, 0.616 and 0.571, respectively. In such a case, compared to the third configuration, the wireless environment in the second configuration is more similar to that in the first configuration.

In each indoor configuration, activities performed by five persons are collected as the dataset, and each person performs seven activities: empty room, sitting down, sitting, standing up, standing, walking, and running. Each activity is performed 200 times in total. The dataset is partitioned into the training dataset and testing dataset. The training data set is used to train the network for recognizing human behaviors in the testing environment. We consider two different training data sets, i.e., "one-shot" and "five-shot", using 1 and 5 samples respectively for each activity from the testing environment, together with the whole data set from the PSE. In the experiments, we achieve robust scaling for the proposed scheme in the following way. Firstly, after the stage of data processing (i.e., the proposed CCFE method), we normalize the input data before putting it into the MatNet architecture. Then, in the training stage, we adopt the Batch Normalization method [38] to normalize the inputs of each layer.

We also briefly summarize the experimental setups of methods for comparison (i.e., RNN [14], EI [22], MatNet [29], and TNNAR [20]). Specifically, the method in [14] is developed for HAR based on RNN architecture, which has four hidden layers. 200 hidden units are contained in each hidden layer. In EI method [22], the three-layer stacked CNNs are adopted to extract the activity features. In each layer of CNNs, 2D kernels are used as the filters. Then, a batch norm layer is applied to normalize the mean and variance of the data at each layer. The method in [29] is based on the traditional MatNet that contains a CNN with 8 convolutional layers. Each layer contains a convolution, a ReLU non-linearity operation, and a max-pooling. The TNNAR method [20] is developed based on transfer learn-

ing. This work uses two convolutional layers with max-pooling layers, one LSTM layer, and two fully-connected layers. The batch size and learning rate for four methods are all set as 64 and 0.001, respectively.

5.2 Performance Evaluation

In this section, we first evaluate the performance of our proposed MatNet-eCSI scheme and compare it with four other state-of-the-art methods (i.e., RNN [14], EI [22], MatNet [29], and TNNAR [20]) considering various parameters and configurations. We then analyze the impact of the proposed CCFE method and other parameters (e.g., size of reference data set) on the performance of MatNet-eCSI.

It is important to note that the proposed MatNet-eCSI has two key differences in comparison with MatNet in [29]. Firstly, our proposed MatNet-eCSI uses CCFE, which enhances the activity-dependent information and decreases the training time. Secondly, MatNet-eCSI uses a tailored novel training strategy that enables better exploration of the properties of MatNet. Through this training strategy, the recognition task can be accomplished using at the minimum, one set of training data from the testing environment.

5.2.1 Performance Comparison for Different Methods

Table 1 $\sim$ Table 3 demonstrate the average recognition accuracy of the five methods for seven activities considering different configurations and parameters. The testing environments in Table 1 $\sim$ Table 3 are the first, second and third configurations, respectively. $\mathbf{PSE}1$, $\mathbf{PSE}2$ and $\mathbf{PSE}3$ denote the first, second and third configurations as PSEs, respectively. "One-shot" and "Five-shot" indicate using 1 and 5 samples respectively for each activity from the testing environment, together with the whole data set from the PSE.

From these tables, we can observe that the proposed MatNet-eCSI significantly outperforms the other four methods in all indoor configurations for both "one-shot" and "five-shot". The reason is that, our proposed CCFE method improves and condenses the activity-dependent information in input signals. Consequently, the activity-related features can be effectively learned and extracted, which is beneficial for distinguishing activities. Moreover, we proposed a tailored training strategy to better utilize the property of MatNet for reliable sensing performance. As a result, the bridge between the PSE and the testing environment can be effectively built using even one sample for each activity from the testing environment. Therefore, our proposed scheme is capable of achieving much higher sensing results with even one sample from the testing environment together with the dataset from one PSE, which is also the main advantage of the proposed MatNet-eCSI. By contrast, TNNAR and MatNet require many samples from the testing environment and numerous PSEs to facilitate the activity recognition. Although EI does not need samples from the testing environment, it requires data from a large number of PSEs. When the number of PSEs is insufficient, all the above methods (i.e., TNNAR, MatNet and EI) fail to obtain reliable recognition performance, as illustrated in Table 1 $\sim$ Table 3. For RNN, it needs huge amounts of data from the testing environment for activity recognition. Since only one or five samples from the testing environment are selected in the considered scenario, it is difficult for RNN to achieve a satisfactory result.

TABLE 1
Average recognition accuracy of the five methods in the first indoor configurations

Method	$\mathbf{PSE}2$		$\mathbf{PSE}3$	
	One-shot	Five-shot	One-shot	Five-shot
Proposed MatNet-eCSI	0.868	0.934	0.822	0.923
MatNet	0.402	0.447	0.398	0.444
RNN	0.206	0.253	0.216	0.268
EI	0.354	0.411	0.351	0.407
TNNAR	0.328	0.393	0.323	0.390

TABLE 2
Average recognition accuracy of the five methods in the second indoor configurations

Method	$\mathbf{PSE}1$		$\mathbf{PSE}3$	
	One-shot	Five-shot	One-shot	Five-shot
Proposed MatNet-eCSI	0.802	0.881	0.761	0.861
MatNet	0.376	0.429	0.401	0.439
RNN	0.153	0.186	0.219	0.236
EI	0.315	0.405	0.345	0.402
TNNAR	0.302	0.373	0.301	0.369

For detailed exam of the performance, we provide the confusion matrix for each method for the case of one-shot learning, as illustrated in Fig. 6. In this figure, the activities are performed under the first experimental configuration. $\mathbf{PSE}2$ is selected as PSE. As can be seen, the performance of the proposed work is greatly better than those of the existing methods. Specifically, for the proposed MatNet-eCSI, each predicted activity matches the corresponding actual activity, meaning that our proposed scheme is able to obtain a reliable recognition result for each activity. By contrast, for the other four sensing methods, the predicted activities are not accordance with the corresponding actual activities. Therefore, from Table 1 $\sim$ Table 3 and Fig. 6, we can conclude that the proposed MatNet-eCSI is able to successfully perform one-short learning to recognize human activities in new/testing environments, using one PSE only. The sensing accuracy of the proposed MatNet-eCSI is notably higher than that of the existing methods.

Although the proposed MatNet-eCSI is able to obtain a reliable sensing result, it is shown to be less robust to some activities which induce similar impacts on CSI. Take the activity "walk" in Fig. 6(a) as an instance, the probability of correctly detecting "walk" is 0.66, while the probabilities of sensing it as "running" is 0.12. This is because when the speed of running is low, its impact on CSI is similar to that of "walk". The robustness can be improved by using more samples from the testing environments (e.g., "five-shot"). To illustrate this, in Fig. 7 we show the confusion matrix of the proposed MatNet-eCSI with "five-shot". As can be seen from the figure, the recognition accuracy of each activity for "five-shot" is higher than that of "one-shot", implying that increasing the number of samples from the testing environment can result in better recognition performance. This is achieved at the cost of increased complexity and samples, as illustrated in Fig. 12(b).

Fig. 8 demonstrates the impact of the used number of receiving antennas, represented as the number of total

TABLE 3
Average recognition accuracy of the five methods in the third indoor configurations

Method	**PSE**2		**PSE**1	
	One-shot	Five-shot	One-shot	Five-shot
Proposed MatNet-eCSI	0.577	0.758	0.461	0.749
MatNet	0.417	0.462	0.374	0.462
RNN	0.163	0.205	0.186	0.214
EI	0.365	0.421	0.333	0.412
TNNAR	0.317	0.388	0.319	0.385

TABLE 4
The number of PSEs required by different methods for the similar sensing accuracy

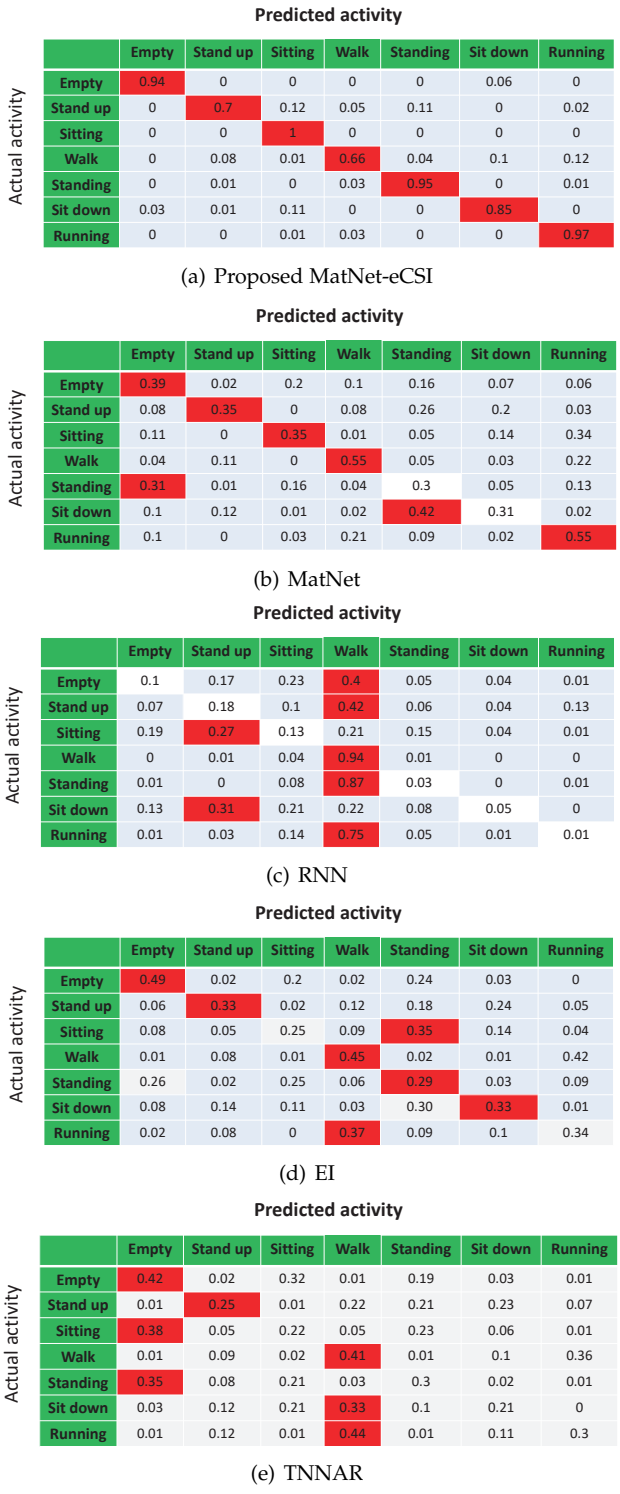
Method	Number of PSE
Proposed MatNet-eCSI	1
MatNet	18
EI	23
TNNAR	25

subcarriers, on the average recognition accuracy in the second experimental configuration. The PSE is **PSE**3. From this figure, it is clear that for both "One-shot" and "Five-shot", increasing the number of subcarriers can result in better average recognition accuracy for each method. The improvement is more obvious in our proposed method, particularly in "one-shot".

In Fig. 9, we illustrate the sensing performance of different methods with the increased number of PSEs. As can be observed from the figure, EI, TNNAR, MatNet, and our proposed MatNet-eCSI all achieve better recognition performance when the number of PSEs increases. This is because, with more PSEs, these four methods are able to better extract common features shared by PSEs and testing environment, which is beneficial for recognizing human activities. On the contrary, sensing accuracy for RNN is not necessarily improved when the number of PSEs increases. The reason is that RNN cannot extract transferable features shared by PSEs and the testing environment. In addition, the proposed MatNet-eCSI is able to achieve a satisfactory sensing accuracy with even one PSE, which is difficult for the other methods to achieve.

In Table 4, the required numbers of PSEs for achieving a recognition accuracy above 80% are shown for four methods. The required number of PSEs (e.g., over 20 PSEs) is obtained by collecting training samples from different rooms with different sizes or layouts. Note that different layouts in the same room are treated as different environments. Five people performed activities in each environment. Since PSEs have no notable impact on the performance of RNN, which requires a large number of samples from the testing environment, we did not present its result in this table. We can observe from this table that to achieve similar accuracy, our proposed MatNet-eCSI only requires the training samples from one PSE. By contrast, MatNet, EI and TNNAR need training samples from 18, 23 and 25 different PSEs, respectively. Since obtaining samples from numerous different PSEs is always impractical or expensive, the proposed MatNet-eCSI is superior compared to the other three methods.

We investigate how well the proposed CCFE affects



(a) Proposed MatNet-eCSI

(b) MatNet

(c) RNN

(d) EI

(e) TNNAR

Fig. 6. Confusion matrix for different human activity recognition methods.

the sensing accuracy for different methods, as shown in Table. 5. In this table, the activities are described in the first configuration, and PSE is **PSE**2. It is obvious that the recognition accuracy for each method with CCFE is better than the case without CCFE. This is because CCFE is able to enhance the activity-related information, thereby contributing to distinguishing different activities. Note that the proposed MatNet-eCSI with CCFE obtains higher accuracy than MatNet with CCFE. This is because the novel training strategy used in our proposed MatNet-eCSI is able to better

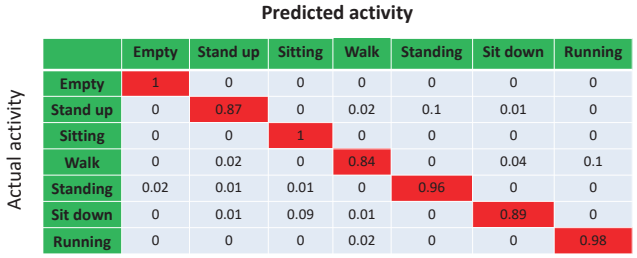
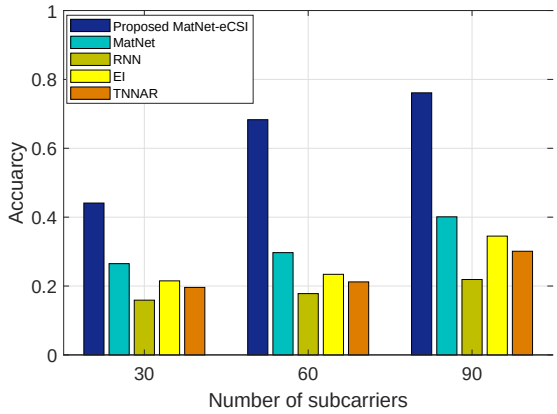
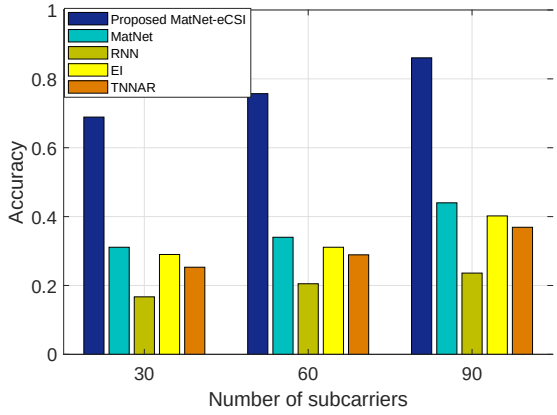


Fig. 7. Confusion matrix of proposed MatNet-eCSI for five-shot



(a) One-shot



(b) Five-shot

Fig. 8. Impact of the used number of receiving antennas, represented as the number of total subcarriers, on the recognition accuracy.

utilize the properties of MatNet for feature extraction.

The sensing results of different methods with sufficient training samples are presented in Fig.10. In this figure, the training dataset is collected by using 200 samples for each activity from the testing environment, together with the whole data set from eight PSEs. We can see that all methods achieve high sensing accuracies given sufficient samples from the testing environment and various PSEs.

TABLE 5
Impact of CCFE on recognition accuracy for different methods

Method	Without CCFE	With CCFE
Proposed MatNet-eCSI	0.616	0.868
MatNet	0.402	0.632
RNN	0.206	0.329
EI	0.354	0.521
TNNAR	0.328	0.502

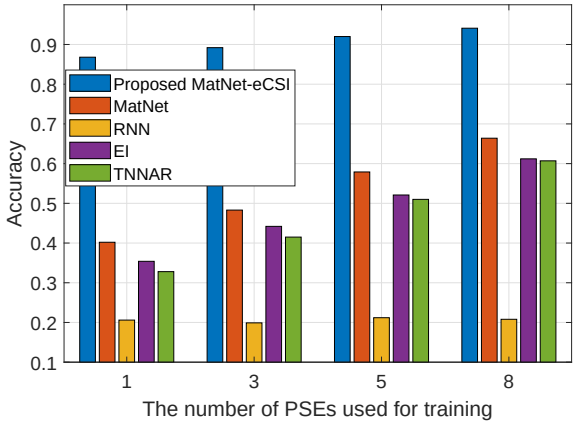


Fig. 9. Recognition accuracy with increased number of PSEs

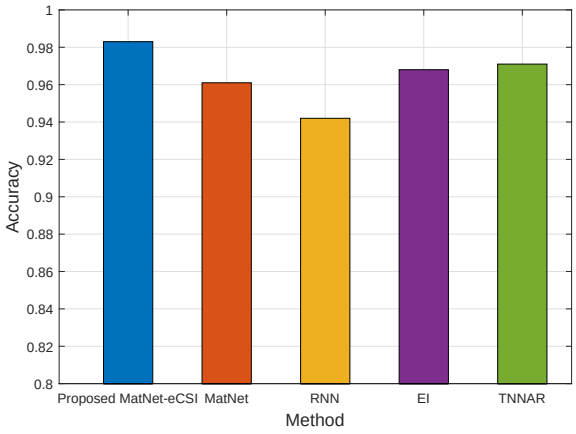


Fig. 10. Recognition accuracy of different methods with sufficient training samples

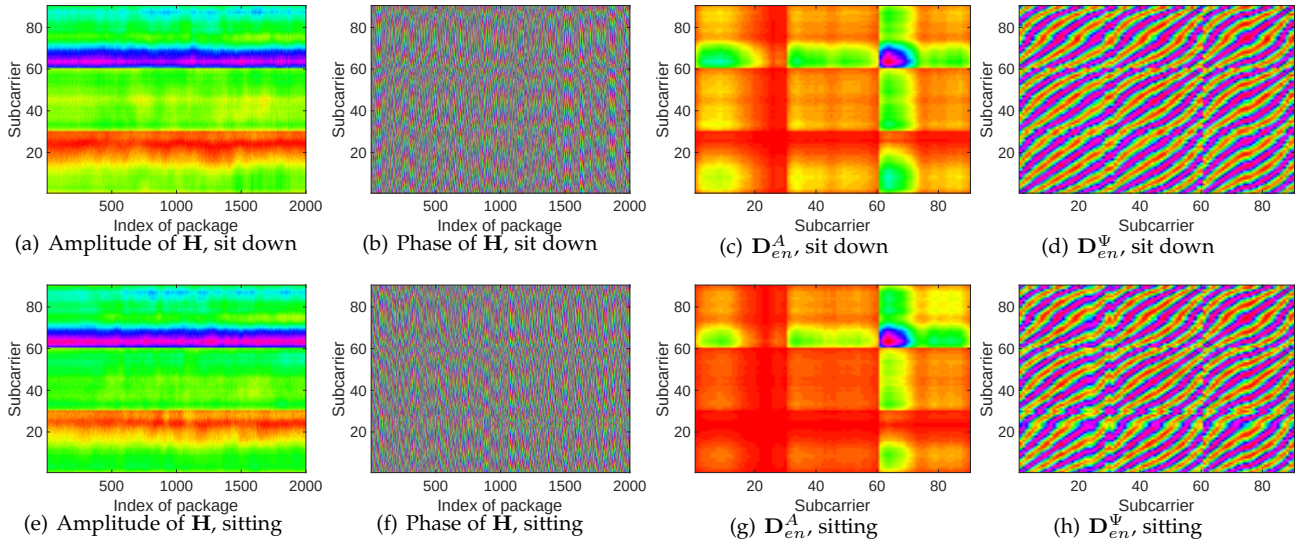
This is because these methods can effectively train their respective models with sufficient samples, thereby achieving reliable sensing performance. Our proposed MatNet-eCSI still outperforms other methods in this case, crediting to the proposed CCFE method.

5.2.2 *Effect of CCFE on MatNet-eCSI*

The impact of CCFE on the performance of the proposed MatNet-eCSI is investigated in this subsection, to demonstrate the importance of CCFE for the whole proposed scheme.

We first use two similar activities, e.g., "sit down" and "sitting", as an instance to illustrate the effect of CCFE on enhancing the feature signals. From Fig. 11(a) and 11(e) (or from Fig. 11(b) and 11(f)), we can see that it is difficult to distinguish between "sit down" and "sitting" by only utilizing the amplitude (or phase) of $\mathbf{H}$. By contrast, it is much easier to differentiate these two activities based on $\mathbf{D}_{en}^{A}$ (or $\mathbf{D}_{en}^{\Psi}$) that enlarges the difference between similar activities. This is because CCFE reduces activity-unrelated information, hence enlarging the difference. Additionally, $\mathbf{D}_{en}^{A}$ (or $\mathbf{D}_{en}^{\Psi}$) reduces the dimensions of output signals, compared to the amplitude (or phase) of $\mathbf{H}$.

Fig. 12 presents how well CCFE can improve the average recognition accuracy and reduce the training time, compared to the case without using it. The activities are performed in the first experimental configuration, and $\mathbf{PSE}3$



(a) Amplitude of $\mathbf{H}$, sit down (b) Phase of $\mathbf{H}$, sit down (c) $\mathbf{D}^{A}_{en}$, sit down (d) $\mathbf{D}^{\Psi}_{en}$, sit down

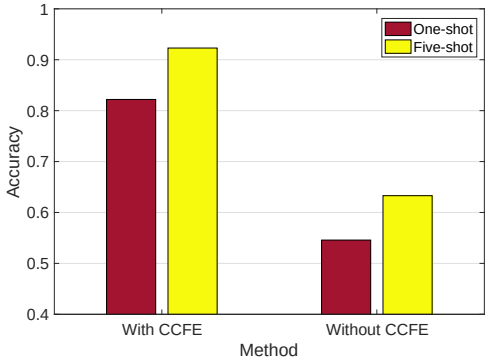
(e) Amplitude of $\mathbf{H}$, sitting (f) Phase of $\mathbf{H}$, sitting (g) $\mathbf{D}^{A}_{en}$, sitting (h) $\mathbf{D}^{\Psi}_{en}$, sitting

Fig. 11. Effect of CCFE on enhancing the feature signals for two similar activities"sit down" and "sitting".

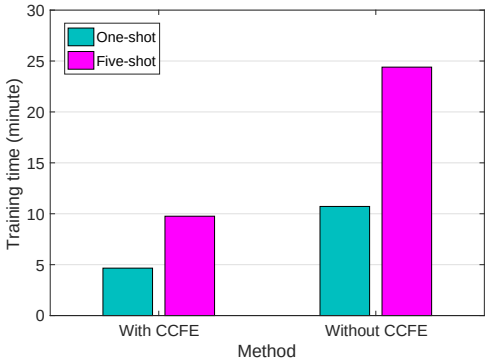
is selected as PSE. As illustrated in Fig.12(a), the average recognition accuracy of MatNet-eCSI with CCFE is shown to be much better than that of without CCFE for both "One-shot" and "Five-shot". This is because the proposed CCFE is capable of enhancing the activity-related features by removing activity-unrelated information. Moreover, the similarities of the enhanced CSI across different environments (i.e., outcomes of CCFE) become higher, in comparison with initial CSI signals, which is beneficial for improving sensing performance. Take the activity "sit down" as a study case. The similarity of the initial CSI signal for this activity across the first and third configurations is 0.559. The initial CSI signal is input to the proposed CCFE for processing, and the final outputs include static components (i.e., static CSI) and dynamic components (i.e., enhanced CSI). The similarities of the static CSI and enhanced CSI across the first and third environments are 0.461 and 0.632, respectively. It is clear that, compared to the initial CSI signal, the similarity of the static CSI across different environments becomes smaller, while the similarity of the enhanced CSI becomes larger. Since the static CSI is mostly removed before the training stage, they have little impact on the sensing performance. On the other hand, the enhanced CSI is fed into the deep learning network for training, which contributes to improved recognition results.

Fig. 13 shows the impact of phase compensation (an important part of CCFE) on the sensing performance. The activities are performed in the first experimental configuration, and $\mathbf{PSE}2$ is selected as PSE. As can be observed from the figure, the sensing accuracy for the case with phase compensation is much higher than that without phase compensation. The reason is that the proposed phase compensation is capable of compensating the phase shift that is caused by timing offset. As a result, the quality of CSI can be improved, which is beneficial for recognizing different activities.

We also investigate the impact of the number of segment K (an important factor in CCFE) on the average recognition accuracy, as presented in Fig. 14. We present the results for the second experimental configuration, and PSE is $\mathbf{PSE}1$. As can be seen from this figure, a larger K leads



(a) Average recognition accuracy



(b) Training time

Fig. 12. Impact of CCFE on the recognition accuracy and required training time.

to higher average recognition accuracy and better sensing performance. The reason is that the proposed CCFE with a larger K can extract more correlation information/features for human activity. Note that the recognition accuracy cannot be infinitely improved with the increase of K. This is because, when K is very large, the difference between adjacent segments is insufficient for providing additional useful signal features for HAR. Additionally, a larger K causes higher computational complexity. Therefore, the user can select the value of K to balance recognition performance and computational complexity.

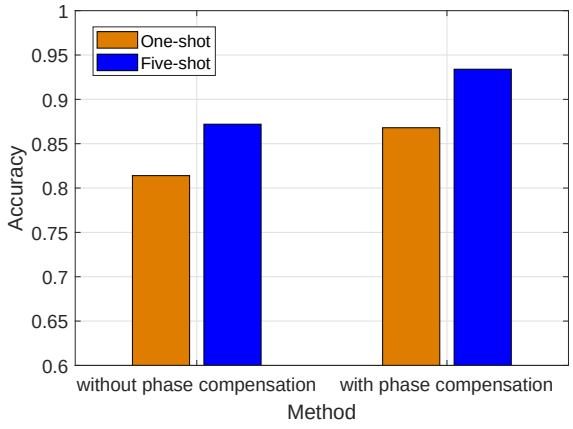


Fig. 13. Impact of phase compensation on recognition accuracy

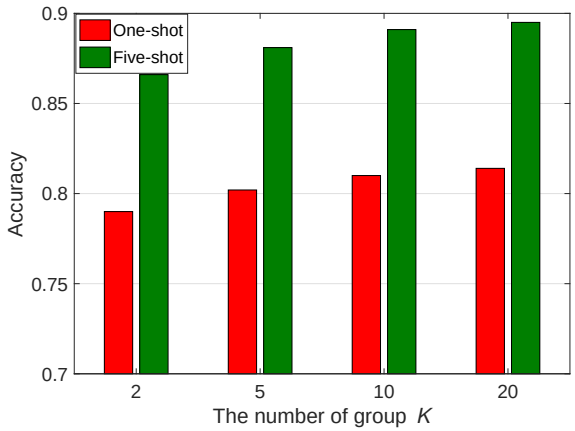


Fig. 14. Impact of the number of segment K on the recognition accuracy.

5.2.3 Impact of Input Signals on MatNet-eCSI

In this subsection, we investigate how the type of input signals and the size of data set from the PSE affect the recognition performance of MatNet-eCSI.

Table 6 illustrates the average recognition accuracy with different input signals in the first indoor configuration. In this table, the PSE is $\mathbf{PSE}2$. In the previous results, MatNet-eCSI uses both the amplitude and phase of $\mathbf{H}$ as inputs. Here, we test MatNet-eCSI-AM and MatNet-eCSI-PH, which indicate that MatNet-eCSI only adopts the amplitude or phase of $\mathbf{H}$ as the input. Table 6 shows that MatNet-eCSI is superior to the other two methods for both "One-shot" and "Five-shot". This is because more essential features for human recognition can be extracted from the combination of amplitude and phase of $\mathbf{H}$. It is also interesting to see that MatNet-eCSI-PH achieves better accuracy than MatNet-eCSI-AM, which suggests that the amplitude of $\mathbf{H}$ is more susceptible to the propagation environment change.

The impact of the size of data set on recognition performance is presented in Fig. 15. The horizontal axis means the number of times collected for each activity in a single environment. In this figure, the activities are performed as per the second configuration, and PSE is $\mathbf{PSE}2$. It is clear that, a larger training data set can result in a higher accuracy for the proposed scheme in both "one-shot" and "five-shot" cases. The improvement in recognition accuracy becomes quite small, after a sufficient number of training samples. Note that more training samples require more time and resource for processing, leading to higher computational complexity. Therefore, it is important to select a proper size of data set, to achieve a good balance between the recognition accuracy and complexity.

TABLE 6
Sensing performance using different input signals in the first configuration with $\mathbf{PSE}2$.

Method	One-shot	Five-shot
MatNet-eCSI	0.868	0.934
MatNet-eCSI-AM	0.79	0.862
MatNet-eCSI-PH	0.823	0.912

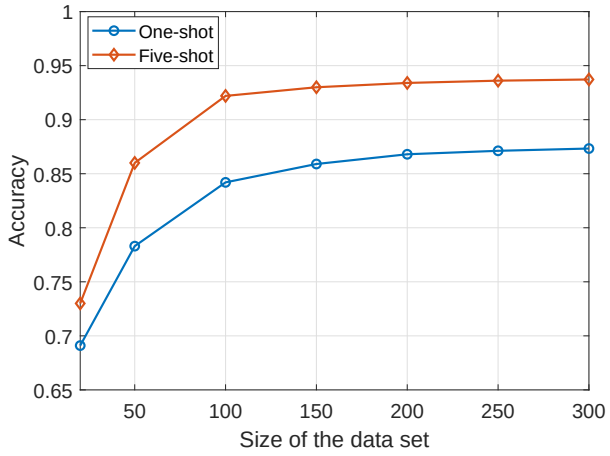


Fig. 15. Impact of the size of data set on the recognition accuracy

5.2.4 Impact of training strategy and human diversity on MatNet-eCSI

In this subsection, we investigate how the sensing performance of MatNet-eCSI varies with the proposed training strategy and different human beings. In this subsection, different activities are performed in the first indoor configuration, and the PSE is $\mathbf{PSE}2$.

Fig. 16 shows the variation in sensing accuracy of MatNet-eCSI with different human subjects. In the figure, two volunteers participate in the training process and the other three are new for the testing. In this figure, we calculate the average sensing accuracy for each person when he/she acts as the testing subject. We can see that the average accuracy varies across different persons, meaning that different persons could have different impact on recognition performance. However, it is important to note that the average accuracy does not show an obvious difference across persons, and the overall accuracy for five persons is still reliable. For instance, for "five-shot", the average accuracy of all volunteers are higher than 88%. Therefore, the proposed scheme demonstrates robustness to human diversity.

Fig. 17 demonstrates the impact of the proposed training strategy on the sensing performance of MatNet-eCSI. From this figure, we can observe that using the novel training strategy enables the proposed scheme to achieve a higher recognition accuracy, compared to the case without using it. This is because via the proposed training strategy, the common features shared by PSE and the testing environment

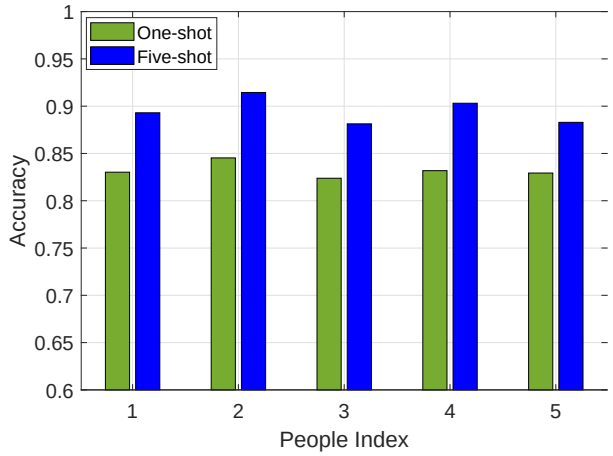


Fig. 16. Average recognition accuracy for different people

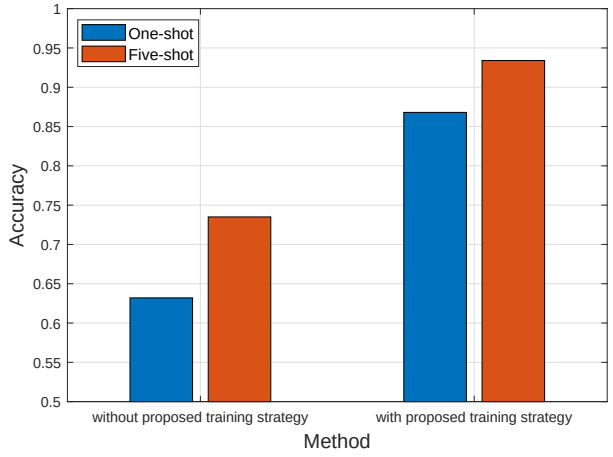


Fig. 17. Impact of training strategy on the recognition accuracy

can be effectively extracted using the training data set from one PSE and at the minimum, one sample, from the new testing environment.

6 Conclusion and Future Work

In this work, we proposed a novel MatNet-eCSI scheme to realize one-shot learning human activity recognition. Our approach includes an innovative CCFE methodology and a novel training strategy. The CCFE method can improve activity-related signals by removing activity-unrelated information. The dimension of input signals is also largely decreased, which reduces the computational complexity and the training time. We developed a better training strategy for recognizing human behaviors using only one sample from the testing environment along with the data set from the PSE. The extensive experimental results confirm that our proposed MatNet-eCSI significantly outperforms the existing related work in notably improving the recognition accuracy and reducing the training time.

It is noteworthy to state that we have only investigated the activity recognition with a single person under the current methodology. We will take the multiple-person activity recognition as a natural extension to this work in the future, which is a considerably challenging task.

References

[1] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with wi-fi!" *IEEE Transactions on Mobile Computing*, vol. 15, no. 11, pp. 2907–2920, Nov 2016.

[2] 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.

[3] Q. Pu, S. Gupta, S. Gollakota, and S. Patel, "Whole-home gesture recognition using wireless signals," in *Proceedings of the 19th Annual International Conference on Mobile Computing & Networking*, ser. MobiCom '13. New York, NY, USA: ACM, 2013, pp. 27–38. [Online]. Available: http://doi.acm.org/10.1145/2500423.2500436

[4] 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 *Proceedings of the 20th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '14. New York, NY, USA: ACM, 2014, pp. 617–628. [Online]. Available: http://doi.acm.org/10.1145/2639108.2639143

[5] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Understanding and modeling of wifi signal based human activity recognition," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '15. New York, NY, USA: ACM, 2015, pp. 65–76. [Online]. Available: http://doi.acm.org/10.1145/2789168.2790093

[6] C. Wu, Z. Yang, Z. Zhou, X. Liu, Y. Liu, and J. Cao, "Non-invasive detection of moving and stationary human with wifi," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 11, pp. 2329–2342, Nov 2015.

[7] 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.

[8] T. Xin, B. Guo, Z. Wang, M. Li, Z. Yu, and X. Zhou, "Freesense: Indoor human identification with wi-fi signals," in *2016 IEEE Global Communications Conference (GLOBECOM)*, Dec 2016, pp. 1–7.

[9] Z. Chen, Q. Zhu, Y. C. Soh, and L. Zhang, "Robust human activity recognition using smartphone sensors via ct-pca and online svm," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 3070–3080, Dec 2017.

[10] 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.

[11] J. Wang, X. Zhang, Q. Gao, H. Yue, and H. Wang, "Device-free wireless localization and activity recognition: A deep learning approach," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 7, pp. 6258–6267, July 2017.

[12] 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*, pp. 1–1, 2018.

[13] H. Zou, Y. Zhou, J. Yang, H. Jiang, L. Xie, and C. J. Spanos, "Deepsense: Device-free human activity recognition via autoencoder long-term recurrent convolutional network," in *2018 IEEE International Conference on Communications (ICC)*, May 2018, pp. 1–6.

[14] 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.

[15] Z. Shi, J. A. Zhang, R. Xu, and Q. Cheng, "Deep learning networks for human activity recognition with csi correlation feature extraction," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, May 2019, pp. 1–6.

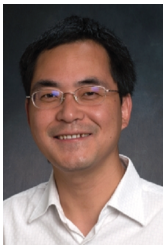
[16] W. Wang, A. X. Liu, M. Shahzad, K. Ling, and S. Lu, "Device-free human activity recognition using commercial wifi devices," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 5, pp. 1118–1131, May 2017.

[17] J. Wan, G. Guo, and S. Z. Li, "Explore efficient local features from rgb-d data for one-shot learning gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1626–1639, Aug 2016.

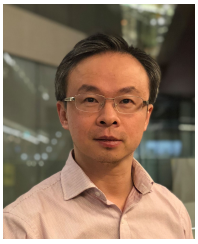
[18] S. Rahman, S. Khan, and F. Porikli, "A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5652–5667, Nov 2018.

[19] M. Rodriguez, C. Orrite, C. Medrano, and D. Makris, "One-shot learning of human activity with an map adapted gmm and simplex-hmm," *IEEE Transactions on Cybernetics*, vol. 47, no. 7, pp. 1769–1780, July 2017.
[20] 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. ICCSE18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi-org.ezproxy.lib.uts.edu.au/10.1145/3265689.3265705
[21] 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. [Online]. Available: http://doi.acm.org/10.1145/3241539.3241570
[22] 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. [Online]. Available: http://doi.acm.org/10.1145/3241539.3241548
[23] Y. Zheng, Y. Zhang, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Zero-effort cross-domain gesture recognition with wi-fi," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '19. New York, NY, USA: ACM, 2019, pp. 313–325. [Online]. Available: http://doi.acm.org/10.1145/3307334.3326081
[24] H. J. Seo and P. Milanfar, "Action recognition from one example," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 867–882, May 2011.
[25] L. Zhang, S. Zhang, F. Jiang, Y. Qi, J. Zhang, Y. Guo, and H. Zhou, "Bomw: Bag of manifold words for one-shot learning gesture recognition from kinect," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2562–2573, Oct 2018.
[26] Y. Yang, I. Saleemi, and M. Shah, "Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1635–1648, July 2013.
[27] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, April 2006.
[28] Z. Guo and Z. J. Wang, "An unsupervised hierarchical feature learning framework for one-shot image recognition," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 621–632, April 2013.
[29] 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. [Online]. Available: http://papers.nips.cc/paper/6385-matching-networks-for-one-shot-learning.pdf
[30] 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. [Online]. Available: http://doi.acm.org/10.1145/1925861.1925870
[31] S. W. Roberts, "Control chart tests based on geometric moving averages," *Technometrics*, vol. 1, no. 3, pp. 239–250, 1959.
[32] F. Meng, H. Liu, Y. Liang, J. Tu, and M. Liu, "Sample fusion network: An end-to-end data augmentation network for skeleton-based human action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5281–5295, Nov 2019.
[33] N. Takahashi, M. Gygli, and L. Van Gool, "Aenet: Learning deep audio features for video analysis," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 513–524, March 2018.
[34] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Computer Vision – ACCV 2010*, R. Kimmel, R. Klette, and A. Sugimoto, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 709–720.
[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[36] 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.
[37] O. Vinyals, S. Bengio, and M. Kudlur, "Order matters: Sequence to sequence for sets," *arXiv preprint arXiv:1511.06391*, 2015.
[38] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

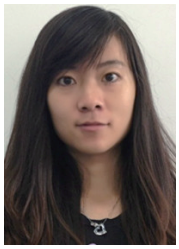
Zhenguo Shi received his M.S.(in 2011) degree and PhD (in 2016) from Harbin Institute of Technology(HIT), China. He was a visiting student at the University of Leeds (2012-2013) and a visiting scholar at the University of Technology Sydney, Australia (2016-2017). He is currently working as a PhD candidate in the School of Electrical and Data Engineering at the University of Technology Sydney, Australia. His research interests include Human activity recognition, IoT, Deep learning, Cognitive radio, Interference alignment, Massive MIMO and UWB wireless communication.



J. Andrew Zhang (M04CSM11) received the B.Sc. degree from Xian JiaoTong University, China, in 1996, the M.Sc. degree from the Nanjing University of Posts and Telecommunications, China, in 1999, and the Ph.D. degree from Australian National University in 2004. He was a Researcher with Data61, CSIRO, Australia, from 2010 to 2016, the Networked Systems, NICTA, Australia, from 2004 to 2010, and ZTE Corp., Nanjing, China, from 1999 to 2001. He is currently an Associate Professor with the School of Electrical and Data Engineering, University of Technology Sydney, Australia. He has published more than 170 papers in leading international journals and conference proceedings. His research interests include the area of signal processing for wireless communications and sensing and autonomous vehicular networks. He has received five best paper awards for his work. He was a recipient of CSIRO Chairmans Medal and the Australian Engineering Innovation Award in 2012 for exceptional research achievements in multi-gigabit wireless communications.



Richard Yi Da Xu is currently an Associate Professor with the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia. He has authored about 50 papers, including the IEEE Transactions on Image Processing, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Neural Networks and Learning Systems, Pattern Recognition, ACM Transactions on Knowledge Discovery from Data, the Association for the Advancement of Artificial Intelligence, and the International Conference on Image Processing. His current research interests include machine learning, computer vision, and statistical data mining.



Qingqing Cheng received her M.E. degree from the Harbin Institute of Technology, China in 2014, the Master of Research (MRes) degree from the Macquarie University, Australia, in 2016, and the Ph.D degree from University of Technology Sydney, Australia, in 2020. She is currently a research associate with the University of New South Wales, Australia. Her research interests include deep learning, 5G security, privacy preservation, cognitive radio, and massive MIMO.