DEVICE-FREE WIFI SENSING FOR HUMAN ACTIVITY RECOGNITION

by
Zhenguo Shi

Dissertation submitted in fulfilment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

under the supervision of

A/Prof. J. Andrew Zhang
A/Prof. Richard Yida Xu

School of Electrical and Data Engineering
University of Technology Sydney
Sydney, Australia

July 2021

ABSTRACT

Human activity recognition (HAR) using WiFi signals (WiFi-based HAR) has drawn considerable interest from the research community. In contrast to traditional device-based sensing techniques, 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.

Recently, applying deep learning (DL) to WiFi-based HAR has received strong research interest. Assisted by signal processing techniques, DL-based HAR methods are able to automatically extract deep features from input signals, contributing to successful recognitions. Despite its effectiveness in improving recognition performance, DL-based HAR methods suffer from several inherent drawbacks. First, feature extraction is a challenging task that always bottlenecks the recognition performance. Second, DL-based HAR requires a large number of training examples from the testing/targeted environment or/and previously seen environments (PSEs) to train the corresponding DL architectures. When the number of required samples is not sufficient, the sensing performance will drop dramatically. Third, the trained model in one environment cannot be directly applied to another environment without additional effort.

My PhD thesis aims to provide novel solutions to the above WiFi-based HAR issues. Specifically, to extract effective features, we propose two advanced methods together with leveraging the property of DL architectures to enhance the quality of input signals of DL networks and extracted repre-

sentative features. For a reliable recognition with limited training samples, we propose a novel HAR scheme by developing innovative signal processing methods and exploring the characteristics of one-shot learning to reduce the number of required training samples. The proposed HAR scheme is able to accomplish successful recognitions 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). To achieve environmental robustness, we propose two novel signal processing algorithms and leverage the features of the matching network. The proposed models are trained once and can be directly applied to various new/testing environments for reliable recognitions without requiring an additional retraining process.

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, ZHENGUO SHI, declare that this thesis, is submitted in fulfilment of the requirements for the award of DOCTOR OF PHILOSOPHY, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Name of Student: Zhenguo Shi

Signature of Student: Production Note:
Signature removed prior to publication.

Date: 29/03/2022

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere appreciation and gratitude to my supervisors, Prof. J. Andrew Zhang and Prof. Richard Yida Xu. It is by their conscientious guidance and unwavering support that I attained all the accomplishments in my PhD study. It is by their generous encouragement and enduring patience that I have developed an open and logical mindset. Their profound knowledge and rigorous attitudes on research helped me to be an independent researcher, and I am honored and pleased to have worked with them.

My special thanks are to Dr. Gengfa Fang, Dr. Forest Zhu and Dr. Yang Yang for their insightful suggestions on my research and kind help in my life. Moreover, I very strongly appreciate all the staff from the School of Electrical and Data Engineering, University of Technology Sydney, for their warmhearted help and support. Many thanks to all my colleagues and friends for their company, help, and support during my PhD study.

Last but most importantly, the deepest appreciation, respect and love are for my family. My beloved wife, Qingqing Cheng, has been accompanying and supporting me to overcome all difficulties and challenges in this long journey. It would have been impossible to accomplish this work without her unconditional love and unwavering encouragement. I wish to enormously thank my parents for raising me and encouraging me to pursue my dream. Their unbounded love and tremendous support will be the source of motivation throughout my life.

Dedicated to My Beloved Family

Contents

Abstract **iii**

Acknowledgments **ix**

Table of Contents **xiii**

List of Figures **xvii**

List of Tables **xxi**

List of Publications **xxiii**

1 Introduction **1**

1.1 Background 1

1.2 Motivations and Contributions 5

1.3 Thesis Organization 8

2 Literature Review **11**

2.1 Evolution of Human Activity Recognition 11

2.1.1 Wearable-device based HAR 11

2.1.2 Vision based HAR 12

2.1.3 Acoustic based HAR 13

2.2 Human Activity Recognition adopting Wireless Signals 15

2.2.1 RSSI-based HAR 16

2.2.2 CSI-based HAR 18
2.3 Deep Learning for CSI-based Human Activity Recognition 19
2.3.1 Sparse Autoencoder-based HAR 20
2.3.2 Convolutional Neural Network-based HAR 20
2.3.3 Recurrent Neural Networking-based HAR 22
2.3.4 Other DL-based HAR 23
2.4 Challenges for DL-based Human Activity Recognition 24
2.5 Summary 27

3 DL-based Human Activity Recognition with Sufficient Training Samples 29

3.1 Introduction 29
3.2 System Model of CSI-based Human Activity Recognition 31
3.3 The DLN-eCSI Scheme 33
3.3.1 Data collection 33
3.3.2 Data Preprocessing 34
3.3.3 Deeper Feature Extraction and Classification 36
3.4 The HAR-AF-DLN Scheme 37
3.4.1 CCE for CSI Preprocessing 37
3.4.2 AF-DLN based Human Activity Recognition 41
3.5 Implementation and Evaluation 43
3.5.1 Experimental Setup 43
3.5.2 Performance Evaluation 44
3.6 Conclusion 52

4 DL-based Human Activity Recognition with Limited Training Samples 55

4.1 Introduction 55
4.2 The MatNet-eCSI Scheme 57

4.2.1 CSI Collection and Preprocessing 57
4.2.2 MatNet based Activity Recognition 59
4.3 CCFE for CSI Preprocessing 59
4.4 MatNet based Human Activity Recognition 65
4.4.1 Architecture of MatNet 65
4.4.2 Training Strategy and Testing procedure 69
4.5 Implementation and Evaluation 70
4.5.1 Experimental Setup 70
4.5.2 Performance Evaluation 73
4.6 Conclusion 87

5 Environment-Robust WiFi-based Human Activity Recognition 89
5.1 Introduction 89
5.2 HAR-MN-EF Scheme 91
5.2.1 CSI-CE based CSI Preprocessing 92
5.2.2 MatNet-based Human Activity Recognition 94
5.3 AFEE-MatNet Scheme 97
5.3.1 AFEE based CSI Preprocessing 98
5.3.2 MatNet-PCC based Human Activity Recognition 102
5.4 Experiment and Evaluation 106
5.4.1 Performance Comparison of Different Methods 108
5.4.2 Performance Evaluation of Proposed Schemes under Various Conditions 112
5.5 Conclusion 117

6 Conclusion and Future Work 119
6.1 Conclusions 119
6.2 Future Work 121

Abbreviations 123

Bibliography **125**

List of Figures

1.1 Thesis organization 8

3.1 Main processes of human activity recognition system. 32
3.2 Influence of human activity on signal propagation. 32
3.3 A sketch of the human activity recognition system based on WiFi 802.11n. 33
3.4 Structure of deeper features extraction and classification 37
3.5 Main processing modules of the HAR-AF-DLN Scheme. 38
3.6 Structure of AF-DLN based activity recognition using CFM as input. 42
3.7 Layout of two indoor experimental areas:(a) $4m \times 6m$ meeting room. (b) $8m \times 10m$ laboratory. 44
3.8 Performance of the proposed CSI feature enhancement scheme, when comparing two similar activities"sitting" and "sit down". 45
3.9 Confusion matrix for different human activity recognition methods . . 46
3.10 Impact of the number of subcarriers on the sensing accuracy 47
3.11 Confusion matrix for different human activity recognition methods. . 49
3.12 Impact of the number of subcarriers on the recognition accuracy. . . . 50
3.13 Impact of CCE on recognition accuracy. 51
3.14 Impact of AF-DLN on the recognition accuracy. 52

4.1 Main processing modules of the MatNet-eCSI Scheme. 58
4.2 Correlation feature extraction in the proposed CCFE. 63

4.3 Structure of MatNet based activity recognition using CFM $\mathbf{D}_{en}^{A}$ and $\mathbf{D}_{en}^{\Psi}$ as the input. 66
4.4 Structure of embedding function g: CNN with bidirectional LSTM. . 68
4.5 Layout of three indoor experimental areas: (a) $3m \times 4m$ office. (b) $4m \times 6m$ meeting room. (c) $6m \times 7m$ laboratory. 71
4.6 Confusion matrix for different human activity recognition methods. . 76
4.7 Confusion matrix of proposed MatNet-eCSI for five-shot 77
4.8 Impact of the used number of receiving antennas, represented as the number of total subcarriers, on the recognition accuracy. 78
4.9 Recognition accuracy with increased number of PSEs 79
4.10 Recognition accuracy of different methods with sufficient training samples 80
4.11 Effect of CCFE on enhancing the feature signals for two similar activities"sit down" and "sitting". 81
4.12 Impact of CCFE on the recognition accuracy and required training time. 82
4.13 Impact of phase compensation on recognition accuracy 84
4.14 Impact of the number of segment K on the recognition accuracy. . . . 84
4.15 Impact of the size of data set on the recognition accuracy 86
4.16 Average recognition accuracy for different people 86
4.17 Impact of training strategy on the recognition accuracy 87

5.1 Main processing modules of the HAR-MN-EF Scheme. 91
5.2 Structure of a MatNet based HAR using CFM $\mathbf{D}_C$ as input. 94
5.3 Main modules for the AFEE-MatNet Scheme. 97
5.4 Simplified state transition diagram for six different activities. 104
5.5 Layout of three indoor experimental areas: (a) $3m \times 4m$ office. (b) $4m \times 6m$ meeting room. (c) $6m \times 7m$ laboratory. 106
5.6 Confusion matrix for different human activity recognition methods. . 110

5.7 Recognition accuracy with increased number of source environments . 111

5.8 Impact of the number of subcarriers on the recognition accuracy. . . . 112

5.9 Average recognition accuracy for different people 114

5.10 Impact of P on the recognition accuracy and training time for proposed AFEE-MatNet. 116

List of Tables

3.1 Sensing performance of different methods in the two indoor configurations . . . 45
3.2 Training time for different methods . . . 48
3.3 Average Sensing Accuracy of the three methods in the two indoor configurations . . . 48
3.4 Training time for different methods . . . 50
3.5 Comparison for the Proposed Methods . . . 52

4.1 Average recognition accuracy of the five methods in the first indoor configurations . . . 74
4.2 Average recognition accuracy of the five methods in the second indoor configurations . . . 75
4.3 Average recognition accuracy of the five methods in the third indoor configurations . . . 75
4.4 The number of PSEs required by different methods for the similar sensing accuracy . . . 79
4.5 Impact of CCFE on recognition accuracy for different methods . . . 80
4.6 Sensing performance using different input signals in the first configuration with **PSE**2 . . . 85

5.1 Average recognition accuracy of different methods in three indoor configurations . . . 109

5.2 Impact of PCC on recognition accuracy for different methods 113

5.3 Impact of CSI-CE on proposed HAR-MN-EF 113

5.4 Recognition accuracy using different input signals. 114

5.5 Impact of AFEE on proposed AFEE-MatNet 115

5.6 Characters of Proposed Methods 117

List of Publications

Journal Publications

- **Zhenguo Shi**, Andrew Zhang, Richard Xu, Qingqing Cheng, Environment-Robust Device-free Human Activity Recognition with Channel-State-Information Enhancement and One-Shot Learning, in IEEE Transactions on Mobile Computing (TMC), doi: 10.1109/TMC.2020.3012433. (Corresponding to Chapter 4)

- **Zhenguo Shi**, Qingqing Cheng, Andrew Zhang, Richard Xu Environment-independent WiFi-based Human Activity Recognition using Enhanced CSI and Deep Learning under review in IEEE Internet of Things Journal (IoT). (Corresponding to Chapter 5)

Conference Publications

- **Zhenguo Shi**, Andrew Zhang, Richard Xu and Gengfa Fang, Human activity recognition using deep learning networks with enhanced channel state information, in 2018 IEEE Globecom Workshops (GC Workshops), Dec 2018, pp. 16. (Corresponding to Chapter 3)

- **Zhenguo Shi**, Andrew Zhang, Richard Xu, Qingqing Cheng, "WiFi-Based Activity Recognition using Activity Filter and Enhanced Correlation with

Deep Learning," 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 2020. (Corresponding to Chapter 3)

- **Zhenguo Shi**, Andrew Zhang, Richard Xu, Qingqing Cheng "Deep learning networks for human activity recognition with CSI correlation feature extraction", in IEEE International Conference on Communications, ICC 2019. (Corresponding to Chapter 4)

- **Zhenguo Shi**, Andrew Zhang, Richard Xu, Qingqing Cheng, Towards Environment-independent Human Activity Recognition using Deep Learning and Enhanced CSI, accepted in IEEE Global Communications Conference (GLOBECOM), 2020. (Corresponding to Chapter 5)

Patent

- Andrew Zhang, Richard Xu and **Zhenguo Shi**. A System and method for event recognition. Under review in Hong Kong Short Term Patent, Application No. 19129029.5

Other Journal Publications

- Qingqing Cheng, **Zhenguo Shi**, Diep N. Nguyen and Eryk Dutkiewicz, Sensing ofdm signal: A deep learning approach, in IEEE Transactions on Communications (TCOM), 2019.

- **Zhenguo Shi**, Zhilu Wu, Zhendong Yin,Zhutian Yang, and Qingqing Cheng.Novel Markov channel predictors for interference alignment in cognitive radio network.Wireless Netw24,19151925 (2018).

Other Conference Publications

- Qingqing Cheng, **Zhenguo Shi**, Jinhong Yuan, Spectrum Sensing in Full-Duplex OFDM Systems using One-Shot Learning in IEEE International Conference on Communications, ICC 2021.

- Qingqing Cheng, **Zhenguo Shi**, Diep N. Nguyen, Eryk Dutkiewicz, "Non-cooperative OFDM Spectrum Sensing Using Deep Learning", International Conference on Computing, Networking and Communications, ICNC 2020.

- Qingqing Cheng, **Zhenguo Shi**, Diep N. Nguyen, Eryk Dutkiewicz, "An OFDM Sensing Algorithm in Full-Duplex Systems with Self-Interference and Carrier Frequency Offset", in IEEE Global Communications Conference (GLOBECOM), 2019.

Chapter 1

Introduction

This chapter provides an overview of the thesis. The background information about human activity recognition is introduced in Section 1.1, including advanced human activity recognition technologies together with their advantages and disadvantages. Section 1.2 describes the motivations and main contributions of the thesis. Finally, the organization of the thesis is provided in Section 1.3.

1.1 Background

Over the past decades, human activity recognition (HAR) has embraced tremendous momentum with the recent advancement of sensing technologies [1]. HAR plays an influential role in humanity, which has been widely applied to assist people's daily life, covering a wide range of compelling applications such as elder care, smart-home appliances, safety surveillance [2]. The main principle of HAR is to recognize the undergoing activities by monitoring and analyzing a person's behaviors. Many tasks need to be accomplished for a successful HAR, including environment and activity monitoring (i.e., data collection), data processing and pattern classification [3]. In this regards, for a successful HAR, it is critical to:

- select appropriate sensors/devices to monitor and collect the information about a person's activities together with the corresponding changes of the environment;

- design behavior models that enable agents (e.g., software systems) to perform manipulation;
- process and manage the collected data through fusion or aggregation for extracting high-level and informative features;
- develop and propose desirable approaches to infer behaviors using the acquired data from deployed sensors;
- determine the conducted behaviors based on pattern classification.

To facilitate HAR, numerous attempts have been developed over the recent years, mainly including wearable device-based HAR, video-based HAR [4], wireless signal-based HAR [5]. For wearable device-based HAR systems, the target individual is required to wear special sensors or devices to acquire data that is influenced by human behaviors. For the stage of feature extraction, the data is either processed by the wearable device locally or sent to the central server. This type of HAR is capable of recognizing human behaviors reliably utilizing the data collected by special devices. However, wearable sensor-based sensing methods need the person involved to wear or be equipped with some devices for data collection, which limits its practical applications. Moreover, these sensor-based sensing solutions need extra devices, resulting in a surge in the cost of deployment and maintenance [6].

Apart from the wearable device-based HAR, widely deployed cameras make the video-based HAR a feasible solution. Cameras have been widely deployed almost everywhere for providing people with a safe and convenient life. The installed cameras can collect a lot of videos and images, which are critical features for facilitating a sensing task [7]. Although a fair HAR can be performed with the information collected by cameras, this type of HAR is vulnerable to light conditions and restricted to line-of-sight (LOS) scenarios. Another drawback of video-based sensing techniques is that they would expose users' face information, raising privacy concerns. Given the above, there is an urgent demand for developing a new sensing technology with features of low-cost, privacy safe, and convenience [8].

In recent years, wireless networks have gained rapid development and been ubiquitously deployed to meet the demand for wireless data traffic. This stimulated

a surge growth for wireless signals, bringing new chances for reliable and non-intrusive wireless detection techniques, including estimation, detection, tracking, and recognition of human behaviors [9]. The fundamental insight of wireless sensing is that the wireless signals would be affected by obstacles when traveling from the transmitter to the receiver. This may induce various changes on its propagation links, including diffraction, attenuation, multipath effects, reflection, and refraction [10]. In other words, wireless signals would carry the environmental variations in the transmission space. In this regards, when a person performs activities, it leads to variations on the wireless signal propagation, and different activities may have their particular influences. Therefore, it is possible to accomplish the sensing task by utilizing the unique impact of human behaviors on wireless signal propagation.

The wireless sensing can offer many appealing advantages, compared to the conventional video-based and sensor-based sensing solutions [11, 12]. First, the sensing approach using wireless signal does not require extra communication infrastructures. It can be compatible with the existing communication architectures, significantly saving the deployment cost and overhead. Second, the wireless sensing technique is a type of non-intrusive and privacy-preservation detection solution, as it does not expose users' private data such as the face information of a person, which is a common concern for video-based sensing methods. Third, it is convenient to deploy the wireless sensing methods which do not need the target individual to wear additional devices, which however is necessary for sensor-based sensing algorithms.

Thanks to the promising features of the wireless sensing technique, it has been commonly applied to a wide range of applications, including activity recognition, action detection, motion estimation, motion tracking, and indoor localization [13, 14]. Among these applications, employing wireless signals for human activity recognition (wireless signal-based HAR) has gained considerable momentum and has been used in a variety of fields, such as sport, health care, and monitoring systems [15].

In the context of wireless signal-based HAR, various wireless signals have been investigated for detecting a person's behaviors by leveraging unique properties of those signals, such as millimeter wave (mmWave) signals [16] and WiFi signals [17]. Among these signals, WiFi signals, one of the most pervasive wireless signals, have

received paramount interest as a promising source for device-free HAR using wireless signals [18,19]. In such a system, WiFi-based devices are placed at different locations in the targeted environment [20]. Human activities would modulate various changes on the propagation of WiFi information, which can be collected and analyzed to detect different behaviors [21]. To collect and quantify the variations of received WiFi signals, some properties of the physical layer over wireless links can be measured and utilized, including received signal strength indicator (RSSI) and channel state information (CSI) [22, 23]. These properties are readily available using modified software and commercial network interface cards such as Atheors 9580 network interface card (NIC) [24] and Intel 5300 NIC [25].

For RSSI-based HAR, the critical task is to detect human behaviors by utilizing the changes on WiFi signals during the propagation, e.g., attenuation. Since it is easy to obtain RSSI in any commercial WiFi device without the requirement of extra devices or hardware, RSSI-based HAR has drawn tremendous attention from both the academic and industrial communities [26]. It is noteworthy that RSSI can only provide coarse-grained information of channel characteristics, such as the value of single-path loss for each packet, dramatically limiting its sensing capabilities. Consequently, RSSI-based HAR can only perform HAR for limited kinds of human behaviors, restricting its applicability in practice. Moreover, it is difficult for RSSI-based HAR to guarantee a statable sensing performance, especially in complex environments [27]. Therefore, it is necessary to utilize more fine-grained information for HAR, so as to achieve better sensing performance. Given that purpose, CSI has been extensively used for HAR, as it can reflect more complex changes of wireless links such as fading, power decay based on distance and scattering [28, 29]. Consequently, CSI is able to provide more fine-grained features, e.g., phase or amplitude information, which is essential for accomplishing reliable HAR. Compared to RSSI-based HAR, CSI-based HAR is able to offer a reliable recognition for more types of behaviors and to work well even in complex environments [30, 31]. Despite the appealing sensing capability, some challenging issues in CSI-based HAR need to be addressed. For instance, how to select and design proper features is critical for CSI-based HAR, which directly influences the sensing performance [32]. To deal with

this concern, many works have made attempts to develop various signal processing techniques, while feature extraction is still an open research problem.

To address the above problem, recent advances in CSI-based HAR have leveraged the properties of various deep learning (DL) networks [33, 34]. With DL networks and advanced signal processing techniques, representative features can be effectively learned from the input signals of DL architectures. Then the extracted information is processed and classified into different domains or types, accomplishing a successful activity recognition [35]. Many DL architectures have proven their effectiveness for CSI-based HAR, such as the sparse autoencoder (SAE), recurrent neural networking (RNN), convolutional neural network (CNN), and one-shot learning [36, 37]. Although DL-based HAR schemes can achieve desirable sensing results, they suffer from some challenges that severely limit their performance. First, DL-based HAR solutions are relying highly on the extracted features which are directly influenced by the selection of input signals. Designing proper signal input for DL networks always requires an elaborate process and is dependent on designers' experiences [38]. Second, most DL-based HAR approaches would undergo a dramatic performance degradation with a limited number of training/labeled samples. In other words, the recognition model can be trained well only with a sufficient number of training data from the required environments, e.g., previously seen environments or the testing/new environment. However, acquiring an adequate number of labeled data is not always accessible in practical scenarios [39]. Third, another drawback of DL-based HAR methods is that they are heavily specific to environments. A sensing model trained in one environment cannot work well if directly applied it to a new environment, resulting in a notable performance drop. Additional efforts or training processes are required to re-train the model, severely restricting the applicabilities of DL-based methods in practice [40].

1.2 Motivations and Contributions

Given these significant and challenging issues, in this thesis, we design and propose various DL-based HAR schemes/methods by leveraging the features of DL networks

and advanced signal processing algorithms. The main work of the thesis concentrates on three critical issues in DL-based HAR using CSI. Specifically, we first investigate the sensing accuracy of DL-based HAR with sufficient training samples from the required environments, e.g., previously seen environments and the target/testing environment. For that, we design two novel schemes to facilitate high-performance recognitions, by leveraging the properties of advanced signal processing methods and DL architectures. Second, we study the DL-based HAR in a scenario with a limited number of training data from the required environments. Towards this goal, we develop an innovative framework/scheme with the help of one-shot learning and signal processing methods. Third, we focus on the environment-independent HAR to explore the environmental robustness. To do that, we propose two novel DL-based HAR schemes to mitigate the impact of environment-dependent but activity-unrelated data; meanwhile, we enhance the quality of behavior-related information, thereby achieving environmental robustness. We summarize the main contributions of the thesis as follows.

- We design two novel HAR schemes leveraging the features of DL networks and advanced signal processing algorithms. To be specific, we first propose a human activity recognition scheme using Deep Learning Networks with enhanced Channel State information (DLN-eCSI). We develop a CSI feature enhancement scheme (CFES), including two modules of background reduction and correlation feature enhancement, for preprocessing the data input to the DL architectures. To further improve the sensing performance, we develop a novel scheme for CSI-based HAR using activity filter-based deep learning network (HAR-AF-DLN) with enhanced correlation features. We first develop a novel CSI compensation and enhancement (CCE) method to compensate for the timing offset between the WiFi transmitter and receiver, enhance activity-related signals and reduce input dimension to DL networks. Then, we design a novel activity filter (AF) to differentiate similar activities (e.g., standing and lying) based on the enhanced CSI correlation features obtained from CCE. Therefore, the proposed HAR-AF-DLN scheme gains the capability of detecting similar activities with reliable sensing results. (Chapter 3)

- We study the DL-based HAR when the number of training samples from the required environment is limited. To achieve that, 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 employ one-shot learning to learn and extract distinguishable features from the input signals. We also propose a novel training strategy that effectively utilizes the data set from the previously seen environments, and bridges an effective connection between features from previously seen environments and from the target environment. In the least, the strategy can effectively realize activity recognition using only one sample for each activity from the testing environment and the data set from one previously seen environment. The extensive experimental results demonstrate the effectiveness of the developed Mat-Net-eCSI, in both sensing accuracy and training complexity. (Chapter 4)

- We investigate the environmental robustness of DL-based HAR and propose two innovative schemes. We first propose an environment-robust CSI-based HAR, drawing support from a matching network and enhanced features (HAR-MN-EF). Under the proposed HAR-MN-EF scheme, an architecture trained with a limited number of previously seen environments (i.e., source environments) can be used to directly identify different activities in a new/testing environment, without the requirement of a re-training process. To further improve the environmental robustness, we propose an activity-related feature extraction and enhancement method (AFEE) and Matching Network (AFEE-MatNet). The proposed method facilitates the "one-fits-all" recognition scheme, meaning that the trained model can be directly applied in new/unseen environments without any re-training. We design the AFEE method to enhance CSI quality by eliminating the impact of noise. Specifically, the approach mitigates environmental noises unrelated to activity while better compressing and preserving the behaviour-related information. Moreover, the

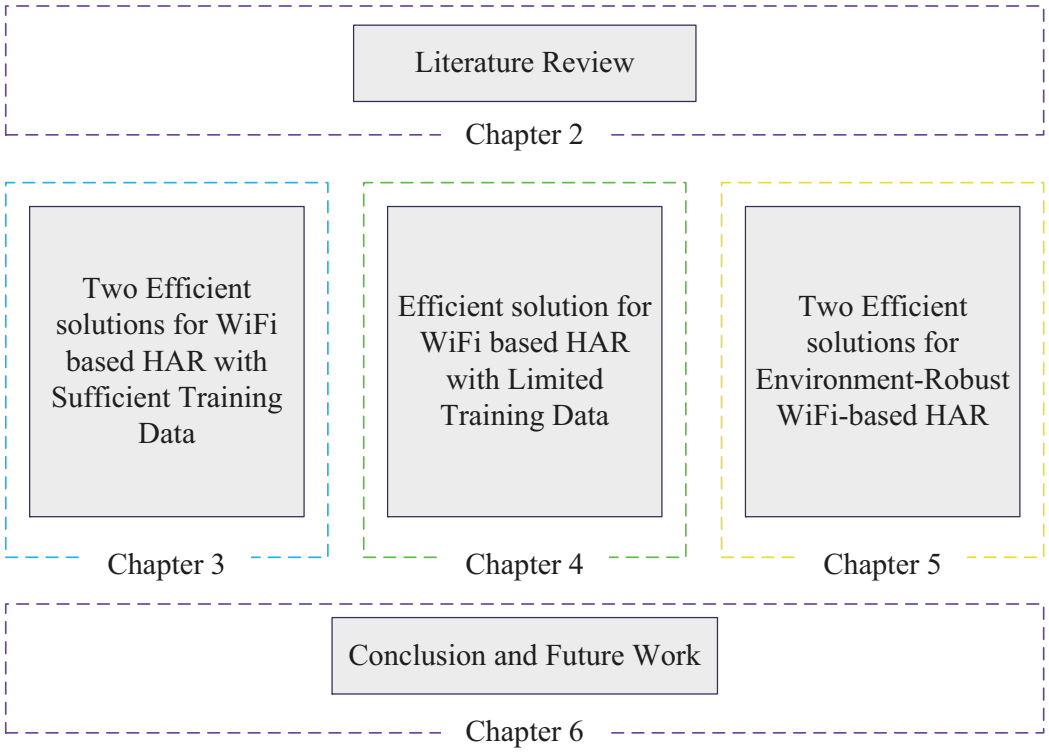


Figure 1.1: Thesis organization

feature signals generated by AFEE are anticipated to have a decreased size, which in turn significantly shortens the training time. For effective feature extraction, we propose to use the matching network to learn transferable features shared among source environments. Further, we introduce a prediction checking and correction scheme to rectify some classification errors that do not abide by the state transition of human behaviours. (Chapter 5)

1.3 Thesis Organization

We structure the remainder of the thesis as follow, which is also shown in Fig. 1.1.

In Chapter 2, we review and study a significant number of previous literature and on-going work to show the up-to-date development of HAR technology. We first present the evolution of HAR including three conventional HAR approaches in Section 2.1. Then, we introduce the development of device-free HAR adopting wireless signals as well as their corresponding features in Section 2.2. Section 2.3 discusses the recent advances in DL-based HAR methods, and the associated

challenges are stated in Section 2.4.

In Chapter 3, we investigate DL-based HAR in a scenario with sufficient required training samples. To achieve accurate sensing results, we propose two novel HAR schemes by leveraging the properties of signal processing methods and DL networks. We first review the cutting-edge work in this area in Section 3.1. The system model of CSI-based HAR and problem formulation are described in Section 3.2. In Section 3.3, we provide the detailed information of our proposed HAR scheme drawing support from DL architectures and enhanced CSI measures. To further improve the sensing accuracy, we propose a second HAR scheme to mitigate the effect of time offset and improve the CSI quality in Section 3.4. The proposed scheme is able to effectively differentiate activities, especially for similar activities. The experiment and evaluation of the designed HAR schemes are stated in Section 3.5. The extensive results demonstrate that the proposed schemes significantly outperform the other relevant HAR works, with notable higher accuracies and reduced complexity. The conclusion of this chapter is summarized in Section 3.6.

In Chapter 4, we investigate DL-based HAR in a scenario with a limited number of required training samples. The up-to-date relevant works in this field are discussed in Section 4.1. Section 4.2 provides an overview of the proposed scheme, including three key stages. In Section 4.3, we present the designed data preprocessing method to improve the CSI quality by enhancing the activity-related features and mitigating activity-unrelated data. Section 4.4 demonstrates the process of recognizing activities using the developed HAR scheme. The performance evaluation of the presented scheme is provided in Section 4.5, demonstrating that the proposed scheme is greatly superior to the other related HAR methods when the number of required training samples is limited. Section 4.6 concludes the main contributions of this chapter.

In Chapter 5, we investigate the environmental robustness of DL-based HAR. For that purpose, we propose two innovative HAR schemes in which the HAR models trained with previously seen environments (i.e., source environments) can be directly applied to the testing environment without extra efforts. In Section 5.1, we introduce state-of-the-art advances of environmental-independent DL-based HAR methods. Then, we state details of the first proposed HAR framework in

Section 5.2, including stages of data collection, feature extraction and behavior classification. To further improve the detection performance, we design a second scheme in Section 5.3, which is able to extract more transferable and generalized features from source environments utilizing the properties of proposed advanced signal processing methods and DL architectures. We conduct numerous experiments to evaluate the performance of developed schemes in Section 5.4, before summarizing the conclusion of this chapter in Section 5.5.

Chapter 2

Literature Review

This chapter reviews and studies a comprehensive number of previous literature and ongoing work to demonstrate the up-to-date development of HAR. To be specific, Section 2.1 provides an overview on the evolution of HAR. Section 2.2 presents the current discussions on HAR schemes utilizing the characteristics of wireless signals. Section 2.3 describes state-of-the-art development of using deep learning (DL) for channel state information-based HAR. Section 2.4 discusses the associated challenges of CSI-based HAR using deep learning.

2.1 Evolution of Human Activity Recognition

Over the past decades, sensing technology has experienced very extensive development and has demonstrated its great potential for HAR, covering a wide range of promising applications, e.g., elder care, safety surveillance, smart home appliances [41–43]. To effectively facilitate HAR, considerable research has been dedicated to detect and identify human activities successfully. As a result, numerous studies for HAR technology have been conducted, mainly concentrating on wearable-device based HAR [44, 45], vision based HAR [46, 47], and acoustic based HAR [48].

2.1.1 Wearable-device based HAR

The critical insight of wearable-device based HAR is that the activity data is first collected by the wearable devices/sensors equipped with the target (e.g., smart-

watches or smart-glasses). Then, the acquired data is processed in the local sensors or transmitted to the central server for feature extraction. As a result, the key features can be extracted, which will be used to detect different behaviors. To effectively facilitate the wearable-device based HAR, plenty of effort has been made from various perspectives. To name a few, a solution proposed in [49] was targeting to realize recognition with the signals acquired by a smartwatch. The basic idea is that human activities or gestures would induce changes of motion energy in the accelerometer and gyroscope of the smartwatch. By studying and classifying these unique variations, the proposed scheme was capable of achieving fair sensing results. The work in [50] proposed a framework to enable the interactions between users' actions and a computing device. By attaching a radio-frequency identification (RFID) tag on the target's fingers, the movements or gestures signals can be interpreted to the computing devices for feature extraction and classification. For sensing performance improvement, the authors in [51] proposed to employ a series of wearable devices to capture the changes of data caused by human actions. Then, the collected data was processed to obtain the distinguishable features before feeding them into the classifier. The authors employed three types of classifiers to cluster extracted features into different categories, achieving a fair recognition. Despite the effectiveness in recognizing behaviors, wearable-device based HAR schemes need the target person to wear specialized sensors or devices anywhere and anytime for data collection, which is not always convenient or feasible in practical scenarios [52]. Moreover, there is extra overheads of wearable-device based HAR with respect to the hardware installation and maintenance requirements, which may severely restrict its applicabilities in practice [53, 54].

2.1.2 Vision based HAR

For the vision based HAR, the core idea is that the video information of human behaviors is acquired via sensors (e.g., cameras). On this basis, the obtained data is then processed into image information for feature extraction and corresponding action classification [55]. To achieve that, a considerable amount of research has been conducted in this direction. To be specific, the authors in [56] proposed a camera-

based HAR scheme and applied it to health services. They first used a camera to capture the video data of human actions in daily life. Then the collected signals were processed to extract the image features, classify the obtained information into different categories and identify human behaviors. For a reliable HAR, recent work in [57] developed a camera-based HAR using multilevel wavelet decomposition. Upon receiving the video recording of human behaviors, the multilevel wavelet decomposition was proposed to extract distinguishable features from activity frames. Next, the captured information was clustered into different types for identifying human actions. To improve detection accuracy, the authors of [58] designed a HAR scheme using data collected from the camera. For that purpose, they processed the sequence of depth images captured by a camera to extract spatiotemporal multi-fused features. These features were used to indicate the characteristics of human activities and differentiate different behaviors. Although the vision based HAR is able to perform activity recognition, it still has some limitations. For instance, these methods heavily rely on visible light sensors or cameras, increasing the cost of installation and maintenance. When this requirement cannot be met, it is difficult to achieve a reliable sensing result. Even with those required devices, the light condition in the sensing environment is another factor restricting the recognition performance. In other words, the vision based HAR can only perform successful recognitions under certain light conditions. In such a case, the sensing results would be easily affected by some issues, e.g., opaque obstructions, fog, illumination condition, or smoke. Another associated drawback of vision based HAR is that it performs HAR intrusively, as it may expose the target's private information (e.g., face information) to others when collecting the video data [59, 60].

2.1.3 Acoustic based HAR

For the acoustic based HAR approaches, they perform HAR by first collecting the audio signals of individuals' activities using some audio sensors such as microphones [61,62]. Then the stored data is processed to extract proper audio features which are essential for classifying different behaviors [63,64]. Recently, more and more research works have focused their attention on the audio based HAR methods. For instance,

a recognition model was designed in [65] by exploring the properties of channel impulse response (CIR) measurements for precise detection. With that purpose, the authors first acquired CSI measurements indicating the essential information of ultrasound signals, such as the strength and propagation paths of reflected signals. The collected CSI measurements were mapped into images, and then classified into different categories for recognition. Another solution in [66] concentrated on active ultrasonic sensing for activity recognition. In particular, a smartphone was used as a receiver to collect ultrasonic signals emitted by the speaker. On this basis, this stored data was processed to extract informative features such as the doppler effect of received signals. Then, the obtained features were used for classifying different activities or gestures. For better sensing performance, the authors of [67] designed an acoustic based HAR scheme to improve the recognition accuracy and robustness. Towards that target, they elaborated on a mechanism to eliminate the effect of signal interference and frequency selective fading. Moreover, data augmentation methods were developed using a small number of received data to increase the robustness of the proposed scheme. As a consequence, the proposed framework is able to achieve reliable sensing results even with a limited amount of collected signals. In spite of the promising capability in differentiating different behaviors, the acoustic based methods are highly susceptible to many factors in our daily life, such as surrounding sound noise and ambient interference [68]. Moreover, since the acoustic signals suffer from fast attenuation when travelling in space, the detection range is quite limited, which severely limits its potential in practical applications. Apart from that, the high cost of specialized devices is another issue that needs to be considered [69, 70].

Based on the discussion mentioned above, it is obvious that the conventional HAR techniques fail to provide non-intrusive, reliable, and low-cost solutions. To overcome the limitations associated with these traditional HAR methods, a more advanced and promising technique is urgently required to meet the increasing demand for activity recognition.

2.2 Human Activity Recognition adopting Wireless Signals

Recent advances in device-free sensing using wireless signals demonstrate its great potential for HAR, drawing considerable interest from academia and industry [71–73]. In contrast to traditional device-based sensing techniques, wireless signal-based HAR (WS-HAR) removes the requirement of equipping the target with any devices. The underlying principle of WS-HAR is that human activities induce different effects on wireless signal propagations or links, causing e.g., shadowing, reflection, diffraction, scattering phenomena of wireless links [74–77]. By detecting the variations of wireless signals, e.g., differences in phase or amplitude of the received data, one can then accomplish the classification task and classify these activities [78–81].

Compared with conventional device-based sensing techniques (e.g., wearable devices), WS-HAR possesses several appealing advantages. To name a few, WS-HAR does not require the targeting users to wear any devices, so the process of recognition becomes more feasible and convenient [82–85]. Moreover, WS-HAR is able to complete the sensing task without exposing the user's private information, such as the face of the user, which is indeed a concern for camera-based HAR. Apart from that, WS-HAR also enjoys some appealing features such as low-cost and wide availability, making it an attractive sensing solution for a wide range of applications in security and safety in home and office, health care, and intelligent monitoring [86–90].

In the context of WS-HAR, various types of wireless signals have been investigated for identifying human behaviors, including mmWave and WiFi signals. For mmWave signal, it has played an influential role in wireless sensing and been treated as a promising solution for WS-HAR [91]. Since mmWave frequency bands can provide more available bandwidths than low frequencies (e.g., 2.4GHz), more information can be involved in the transmitted signals. In this regards, more discriminative features can be effectively captured and extracted, contributing to an accurate sensing [92, 93]. Despite the promising features of the mmWave signal,

it suffers from severe attenuation due to its short carrier wavelength. In other words, it is difficult for mmWave signal to bypass objects which are larger than its wavelength. Moreover, the transmission range of the mmWave signal is quite short, restricting its applicabilities in practical scenarios. Apart from that, mmWave-based sensing is heavily relying on specialized devices which are always very expensive. Given the above, mmWave-based sensing does not suit recognition in large-scale deployments [94].

Recently, WiFi-based HAR is receiving particular attention because of its promising features. Compared to mmWave signal, WiFi signal enjoys many unique advantages, such as wider availability, lower cost, longer transmission coverage, stronger capability of passing through objects [95–97]. To boost WiFi-based HAR, numerous studies have been conducted to capture and quantify the differences in WiFi signal propagation induced by different human activities. The majority of the existing techniques can be categorized into two main branches: leveraging the RSSI of WiFi signals [98–100] and the CSI of WiFi physical layer information [101–103].

2.2.1 RSSI-based HAR

Utilizing RSSI for WiFi-based HAR (RSSI-based HAR) has drawn considerable interest from the research community. The underlying principle of RSSI-based HAR is that the movements of a human body within the area of the WiFi network would induce various signal attenuations, resulting in unique fluctuation patterns on RSSI. Through analyzing the differences in RSSI, e.g., means and peak-to-peak values of RSSI, it is possible to achieve a fair sensing accuracy [104, 105]. A recent work [106] investigated WiFi-based HAR by leveraging the variations of RSSI. To perform HAR, the authors in this work first extracted several features empirically from RSSI signals, such as median signal strength and the highest signal peak. Then, they adopted a k nearest neighborhood (KNN) classifier to classify some types of activities, including walking, standing, crawling, and lying. With the same purpose, another work [107] developed a system to detect human motions, specifically for jointly detecting the fall activity and human localization. For detecting the falling behavior, a hidden Markov model is employed to collect and track RSSI values, so as

to discern special features of different motions. Regarding human localization, the impact on wireless signals exerted by human beings, such as reflection, scattering and diffraction phenomena of signals, is learned and leveraged to determine the location of people. While focusing on improving the applicability of HAR techniques, the authors in [108] put their attention on detecting both dynamic and static activities in the non-ad-hoc, active and passive HAR systems, respectively. They put key features (e.g., the maximum peak of the signal amplitude) into KNN and decision tree classifier, respectively, to identify different motions or behaviors. The work in [109] proposed a two-stage approach to achieve a reliable HAR result. Specifically, the authors first collected and stored WiFi signals modulated by human motions, and then extracted the dominant frequency component and time-difference variant of received signals. With the extracted features, their proposed scheme was able to classify different movements drawing support from the random forest algorithm.

The above studies demonstrate that RSSI-based schemes/methods are capable of achieving a fair HAR performance, while they encounter some limitations which severely restrict the sensing accuracy and stability. For instance, RSSI values can provide coarse-grained information of wireless channels, such as single path loss for each packet. As a consequence, HAR methods based on RSSI can only recognize a limited range of behaviors. Moreover, it is impossible to guarantee stable RSSI values even within a static indoor environment, leading to unreliable sensing results and thus limiting its potential in practical applications. Another drawback of RSSI-based HAR methods is that they are vulnerable to noise, shadow fading and the multi-path effect, so the sensing accuracy will degrade severely if the environment is complex [110, 111]. To overcome these challenges, it is essential to obtain and utilize fine-grained information of WiFi signals to detect human behaviors. Different from RSSI, CSI contains more complex values of wireless signals, e.g., phase and amplitude information of the received signals. Thus, CSI has been treated as a promising candidate to provide fine-grained information of WiFi signals, attracting a considerable increase of attention from researchers.

2.2.2 CSI-based HAR

For CSI-based human activity recognition (CSI-based HAR), the sensing task can be accomplished by analyzing the different characteristics of the CSI that represent more fine-grained information, such as amplitude, phase, and frequency diversity [112–114]. In such a case, CSI-based HAR is more promising for achieving better recognition performance, compared to RSSI-based HAR [115, 116]. A lot of existing works on CSI-based HAR have devoted to improving the sensing accuracy by leveraging signal processing measurements [117,118]. For example, a recent solution in [119] investigated the HAR problem and developed two models, e.g., CSI-speed based model and CSI-activity based model, to facilitate a successful detection. To be specific, in the CSI-speed based model, the authors built an effective connection between the speed of human movements and the frequencies of CSI power variations. The CSI-activity based model describes the relationship between human behaviors and the movement speed corresponding to different parts of the human body. With the proposed two models, a successful HAR with reliable sensing accuracy can be achieved. Towards the same goal, the authors in [120] presented a Hilbert-Huang Transform (HHT)-based scheme to facilitate HAR. They learnt the discriminative features of CSI using the HHT method and also tried to eliminate the differential interference exerted by repeating the same activity. Using the proposed method, the connection among the CSI signal feature, activity interval time and activity duration can be effectively bridged, which plays a significant role in classifying activities. To further improve the CSI quality, the work in [121] investigated the co-channel interference of CSI and developed corresponding solutions to mitigate the effect of co-channel interference on sensing performance. To that end, the authors exploited the phase component that is independent of co-channel interference to eliminate its impact on the quality of CSI. Consequently, the recognition accuracy can be improved using the enhanced CSI generated from the former step. Apart from the above works, the authors of [122] accomplished a desirable HAR using the principal component analysis (PCA) method. In particular, the designed framework is able to detect stationary and moving people simultaneously. For the detection of stationary persons, the authors took human breathing into consideration and treated it as

an intrinsic indicator to capture the presence of people. To identify the activities of moving people, both the phase and amplitude features of CSI were explored to distinguish various behaviors. The works in [123] and [124] also investigated CSI-based techniques and paid attention to dealing with the undesirable factors in the recognition process. Specifically, these two works mainly employed discrete wavelet transform (DWT) to remove the background noise and improve the quality of CSI.

Based on those works mentioned above, various pioneering approaches for CSI-based HAR have been proposed by exploring the properties of signal processing techniques [125–128]. While they have achieved some promising results, these methods often face challenges such as feature selections and feature fusions. To be specific, the performance of these methods nonetheless heavily depends on the precursor step of feature selection. Therefore, should the precursor steps fail to achieve their goal, the recognition accuracy may degrade significantly [129]. However, selecting and extracting a proper feature always requires elaborated designs with a complex process, increasing the sensing overhead. Moreover, feature fusion is another challenging issue that may bottleneck the detection performance. The reason is that, as a general rule, it is difficult to achieve a fair sensing result using a single type of feature, especially in complex scenarios. Consequently, fusing various features becomes an effective way of improving recognition accuracy. However, feature fusion is an open research problem for the existing CSI-based HAR approaches. Although some works have investigated this issue and achieved some progress to a certain extent, more effort and attention are still required.

2.3 Deep Learning for CSI-based Human Activity Recognition

To overcome the above challenges, DL networks have been widely utilized in CSI-based HAR [130, 131]. Assisted by signal processing techniques, DL-based HAR methods are able to automatically extract and transform deep features from input signals. As a consequence, proper features for recognition can be effectively learned and extracted, significantly improving recognition performance [132–134]. To facili-

tate DL-based HAR, numerous schemes have been proposed using a variety of deep learning networks (DLNs) together with unique signal processing methods.

2.3.1 Sparse Autoencoder-based HAR

The sparse autoencoder (SAE) is a kind of fully connected DL network including the input layer, hidden layers, and the corresponding output layer. It follows an unsupervised manner to train the HAR model using received signals. Since SAE is an easy-to-train and lightweight DL architecture, it has been treated as a promising option for HAR [135]. The authors in [136] transformed CSI measurements from multiple wireless links into radio images, followed by learning the texture and color information from those radio images. Then, they adopted an SAE architecture to extract representative features from CSI signals to identify different activities. Using the same SAE architecture, the authors of [137] developed a recognition method by transferring the CSI measurements into radio images before using the SAE network for feature extraction. The extracted information from radio images was then processed by the softmax regression method for activity recognition. However, the sensing performance of the above methods is susceptible to the quality of input features. To achieve higher sensing accuracies, a HAR solution was proposed in [138] leveraging the property of stacking denoising autoencoder (SDAE). By leveraging the property of SDAE, the noise was first removed from the raw received signals, then the data was processed to extract distinguishable features for HAR following an unsupervised learning manner.

2.3.2 Convolutional Neural Network-based HAR

The convolutional neural network (CNN) generally includes multiple layers, with two main processing modules at each layer, e.g., the convolution module and pooling module. Since CNN show great potential in dealing with two-dimensional data, it has been regarded as an effective architecture for DL-based HAR, thereby gaining very considerable momentum. The authors in [139] applied CNN and long-short term memory (LSTM) for behavior recognition, by exploiting the characteristics

of spatial information collected from multiple antenna pairs. With the proposed scheme, the quality of extracted features can be improved, contributing to a reliable sensing result. Another work in [140] accomplished the sensing task drawing support from CNN and unique signal processing methods. They first converted the received signals into CSI images containing spatial, frequency and time data. Then, to extract effective features, they designed an innovative CNN architecture to automatically learn and extract inherent features from CSI images, which plays a vital role in classifying different behaviors. For an improved sensing performance, a DL-based framework is proposed in [141] that targets CSI-based HAR in resource-constrained edge devices. The authors adopted CNN to adaptively fuse the information in both frequency and time domains, and then they extracted the informative features from the fused data. Moreover, the point-wise grouped convolution and depth-wise separable convolution were employed to confine the scale of the proposed framework, speeding up the interference execution time and saving the detection overhead. Focusing on the same goal, the authors in [142] explored CNN and transfer learning architectures for feature extraction. They first used a stacked CNN to capture the activity-related information from the CSI recordings acquired within three environments. For a successful sensing result, a variety of features were taken into consideration, such as characteristics of non-line-of-sight (Non-LOS) links and multilevel dwellings. To minimize the training overhead, transfer learning was employed to further simplify the calibration process. Also applying CNN and transfer learning, the work in [143] investigated the generalization of the sensing model. For feature extraction, a CNN architecture was adopted to capture different information related to HAR, such as sampling rate, environment and sensor modality. For further improving the sensing performance, the authors employed transfer learning to learn features commonly shared among different datasets and to speed up the training process. In order to improve the domain adaptation, a novel recognition scheme was developed by [144], in which the combinations of Bidirectional Gated Recurrent Units (BiGRU) and CNN were employed to learn users' independent spatial-temporal features. Moreover, multiple domain discriminators were used to mitigate the distribution discrepancy, which is beneficial for performing domain

adaptation.

2.3.3 Recurrent Neural Networking-based HAR

Apart from the above DLNs, recurrent neural networking (RNN) is another typical type of DL architecture extensively applied for HAR, which is composed of the input layer, hidden layers, and the output layer. The unique feature of RNN is that it enjoys appealing features in dealing with time-sequential signals. In other words, the decision of one time instant can be used to estimate or predict that of new time instants, by utilizing the relationship contained in the time sequence [145]. Therefore, it plays a significant role in DL-based HAR, and a considerable number of works have been proposed. For example, the long-short term memory recurrent neural networking (LSTM-RNN) has been adopted in [146] for feature extraction, through which the informative features in CSI signals can be effectively learned and extracted. Note that using the raw CSI as inputs relies completely on RNN to separate them and extract feature signals for activities, and the volume of the input is also large, which makes it time-consuming to train the RNN. In [147], a DeepSense method is presented, which combines three DLNs, e.g., autoencoder, CNN module and LSTM, for activity recognition. With the employed DL architectures, the proposed scheme was able to sanitize the annoying factors in the received signals, such as the noise in raw CSI recordings. Moreover, distinguishable high-level features can be effectively captured with the help of DL networks. Although it can achieve a better performance than [146], its complexity is significantly higher. Another drawback of [147] is that it is susceptible to the phase shift in CSI caused by timing offset between the WiFi transmitter and receiver. A small mismatch in the phase domain of CSI can result in notable performance degradation. To estimate and compensate for the timing offset, a few solutions have been proposed. For example, a linear fitting method was developed in [148] and a phase calibration approach is proposed in [149], but both schemes have high computational complexity. The bi-directional long short-term memory (ABLSTM) was adopted in [150] to accomplish reliable recognitions. To process sequential CSI data, an advanced LSTM architecture was used to extract representative features from forward and

backward directions. As a result, a connection between the past and future data can be built to determine the current state of LSTM, generating representative information for feature extraction. On top of that, the authors also proposed an attention mechanism to further improve the quality of extracted information, by assigning different values of weights to activities given their importance on final HAR performance. To detect continuous motions or actions, the authors of [151] leveraged the property of LSTM to extract special features for HAR. To improve the quality of CSI, the authors proposed a feature enhancement scheme and employed RNN for feature extraction. In particular, they designed a fusion layer to fuse the raw received radar signals, so as to avoid losing information of performing data presentation. After that, they employed LSTM to extract temporal features from the raw radar data, and then conducted HAR using the extracted information.

2.3.4 Other DL-based HAR

Apart from SAE, CNN and RNN networks, other DL architectures, e.g., deep adversarial network, deep reinforcement learning (DRL), matching network, and transfer learning, are also widely applied for CSI-based HAR [152]. For instance, a HAR solution was proposed in [153] to facilitate subject-independent HAR leveraging the property of the adversarial network, through which the subject-dependent information in the received CSI data can be removed. At the same time, the activity-related features can be extracted, contributing to an improved sensing result. With the same target, the work in [154] proposed a cross-scenario HAR with the help of deep adversarial networks. To that end, the authors developed a maximum-minimum adversarial scheme to bridge the connection between the source features to the target features. Moreover, a strategy of center alignment was designed to further improve the feature quality in the source domain, which is beneficial for performance improvement. Focusing on improving sensing accuracies, the authors in [155] presented a DRL-based HAR. To that end, two challenges need to be jointly addressed, e.g., the feature patterns of the received signals and the mapping of the received data to the detection results. Given that, they formulated a joint optimization problem of the above two issues, and elaborated a DRL-based

scheme to obtain the optimal solution by minimizing the cross-entropy loss of the recognition results. For better sensing accuracies, a HAR solution was proposed in [156] leveraging the property of metric learning and matching network. To extract good feature representation, the authors adopted a matching network to map the received signal to a high dimensional space. Then, they built an effective relationship between the testing dataset and the previously given dataset adopting the idea of metric learning. As a result, more informative features could be extracted and learned, significantly improving the detection accuracy. To realize the location-independent HAR, recent work in [157] presented a HAR framework using transfer learning and CNN. For feature extraction, the data distribution of the received signals and its characteristics were analyzed. Then, transfer learning and CNN were adopted to effectively learn discriminative information, by transferring the features from source environments to the testing environment. Consequently, sensing performance can be improved with the extracted features.

2.4 Challenges for DL-based Human Activity Recognition

Despite the effectiveness in improving recognition performance, existing DL-based methods still suffer from several inherent drawbacks, as summarized below.

- First, the feature design and selection, which are treated as input signals of DL networks, is a challenging issue for the current DL-based HAR schemes [158, 159]. The type and quality of selected features directly influence the output of DL architectures, resulting in a great impact on final recognition performance. However, the selection of input signals for DL networks heavily relies on the designers' experience, and more research effort is required to address this issue.

- The second challenging problem for DL-based HAR schemes is that most of the existing DL-based schemes require a large number of training examples from the particular environments (e.g., previously seen environments or the

testing/new environment) to train the corresponding DL networks [160–162]. Consequently, the performance is dependant upon the number of training samples, which becomes particularly problematic when large amounts of required training samples from environments are not accessible;

- Third, the problem is further exacerbated by the fact that the current DL-based recognition solutions are highly specific to environments where the HAR model is trained [163, 164]. As a result, the recognition accuracy usually drops dramatically if using the classifier trained with primitive features in sources/previously seen environments to recognize activities in new/unseen environments. In other words, well-trained schemes in the above works cannot be directly used for HAR in unseen environments. This may severely restrict the applicabilities of DL-based methods in practice [165].

To solve the above problems, plenty of effort have been made from a variety of perspectives. In particular, to address the first challenge, many works devoted their effort to design various signal processing methods for improving the quality of input signals [166, 167]. For example, recent work in [168] proposed an activity segmentation method to improve the quality of input data for DL networks. For that purpose, the authors first discretized the continuous CSI sequences into discrete bins, and then classified them into four states. As a result, the starting point and ending point of each activity can be identified using these state labels, enhancing the input signals of DL networks. With the same target, the authors of [169] proposed to utilize the property of different angle of arrival to mitigate the impact of the background environment. Then, they adopted the PCA algorithm to eliminate the noise and reduce the dimension of the input signal. As a result, the input signal of DL networks is enhanced with more useful information and a smaller dimension. Although some existing schemes can improve the quality of input signal to a certain extent, more research attempts are required for further performance improvement.

To deal with the second and third challenges, recent DL-based HAR schemes have attempted various advanced signal processing techniques and the corresponding learning methodologies to reduce the number of required training samples in order

to improve the recognition performance in an environment-invariant fashion [170]. To name a few, a recent solution in [171] exploited transfer learning to realize environment-robust recognition. The authors in [172] 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. Although these models can facilitate reliable sensing results, they require 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 [173] does not need multiple PSEs for training, while it still requires several hundreds of samples from the testing environment to perform network refinement. Apart from the above, in [174] the authors developed a cross-environment recognition model by extracting environment-independent features.

Although environment-independent recognition has been achieved to a certain extent, the above methods have some limits. To be specific, their recognition accuracies rely heavily on the number of different source environments (i.e., PSEs). 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 (e.g., [171, 172]). It is also challenging for them to extract high-quality and informative features across different environments due to the limitation of feature extraction processes and deep learning architectures. Additionally, some works do not require multiple PSEs or many samples from the testing environment (e.g., [174]); they concentrated on recognizing intensive (i.e., highly dynamic) behaviors only, so it is difficult to effectively identify the light activities (e.g., standing and laying)

One-shot learning can be considered as a promising candidate to help address the above challenges. It has been successfully applied in many vision-based activity recognition and object classification problems [175–177], and this makes it a plausible technique to solve CSI-based 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 [178, 179]. In other words, only one sample is enough to learn/extract discriminating features about the environment by bridging the gap between this environment and PSEs, no matter how different they may be. To the best of our knowledge, most of the one-shot learning approaches have been focusing on vision-based scenarios in which video signals are analyzed for recognition. For CSI-based HAR, very little has been investigated so far. In particular, a shortcoming of one-shot learning is that although it only needs one sample from the testing environment, it still requires a large number of samples from a wide variety of PSEs. This may not be accessible under many CSI-based HAR settings, as obtaining samples from diverse environments is usually expensive or impractical.

2.5 Summary

In this chapter, we reviewed and studied the state-of-the-art works in HAR. First, we introduced the evolution of HAR including three types of conventional HAR schemes and their associated features. Second, we presented recent advances of device-free HAR using wireless signals together with their advantages and disadvantages. Third, the discussions on applying DL networks for CSI-based HAR were provided. Finally, we analyzed the associated challenges of DL-based HAR.

Chapter 3

DL-based Human Activity Recognition with Sufficient Training Samples

CSI-based human activity recognition (CSI-based HAR) is attracting significant research interest. Compared with conventional device-based sensing techniques (e.g., wearable devices), CSI-based HAR does not require the targeted users to wear any devices. Moreover, CSI-based HAR is able to complete the sensing task without exposing the user's private information, such as the face of a user. In this chapter, we investigate the DL-based HAR in a scenario with sufficient training samples and try to improve the sensing performance. To achieve that, two novel HAR schemes are proposed and analyzed in this chapter.

3.1 Introduction

Recently, applying DL to CSI-based HAR has received strong research interest. With a sufficient number of required training samples, DL-based HAR methods are able to automatically extract deep features from input signals, which is critical for realizing reliable recognitions [132,139]. In this field, a variety of research attempts have been dedicated to improve recognition accuracy. For instance, the authors in [146] used RNN to extract hidden features from the raw CSI. Consequently, representative

characteristics in CSI measurements were captured and extracted effectively, resulting in accurate detections. Targeting the same goal, the authors in [147] adopted an Autoencoder Long-term Recurrent Convolutional Network framework (AE-LRCN) to extract high-level representative features in CSI. Moreover, the impact of noise involved in the CSI sequences could also be sanitized, which was beneficial for performance improvement. However, the performance of these methods can still be improved in both sensing accuracy and training complexity. Moreover, they are susceptible to phase shift in CSI caused by timing offset between the WiFi transmitter and receiver. A small mismatch in the phase domain of CSI can result in notable performance degradation.

Given the above challenges, in this chapter, we propose two novel DL-based HAR schemes by leveraging the properties of deep learning architectures and advanced signal processing methods. The main contributions of this chapter are summarized as follows:

- We propose a HAR scheme using Deep Learning Networks with enhanced CSI (DLN-eCSI), which can achieve significantly improved sensing performance with reduced training complexity. We develop a CSI feature enhancement scheme (CFES) for cleaning and compressing signals input to the DL networks. CFES includes two modules: background reduction and correlation feature enhancement. In background reduction, we propose two methods for removing activity-unrelated information from CSI. In correlation feature enhancement, we compute correlation signals over all subcarriers and streams to improve the reliability of feature signals, as well as compressing the signals. We then use LSTM-RNN to extract the deeper features from the enhanced correlation signals. Therefore, the proposed DLN-eCSI is capable of achieving accurate detection results with less training complexity.

- To further improve the detection performance, we propose a novel scheme for CSI-based HAR using activity filter-based deep learning network (HAR-AF-DLN) with enhanced correlation features. Our scheme can effectively solve the phase mismatch problem caused by timing offset and significantly improve the

identification accuracy for similar activities. Two major innovations in HAR-AF-DLN include CSI compensation and enhancement (CCE) and activity filter (AF). The CCE method is proposed to compensate for the timing offset between the WiFi transmitter and receiver, so as to improve the quality of CSI. Besides, CCE can also enhance activity-related signals and reduce the dimension of signals input to DL networks, thereby increasing recognition accuracy with less complexity. The AF method is designed to distinguish similar activities (e.g., lying and standing) using the enhanced CSI correlation features obtained from CCE.

- We design and perform numerous experiments to verify the performance of DLN-eCSI and HAR-AF-DLN with regard to both sensing accuracies and training time. Extensive experimental results demonstrate that our proposed DLN-eCSI and HAR-AF-DLN schemes are superior to state-of-art HAR schemes, with less training time and higher recognition accuracy.

The remainder of this chapter is structured as follows. In Section 3.2, the system model of WiFi based HAR is provided. In Section 3.3 and Section 3.4, we propose two novel HAR schemes, respectively. The experiment and performance evaluation are discussed in Section 3.5, followed by the conclusions of this chapter in Section 3.6.

3.2 System Model of CSI-based Human Activity Recognition

In this section, we introduce the system architecture for CSI-based HAR, as shown in Fig. 3.1. There are three main stages: data collection, data preprocessing, and activity classification.

In the stage of data collection, from wireless communication links between the transmitters and receivers, the physical layer information (e.g., CSI), which reflects the variations of the wireless environment caused by human activities, is collected for processing at the receivers. As shown in Fig. 3.2, when a person performs certain

Figure 3.1: Main processes of human activity recognition system.

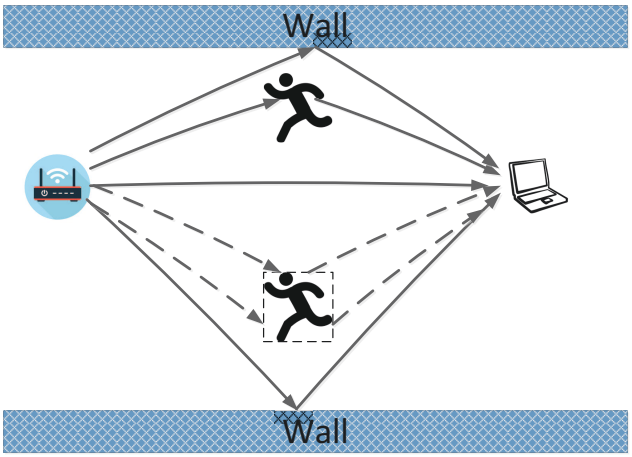


Figure 3.2: Influence of human activity on signal propagation.

activities such as running in an indoor environment covered by a wireless network, the wireless signal propagation is distorted with changes in both the number of multipath signals and their amplitudes and phases. These distortions will cause the variation of CSI. The receivers are typically WiFi access points in our considered setup. In this paper, we use the widely adopted Intel 5300 802.11n network interface card (NIC) for CSI acquisition [25]. The CSI at 30 subcarriers from all three antennas is used. More details are provided in Section 3.5.1.

In the stage of data preprocessing, activity-unrelated information is removed and unique features are extracted from CSI for detecting human activities. In this stage, we use the proposed different algorithms to filter out activity-unrelated components in the raw CSI and obtain enhanced CSI with information ideally solely related to human activities. Moreover, unique correlation features from the enhanced CSI at the OFDM subcarrier level are also extracted. In this stage, all the subcarriers are employed to calculate the correlation feature matrix which will be used for identifying distinctive feature patterns of different activities.

In the last stage of deeper feature extraction and classification, the deeper features are automatically extracted from the output in the previous stage. To

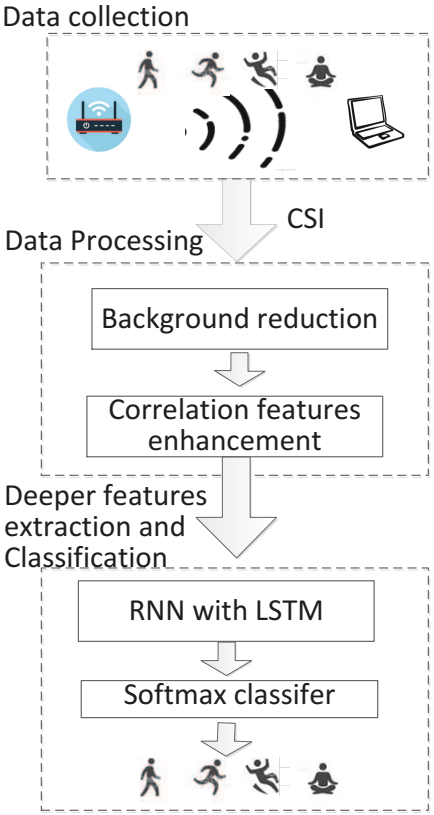


Figure 3.3: A sketch of the human activity recognition system based on WiFi 802.11n.

that end, a deep learning network (e.g., LSTM-RNN) is employed as the features extractor, and the softmax regression algorithm [136] is adopted as the classifier. The deep learning network can be trained during the training process using the training data offline. Then during the online sensing phase, the proposed scheme classifies the human activities using the trained network coefficients.

3.3 The DLN-eCSI Scheme

In this section, we present the process of the proposed DLN-eCSI in details. The structure of proposed scheme is shown in Fig. 3.3.

3.3.1 Data collection

Let N_t and N_r stand for the number of antennas at the transmitter and receiver, respectively. Thus, there are $N = N_t \times N_r$ streams (links) contained in a CSI packet,

which can be expressed as

$$\mathbf{h}(i)=[h_{1,1}(i),\ldots,h_{1,m}(i),\ldots,h_{n,m}(i),\ldots,h_{N,M}(i)]^T, \tag{3.1}$$

where $\mathbf{h}(i)$ represents the CSI vector obtained at time i, n is the indicator of the nth stream, M denotes the total number of available subcarriers in each stream, m denotes the mth subcarrier in the stream, and T stands for the transposition operation. Then the CSI matrix within a time period (e.g., $i = 1, 2, \ldots, I$) is adopted to sense the human activities, and is given by

$$\mathbf{H} = [\mathbf{h}(1), \ldots, \mathbf{h}(i) \ldots, \mathbf{h}(I)]. \tag{3.2}$$

3.3.2 Data Preprocessing

Note that the CSI matrix $\mathbf{H}$ in (3.2) is only the raw information, which is not suitable for HAR directly for the following reasons. First, $\mathbf{H}$ contains too much activity-unrelated information, which will degrade the quality of extracted features. Second, applying $\mathbf{H}$ directly to detect human activities (such as in [146]) is time-consuming and increases the system overhead due to the large size of $\mathbf{H}$. We use CFES, including background reduction and correlation feature enhancement modules, to overcome these problems.

Background reduction

The core task in the background reduction module is to filter out the activity-unrelated information while retaining the activity-related information. Although human activities may cause the distortion of some multipath signals, there could generally be more multipath largely unchanged. Thus $\mathbf{h}(i)$ can be divided into two parts: dynamic CSI and static CSI, which can be expressed as

$$\mathbf{h}(i) = \mathbf{h}^{st}(i) + \mathbf{h}^{dy}(i), \tag{3.3}$$

where $\mathbf{h}^{st}(i)$ stands for the static CSI vector which is activity-unrelated; and $\mathbf{h}^{dy}(i)$ represents the dynamic CSI vector that is related to human activities. Notably, $\mathbf{h}^{st}(i)$ is the dominating component in $\mathbf{h}(i)$, and has much larger impact on $\mathbf{h}(i)$ than

$\mathbf{h}^{dy}(i)$. The reason is that the influence of human behavior on the whole environment is generally limited, which is especially true when a person performs some minor actions, e.g., raising hands, sitting, standing, etc. Under such a situation, applying $\mathbf{h}(i)$ directly to HAR could degrade the recognition accuracy (refer to Fig. 3.13). Therefore, it is worthwhile to subtract the static information $\mathbf{h}^{st}(i)$ from $\mathbf{h}(i)$. To that end, we propose two methods for mitigating $\mathbf{h}^{st}(i)$ from $\mathbf{h}(i)$.

We call the first one as the *local mean (LM)* method which estimates $\mathbf{h}^{st}(i)$ over a period of time using the exponentially weighted moving average (EWMA) approach [180]. The estimated value of $\mathbf{h}^{st}(i)$ is

$$\widehat{\mathbf{h}}^{st}(i) = \delta\mathbf{h}(i) + (1-\delta)\widehat{\mathbf{h}}^{st}(i-1), \tag{3.4}$$

where $\widehat{\mathbf{h}}^{st}(i)$ denotes the recursive static CSI estimation at the i-th packet, and δ is the forgetting factor and is set with small value such as 0.01. The initial value of $\widehat{\mathbf{h}}^{st}(1)$ is set as $\mathbf{h}(1)$. Then the estimated dynamic CSI, $\widehat{\mathbf{h}}^{dy}(i)$, is equal to

$$\widehat{\mathbf{h}}^{dy}_{LM}(i) = \mathbf{h}(i) - \widehat{\mathbf{h}}^{st}(i). \tag{3.5}$$

Thus, the whole estimated dynamic CSI matrix is

$$\widehat{\mathbf{H}}^{dy}_{LM} = [\widehat{\mathbf{h}}^{dy}_{LM}(1), \ldots, \widehat{\mathbf{h}}^{dy}_{LM}(i), \ldots, \widehat{\mathbf{h}}^{dy}_{LM}(I)], \tag{3.6}$$

We call the second one as the *differential method (DM)*, which extracts dynamic CSI from $\mathbf{h}(i)$.The DM method computes the difference of $\mathbf{h}$ between two time slot and is simpler to implement. The estimated dynamic CSI vector using DM is expressed as

$$\widehat{\mathbf{h}}^{dy}_{DM}(i) = \mathbf{h}(i) - \mathbf{h}(i-1), \tag{3.7}$$

where $\widehat{\mathbf{h}}^{dy}_{DM}(i)$ denotes the estimated dynamic CSI, and $\widehat{\mathbf{h}}^{dy}_{DM}(1) = \mathbf{0}$.

From (3.5) and (3.7), we can see that both LM and DM can extract the activity-related information contained in CSI. LM is capable of providing a high level of accuracy at the expense of relatively high complexity. The forgetting factor δ affects the estimation performance, and its optimization is yet to be investigated. DM's computation complexity is lower, but with sacrificed estimation performance. The impacts of LM and DM on the performance of human activity recognition will be provided in more details in Section 3.5.

Correlation Feature Enhancement

Let $\widehat{\mathbf{H}}^{dy}$ denote the estimated dynamic CSI matrix hereafter, unless stated otherwise. The size of $\widehat{\mathbf{H}}^{dy}$ is $MN \times I$ and will lead to high computational complexity if directly used for training and running in DL architectures. Here, we propose a novel method for significantly reducing the dimension of the input to DL networks. This method is capable of extracting distinctive features from $\widehat{\mathbf{H}}^{dy}$ and hence improving the recognition performance as well.

Different from existing methods which only leverage the correlation features of multiple subcarriers within one stream, e.g., [181], we compute the correlation between signals at all subcarriers from all streams, given by

$$\mathbf{C}_D = \widehat{\mathbf{H}}^{dy} \times (\widehat{\mathbf{H}}^{dy})^T, \tag{3.8}$$

where $\mathbf{C}_D$ denotes the $MN \times MN$ correlation matrix. Such correlation information compresses the signals more effectively and provides more reliable features for behavior recognition. More specifically, the number of correlation features is significantly decreased from $MN \times I$ (i.e., the size of $\mathbf{H}^{dy}$) to $MN \times MN$.

3.3.3 Deeper Feature Extraction and Classification

In this section, we first extract the deeper features from $\mathbf{C}_D$ using LSTM-RNN, then classify human activities based on these extracted deeper features using the softmax regression algorithm. This process is demonstrated in Fig. 3.4.

The signal $\mathbf{C}_D$ contains compressed discriminative patterns for different human activities. We now feed it into LSTM-RNN to extract deeper features as shown in Fig. 3.4. The LSTM-RNN has the capability of extracting the deeper features of input data automatically. These extracted deeper features are then used to recognize human activities, through the softmax classifier. Note that the LSTM is able to distinguish similar behaviors, which can improve the recognition performance by distinguishing similar activities, such as "standing" and "stand up". Moreover, compared to other neural networks, RNN with LSTM can achieve much more reliable HAR performance, as it fits for processing continuous signals such as human activities. Specifically, when a person performs a set of different activities continuously,

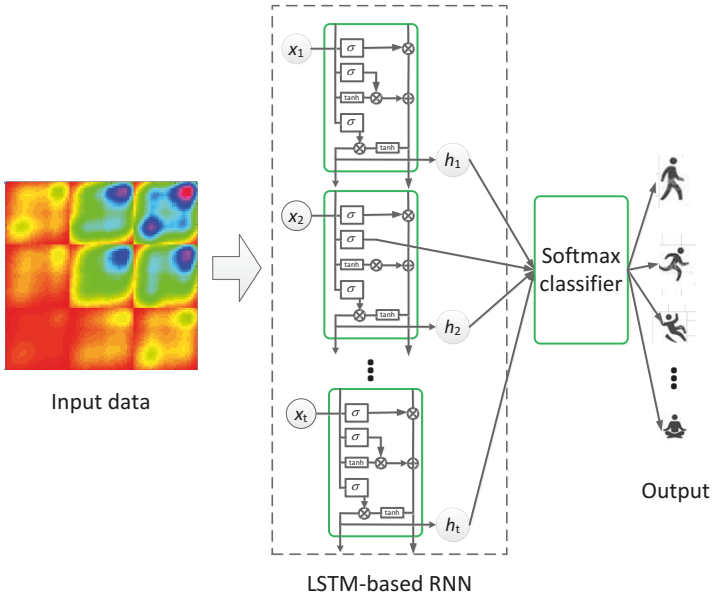


Figure 3.4: Structure of deeper features extraction and classification

these behaviors are not independent but belong to a circle of continuous states. RNN with LSTM can extract effective information from those states, contributing to high recognition accuracy.

Conventional RNN-based sensing methods usually have a time-consuming training process, due to the large quantities of training data. By using overall correlation matrices with dramatically reduced volume of the input data, our DLN-eCSI scheme achieves significantly reduced training overhead.

3.4 The HAR-AF-DLN Scheme

We illustrate the diagram of the proposed HAR-AF-DLN scheme in Fig. 3.5, including three main modules: CSI Collection, CSI Preprocessing and Activity Recognition. In section 3.4.1 and Section 3.4.2, we provide the details of the last two modules.

3.4.1 CCE for CSI Preprocessing

In this section, we provide details of the designed CCE for CSI preprocessing. We will first discuss the first two steps, i.e., *timing offset compensation* and *activity-*

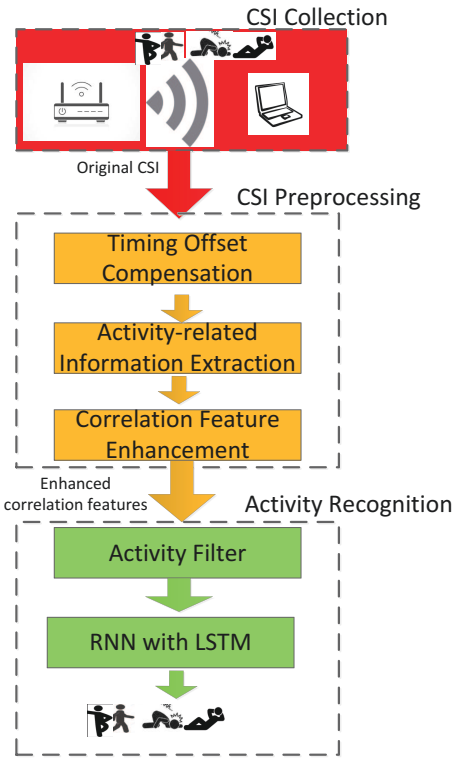


Figure 3.5: Main processing modules of the HAR-AF-DLN Scheme.

related information extraction. Then we will describe the last step: *correlation feature extraction*.

Timing Offset Compensation and Activity-related Information Extraction

As aforementioned, these two steps aim to calibrate the CSI phase, reduce activity-unrelated information while retaining activity-related information. Consequently, it is possible to extract feature signals which are more activity-related and less environment-dependent. To achieve that, $\mathbf{h}(i)$ in (3.2) is partitioned into two parts: dynamic CSI $\mathbf{h}^{dy}(i)$ and static CSI $\mathbf{h}^{st}(i)$. $\mathbf{h}^{dy}(i)$, which is more related to human activities, can be obtained by applying LM algorithm proposed in Section 3.3.2. However, one major problem here needs to be first resolved: the timing offsets between the WiFi transmitters and receivers are not clock-wise synchronized, which can vary over packets and cause linear phase shift of CSI. Therefore, before applying the recursive operation, estimation and compensation for the timing offset are required.

Let the recursive static CSI estimation at the i-th packet be $\widehat{\mathbf{h}}^{st}(i)$. The recursive operation over consecutive packets can be written as

$$\widehat{\mathbf{h}}^{st}(i) = \delta(\widehat{\mathbf{\Phi}}^{*}(i) \otimes \mathbf{I}_N)\mathbf{h}(i) + (1-\delta)\widehat{\mathbf{h}}^{st}(i-1), \tag{3.9}$$

where δ is the forgetting factor and is set with small value such as 0.01 empirically, the superscript $*$ represents conjugate of a matrix/vector, $\mathbf{I}_N$ stands for an $N \times N$ identity matrix, $\otimes$ denotes the Kronecker product, $\widehat{\mathbf{\Phi}}(i) = \text{diag}\{\exp(j\widehat{\phi}_{m,i})\}$ depicts a diagonal matrix with the m-th diagonal 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. Note that a common local clock is typically used for signals from/to all antennas, therefore, the timing offset due to clock offset is the same for all antennas. The phase shift $\phi_{m,i}$ can be represented as $\phi_{m,i} = m\psi_i + \theta_i$, where ψ_i and θ_i stand for phase shifts related to the timing offset.

To estimate ψ_i and θ_i, we first compute the dot product $\odot$ between $\mathbf{h}(i)$ and $(\widehat{\mathbf{h}}^{st}(i-1))^{*}$, by

$$\begin{aligned}
\mathbf{r}(i) \triangleq & \quad \mathbf{h}(i) \odot (\widehat{\mathbf{h}}^{st}(i-1))^{*} \\
= & \quad (\mathbf{h}^{st}(i) + \mathbf{h}^{dy}(i)) \odot (\widehat{\mathbf{h}}^{st}(i-1))^{*} \\
\approx & \quad \mathbf{h}^{st}(i) \odot (\widehat{\mathbf{h}}^{st}(i-1))^{*} \\
\approx & \quad (\mathbf{\Phi}(i) \otimes \mathbf{I}_N)|\widehat{\mathbf{h}}^{st}(i-1)|^2,
\end{aligned} \tag{3.10}$$

where $|\widehat{\mathbf{h}}^{st}(i-1)|^2$ means element-wise square of the absolute value. In (3.10), the first approximation is obtained based on the fact that the power of static paths are typically much more significant 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.

Then, $\mathbf{r}(i)$ is stacked into an $M \times N$ array, and each column contains CSI for one antenna. The mean over each row is computed, getting a new $M \times 1$ vector $\bar{\mathbf{r}}(i)$. Next, we compute the cross-correlation for neighbouring elements with equal spaced subcarrier indices in $\bar{\mathbf{r}}(i)$ and then compute the mean of the output, obtaining a sample denoted by γ_i. Then the estimate for ψ_i is given by

$$\widehat{\psi}_i = \angle(\gamma_i)/K_s, \tag{3.11}$$

where K_s represents the index intervals between the selected subcarriers which are equally spaced. In this paper, we use the Intel NIC 5300 card in the experiments, $K_s = 2$. Let $\bar{r}_{m,i}$ denote the m-th element in $\bar{\mathbf{r}}(i)$, then we estimate the parameter θ_i by $\widehat{\theta}_i = \angle\left(\sum_m \bar{r}_{m,i} e^{-jm\hat{\psi}_i}\right)$, where the sum operation is conducted over a selected number of samples with larger energy for mitigating the noise.

Note that the proposed timing offset compensation approach in the thesis is superior to the existing related methods. First, our approach estimates timing offset in signals for compensation and does not change the original data structure. Thus, the key information contained in original signals can be remained, such as amplitude or phase features, contributing to reliable recognition results. Second, the proposed compensation method in the thesis can be directly applied to other HAR methods for compensating timing offset, which is difficult for the existing methods to achieve.

With the above process, the estimate $\widehat{\mathbf{\Phi}}(i)$ and the recursive output $\widehat{\mathbf{h}}^{st}(i)$ can be obtained, respectively. Notably, we obtain the initial value of $\widehat{\mathbf{h}}^{st}(1)$ using (3.9) in a quiet environment.

As a result, the estimated value of dynamic CSI $\widehat{\mathbf{h}}^{dy}(i)$, obtained at the ith packet, can be expressed as

$$\widehat{\mathbf{h}}^{dy}(i) = (\widehat{\mathbf{\Phi}}^*(i) \otimes \boldsymbol{I}_N)\mathbf{h}(i) - \widehat{\mathbf{h}}^{st}(i). \tag{3.12}$$

The whole estimated dynamic CSI matrix $\widehat{\mathbf{H}}^{dy}$, over I packets, is represented as

$$\widehat{\mathbf{H}}^{dy} = [\widehat{\mathbf{h}}^{dy}(1), \ldots, \widehat{\mathbf{h}}^{dy}(i), \ldots, \widehat{\mathbf{h}}^{dy}(I)]. \tag{3.13}$$

Note that, the information contained in $\widehat{\mathbf{H}}^{dy}$ is mostly activity-related, hence, it can be utilized to extract more distinctive features that are less environment-dependent for HAR.

Correlation Feature Extraction

It is important to note that a person's activity can be divided into different stages. We can use a feature signal to represent each stage, and different stages are mutually dependent. Take the activity "stands up" as an example: a series of stages are involved during this process, from static, standing with accelerating, standing up

with decelerating to standing still. The features of different stages, e.g., speed and spatial positions, are different but mutually correlated. Notably, all the feature signals for such an activity are hidden in $\widehat{\mathbf{H}}^{dy}$. Besides, there are also correlations between $\widehat{\mathbf{H}}^{dy}$ in different subcarriers, which provides additional information for HAR.

We can hence compute the correlation between signals at all subcarriers from all wireless links, given by

$$\mathbf{D}^{dy} = \widehat{\mathbf{H}}^{dy} \times (\widehat{\mathbf{H}}^{dy})^T, \tag{3.14}$$

where $\mathbf{D}^{dy}$ represents the correlation feature matrix. Note that, the size of $\mathbf{D}^{dy}$ $(MN \times MN)$ is significantly smaller than the size of $\widehat{\mathbf{H}}^{dy}$ $(MN \times I)$. Consequently, inputting $\mathbf{D}^{dy}$, instead of $\widehat{\mathbf{H}}^{dy}$, into DL network for training process can considerably reduce the computational complexity.

3.4.2 AF-DLN based Human Activity Recognition

In this section, we present details of the proposed AF-DLN method, as depicted in Fig. 3.6. The method includes two main steps: activity filter (AF), and deeper feature extraction and classification. Note that, the first step divides similar activities into the same group, which allows DLNs (in the second step) to focus on the feature extraction of similar motions. Consequently, more distinctive characteristics of similar movements can be extracted, compared with obtaining features from all the activities, which is beneficial to classify these similar behaviors. The details of performance assessment for AF-DLN is provided in Fig. 3.14.

Step 1: Activity filter (AF)

According to the intensity and range of motions, human activities can be divided into two main groups: light activity and intensive activities. The former group refers to the activities with low intensity and small movement range, including *lying, standing, empty room, sitting*, which cause less CSI variation. The latter group refers to the activities with high intensity and large movement ranges, including, e.g., *walk, fall, running*, which cause larger CSI variation. The key task of AF is to determine which group the input signals $\mathbf{D}^{dy}$ belongs to (i.e., "light activity" or "intensive

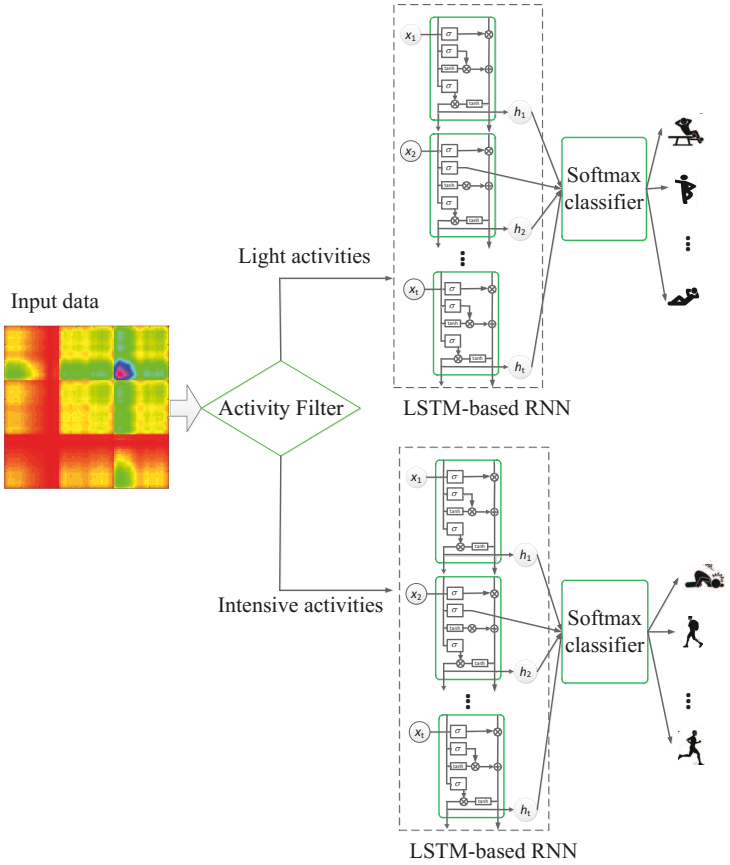


Figure 3.6: Structure of AF-DLN based activity recognition using CFM as input.

activity"). To achieve that, we apply a singular value decomposition (SVD) to $\mathbf{D}^{dy}$ and obtain its singular values, by

$$\Lambda = \text{svd}(\mathbf{D}^{dy}), \tag{3.15}$$

where svd(.) stands for the SVD operation, and $\Lambda \triangleq [\lambda_1, \lambda_2, \ldots, \lambda_{MN}]$ represents the vector containing singular values of $\mathbf{D}^{dy}$ in the descending order. Since the first two singular values (i.e., λ_1 and λ_2) contain most environment-dependent information [182], we adopt λ_3 as the metric for dividing human activities into two groups. To be specific, if λ_3 is smaller than a threshold β that is obtained empirically, the signal $\mathbf{D}^{dy}$ is divided into the "light activity" group, otherwise the "intensive activity" group.

Step 2: Deeper Feature Extraction and Classification

In this step, for each group (i.e., "light activity" or "intensive activity"), we train one DL architecture to distinguish human activities. For each DL architecture, we first apply RNN with LSTM to automatically learn and extract the hidden features from $\mathbf{D}^{dy}$. We then utilize the softmax regression algorithm for classification using the extracted deeper features from $\mathbf{D}^{dy}$. The process of the proposed AF-DLN is illustrated in Fig. 3.6.

It is noteworthy that conventional RNN-based sensing methods generally have time-consuming training processes due to the large size of training data. In contrast, our proposed HAR-AF-DLN scheme can effectively complete the training process with significantly less training overhead by using $\mathbf{D}^{dy}$ with largely reduced size of input data.

3.5 Implementation and Evaluation

In this section, we present the experimental results for evaluating the performance of the proposed DLN-eCSI scheme and HAR-AF-DLN Scheme.

3.5.1 Experimental Setup

To implement our proposed DLN-eCSI and HAR-AF-DLN, two computers equipped with Intel WiFi 5300 NIC are adopted as the transmitter and receiver, respectively. The transmitter continuously sends its packets with its single antenna ($N_t = 1$) at 5.32 GHz frequency band, which follows the protocol of IEEE 802.11n. The receiver, which uses the CSI tools [25], collects and stores CSI with three antennas ($N_r = 3$) for 30 subcarriers ($M = 30$). Five persons perform six activities in total. The each target person performs activities randomly in the experimental environment. We demonstrate the average HAR performance by considering different locations. Since the rate of samples is 1KHz, the CSI matrix ($\mathbf{H}$) has size of 90×1000. We use a 3.4GHz PC with Nvidia P5000 graphic card (16GB memory) to train the presented DLN-eCSI and HAR-AF-DLN. The total number of training iterations is 2000. We

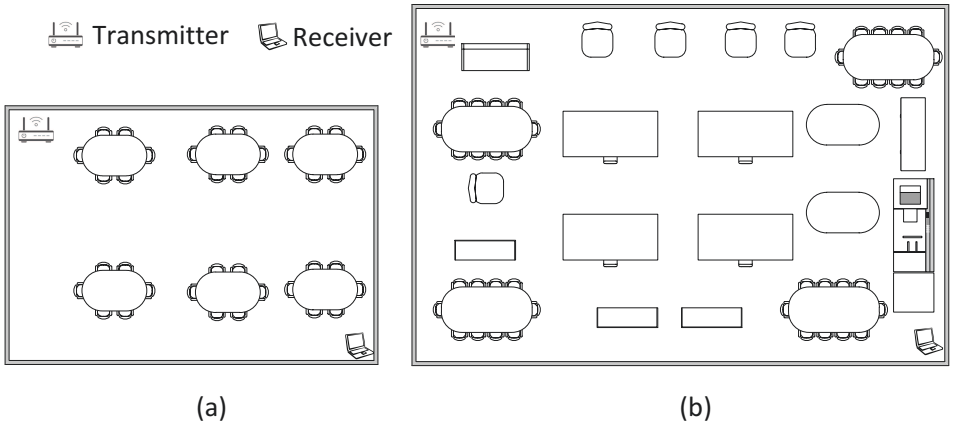


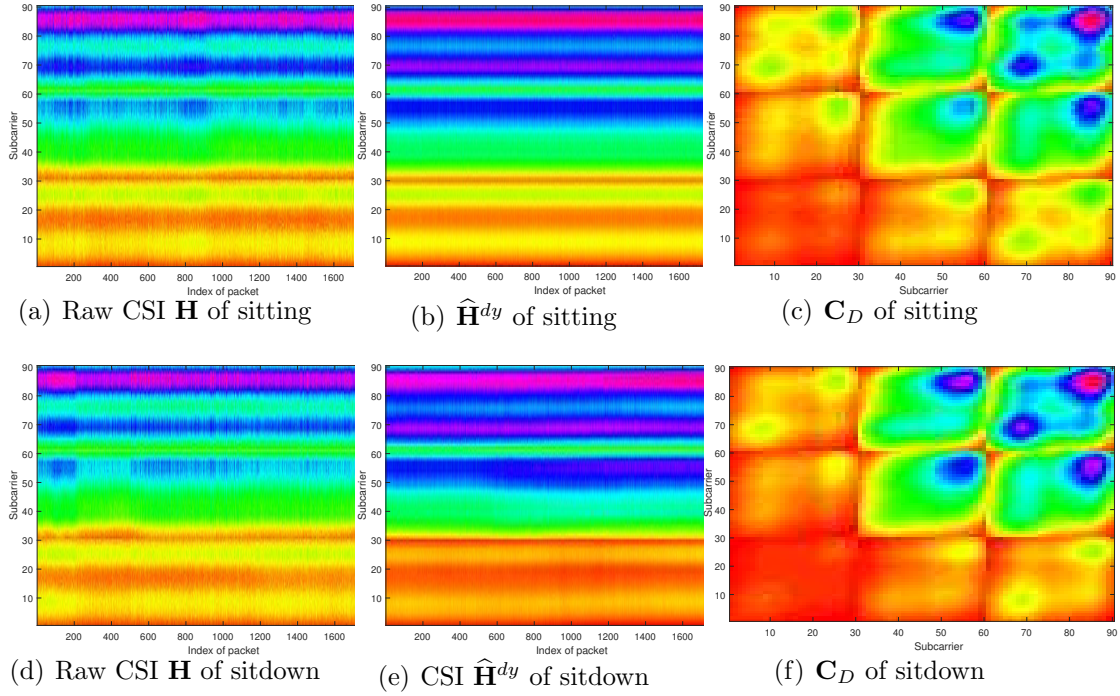
Figure 3.7: Layout of two indoor experimental areas:(a) $4m \times 6m$ meeting room. (b) $8m \times 10m$ laboratory.

use LSTM-RNN with three hidden layers, and the hidden units for each layer are 200. We set the batch size and learning rate as 64 and 0.001, respectively. The above hyper-parameters are obtained by using grid search method. We empirically set the value of the threshold β of AF method to 0.6.

We conduct the experiments of the designed DLN-eCSI and HAR-AF-DLN in two indoor configurations with different environmental complexities. Fig. 3.7 illustrates the layout of each indoor configurations. The first, with several obstacles between the transmitter and receiver, is a 4m$\times$ 6m meeting room. The second, with many obstacles between the transmitter and receiver, is a 8m$\times$ 10m laboratory room. Both the training and testing data sets of each indoor configuration include six activities, with 300 times for each activity.

3.5.2 Performance Evaluation

In this section, we evaluate the performance of our proposed DLN-eCSI and HAR-AF-DLN by comparing them with other state-of-the-art methods. Various parameters and methods are used to provide a comprehensive comparison. Firstly, the recognition performance of proposed DLN-eCSI and HAR-AF-DLN are discussed, respectively. Then the comparison of these two schemes is provided.



(a) Raw CSI $\mathbf{H}$ of sitting (b) $\widehat{\mathbf{H}}^{dy}$ of sitting (c) $\mathbf{C}_D$ of sitting

(d) Raw CSI $\mathbf{H}$ of sitdown (e) CSI $\widehat{\mathbf{H}}^{dy}$ of sitdown (f) $\mathbf{C}_D$ of sitdown

Figure 3.8: Performance of the proposed CSI feature enhancement scheme, when comparing two similar activities"sitting" and "sit down".

Table 3.1: Sensing performance of different methods in the two indoor configurations

Methods	1st Exp.	2nd Exp.
RNN [146]	87.5%	78.4%
Proposed DLN-eCSI-L	94.2%	90.7%
Proposed DLN-eCSI-D	91.2%	88.9%

Performance of DLN-eCSI

Fig. 3.8 compares the performance for two similar activities (i.e., sitting and sit down), using the original CSI matrix $\mathbf{H}$, the estimated dynamic CSI $\widehat{\mathbf{H}}^{dy}$ and the correlation feature matrix $\mathbf{C}_D$, respectively. As can be seen from Figs. 3.8(a) and 3.8(d), it is challenging to differentiate between "sitting" and "sit down" based on $\mathbf{H}$. However, they can be readily separated by using $\mathbf{C}_D$ as is clear from Figs. 3.8(c) and 3.8(f), because $\mathbf{C}_D$ significantly enhance the differences between these two activities. Moreover, the size of $\mathbf{C}_D$ (i.e., 90×90) is much smaller than $\mathbf{H}$ (i.e., 90×1000). Therefore, the training complexity is notably reduced in LSTM-RNN.

Predicted activity

Actual activity	Stand up	Lying	Walk	Standing	Fall	Empty
Stand up	0.959	0.01	0.005	0.026	0	0
Lying	0	0.912	0	0.01	0.014	0.064
Walk	0.01	0	0.964	0.006	0.02	0
Standing	0.036	0.012	0.001	0.941	0	0.01
Fall	0.01	0.012	0.061	0	0.917	0
Empty	0	0.031	0	0.014	0.006	0.949

(a) Proposed DLN-eCSI-L

Predicted activity

Actual activity	Stand up	Lying	Walk	Standing	Fall	Empty
Stand up	0.914	0.008	0.032	0.036	0.01	0
Lying	0	0.877	0.001	0.045	0.004	0.073
Walk	0.021	0.002	0.925	0.027	0.025	0
Standing	0.054	0.039	0	0.896	0	0.011
Fall	0.008	0.036	0.042	0	0.914	0
Empty	0.004	0.034	0	0.016	0	0.946

(b) Proposed DLN-eCSI-D

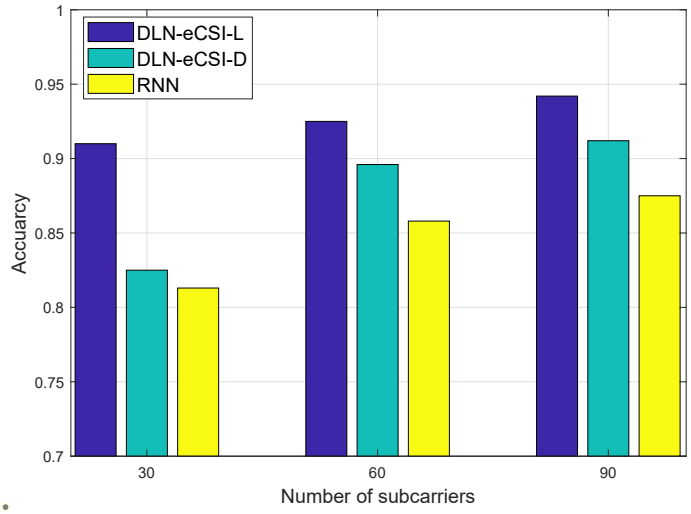
Predicted activity

Actual activity	Stand up	Lying	Walk	Standing	Fall	Empty
Stand up	0.897	0.034	0.017	0.009	0.043	0
Lying	0.015	0.92	0.007	0.058	0	0
Walk	0.032	0.004	0.931	0.027	0.006	0
Standing	0.011	0.148	0.027	0.791	0.014	0.009
Fall	0.008	0	0.053	0	0.939	0
Empty	0	0	0.019	0.207	0	0.774

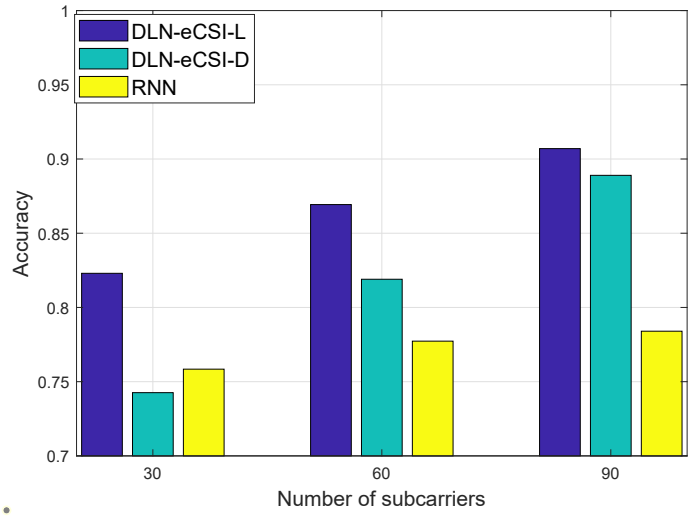
(c) RNN [146]

Figure 3.9: Confusion matrix for different human activity recognition methods

Table 3.1 shows the average sensing accuracy for LM and DM for six different activities. DLN-eCSI-L and DLN-eCSI-D are used to denote the cases when DLN-eCSI adopts LM and DM to extract the dynamic CSI, respectively. DLN-eCSI-L clearly outperforms the other two methods in both indoor configurations. Specifically, in the first configuration, the sensing accuracy of DLN-eCSI-L and DLN-eCSI-D are 94.2% and 91.2%, respectively. However, the sensing accuracy of RNN [146] is only 87.5%. In addition, for each method, the sensing accuracy in the first configuration is better than that in the second one, because the environment in the first configuration is simpler and hence it is easier for human activity recognition.



(a) The first experimental configuration



(b) The second experimental configuration

Figure 3.10: Impact of the number of subcarriers on the sensing accuracy

To further analyze the performance of different methods, we provide the confusion matrix for different activities in the first experimental configuration in Fig 3.9. DLN-eCSI-L achieves much higher sensing accuracy than the RNN method [146] for all the six activities, so does DLN-eCSI-D. For instance, the overall sensing accuracy of DLN-eCSI-L and DLN-eCSI-D for each activity is all above 0.912 and 0.877, respectively, but the accuracy for RNN is only 0.774.

Fig. 3.10 illustrates the impact of the number of subcarriers on the sensing accuracy of various methods in two configurations. Obviously, in both config-

Table 3.2: Training time for different methods

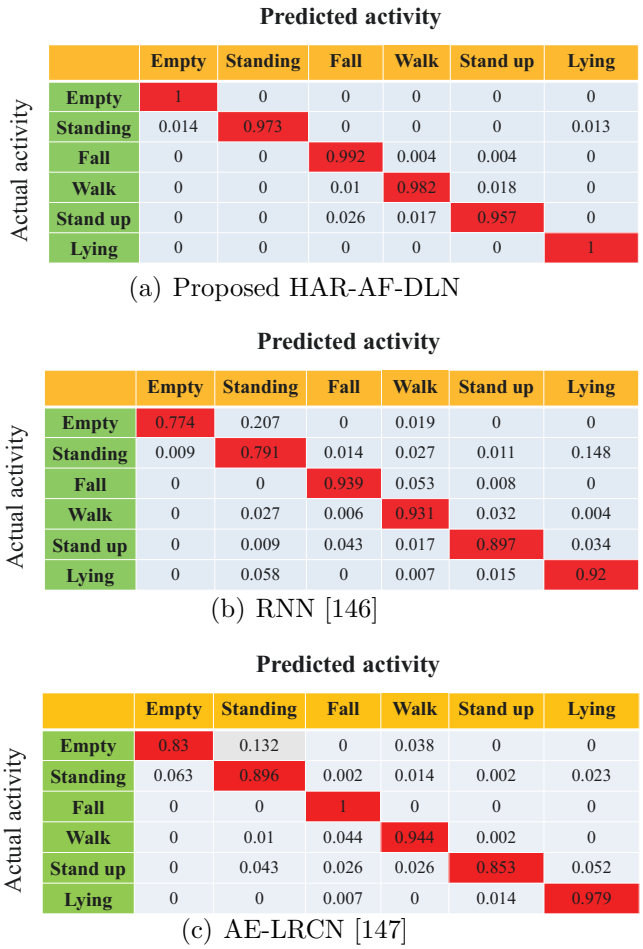
Methods \ Hidden units	300	500
RNN [146]	3822.8s	7837.1s
Proposed DLN-eCSI-L	577.3s	1631.4s
Proposed DLN-eCSI-D	572.7s	1603.6s

Table 3.3: Average Sensing Accuracy of the three methods in the two indoor configurations

Methods	1st Exp.	2nd Exp.
Proposed HAR-AF-DLN	98.4%	93.4%
RNN [146]	87.5%	78.4%
AE-LRCN [147]	92.4%	89.7%

urations, DLN-eCSI-L achieves the best sensing performance among these three methods increasing the number of subcarriers. For instance, when the number of subcarriers is 60, the sensing accuracy for DLN-eCSI-L and DLN-eCSI-D in the second configuration are 0.869 and 0.819, respectively. By contrast, the sensing accuracy for RNN in [146] is only 0.777. Notably, when the number of subcarriers is less than 60, the sensing accuracy of DLN-eCSI-D is lower than that of RNN in [146], while the training time in LSTM-RNN of DLN-eCSI-D is much smaller than RNN (refer to Table 3.2).

Table 3.2 compares the training time of LSTM-RNN for various methods with different Hidden units. We utilize a 3.4GHz PC with Nvidia P5000 graphic card (16GB memory) to train the LSTM-RNN. The number of training iteration is 2000 and the training data set contains 1200 samples. It is clear that our proposed methods are superior to the RNN method in [146]. Specifically, when the number of hidden units is 500, the training time of LSTM-RNN for our proposed two methods are less than about one-quarter of that for RNN in [146].



(a) Proposed HAR-AF-DLN

(b) RNN [146]

(c) AE-LRCN [147]

Figure 3.11: Confusion matrix for different human activity recognition methods.

Performance of HAR-AF-DLN

In this section, we first compare the performance of our proposed HAR-AF-DLN scheme with other state-of-the-art methods (i.e., RNN [146] and AE-LRCN [147]), taking various parameters and configurations into consideration. We then provide in-depth evaluations of the effect of CCE and AF-DLN on our proposed scheme, respectively.

Table 3.3 illustrates the average recognition accuracy of three methods for the six activities with different configurations and parameters. As can be seen, the proposed HAR-AF-DLN clearly outperforms the other two methods in both indoor configurations. Take the second configurations as an instance, our proposed HAR-AF-DLN can achieve the average accuracy at 93.4%. In contrast, the corresponding sensing accuracies for the other three methods (i.e., RNN [146] and AE-LRCN [147])

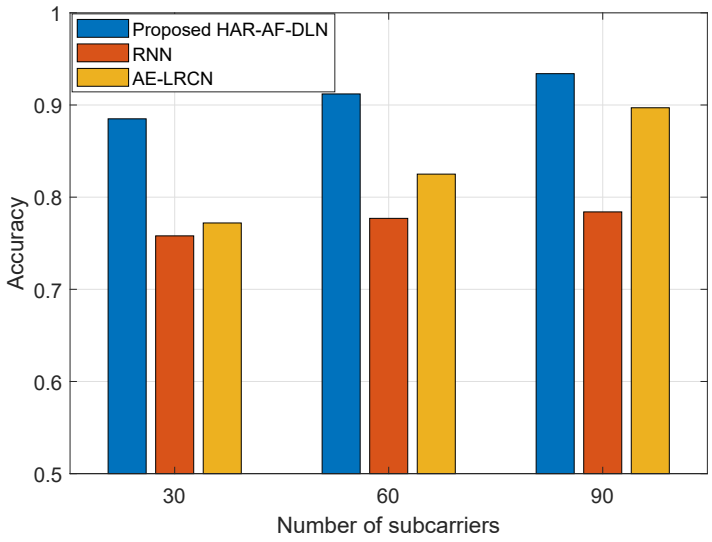


Figure 3.12: Impact of the number of subcarriers on the recognition accuracy.

Table 3.4: Training time for different methods

Methods \ Hidden units	300	500
Proposed HAR-AF-DLN	632.2s	1753.1s
RNN [146]	3822.8s	7837.1s
AE-LRCN [147]	5591.4s	8956.1s

are lower than 90%.

To examine the performance of each method in detail, we present the confusion matrix for six activities in the first configuration in Fig. 3.11. The proposed HAR-AF-DLN performs much better than the other two methods in identifying these activities, particularly in differentiating similar activities such as lying and standing.

Fig. 3.12 demonstrates the impact of the number of subcarriers on average sensing accuracy. The six activities are performed in the second experimental configuration. Clearly, with an increasing number of subcarriers, each sensing method can achieve better average recognition accuracy. In all the cases with different numbers of subcarriers, HAR-AF-DLN achieves higher sensing accuracy than the others.

We provide Table 3.4 to compare the training time for the three methods. The DLNs are trained using a 3.4GHz workstation with Nvidia P5000 graphic card

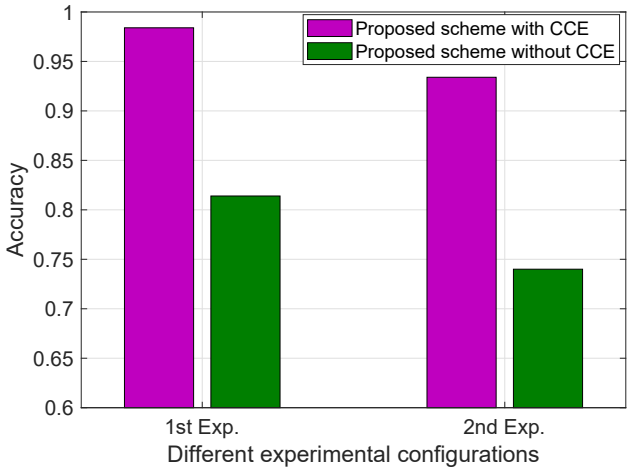


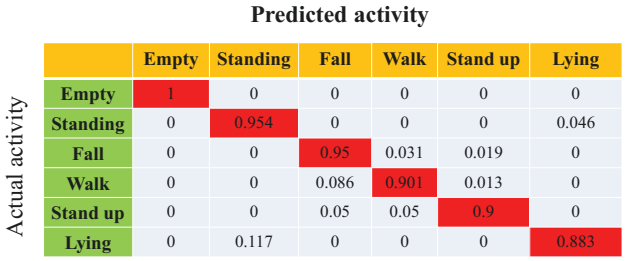
Figure 3.13: Impact of CCE on recognition accuracy.

(16GB memory). The numbers of training samples and iterations are 1200 and 2000, respectively. Our proposed HAR-AF-DLN scheme is shown to have much less training time. This largely credits to the notably reduced dimension of the input matrix $\mathbf{D}^{dy}$.

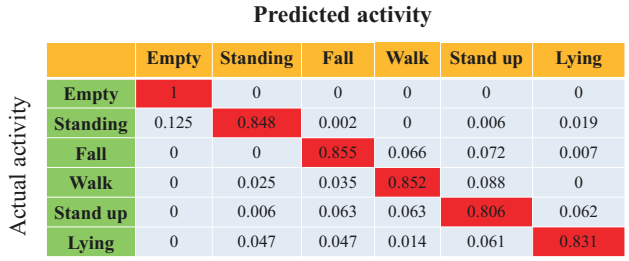
We present Fig. 3.13 to show the impact of the proposed CCE method on sensing performance in both experimental configurations. The average sensing accuracy of the proposed method with CCE is significantly higher than that of the method without CCE. This is because the CCE method can compensate for the timing offset, reduce activity-unrelated information and enhance activity-related information, which is beneficial to extract proper features for HAR.

The impact of the proposed AF method on recognition performance is shown in Fig. 3.14, under the second experimental configuration. From this figure, the sensing accuracy of the proposed method with AF in recognizing similar behaviors (e.g., standing and lying) is notably higher than that without AF. This is because the AF method can effectively distinguish similar activities, improving recognition performance.

In table 3.5, we compare the performance of the proposed DLN-eCSI and the proposed HAR-AF-DLN from various perspectives. Specifically, as observed from this table, the proposed HAR-AF-DLN outperforms the proposed DLN-eCSI in sensing accuracies. This is because, the proposed HAR-AF-DLN utilizes phase features and improve their quality by compensating for timing offset. Moreover,



(a) Proposed method with AF-DLN



(b) Proposed method without AF-DLN

Figure 3.14: Impact of AF-DLN on the recognition accuracy.

Table 3.5: Comparison for the Proposed Methods

Methods	Accuracy	Training Time	Hardware Consumption
Proposed DLN-eCSI	Fair	Short	Low
Proposed HAR-AF-DLN	High	Fair	High

the proposed AF method helps to identify similar activities, contributing to better sensing results. However, compared to the developed DLN-eCSI, the proposed HAR-AF-DLN requires a longer time and more hardware resources to train the model well. In a word, the user can make a selection of these two schemes based on their requirements on sensing accuracies, training time and hardware resource.

3.6 Conclusion

In this paper, we developed two novel HAR schemes for device-free human activity recognition, i.e., DLN-eCSI and HAR-AF-DLN. We first proposed a DLN-eCSI with the key innovative CFES method for data preprocessing. The CFES method enhances activity-related signals via using a recursive algorithm and condenses the

enhanced signals via computing the correlation across segmented signals over time and frequency domains. The CFES scheme is hence an efficient pre-processing tool for device-free WiFi sensing, and can be potentially used with many other sensing schemes. To further improve the sensing accuracy, we developed a HAR-AF-DLN scheme for human activity recognition, which consists of novel CCE and AF methods. The CCE method can compensate for the timing offset, enhance the activity-related signals and reduce the dimension of input signals to DL networks. The AF method is able to distinguish similar activities based on the enhanced CSI correlation features achieved from CCE. Through extensive experimental results, we validate that the proposed DLN-eCSI and HAR-AF-DLN schemes are superior to state-of-the-art methods in terms of recognition accuracy and training complexity.

Chapter 4

DL-based Human Activity Recognition with Limited Training Samples

In the previous chapter, we investigated the DL-based HAR in a scenario with a large number of required training samples. However, those two proposed schemes are deficient for the scenario in which the training samples are limited. In this chapter, to address the above concern, we propose a novel HAR scheme drawing support from the DL networks and innovative signal processing techniques. The proposed HAR scheme aims to facilitate a successful HAR using limited training samples, e.g., the dataset from one previously seen environment (PSE) and, at the minimum, one sample for each activity from the testing environment.

4.1 Introduction

Numerous DL-based approaches have been developed to facilitate HAR in the scenario with a limited number of training samples. Recent work in [171] leveraged the property of transfer learning to accomplish environment-robust recognition. Another work in [172] proposed to employ adversarial learning to realize environment-independent recognition. Although these two methods can identify human behaviors, they need many PSEs for training. The HAR model designed in [173] can

accomplish HAR and does not require multiple PSEs for training. However, a large number of training samples from the testing environment (e.g., several hundreds of samples for each activity) is still needed for performance refinement. 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.

To address the challenging issues aforementioned, we propose a novel HAR scheme to realize a reliable HAR using the dataset from one PSE and, at the minimum, one sample for each activity from the testing environment. The major contributions of this chapter are as follows.

- We propose a HAR scheme using Matching Network with enhanced CSI (MatNet-eCSI) to successfully perform one-shot 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 dimensions 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.

We organize the remainder of this chapter as follows. In Section 4.2, we provide an overview of the proposed MatNet-eCSI scheme. The details of the designed CCFE is described in Section 4.3. The process of detecting activities with the developed HAR scheme is presented in Section 4.4. Section 4.5 shows the performance evaluation of the proposed MatNet-eCSI scheme. Section 4.6 summarizes conclusions of this chapter.

4.2 The MatNet-eCSI Scheme

The diagram of the proposed MatNet-eCSI scheme is shown in Fig. 4.1, 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 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 4.3 and Section 4.4, respectively.

4.2.1 CSI Collection and Preprocessing

For the part of CSI collection, we adopt the Intel 5300 NIC, as stated in Section 3.2, to acquire and collect CSI.

The CSI vector $\mathbf{h}(i)$, acquired from the i-th packet, is written as $\mathbf{h}(i) = [H_{1,1}(i), \ldots, H_{1,m}(i), \ldots, H_{n,m}(i), \ldots, H_{N,M}(i)]^T$, 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 $\mathbf{H}$,

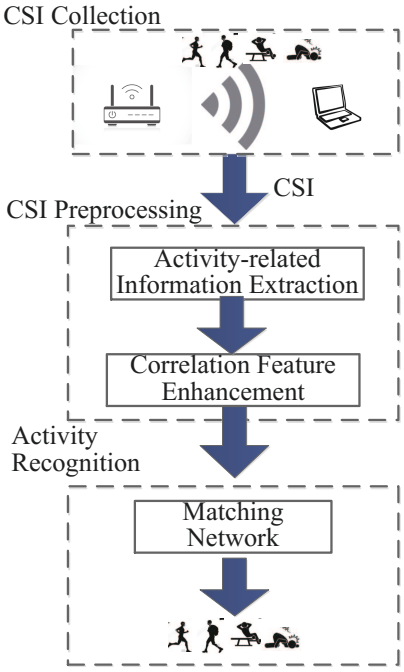


Figure 4.1: Main processing modules of the MatNet-eCSI Scheme.

made up of CSI vectors obtained from I packets, is

$$\mathbf{H} = [\mathbf{h}(1), \ldots, \mathbf{h}(i) \ldots, \mathbf{h}(I)]. \tag{4.1}$$

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 the original CSI matrix $\mathbf{H}$.

4.2.2 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.

4.3 CCFE for CSI Preprocessing

In this section, we present a 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.

One of the core tasks 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 (4.1) into two parts: dynamic CSI and static CSI, given by

$$\mathbf{h}(i) = \mathbf{h}^{st}(i) + \mathbf{h}^{dy}(i), \tag{4.2}$$

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. 4.12). 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 EWMA approach [180].

There is one major problem here: the timing offset between the WiFi transmitter and receiver, which are not clock-wise 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 $\widehat{\mathbf{h}}^{st}(i) = \delta(\widehat{\mathbf{\Phi}}^{*}(i) \otimes \mathbf{I}_N)\mathbf{h}(i) + (1-\delta)\widehat{\mathbf{h}}^{st}(i-1)$, where $\mathbf{I}_N$ represents an $N \times N$ identity matrix, $\otimes$ represents the Kronecker product, δ stands for the forgetting factor, the superscript $*$ denotes conjugate, $\widehat{\mathbf{\Phi}}(i) = \text{diag}\{\exp(j\widehat{\phi}_{m,i})\}$ 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

$$\phi_{m,i} = m\psi_i + \theta_i, \tag{4.3}$$

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

$$\begin{aligned} \mathbf{r}(i) \triangleq \quad & \mathbf{h}(i) \odot (\widehat{\mathbf{h}}^{st}(i-1))^{*} \\ \approx \quad & (\mathbf{\Phi}(i) \otimes \mathbf{I}_N)|\widehat{\mathbf{h}}^{st}(i-1)|^2 \end{aligned} \tag{4.4}$$

where $|\widehat{\mathbf{h}}^{st}(i-1)|^2$ denotes element-wise square of the absolute value. In (4.4), 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 $\widehat{\psi_i} = \angle(\gamma_i)/K_s$, where K_s is the index intervals between the used subcarriers that are equally spaced. K_s is set to 2 for using the Intel NIC 5300 card in the experiments .

Let $\bar{r}_{m,i}$ be the m-th element in $\bar{\mathbf{r}}(i)$. The parameter θ_i in (4.3) can then be estimated as

$$\widehat{\theta_i} = \angle \left(\sum_m \bar{r}_{m,i} e^{-jm\hat{\psi}_i} \right), \tag{4.5}$$

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

$$\widehat{\mathbf{h}}^{dy}(i) = (\widehat{\mathbf{\Phi}}^*(i) \otimes \boldsymbol{I}_N)\mathbf{h}(i) - \widehat{\mathbf{h}}^{st}(i). \tag{4.6}$$

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

$$\begin{aligned} \widehat{\mathbf{A}}^{dy} &= [\widehat{\mathbf{a}}^{dy}(1), \ldots, \widehat{\mathbf{a}}^{dy}(i), \ldots, \widehat{\mathbf{a}}^{dy}(I)], \\ \widehat{\mathbf{\Psi}}^{dy} &= [\widehat{\boldsymbol{\psi}}^{dy}(1), \ldots, \widehat{\boldsymbol{\psi}}^{dy}(i), \ldots, \widehat{\boldsymbol{\psi}}^{dy}(I)]. \end{aligned} \tag{4.7}$$

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.

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 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. 4.2.

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

$$\mathbf{U}^A = [\mathbf{U}^A(1), \mathbf{U}^A(2), \ldots, \mathbf{U}^A(k), \ldots, \mathbf{U}^A(K)], \tag{4.8}$$

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

$$\mathbf{C}^A_{i,j} = \mathrm{E}\big[(\mathbf{U}^A(i) - \mathrm{E}[\mathbf{U}^A(i)])(\mathbf{U}^A(j) - \mathrm{E}[\mathbf{U}^A(j)])^T\big], \tag{4.9}$$

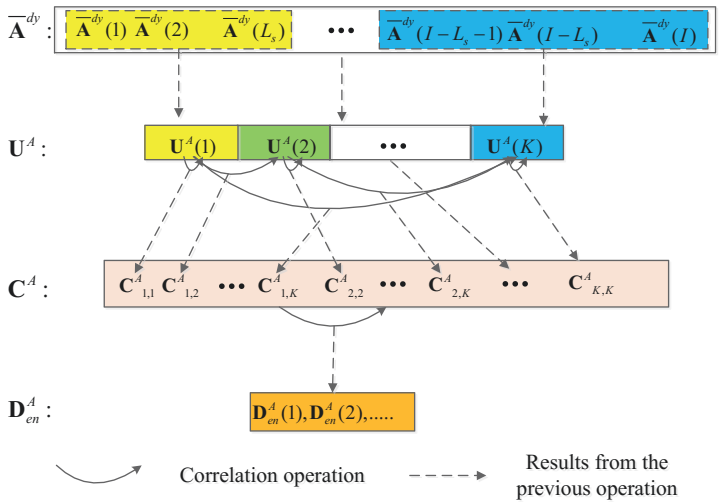


Figure 4.2: Correlation feature extraction in the proposed CCFE.

where 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

$$\mathbf{C}^A = [\mathbf{C}^A_{1,1}, \mathbf{C}^A_{1,2}, \ldots, \mathbf{C}^A_{1,K}, \mathbf{C}^A_{2,2}, \mathbf{C}^A_{2,3}, \ldots, \mathbf{C}^A_{K,K}]. \tag{4.10}$$

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

$$\mathbf{D}^A_{en} = \mathbf{C}^A \times (\mathbf{C}^A)^T, \tag{4.11}$$

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. 4.12).

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 $1 \leq k \leq K$
    $\mathbf{U}^A(k) = [\widehat{\mathbf{A}}^{dy}((k-1)L_s+1), \ldots, \widehat{\mathbf{A}}^{dy}(kL_s)]$;
    $\mathbf{U}^\Psi(k) = [\widehat{\mathbf{\Psi}}^{dy}((k-1)L_s+1), \ldots, \widehat{\mathbf{\Psi}}^{dy}(kL_s)]$;
  end
  Construct $\mathbf{U}^A$ and $\mathbf{U}^\Psi$ based on Eq. (4.8);
  For $1 \leq i \leq K$
    For $i \leq j \leq K$
      Compute $\mathbf{C}^A_{i,j}$ and $\mathbf{C}^\Psi_{i,j}$ based on Eq. (4.9);
    end
  end
  Construct $\mathbf{C}^A$ and $\mathbf{C}^\Psi$ based on Eq. (4.10);
  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

4.4 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 the DL network with data from massive different environment, which is however very costly. Although some data processing techniques, e.g., data augmentation and regularization [183, 184], 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.

4.4.1 Architecture of MatNet

The architecture of MatNet based HAR is illustrated in Fig. 4.3. 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

$$R = \{(x_i, y_i)\}_{i=1}^{N_k}. \tag{4.12}$$

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

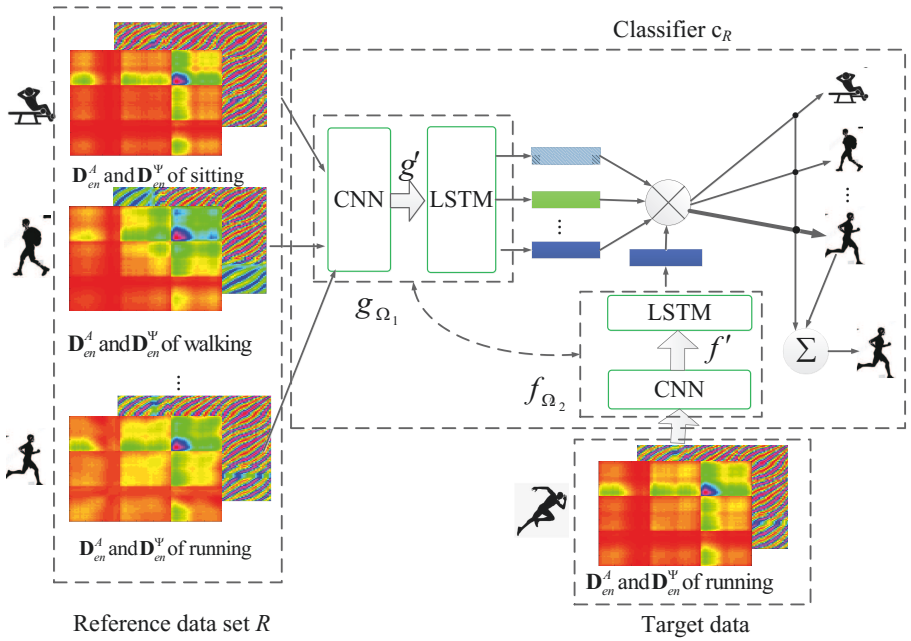


Figure 4.3: Structure of MatNet based activity recognition using CFM $\mathbf{D}^A_{en}$ and $\mathbf{D}^\Psi_{en}$ as the input.

defined as

$$P(\hat{y}|\hat{x}, R) \triangleq R \rightarrow c_R(\hat{x}), \tag{4.13}$$

where P stands for the probability distribution, which is parameterised by the CNN and LSTM (shown in Fig. 4.3). 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

$$\hat{y} = \arg\max_{y} P(y|\hat{x}, R). \tag{4.14}$$

One simple way to estimate $\hat{y}$ is calculating the linear combination of y in the reference data set R, so (4.14) is equal to

$$\hat{y} = \sum_{i=1}^{N_k} a(\hat{x}, x_i) y_i, \tag{4.15}$$

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

$$a(\hat{x}, x_i) = \frac{e^{\cos(f(\hat{x}), g(x_i))}}{\sum_{j=1}^{N_k} e^{\cos(f(\hat{x}), g(x_j))}}, \tag{4.16}$$

where $\cos(\alpha, \beta)$ is the cosine similarity function [185], defined as

$$\cos(\alpha, \beta) = \frac{\alpha \cdot \beta}{\| \alpha \| \| \beta \|}. \tag{4.17}$$

In (4.16), 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. 4.3, 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 (4.15).

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.

The structure of g is shown in Fig. 4.4, which consists of a CNN with a bidirectional LSTM [186]. 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

$$g(x_i, R) = \vec{h}_i + \overleftarrow{h}_i + g'(x_i), \tag{4.18}$$

$$\vec{h}_i, \vec{c}_i = \text{LSTM}(g'(x_i), \vec{h}_{i-1}, \vec{c}_{i-1}), \tag{4.19}$$

$$\overleftarrow{h}_i, \overleftarrow{c}_i = \text{LSTM}(g'(x_i), \overleftarrow{h}_{i+1}, \overleftarrow{c}_{i+1}), \tag{4.20}$$

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 [187]. 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.

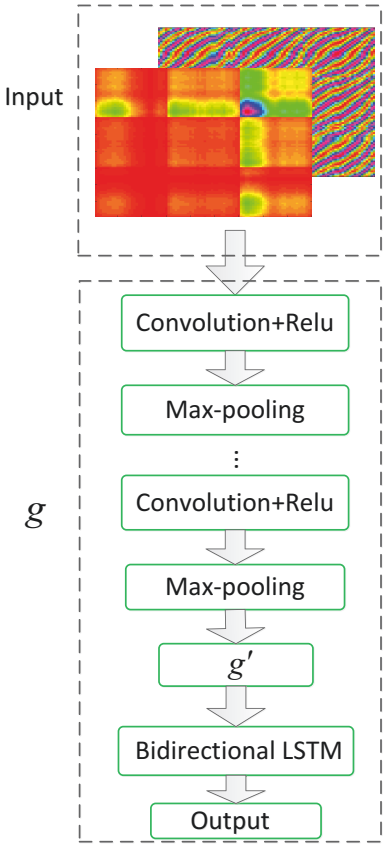


Figure 4.4: Structure of embedding function g: CNN with bidirectional LSTM.

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 [188]. 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

$$f(\hat{x}, R) = \text{attLSTM}(f'(\hat{x}), g(R), N_p), \tag{4.21}$$

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 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:

$$h_{n_p} = \hat{h}_{n_p} + f'(\hat{x}), \tag{4.22}$$

$$\hat{h}_{n_p}, c_{n_p} = \text{LSTM}(f'(\hat{x}), [h_{n_p-1}, r_{n_p-1}], c_{n_p-1}), \tag{4.23}$$

where $\text{LSTM}(f'(\hat{x}), [h_{n_p-1}, r_{n_p-1}], c_{n_p-1})$ follows the implementation described in [187]; 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

$$r_{n_p-1} = \sum_{i=1}^{N_s} a(h_{n_p-1}, g(x_i))g(x_i), \tag{4.24}$$

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

$$a(h_{n_p-1}, g(x_i)) = \text{softmax}(h_{n_p-1}^T g(x_i)). \tag{4.25}$$

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 (4.22).

4.4.2 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 [189]. However, the training process in [189] 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

$$\mathcal{L} = -\text{E}_{L\sim\mathcal{T}}\left[\text{E}_{\mathcal{R},\mathcal{B}}\left[\sum_{(x,y)\in B} \log P_\Omega\left(y|x,\mathcal{R}\right)\right]\right], \tag{4.26}$$

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

$$\Omega = \arg\min_{\Omega} \mathcal{L}(\Omega). \tag{4.27}$$

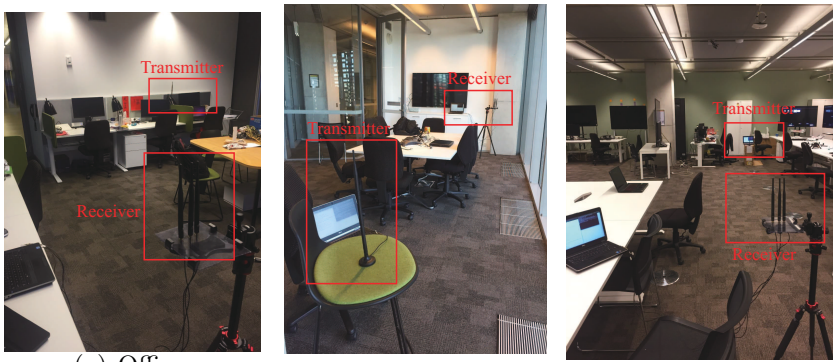
It is important to note that, for each episode, our proposed training strategy includes two key steps with different data in R and B. Specifically, in the first step, the samples in R are only from the PSE, while the samples in B are from both the testing environment and the PSE. Notably, there is no 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.

4.5 Implementation and Evaluation

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

4.5.1 Experimental Setup

To implement the proposed MatNet-eCSI, we use two computers with Intel WiFi NIC 5300 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 [25]. 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



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

Figure 4.5: Layout of three indoor experimental areas: (a) $3m \times 4m$ office. (b) $4m \times 6m$ meeting room. (c) $6m \times 7m$ laboratory.

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. 5.5. Specifically, the first configuration is a $3m \times 4m$ square area. The second one is a $4m \times 6m$ meeting room, and the 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 a 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 between different environments, via calculating the similarity of wireless environments involved in different configurations. To do this, we compute the cosine similarity function [185] 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 each target person performs activities randomly in the experimental environment. We demonstrate the average HAR performance by considering different locations. 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 [190] to normalize the inputs of each layer.

We also briefly summarize the experimental setups of methods for comparison (i.e., RNN [146], TNNAR [171], EI [172], and MatNet [189]). Specifically, the method in [146] is developed for HAR based on RNN architecture, which has four hidden layers. 200 hidden units are contained in each hidden layer. The TNNAR method [171] is developed based on transfer learning. 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. In EI method [172], 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 [189] 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.

4.5.2 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 [146], TNNAR [171], EI [172], and MatNet [189]) 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 [189]. Firstly, our proposed MatNet-eCSI uses C-CFE, 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.

Performance Comparison for Different Methods

Table 4.1 $\sim$ Table 4.3 demonstrate the average recognition accuracy of the five methods for seven activities considering different configurations and parameters. The testing environments in Table 4.1 $\sim$ Table 4.3 are the first, second and third configurations, respectively. **PSE**1, **PSE**2 and **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,

Table 4.1: Average recognition accuracy of the five methods in the first indoor configurations

Method	**PSE**2		**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

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 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 4.1 $\sim$ Table 4.3. 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.

For detailed exam of the performance, we provide the confusion matrix for each

Table 4.2: Average recognition accuracy of the five methods in the second indoor configurations

Method	**PSE**1		**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

Table 4.3: 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

method for the case of one-shot learning, as illustrated in Fig. 4.6. In this figure, the activities are performed under the first experimental configuration. **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 in accordance with the corresponding actual activities. Therefore, from Table 4.1 $\sim$ Table 4.3 and Fig. 4.6, 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.

Predicted activity

Actual activity	Empty	Stand up	Sitting	Walk	Standing	Sit down	Running
Empty	0.94	0	0	0	0	0.06	0
Stand up	0	0.7	0.12	0.05	0.11	0	0.02
Sitting	0	0	1	0	0	0	0
Walk	0	0.08	0.01	0.66	0.04	0.1	0.12
Standing	0	0.01	0	0.03	0.95	0	0.01
Sit down	0.03	0.01	0.11	0	0	0.85	0
Running	0	0	0.01	0.03	0	0	0.97

(a) Proposed MatNet-eCSI

Predicted activity

Actual activity	Empty	Stand up	Sitting	Walk	Standing	Sit down	Running
Empty	0.39	0.02	0.2	0.1	0.16	0.07	0.06
Stand up	0.08	0.35	0	0.08	0.26	0.2	0.03
Sitting	0.11	0	0.35	0.01	0.05	0.14	0.34
Walk	0.04	0.11	0	0.55	0.05	0.03	0.22
Standing	0.31	0.01	0.16	0.04	0.3	0.05	0.13
Sit down	0.1	0.12	0.01	0.02	0.42	0.31	0.02
Running	0.1	0	0.03	0.21	0.09	0.02	0.55

(b) MatNet

Predicted activity

Actual activity	Empty	Stand up	Sitting	Walk	Standing	Sit down	Running
Empty	0.1	0.17	0.23	0.4	0.05	0.04	0.01
Stand up	0.07	0.18	0.1	0.42	0.06	0.04	0.13
Sitting	0.19	0.27	0.13	0.21	0.15	0.04	0.01
Walk	0	0.01	0.04	0.94	0.01	0	0
Standing	0.01	0	0.08	0.87	0.03	0	0.01
Sit down	0.13	0.31	0.21	0.22	0.08	0.05	0
Running	0.01	0.03	0.14	0.75	0.05	0.01	0.01

(c) RNN

Predicted activity

Actual activity	Empty	Stand up	Sitting	Walk	Standing	Sit down	Running
Empty	0.49	0.02	0.2	0.02	0.24	0.03	0
Stand up	0.06	0.33	0.02	0.12	0.18	0.24	0.05
Sitting	0.08	0.05	0.25	0.09	0.35	0.14	0.04
Walk	0.01	0.08	0.01	0.45	0.02	0.01	0.42
Standing	0.26	0.02	0.25	0.06	0.29	0.03	0.09
Sit down	0.08	0.14	0.11	0.03	0.30	0.33	0.01
Running	0.02	0.08	0	0.37	0.09	0.1	0.34

(d) EI

Predicted activity

Actual activity	Empty	Stand up	Sitting	Walk	Standing	Sit down	Running
Empty	0.42	0.02	0.32	0.01	0.19	0.03	0.01
Stand up	0.01	0.25	0.01	0.22	0.21	0.23	0.07
Sitting	0.38	0.05	0.22	0.05	0.23	0.06	0.01
Walk	0.01	0.09	0.02	0.41	0.01	0.1	0.36
Standing	0.35	0.08	0.21	0.03	0.3	0.02	0.01
Sit down	0.03	0.12	0.21	0.33	0.1	0.21	0
Running	0.01	0.12	0.01	0.44	0.01	0.11	0.3

(e) TNNAR

Figure 4.6: Confusion matrix for different human activity recognition methods.

Predicted activity

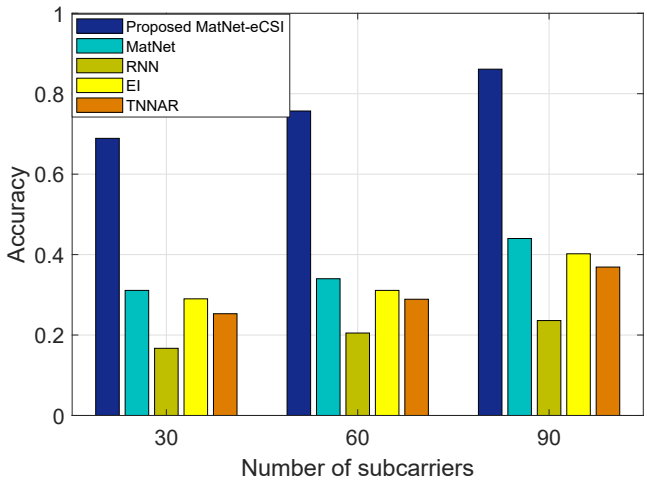
Actual activity	Empty	Stand up	Sitting	Walk	Standing	Sit down	Running
Empty	1	0	0	0	0	0	0
Stand up	0	0.87	0	0.02	0.1	0.01	0
Sitting	0	0	1	0	0	0	0
Walk	0	0.02	0	0.84	0	0.04	0.1
Standing	0.02	0.01	0.01	0	0.96	0	0
Sit down	0	0.01	0.09	0.01	0	0.89	0
Running	0	0	0	0.02	0	0	0.98

Figure 4.7: Confusion matrix of proposed MatNet-eCSI for five-shot

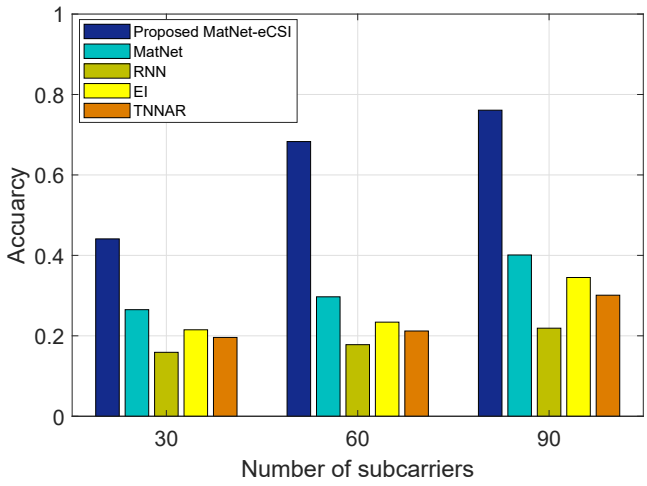
Take the activity "walk" in Fig. 4.6(a) as an instance, the probability of correctly detecting "walk" is 0.66, while the probabilities of sensing it as "running" is 0.12. This is because when the speed of running is low, its impact on CSI is similar to that of "walk". The robustness can be improved by using more samples from the testing environments (e.g., "five-shot"). To illustrate this, in Fig. 4.7 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. 4.12(b).

Fig. 4.8 demonstrates the impact of the used number of receiving antennas, represented as the number of total 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. 4.9, 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 the 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



(a) One-shot



(b) Five-shot

Figure 4.8: Impact of the used number of receiving antennas, represented as the number of total subcarriers, on the recognition accuracy.

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 4.4, 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

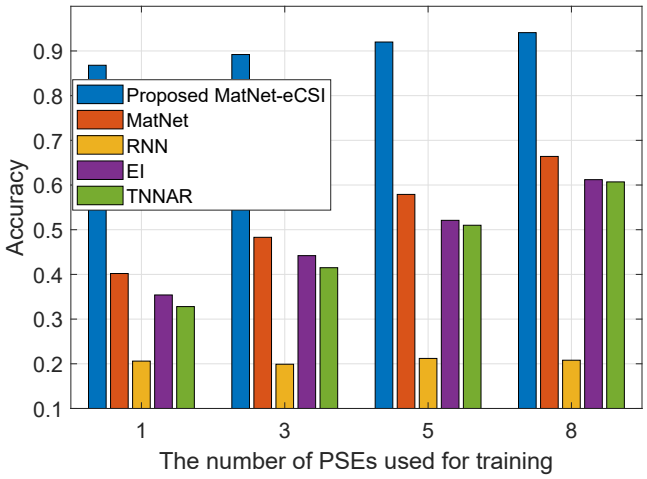


Figure 4.9: Recognition accuracy with increased number of PSEs

Table 4.4: 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

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 the sensing accuracy for different methods, as shown in Table. 4.5. 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

Table 4.5: 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

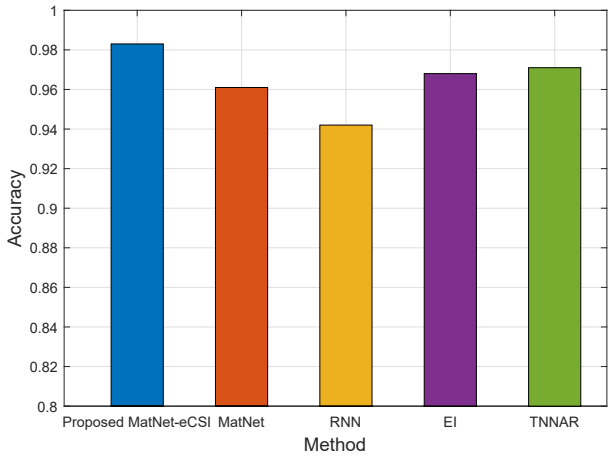
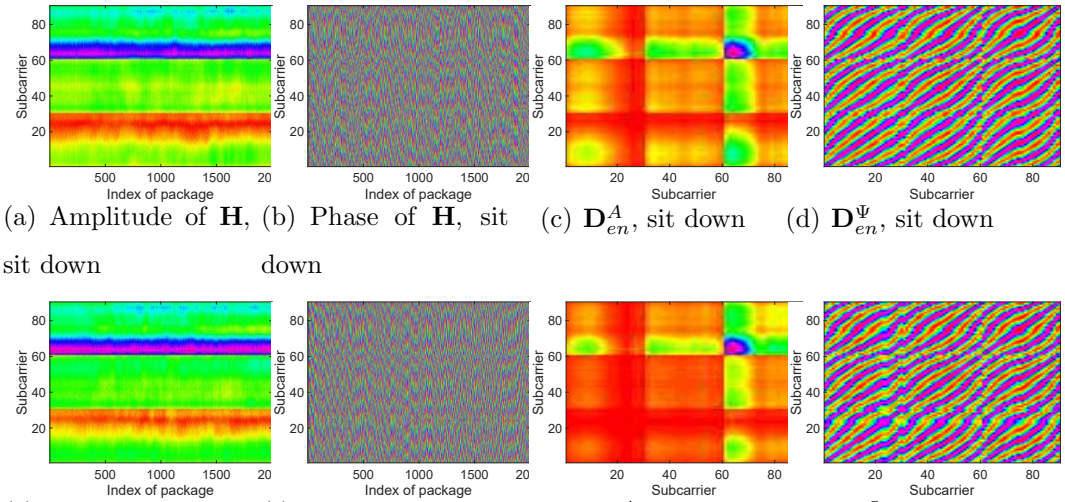


Figure 4.10: Recognition accuracy of different methods with sufficient training samples

strategy used in our proposed MatNet-eCSI is able to better utilize the properties of MatNet for feature extraction.

The sensing results of different methods with sufficient training samples are presented in Fig.4.10. 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. 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.



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

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

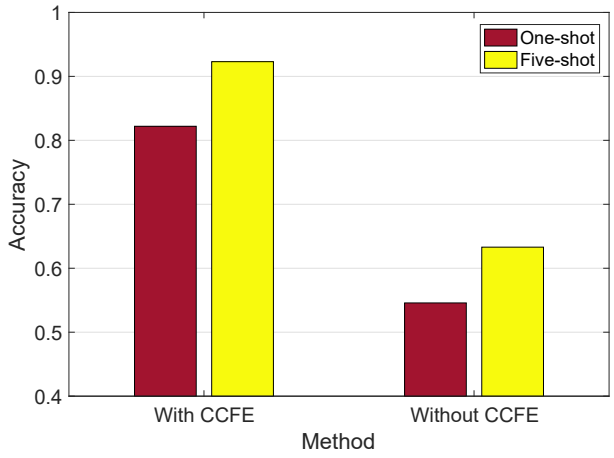
Figure 4.11: Effect of CCFE on enhancing the feature signals for two similar activities"sit down" and "sitting".

Effect of CCFE on MatNet-eCSI

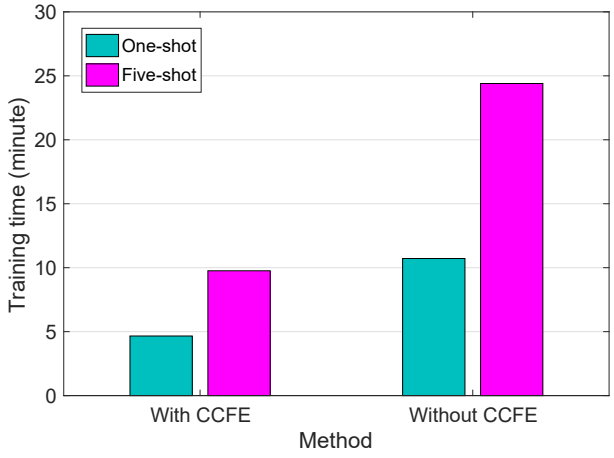
The impact of CCFE on the performance of the proposed MatNet-eCSI is investigated in this subsection, to demonstrate the importance of CCFE for the whole proposed scheme.

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. 4.11(a) and 4.11(e) (or from Fig. 4.11(b) and 4.11(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. 4.12 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 **PSE**3 is selected as PSE. As illustrated in Fig.4.12(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"



(a) Average recognition accuracy



(b) Training time

Figure 4.12: Impact of CCFE on the recognition accuracy and required training time.

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. 4.13 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 **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. 4.14. We present the results for the second experimental configuration, and PSE is **PSE**1. As can be seen from this figure, a larger K leads 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.

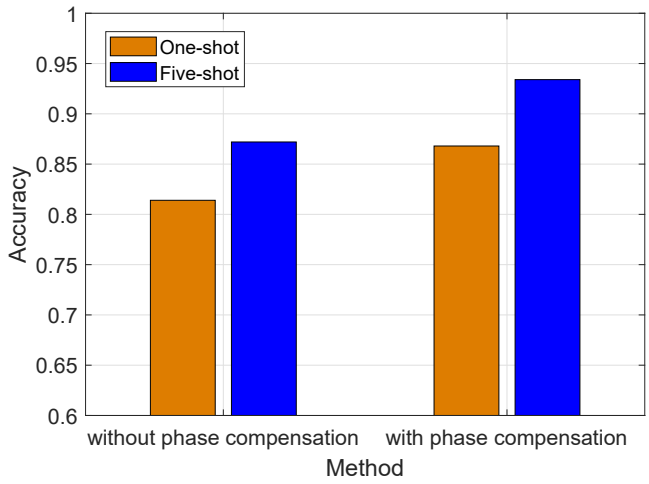


Figure 4.13: Impact of phase compensation on recognition accuracy

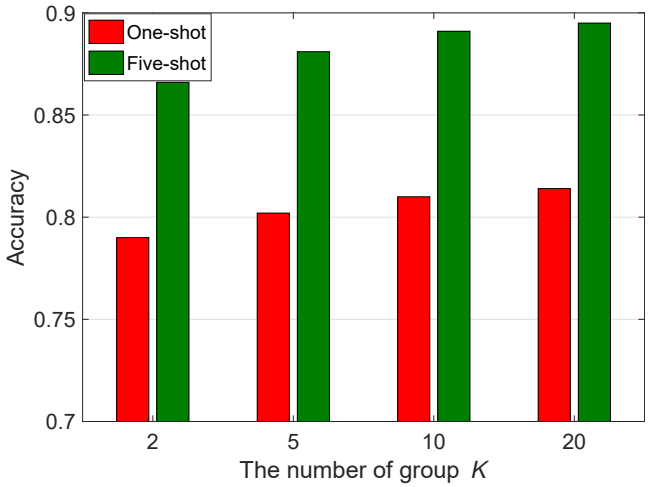


Figure 4.14: Impact of the number of segment K on the recognition accuracy.

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.

Impact of Input Signals on MatNet-eCSI

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

Table 4.6 illustrates the average recognition accuracy with different input signals in the first indoor configuration. In this table, the PSE is **PSE**2. In the previous results, MatNet-eCSI uses both the amplitude and phase of $\mathbf{H}$ as inputs. Here, we

Table 4.6: Sensing performance using different input signals in the first configuration with **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

test MatNet-eCSI-AM and MatNet-eCSI-PH, which indicate that MatNet-eCSI only adopts the amplitude or phase of **H** as the input. Table 4.6 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 **H**. It is also interesting to see that MatNet-eCSI-PH achieves better accuracy than MatNet-eCSI-AM, which suggests that the amplitude of **H** is more susceptible to the propagation environment change.

The impact of the size of data set on recognition performance is presented in Fig. 4.15. 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 **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.

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 **PSE**2.

Fig. 4.16 shows the variation in sensing accuracy of MatNet-eCSI with different human subjects. In the figure, two volunteers participate in the training process and

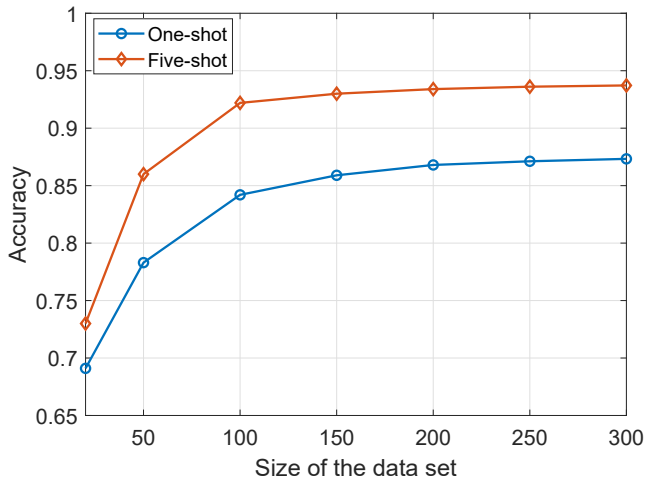


Figure 4.15: Impact of the size of data set on the recognition accuracy

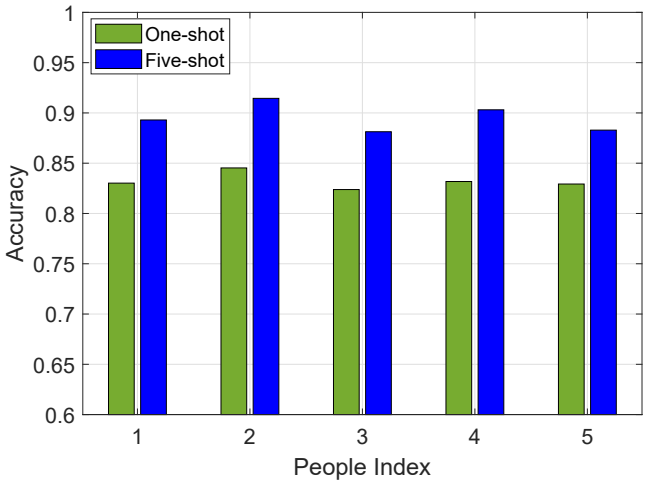


Figure 4.16: Average recognition accuracy for different people

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. 4.17 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

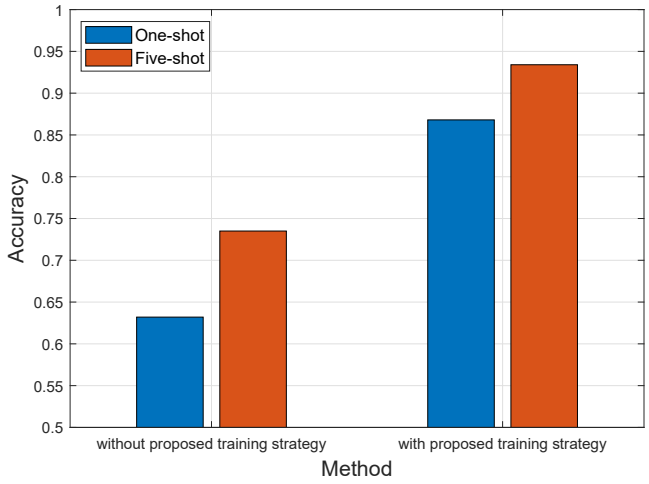


Figure 4.17: Impact of training strategy on the recognition accuracy

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

4.6 Conclusion

In this chapter, 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 novel 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.

Chapter 5

Environment-Robust WiFi-based Human Activity Recognition

In this chapter, we investigate the CSI-based HAR to achieve environmental robustness. To that end, we propose two novel HAR schemes which can remove environment-dependent but activity-unrelated data. At the same time, the activity-related information is effectively enhanced. The proposed two methods are able to achieve reliable recognition accuracy in a new test environment without retraining the networks.

5.1 Introduction

Many works have investigated the generalization ability of DL-based HAR by exploring various DLNs. To name a few, a transfer neural network is used by [171] to capture the common features shared by the testing and source environments. As a result, the transferable knowledge in time and spatial domains can be extracted, which contributes to environment-robust HAR. For the same purpose, the authors in [172] attempted to mitigate the environment-dependent data involved in the action and to learn the activity-related information by leveraging the properties of the adversarial network. Apart from the aforementioned works, the authors of [174] first proposed a method to learn environment-robust features, then they input these features into CNN and RNN for cross-environment HAR.

Although environment-independent recognition has been achieved to some extent, the above methods encounter some limits. To be specific, their recognition accuracies rely heavily on the number of different source environments, and the performance would undergo a dramatic drop if the diversity of source environments is insufficient (e.g., [171,172]). It is also challenging for them to extract high-quality and discriminative features across different environments due to the associated drawbacks of feature collection processes and deep learning architectures. Additionally, some works (e.g., [174]) concentrated on recognizing intensive (i.e., highly dynamic) behaviors only, failing to identify light activities such as standing and laying.

To address the above problems, we propose two schemes to accomplish a cross-environment HAR. We summarize the major contributions of this chapter as follows:

- We propose an environment-robust CSI-based HAR, drawing support from a matching network (MatNet) and enhanced features (HAR-MN-EF). To improve the quality of the CSI, we develop 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. 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.

- To further improve the sensing performance, we propose a scheme using the activity-related feature extraction and enhancement (AFEE) method and matching network (AFEE-MatNet). We propose AFEE to eliminate behavior-unrelated but environment-specific signals, mitigate noise, and enhance activity-related information. The proposed AFEE is composed of two steps: CSI cleaning and enhancement, and frequency domain feature extraction and signal compression. The proposed AFEE is capable of decreasing the size of feature signals, considerably reducing the training time.

- Moreover, for AFEE-MatNet, we propose a MatNet with a prediction check-

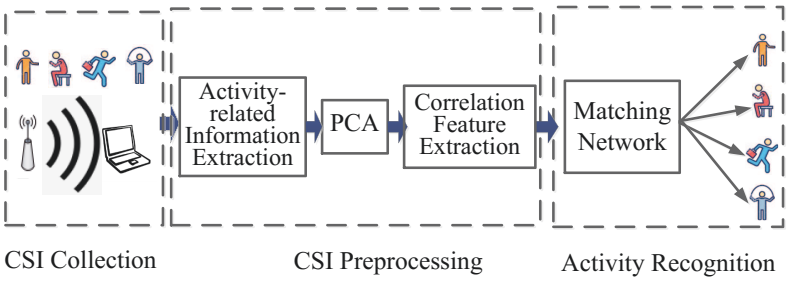


Figure 5.1: Main processing modules of the HAR-MN-EF Scheme.

ing and correction (MatNet-PCC) scheme to realize environment-independent HAR. We first employ the MatNet architecture to extract transferable features across source environments. To achieve better sensing results, we then design a prediction checking and correction (PCC) scheme to further rectify some recognition errors that do not follow the state transition of human activities.

- We design and perform numerous experiments under various conditions and scenarios. The results demonstrate that our proposed HAR-MN-EF and AFEE-MatNet gains a vast advantage over state-of-the-art HAR methods, in ameliorating recognition accuracies and reducing the training time.

The rest of this chapter is structured as follows. Section 5.2 describes the details of the first proposed HAR-MN-EF scheme. Then, we present the information of the designed AFEE-MatNet in Section 5.3. The performance evaluation is performed in Section 5.4, and the conclusions of this chapter is stated in Section 5.5.

5.2 HAR-MN-EF Scheme

The architecture of the proposed HAR-MN-EF scheme, illustrated in Fig. 5.1, includes three main modules: CSI collection, CSI preprocessing and activity recognition. The detailed introduction of the last two modules are presented in Section 5.2.1 and Section 5.2.2, respectively.

5.2.1 CSI-CE based CSI Preprocessing

The purpose of the proposed CSI-CE is to improve the quality of the CSI matrix, by mitigating the activity-unrelated information and condensing activity-dependent features. There are two main stages of CSI-CE. 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 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.

Let $\mathbf{h}(i)$ be the magnitude of CSI vector at the i-th received packet, which can be given by $\mathbf{h}(i) = [H_{1,1}(i),\ldots,H_{1,m}(i),\ldots,H_{n,m}(i),\ldots,H_{N,M}(i)]^T$, where $H_{n,m}(i)$ is the CSI information in the nth wireless link for the mth subcarrier; M stands for the total number of subcarriers in each wireless link; N denotes the number of wireless links in total, and $N = N_t \times N_r$, N_t and N_r represents the number of transmitter and receiver antennas, respectively; T represents the transpose operation.

The key task of CSI-CE 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

$$\mathbf{h}(i) = \mathbf{h}^{st}(i) + \mathbf{h}^{dy}(i), \tag{5.1}$$

where $\mathbf{h}^{dy}(i)$ represents the dynamic CSI vector induced by human behaviors; $\mathbf{h}^{st}(i)$ stands for the static CSI vector that is unrelated to human activities. Note that, although $\mathbf{h}^{st}(i)$ does not present characters of human activities, it has greater impact on the sensing performance than $\mathbf{h}^{dy}(i)$. 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 $\mathbf{h}(i)$ for HAR. Therefore, it

is necessary to remove the static CSI vector $\mathbf{h}^{st}(i)$ from $\mathbf{h}(i)$, in order to significantly improve the quality of extracted feature signals and simplify the signal structure.

To filter out $\mathbf{h}^{st}(i)$ from $\mathbf{h}(i)$, we propose a recursive algorithm by leveraging the exponentially weighted moving average approach [180]. In such a case, $\mathbf{h}^{st}(i)$ from the i-th recursion can be estimated as $\widehat{\mathbf{h}}^{st}(i) = \delta\mathbf{h}(i) + (1-\delta)\widehat{\mathbf{h}}^{st}(i-1)$, where $\widehat{\mathbf{h}}^{st}(i)$ stands for the estimated value of $\mathbf{h}^{st}(i)$, and δ is the forgetting factor. Under this situation, the estimated dynamic CSI, $\widehat{\mathbf{h}}^{dy}(i)$, can be expressed as $\widehat{\mathbf{h}}^{dy}(i) = \mathbf{h}(i) - \widehat{\mathbf{h}}^{st}(i)$.

Note that, the timing offset between the WiFi transmitters and receivers can be estimated. In this regard, the estimation of the dynamic CSI matrix for I packets can be given by

$$\widehat{\mathbf{H}}^{dy} = [\widehat{\mathbf{h}}^{dy}(1), \ldots, \widehat{\mathbf{h}}^{dy}(i), \ldots, \widehat{\mathbf{h}}^{dy}(I)]. \tag{5.2}$$

It should be noted that $\widehat{\mathbf{H}}^{dy}$ 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 to use PCA operation to acquire more reliable features for HAR.

We conduct a PCA operation on $\widehat{\mathbf{H}}^{dy}$ 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

$$\mathbf{C}_p = \widehat{\mathbf{H}}^{dy} \times \mathrm{PCA}(\widehat{\mathbf{H}}^{dy}), \tag{5.3}$$

where $\mathbf{C}_p$ denotes the extracted features via the PCA operation, with size of $MN \times I$; 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

$$\mathbf{D}_C = \mathbf{C}_p \times \mathbf{C}_p^T. \tag{5.4}$$

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

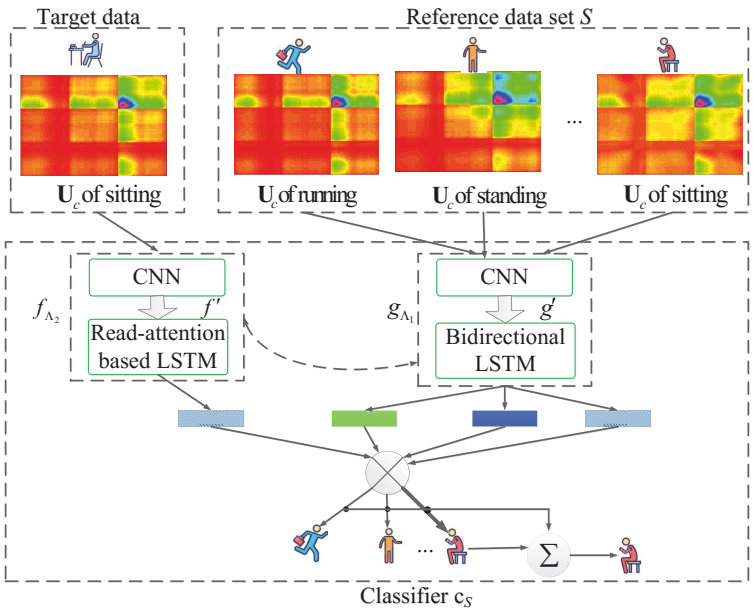


Figure 5.2: Structure of a MatNet based HAR using CFM $\mathbf{D}_C$ as input.

5.2.2 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 MatNet 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.

The architecture of HAR using a MatNet is illustrated in Fig 5.2. 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

$$S = \{(x_i, y_i)\}_{i=1}^{I}, \tag{5.5}$$

where (x, y) stands for the CFM-label pairs; $x = \{\mathbf{D}_c\}$ with a size of $MN \times MN$ 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 $P(\hat{y}|\hat{x}, S) \triangleq S \rightarrow c_S(\hat{x})$, where P denotes the probability distribution that is parameterized by the CNN and LSTM, as demonstrated in Fig. 5.2. Following this,

we can obtain the estimated output label $\hat{y}$ given a source data set S and input $\hat{x}$, by

$$\hat{y} = \arg\max_{y} P(y|\hat{x}, S). \tag{5.6}$$

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 (5.6) is equal to

$$\hat{y} = \sum_{i=1}^{N} \frac{e^{\cos(f(\hat{x}), g(x_i))}}{\sum_{j=1}^{N} e^{\cos(f(\hat{x}), g(x_j))}} y_i, \tag{5.7}$$

where $\cos(\alpha, \beta)$ denotes the cosine similarity function.

In equation (5.7), 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. 5.2, both f and g act as a bridge to input data for obtaining the maximum performance with the classifier as discussed in equation (5.7). To learn the discriminative and transferable features, 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.

For the embedding function f, it is composed by a CNN with LSTM. The read-attention based LSTM [188] is used as LSTM structure. 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

$$f(\hat{x}, S) = \text{attLSTM}(f'(\hat{x}), g(S), N_p), \tag{5.8}$$

where $f'(\hat{x})$ denotes the input data for the read-attention based LSTM, which is extracted from the CNN; $g(S)$ 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

$$h_{n_p} = \hat{h}_{n_p} + f'(\hat{x}), \tag{5.9}$$

$$\hat{h}_{n_p}, c_{n_p} = \text{LSTM}(f'(\hat{x}), [h_{n_p-1}, r_{n_p-1}], c_{n_p-1}), \tag{5.10}$$

$$r_{n_p-1} = \sum_{i=1}^{N_s} \text{softmax}(h_{n_p-1}^T g(x_i)) g(x_i). \tag{5.11}$$

where $\text{LSTM}(f'(\hat{x}), [h_{n_p-1}, r_{n_p-1}], c_{n_p-1})$ follows the implementation described in [187].

According to Fig. 5.2, the embedding function g consists of a CNN with a bidirectional LSTM [186]. 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

$$g(x_i, S) = \vec{h}_i + \overleftarrow{h}_i + g'(x_i), \tag{5.12}$$

where $\vec{h}_i$ and $\overleftarrow{h}_i$ stand for the output of the forward and backward LSTM, respectively.

Let $\vec{c}_i$ and $\overleftarrow{c}_i$ denote the cell of the forward and backward LSTM. The state of LSTM can be written as

$$\vec{h}_i, \vec{c}_i = \text{LSTM}(g'(x_i), \vec{h}_{i-1}, \vec{c}_{i-1}), \tag{5.13}$$

$$\overleftarrow{h}_i, \overleftarrow{c}_i = \text{LSTM}(g'(x_i), \overleftarrow{h}_{i+1}, \overleftarrow{c}_{i+1}), \tag{5.14}$$

where $\text{LSTM}(g', h, c)$ is in accordance with the definition in [187]. 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 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.

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

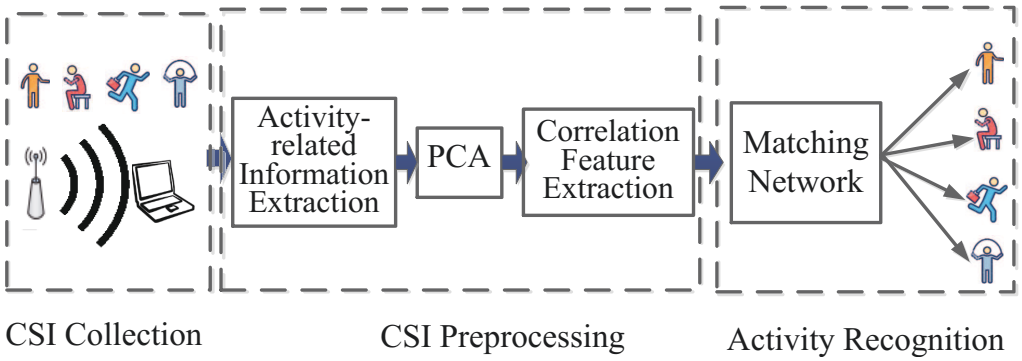


Figure 5.3: Main modules for the AFEE-MatNet Scheme.

loss function of MatNet based HAR, $\mathcal{L}$, can be obtained by

$$\mathcal{L} = -\mathrm{E}_{L\sim\mathcal{T}}\left[\mathrm{E}_{\mathcal{S},\mathcal{B}}\left[\sum_{(x,y)\in B}\log P_{\Lambda}\left(y|x,\mathcal{S}\right)\right]\right], \tag{5.15}$$

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

$$\Lambda = \arg\min_{\Lambda}\mathcal{L}(\Lambda). \tag{5.16}$$

5.3 AFEE-MatNet Scheme

To design a "one-fits-all" recognition model, we propose the AFEE-MatNet scheme by leveraging the discriminative features extracted from CSI and the developed MatNet-PCC scheme. As Fig. 5.3 depicts, the architecture of the proposed AFEE-MatNet consists of three main modules: CSI Collection, CSI Preprocessing and MatNet-PCC based Activity Recognition. The first module is to collect and store the CSI that reflects the changes of wireless signal propagations caused by human behaviors. The second module aims to clean and enhance the acquired CSI matrix through the processes in both the frequency domain and the time domain. The last module clarifies various human activities, drawing support from the enhanced CSI and MatNet.

MatNet-PCC based Activity Recognition The purpose of this module is to distinguish different behaviors by leveraging the enhanced CSI from the for-

mer module and MatNet-PCC method. In particular, we first employ MatNet to automatically learn hidden features from the enhanced CSI, as so to extract distinctive features commonly shared among different environments. As a result, the information, which is activity-related and environment-independent, can be effectively extracted for HAR. After that, we propose a PCC scheme to further fix recognition errors and improve sensing accuracy. It is noteworthy that MatNet architecture is trained using samples from source environments in an offline manner. The well-trained model is then applied to identify various activities in an online manner.

5.3.1 AFEE based CSI Preprocessing

In this section, we will describe the design of AFEE to improve the quality of the CSI matrix. We first present the CSI cleaning and enhancement method, followed by the discussion of frequency domain feature extraction method.

CSI Cleaning and Enhancement

Let N_r and N_t indicate the amount of antennas at the receiver and transmitter, respectively. The CSI vector $\mathbf{h}(i)$ acquired from the m-th received packet can be represented as $\mathbf{h}(i) = [H_{1,1}(i), H_{1,2}(i), \ldots, H_{n,m}(i), \ldots, H_{N,M}(i)]^T$, where $H_{n,m}(i)$ indicates the CSI data collected in the nth wireless link at the mth subcarrier; $N = N_t \times N_r$ represents the total number of wireless links; M denotes the total number of subcarriers in the wireless link; and T indicates the transpose operation. The CSI matrix $\mathbf{H}$, which is composed of CSI vectors collected from I packets, can be given by $\mathbf{H} = [\mathbf{h}(1), \ldots, \mathbf{h}(i) \ldots, \mathbf{h}(I)]$.

Although putting the original CSI matrix $\mathbf{H}$ into a deep learning network directly can realize behavior recognition, it is not a good option. The reason is that $\mathbf{H}$ contains much activity-unrelated information that could severely influence the recognition result. Moreover, the noise and the phase offset in $\mathbf{H}$ also affect the recognition performance. To address these problems, we first propose to apply a conjugate multiplication (CM) method [191] to improve the quality of the CSI. The key insight of CM is to take the acquired CSI with the best quality as a reference

$\mathbf{h}_{\text{ref}}$, and then calculate a conjugate multiplication of $\mathbf{h}_{\text{ref}}$ and $\mathbf{h}$. The criterion for selecting $\mathbf{h}_{\text{ref}}$ is to choose a CSI vector collected from the antenna with a maximum ratio of amplitudes and standard deviations (MRASD). To obtain $\mathbf{h}_{\text{ref}}$, we first calculate the wireless link with MRASD, by

$$N_{\text{ref}} = \arg\max_{n \in N} \frac{1}{M} \sum_{m=1}^{M} \frac{\text{mean}(|\mathbf{h}_{n,m}|)}{\text{std}(|\mathbf{h}_{n,m}|)}, \tag{5.17}$$

where N_{ref} represents the index of wireless link with MRASD; mean(.) and std(.) stand for the mean operation and the standard deviation operation, respectively. Then $\mathbf{h}_{\text{ref}}$, acquired from the i-th packet, can be achieved in equation (5.18).

$$\begin{aligned} \mathbf{h}_{\text{ref}}(i) \;=\; [&\underbrace{H_{N_{\text{ref}},1}(i),\ldots,H_{N_{\text{ref}},M}(i)}_{1}, \underbrace{H_{N_{\text{ref}},1}(i),\ldots,H_{N_{\text{ref}},M}(i)}_{2},\ldots, \\ &\underbrace{H_{N_{\text{ref}},1}(i),\ldots,H_{N_{\text{ref}},M}(i)}_{N}]^T. \end{aligned} \tag{5.18}$$

Upon obtaining $\mathbf{h}_{\text{ref}}$, the reference CSI matrix $\mathbf{H}_{\text{ref}}$ for I packets can be expressed as

$$\mathbf{H}_{\text{ref}} = [\mathbf{h}_{\text{ref}}(1),\ldots,\mathbf{h}_{\text{ref}}(i)\ldots,\mathbf{h}_{\text{ref}}(I)]. \tag{5.19}$$

Based on equation (5.19), the conjugate multiplications between all the wireless links and reference links can be obtained by

$$\mathbf{C} = \mathbf{H}_{\text{ref}} \odot \mathbf{H}^*, \tag{5.20}$$

where $\odot$ stands for dot product, and $*$ denotes the Hermitian.

Through the above operations, the output $\mathbf{C}$, with size $MN \times I$, is expected to overcome the effect of phase offset [191]. However, it still contains some activity-unrelated information and random noise, negatively affecting recognition results. In order to tackle that concern, we propose to perform PCA to retain the activity-related information and eliminate noise contained in $\mathbf{C}$. Specifically, we divide $\mathbf{C}$ into N sub-matrices, written as

$$\mathbf{C} = [\overline{\mathbf{C}}_1, \ldots, \overline{\mathbf{C}}_n, \ldots, \overline{\mathbf{C}}_N]^T, \tag{5.21}$$

$$\overline{\mathbf{C}}_n = \begin{bmatrix} \mathbf{C}_{n,1}(1) & \dots & \mathbf{C}_{n,1}(i) & \dots & \mathbf{C}_{n,1}(I) \\ \mathbf{C}_{n,2}(1) & \dots & \mathbf{C}_{n,2}(i) & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ \mathbf{C}_{n,M}(1) & \dots & \mathbf{C}_{n,M}(i) & \dots & \mathbf{C}_{n,M}(I) \end{bmatrix}. \tag{5.22}$$

Next, we conduct a PCA operation on each $\overline{\mathbf{C}}_n$ to obtain the top $p+1$ principal components. Note that most of the noise and environment-specific information are involved in the first principal component. To mitigate them, we construct the principal component matrix $\widehat{\mathbf{C}}_n$ by discarding the first principal component and keeping the rest principal components, represented as

$$\widehat{\mathbf{C}}_n = \overline{\mathbf{C}}_n \times \mathbf{\Phi}_n^{\{2:p\}}, \tag{5.23}$$

where $\mathbf{\Phi}_n^{\{2:p\}}$ stands for the matrix constructed by elements from the 2nd to the pth eigenvectors. The new feature matrix based on principal components of all wireless links can be obtained by

$$\widehat{\mathbf{C}} = [\widehat{\mathbf{C}}_1, \widehat{\mathbf{C}}_2, \dots, \widehat{\mathbf{C}}_N]^T. \tag{5.24}$$

Note that, the size of $\widehat{\mathbf{C}}$ is $P \times I$, $P = pN$, and P is used to achieve a tradeoff between computational complexity and recognition accuracy. The impact of P on the recognition performance is discussed in Fig.5.10. Although $\widehat{\mathbf{C}}$ can be used as the input signal to train the DL network, feeding $\widehat{\mathbf{C}}$ into the DL network directly would cause a significant increase in the training complexity due to its large dimension. For instance, in this paper, the time window for each activity is approximately 1s, the rate of samples f_s is 1KHz, and we set $P = 60$ empirically in the experiments, so $\widehat{\mathbf{C}}$ is a matrix with size 60×1000. Under this situation, it would cause extremely high training complexity if taking $\widehat{\mathbf{C}}$ as the input signal for the DL network. Consequently, it is important to decrease the size of the $\widehat{\mathbf{C}}$, thereby lowering the training complexity.

Frequency Domain Feature Extraction

Through the above operations, $\widehat{\mathbf{C}}$ is anticipated to present unique characters for different human activities. However, it has a large dimension that significantly increases the training overhead. Moreover, $\widehat{\mathbf{C}}$ still contains residual activity-unrelated

information and residual noise, leading to performance degradation. To deal with these problems, a frequency domain feature extraction method is proposed below to learn reliable information from the frequency domain.

We conduct Fast Fourier Transform (FFT) for each row of $\widehat{\mathbf{C}}$ to get the frequency domain feature matrix, by

$$\mathbf{C}_F = \mathrm{FFT}(\widehat{\mathbf{C}}), \tag{5.25}$$

where FFT(.) stands for the Fast Fourier Transform operation. $\mathbf{C}_F$ indicates the extracted frequency domain feature matrix.

It is notable that, most of CSI variations caused by human activity in daily life are in a relatively low frequency range (less than 100Hz), due to the limited speed and space of movements. On this basis, we discard the data in the high frequency range contained in $\mathbf{C}_F$ to remove activity-unrelated information whilst retraining the activity-related features. To further enhance the activity-related CSI, we remove the zero frequency component (i.e., the first column of $\mathbf{C}_F$) that are mainly environment-specific but behavior-unrelated. Let $\widehat{\mathbf{C}_F}$ be the compressed feature matrix, and q is the cutoff frequency used to filter out activity-unrelated features. The value of q is determined based on the types of activity to be recognized. In this paper, we intend to recognize six activities $\{laying, standing, walk, fall, standup, empty\}$. The maximum frequency for these behaviors is about 80Hz [119], so we set $q = 80$Hz. Through the aforementioned operations, the size of $\widehat{\mathbf{C}_F}$ is $P \times qI/f_s$ which is much smaller than that of $\mathbf{C}_F$. Thus, using $\widehat{\mathbf{C}_F}$ as the input signal to train the DL network can result in a notable decrease in the training complexity.

It is noteworthy that the correlation features between different wireless links of transmitter-receiver pairs can provide distinguished information for identifying different behaviors. To provide more information for input signals to DL network, we also compute the correlation feature of $\widehat{\mathbf{C}_F}$, by

$$\mathbf{U}_C = \widehat{\mathbf{C}_F} \times (\widehat{\mathbf{C}_F})^T. \tag{5.26}$$

Note that, $\mathbf{U}_C$ and $\widehat{\mathbf{C}_F}$ can provide correlated and complementary features for behavior recognition. On the one hand, $\widehat{\mathbf{C}_F}$ contains features in different wireless

links, while it fails to present other types of information such as the correlation features between different links. On the other hand, $\mathbf{U}_C$ highlights the correlation information between different wireless links, but it losses some information during correlation operations. Therefore, we take $\mathbf{\Theta} = \{\widehat{\mathbf{C}_F}, \mathbf{U}_C\}$ as the input signal of MatNet for extracting reliable and distinguished features.

5.3.2 MatNet-PCC based Human Activity Recognition

From the previous section, the output of our AFEE method is expected to contain significantly enhanced CSI, by retaining the activity-related information whilst mitigating activity-unrelated data. However, it is difficult to fully eliminate the impact of the environment. To overcome this problem, we propose to use MatNet to learn and extract transferable features shared among different environments, thereby these features are robust to environments. Moreover, we propose a prediction checking and correction method to further improve recognition accuracy.

Architecture of MatNet

To realize activity classification, we propose to employ MatNet to automatically learn and extract hidden features from the enhanced CSI obtained from Module 2. The applied MatNet has the similar structure shown in Fig. 5.2. Note that, in this section, $\mathbf{\Theta}$ is used as the input feature signal. Given a source data set S, MatNet is able to build a classier c_s for each S, mapping S to c_S, $S \rightarrow c_S(.)$.

We define a task, denoted as $\mathcal{T}$, as the distribution for potential label sets of human behaviors. In each episode, a set of human activities $\mathcal{L}$ are sampled from $\mathcal{T}$, $\mathcal{L} \sim \mathcal{T}$, consisting of six different behaviors: $\{standing, laying, walk, standup, empty, fall\}$. Next, $\mathcal{L}$ is used for sampling both the source data set S and the batch of target set B, achieving $\mathcal{S} = S \sim \mathcal{L}$ and $\mathcal{B} = B \sim \mathcal{L}$. The purpose of training MatNet is minimizing the error between the estimated and the actual labels in $\mathcal{B}$ under the condition of $\mathcal{S}$, which is

$$\Lambda = \arg\min_{\Lambda} \left\{ - \mathrm{E}_{\mathcal{L}\sim\mathcal{T}} \left[\mathrm{E}_{\mathcal{S},\mathcal{B}} \left[\sum_{(x,y)\in B} \log P_{\Lambda}\left(y|x,\mathcal{S}\right) \right] \right] \right\}, \tag{5.27}$$

where $\Lambda = \{\Lambda_1, \Lambda_2\}$, Λ_1 and Λ_2 stand for parameter sets of embedding functions g and f, respectively.

Note that, the training process is conducted fully conditional on the whole data set S that includes different source environments. Under this situation, the relationship between different source environments can be built drawing support from g and f. As a result, the generalized features among different environments can be learned and extracted for HAR. In other words, the impact of a specific environment on recognition performance is significantly reduced. Therefore, the proposed AFEE-MetNet scheme is robust to environments, contributing to environment-independent recognition.

Prediction Checking and Correction

The core of the proposed PCC scheme is to rectify certain recognition errors which do not match the state transition of human behaviors, so as to further improve the recognition accuracy. When a person performs a set of different activities continuously, these behaviors are not independent but belong to a circle of continuous states. Fig. 5.4 shows the simplified state transition diagram for the case when people perform six different activities $\{laying, standing, walk, fall, standup, empty\}$. To be specific, when a person conducts "stand up" at the current time slot, he/she may perform "fall", "standing" or "stands up" at next time slot. In such a case, if the output of MatNet at the next time slot is "fall", "standing" or "stands up", the result obeys to the state transition diagram. Under this situation, we treat the output of MatNet as a logical result and keep it as the final output. Otherwise, the output of MatNet at the next time is regarded as an incorrect result, and we will correct it using the detailed scheme as follows.

We let N_a stand for the total number of activities to be recognized, and N_a is set to 6, including $\{laying, standing, walk, fall, standup, empty\}$; $Y(t_n - 1)$ denotes the final output of our proposed method at time slot $t_n - 1$, and $\hat{y}(t_n)$ represents the output of MatNet at time slot t_n; $Y(t_n - 1), \hat{y}(t_n) \in [1, N_a]$; Υ is the state transition diagram for different activities. If $Y(t_n - 1)$ and $\hat{y}(t_n)$ abide by the state transition diagram, i.e., $[Y(t_n - 1), \hat{y}(t_n)] \sim \Upsilon$, we set $\eta = 1$, otherwise, $\eta = 0$.

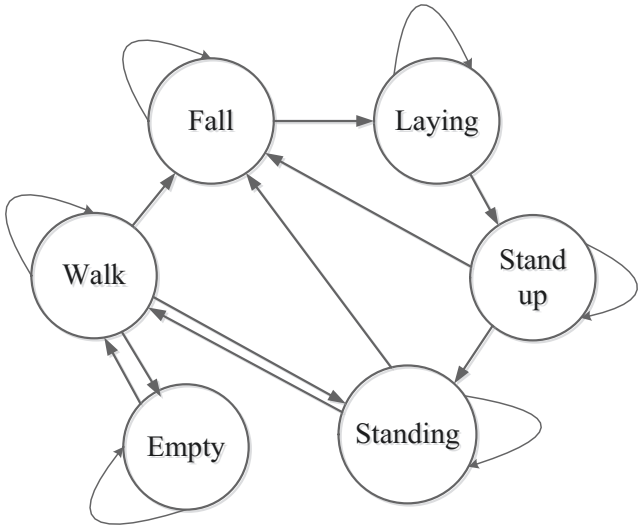


Figure 5.4: Simplified state transition diagram for six different activities.

If $\eta = 0$, it means that the output of the MatNet is incorrect and should be rectified. We let $\mathbf{P}_m$ with size $N_a \times N_a$ be the confusion matrix; $\mathbf{P}_m(i,j)$ stands for the probability of the activity i being recognized as the activity j; $i, j \in [1, N_a]$. We will rectify the incorrect sensing result based on Υ and $\mathbf{P}_m$. Specifically, we first seek out the possible activity set for the time slot t_n conditioned on $Y(t_n - 1)$, denoted as A_{t_n}, which is given by

$$A_{t_n} \triangleq \{j | [Y(t_n - 1), j] \sim \Upsilon, j \in [1, N_a]\}. \tag{5.28}$$

After that, the probability of misjudging A_{t_n} as $\hat{y}(t_n)$ can be obtained with the help of $\mathbf{P}_m$. Moreover, we can get the activity j^* that holds the highest probability of incorrect classification in $\mathbf{P}_m$, expressed as

$$j^* = \arg \max_{j \in A_{t_n}} \mathbf{P}_m(j, \hat{y}(t_n)), \tag{5.29}$$

where j^* can be treated as the final output of our proposed scheme at time slot t_n. Thus the final output of our proposed method at time slot t_n can be expressed as

$$Y(t_n) = \begin{cases} \hat{y}(t_n), & \eta = 1, \\ j^*, & \eta = 0. \end{cases} \tag{5.30}$$

To this end, we can see that the value of $Y(t_n)$ heavily relies on $Y(t_n - 1)$. In other words, the accuracy of $Y(t_n - 1)$ significantly affects $Y(t_n)$, restricting the

Algorithm 1: Prediction Checking and Correction.

begin
  Initialize: the final output $Y(t_n - 1)$,
  the outputs of MatNet $\hat{y}(t_n), \hat{y}(t_n - 1), \ldots, \hat{y}(t_n - \tau)$,
  the state transition diagram Υ;
  the confusion matrix $\mathbf{P}_m$;
  if $[\hat{y}(t_n - 1), \ldots, \hat{y}(t_n - \tau)] \nsim \Upsilon$
    PCC method is inactivated;
    $Y(t_n) = \hat{y}(t_n)$;
  else
    if $[\hat{y}(t_n), Y(t_n - 1)] \sim \Upsilon$
      $\eta = 1$;
    else
      $\eta = 0$;
      Compute A_{t_n} according to equation (5.28);
      Compute j^* according to equation (5.29);
    end
    Compute $Y(t_n)$ according to equation (5.30);
  end
  Update $\mathbf{P}_m$;
end

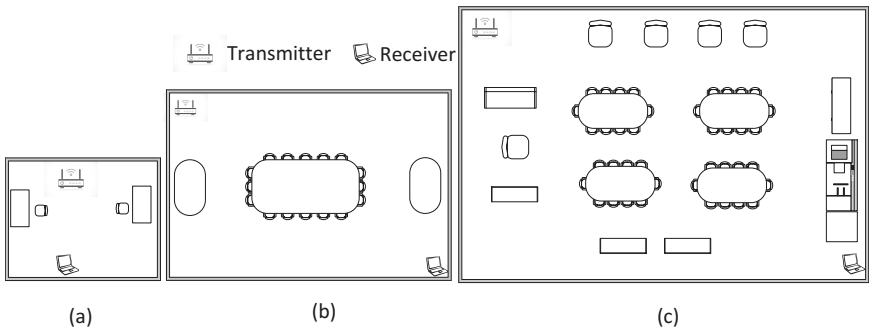


Figure 5.5: Layout of three indoor experimental areas: (a) $3m \times 4m$ office. (b) $4m \times 6m$ meeting room. (c) $6m \times 7m$ laboratory.

performance of PCC. To address this issue, we propose an activation mechanism for PCC to guarantee its reliability, i.e., preventing over-correction cases. To be specific, the proposed PCC method can be activated to correct $\hat{y}(t_n)$, only if the outputs of MatNet $[\hat{y}(t_n - 1), \ldots, \hat{y}(t_n - \tau)]$ abide by Υ but $\hat{y}(t_n)$ does not. In this paper, the value of τ is empirically set as 3. The details of proposed PCC are summarized in Algorithm 1.

5.4 Experiment and Evaluation

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

To validate the recognition performance, we implement the proposed schemes in seven indoor environments with various wireless environmental complexities. Notably, the complexity of the wireless environment is dependent on many factors, e.g., distance and obstacles between transmitter and receiver. In each configuration/environment, each of five people performs six types of behaviors, including falling, laying, standing up, standing, walking, and empty room. Each behavior is conducted 200 times by each person. The each target person performs activities randomly in the experimental environment. We demonstrate the average HAR performance by considering different locations. Among these seven environments, the first four environments are taken as source environments, and the collected data in them are regarded as the training dataset, which is used for training the proposed schemes. The rest three configurations are treated as testing environments, and the

acquired data in these environments is the testing dataset that is for evaluating the recognition performance of the proposed schemes. Due to limited space, we illustrate the layouts of three testing environments in Fig. 5.5. In particular, the first configuration/environment is a $6m \times 7m$ laboratory room, the second one is a $4m \times 6m$ meeting room, and the third one is a $3m \times 4m$ square area.

In the experiments, two computers are employed as transmitter and receiver, respectively, and each equips with Intel WiFi NIC 5300 network card which operation abides by the 802.11n standards. Both the transmitter and receiver conduct data transmission at the operating frequency 5.32 GHz. The transmitter, having one antenna ($N_t = 1$), keeps emitting signals, and the receivers continuously collect data via three antennas ($N_r = 3$). The CSI tool in [25] is adopted to acquire and store signals (i.e., CSI). There are 30 subcarriers ($M = 30$) available for each pair of the transmitter-receiver antennas. We propose to use a sliding window to collect data/samples for each behavior from raw CSI streams, and the time length for this window is set as 1s. In the training process, if multiple behaviors are involved in one time window, its label will be the activity with the largest ratio. We leave the work of designing a sliding window with an adaptive time length based on different activities to be future task. The sample rate is 1 KHz, thus the dimension of CSI matrix ($\mathbf{H}$) is 90×1000.

In the training stage, each embedding function of the proposed schemes includes a CNN with 6 convolutional layers. In each layer, there are 3×3 convolution, a ReLU non-linearity operation, and a 2×2 max-pooling. We employ a 2.3 GHz PC with Nvidia GeForce GTX 1070Ti graphic card (8GB memory) to train the proposed schemes. We select 64 and 0.001 as the batch size and learning rate, respectively. Note that, we obtain robust scaling for the developed schemes in experiments. Specifically, we perform a normalization operation on the input signal before feeding it into the MatNet architecture. Next, in the training process, we utilize the Batch Normalization method [190] for normalizing the input signal of each layer.

5.4.1 Performance Comparison of Different Methods

In this section, numerous simulations are designed and conducted to verify the recognition performance of the proposed HAR-MN-EF and AFEE-MatNet. To do that, we first present extensive simulation results considering various conditions and parameters, to analyze the performance of the proposed schemes (i.e., HAR-MN-EF and AFEE-MatNet) and other four recognition schemes (i.e., RNN [146],TNNAR [171], EI [172], and BVP [174]). Then, we comprehensively validate the sensing capability of our developed HAR-MN-EF and AFEE-MatNet from a wide range of aspects. 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.

Table 5.1 compares the average recognition accuracy of six activities for different sensing methods under different testing environments. The training dataset from four source environments are collected to train each scheme, and each trained model is applied to recognize different behaviors in testing environments. In this table, as can be seen, both the proposed HAR-MN-EF and the proposed AFEE-MatNet notably outperform the other four sensing methods in each testing environment (i.e., RNN [146], TNNAR [171], EI [172], and BVP [174]). This is because, for the proposed HAR-MN-EF, 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. For the proposed AFEE-MatNet, its promising performance credits to the property of the proposed AFEE scheme and MatNet. Specifically, the AFEE is proposed to mitigate most activity-unrelated data and condense the behavior-related information. As a result, the impact of activity-unrelated components (e.g., caused by noise or environment) on feature signals is significantly reduced. Moreover, we propose to employ MatNet architecture to automatically build a relationship among different source environments and extract generalized features for HAR. In other words, the features commonly shared among different source environments are extracted for HAR, and the information subject to a certain environment would be discarded. This enables

Table 5.1: Average recognition accuracy of different methods in three indoor configurations

Method	1st Exp.	2nd Exp.	3rd Exp.
Proposed HAR-MN-EF	0.698	0.711	0.734
Proposed AFEE-MatNet	0.734	0.763	0.803
RNN	0.331	0.394	0.426
EI	0.531	0.577	0.605
TNNAR	0.485	0.511	0.544
BVP	0.687	0.699	0.715

our proposed AFEE-MatNet to achieve robustness to environments. For RNN [146], it cannot learn effective features commonly shared among source environments, degrading the sensing accuracies dramatically. Regarding TNNAR [171] and EI [172], the number of source environment limits their recognition accuracies. When the number is not sufficient, it is hard for these methods to achieve accurate sensing results. BVP [174] fails to reliably classify some light activities (such as laying), degrading its detection accuracy.

To detail the performance verification, we present the confusion matrix of six methods in Fig. 5.6. In this figure, we select the third testing environment to examine the performance of each method. We can observe that the sensing accuracy of the proposed AFEE-MatNet and the proposed HAR-MN-EF are greatly better than those of the other four methods, in terms of identifying different behaviors. To be specific, for the proposed AFEE-MatNet and the proposed HAR-MN-EF, each estimated behavior is in accordance with the corresponding actual one with high probability, implying that the proposed schemes are capable of accomplishing reliable detections. By contrast, for RNN [146], TNNAR [171] and EI [172], their prediction results cannot match the corresponding actual ones, alluding to the fact that those methods have difficulties predicting activities correctly. Although the prediction activity of BVP [174] is consistent with the actual behaviors, the accuracy for light activities are low (e.g., laying and standing).

In Fig. 5.7, we demonstrate the number of source environments on the average

Predicted activity

Actual activity	Walk	Standing	Fall	Laying	Stand up	Empty
Walk	0.78	0.01	0.161	0.001	0.044	0.004
Standing	0.008	0.719	0.001	0.069	0.041	0.162
Fall	0.226	0	0.701	0.043	0.016	0.014
Laying	0.003	0.124	0.046	0.682	0.007	0.138
Stand up	0.054	0.062	0.111	0.041	0.732	0
Empty	0.007	0.052	0	0.141	0	0.8

(a) Proposed HAR-MN-EF

Predicted activity

Actual activity	Walk	Standing	Fall	Laying	Stand up	Empty
Walk	0.836	0.017	0.086	0	0.042	0.019
Standing	0.001	0.819	0.003	0.081	0.058	0.038
Fall	0.046	0.006	0.842	0.062	0.039	0.005
Laying	0	0.08	0.017	0.703	0.063	0.137
Stand up	0.03	0.106	0.06	0	0.804	0
Empty	0.081	0	0.003	0.105	0	0.811

(b) Proposed AFEE-MatNet

Predicted activity

Actual activity	Walk	Standing	Fall	Laying	Stand up	Empty
Walk	0.834	0.035	0.063	0.003	0.044	0.021
Standing	0.011	0.617	0.001	0.138	0.028	0.205
Fall	0.07	0	0.874	0.012	0.04	0.004
Laying	0	0.101	0.011	0.505	0.014	0.369
Stand up	0.001	0.097	0.054	0.003	0.844	0.001
Empty	0.048	0.13	0.001	0.207	0	0.614

(c) BVP

Predicted activity

Actual activity	Walk	Standing	Fall	Laying	Stand up	Empty
Walk	0.724	0.081	0.1	0.001	0.022	0.072
Standing	0.003	0.641	0.014	0.112	0.13	0.1
Fall	0.223	0.001	0.572	0.009	0.194	0.001
Laying	0.001	0.144	0.009	0.415	0	0.431
Stand up	0.042	0.078	0.214	0.083	0.583	0
Empty	0.005	0.09	0.001	0.208	0.001	0.695

(d) EI

Predicted activity

Actual activity	Walk	Standing	Fall	Laying	Stand up	Empty
Walk	0.512	0.018	0.315	0.001	0.122	0.032
Standing	0.006	0.351	0.001	0.412	0.03	0.2
Fall	0.523	0.001	0.381	0.089	0.006	0
Laying	0.001	0.284	0.049	0.336	0.007	0.323
Stand up	0.254	0.127	0.171	0.004	0.464	0
Empty	0.035	0.242	0	0.208	0.003	0.512

(e) RNN

Predicted activity

Actual activity	Walk	Standing	Fall	Laying	Stand up	Empty
Walk	0.651	0.034	0.175	0.002	0.123	0.015
Standing	0.001	0.415	0.004	0.102	0.002	0.476
Fall	0.107	0.173	0.581	0.125	0.009	0.005
Laying	0	0.1	0.003	0.394	0.001	0.502
Stand up	0.185	0.092	0.122	0.003	0.597	0.001
Empty	0.045	0.141	0	0.185	0.001	0.628

(f) TNNAR

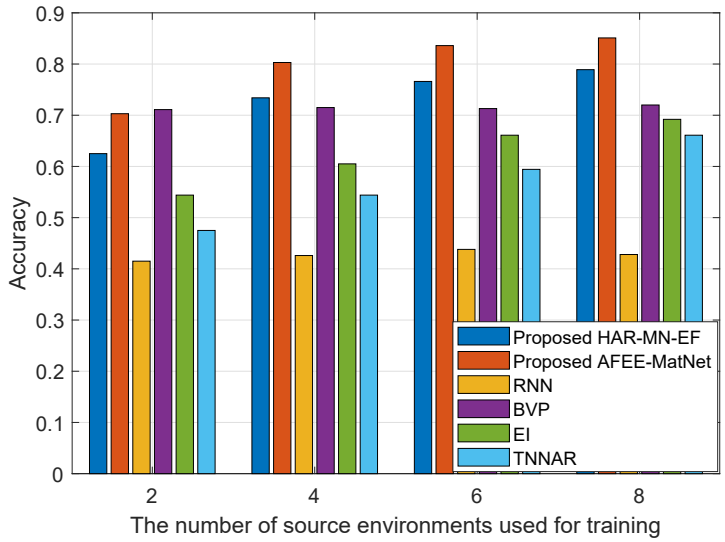


Figure 5.7: Recognition accuracy with increased number of source environments

recognition accuracy for six sensing schemes. The third testing environment is selected for performance evaluation in this figure. As can be seen, the proposed schemes (i.e., HAR-MN-EF and AFEE-MatNet), TNNAR [171], and EI [172] can achieve improved detection results when increasing the number of source environments. The reason is that, with more source environments, these four schemes are able to extract more transferable features shared among these environments, which is beneficial in distinguishing different activities and results in better recognition performance. On the contrary, more source environments do not help the RNN in [146] to achieve a better sensing accuracy. This is because, the RNN in [146] does not have the capability to extract common features shared by source environments and the testing environment. Moreover, increasing the number of source environments does not necessarily lead to an improved result for BVP [174]. The reason is that, the authors in BVP proposed velocity-related data for HAR, which has little relationship with the number of source environments.

In Fig. 5.8, we present how the number of receiving antennas, demonstrated as the number of total subcarriers, affects the average recognition accuracy. We examine the sensing performance of different sensing schemes in the third testing environment. We can observe from this figure that all methods can get improved recognition results when the number of subcarriers increases. Moreover, the proposed schemes are able to achieve larger improvement with more subcarriers, compared to the other

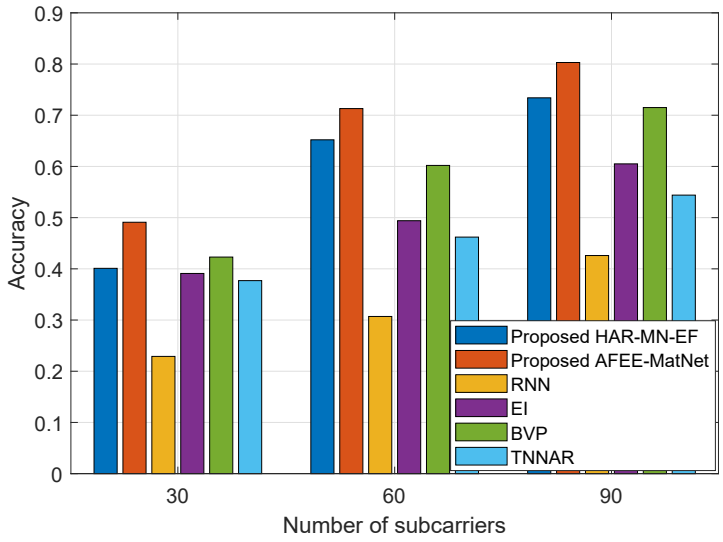


Figure 5.8: Impact of the number of subcarriers on the recognition accuracy.

four methods.

In Table. 5.2, we also discuss how well the proposed PCC influences the classification results of different methods. The recognition stage is conducted in the third testing environment. From this table, it is clear that the proposed PCC can be applied to other methods, enabling these methods to achieve higher sensing accuracies. The reason is that the proposed PCC can correct some detection errors that are not in accordance with the state transition of human behaviors. Moreover, we can see that our proposed work can achieve a more obvious improvement than the other methods. This is because our proposed work achieves higher sensing accuracies, thereby having more opportunities to activate the PCC scheme (refer to the activation mechanism of PCC in Section 5.3.2) to further improve detection performance.

5.4.2 Performance Evaluation of Proposed Schemes under Various Conditions

In this section, the characters of proposed HAR-MN-EF and AFEE-MatNet are analyzed and evaluated, respectively, considering various parameters and conditions. Specifically, we first analyze how the proposed CSI-CE method impacts the sensing accuracy of the proposed HAR-MN-EF. Then, we verify how the recognition accura-

Table 5.2: Impact of PCC on recognition accuracy for different methods

Method	Without PCC	With PCC
Proposed HAR-MN-EF	0.734	0.779
Proposed AFEE-MatNet	0.752.	0.803
RNN	0.426	0.438
BVP	0.715	0.758
EI	0.605	0.621
TNNAR	0.544	0.559

Table 5.3: Impact of CSI-CE on proposed HAR-MN-EF

Method	Accuracy	Training Time
With CSI-CE	0.714	34.5 min
Without CSI-CE	0.412	89.7 min

cy of the proposed AFEE-MatNet varies under different conditions, including input signals, human beings, with/without proposed AFEE method, and the number of principal components for all wireless links (P).

The impact of the proposed CSI-CE method on the performance of HAR-MN-EF is shown in Table 5.3, in the context of sensing accuracy and training time. In this table, the average sensing accuracy of three testing environments is provided. 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.

To examine the performance of our AFEE-MatNet, we discuss how its performance changes with various input signals. In Table 5.4, we investigate changes in recognition performance with different input signals in three testing environments. "AFEE-MatNet" refers to the case of using the output of AFEE (i.e., $\mathbf{\Theta}$) to train MatNet for HAR. "AFEE-MatNet-C" and "AFEE-MatNet-U" stand for cases of putting $\widehat{\mathbf{C}_F}$ and $\mathbf{U}_C$ (refer to Section 5.3.1) into the MatNet architecture for training, respectively. From this table, it is obvious that AFEE-MatNet performs much better

Table 5.4: Recognition accuracy using different input signals.

Method	1st Exp.	2nd Exp.	3rd Exp.
AFEE-MatNet	0.734	0.763	0.803
AFEE-MatNet-C	0.698	0.724	0.766
AFEE-MatNet-U	0.669	0.715	0.749

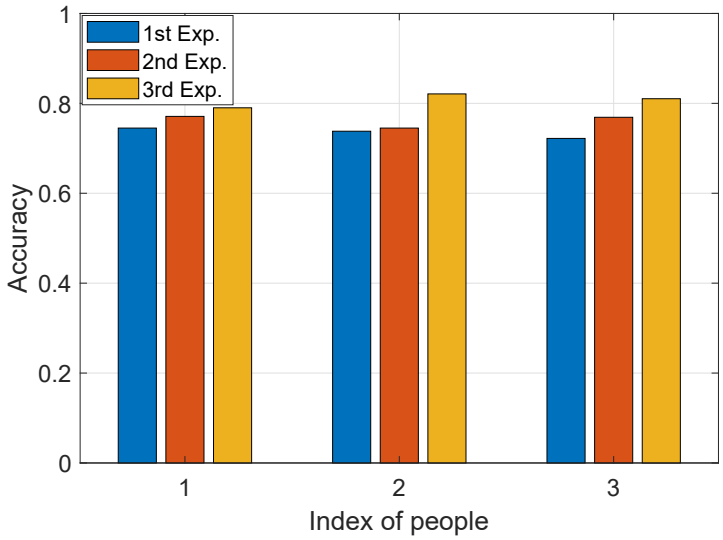


Figure 5.9: Average recognition accuracy for different people

than both "AFEE-MatNet-C" and "AFEE-MatNet-U", because it can extract more distinguishable and generalized feature for recognition.

In Fig. 5.9, we demonstrate how detection accuracies vary with different human beings in each testing environment. Two persons perform different activities in the training stage to build the training dataset. In the testing process, the trained model is used to recognize activities performed by the other three volunteers. In this figure, we provide the average recognition accuracy for three persons. As can be seen, the recognition accuracy changes differently with various testing persons in each testing environment, implying that different persons have diverse impacts on sensing performance. Another observation is that the average recognition results across various persons are still reliable (e.g., above 72%) in all testing environments. This demonstrates that the proposed AFEE-MatNet gains robustness to human diversity.

In this part, we investigate the significance of AFEE on the proposed AFEE-

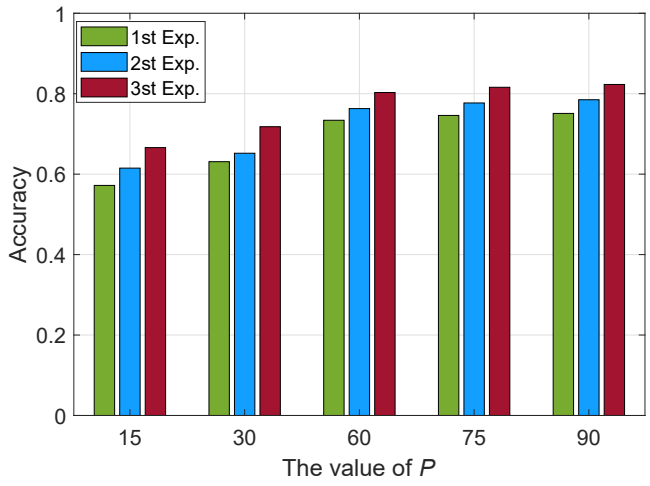
Table 5.5: Impact of AFEE on proposed AFEE-MatNet

Method	Accuracy	Training Time
With AFEE	0.767	131.7 mins
Without AFEE	0.464	531.2 mins

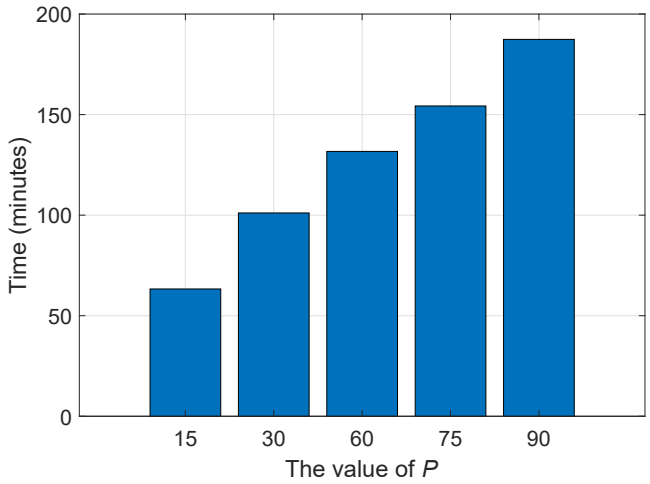
MatNet scheme. In Table 5.5, we demonstrate how AFEE affects the performance of the proposed scheme in the sense of recognition accuracy and training time. In this table, we show the average sensing accuracy of three testing environments. As can be seen, the setup with AFEE achieves much higher training accuracy and less training time, compared to that without AFEE. The reason is that the proposed AFEE has the capability of significantly suppressing activity-unrelated data and enhancing activity-related information. As a result, the impact subject to a specific environment but not helpful for HAR can be effectively reduced, making the proposed scheme more robust to variations of environments. Moreover, we can find that the training time in the case with AFEE is much less than that without AFEE. This is because AFEE can greatly decrease the size of the CSI matrix that are the input signals for the training stage, shortening the training time.

Fig. 5.10 illustrates how the number of principal components for all wireless links (P) in AFEE influences the sensing accuracy and training time of the proposed AFEE-MatNet. As Fig. 5.10 depicts, a larger P can result in an improvement in recognition accuracy for each testing environment, while the speed of improvement slows down when P is sufficiently large. The reason is that a larger P, i.e., more principal components, can provide more useful information for HAR, contributing to better recognition results. However, the proportion of distinctive features for HAR contained in the principal component with a larger index become less. Additionally, the training time becomes longer with a larger P. Therefore, the user can determine the value of P to make a balance between recognition accuracy and training time.

Table 5.6 compares the key features of the proposed HAR-MN-EF and the proposed AFEE-MatNet from various aspects. To be specific, the proposed AFEE-MatNet is capable of achieving higher sensing accuracies using a smaller number of source environments than the proposed HAR-MN-EF, which credits the AFEE



(a) Recognition Accuracy



(b) Training time

Figure 5.10: Impact of P on the recognition accuracy and training time for proposed AFEE-MatNet.

method and the PCC process proposed in AFEE-MatNet. Note that, compared to the proposed HAR-MN-EF, the proposed AFEE-MatNet requires a longer time to train the model well, due to more complex signal processing stages. Given that, a user can select a scheme by balancing their requirements on sensing accuracies, the number of source environments and training time.

Table 5.6: Characters of Proposed Methods

Methods	Accuracy	Training Time	Number of required source environments
Proposed HAR-MN-EF	Fair	Low	Large
Proposed AFEE-MatNet	High	High	Small

5.5 Conclusion

In this chapter, we proposed two novel schemes to achieve environmental-robust recognition. We first proposed 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 was adopted to process features extracted by CSI-CE.

For better detection performance, we proposed an innovative CSI-based HAR scheme, represented as AFEE-MatNet, to accomplish the "one-fits-all" human activity recognition. The designed scheme only requires an initial training, and then it can be directly applied in new environments without an extra re-training process. To achieve this, we proposed a novel AFEE scheme and MatNet-PCC method. The AFEE scheme is able to mitigate noise, eliminate the impact of behavior-unrelated elements subject to the specific environment, and retain the activity-related information. It can also significantly decrease the size of input signals, lowering the computational complexity and shortening the training time. We proposed to employ the MatNet architecture to extract generalized features shared among source environments, facilitating cross-environment recognitions. To further improve sensing accuracy, we proposed a prediction checking and correction scheme to rectify detection errors that do not abide by the state transition of behaviors. We designed and conducted numerous experiments to validate the performance of our proposed HAR-MN-EF and AFEE-MatNet from a wide range of aspects. The

extensive results verified that our proposed schemes vastly outperform existing state-of-the-art techniques, in terms of improving the recognition accuracy and lowering the training time.

Chapter 6

Conclusion and Future Work

In this thesis, we have developed effective solutions for channel state information (CSI) based human activity recognition (CSI-based HAR) using deep learning (DL) networks. The main contributions of this thesis and future research directions are provided in this chapter.

6.1 Conclusions

In Chapter 3, we proposed two DL-based HAR for performance improvement in the scenario with a sufficient number of required training samples. First, we developed a human activity recognition scheme using Deep Learning Networks with enhanced Channel State information (DLN-eCSI). The proposed DLN-eCSI gains appealing features of higher sensing accuracy and lower training complexity. To achieve that, we designed a CSI feature enhancement scheme (CFES) to clean and compress the input signals of DL networks. The signals enhanced by CFES are fed into long-short term memory based recurrent neural networking (LSTM-RNN) to extract deep and inherent features, achieving successful HAR. To further improve the detection accuracy, we developed HAR using activity filter-based deep learning network (HARAF-DLN) with enhanced correlation features. In particular, we proposed two main innovations in HARAF-DLN including CSI compensation and enhancement (CCE) and activity filter (AF). The proposed CCE method is able to address the phase mismatch issue caused by timing offset, thereby improving the

CSI quality. The dimension of input signals for DL networks can also be reduced. The AF method is designed to differentiate similar activities using the enhanced CSI features obtained from CCE. Then, we employed LSTM-RNN to classify extracted features into different categories, thereby effectively identifying different behaviors, especially for similar actions. Extensive experimental results demonstrated that the two proposed schemes show vast superiority in improving sensing accuracy and lowering training complexity.

In Chapter 4, we studied DL-based HAR for the scenario with a limited number of training samples from the testing environment or/and previously seen environments (PSEs). To facilitate an effective HAR, we proposed a novel HAR scheme using Matching Network with enhanced channel state information (MatNet-eCSI). A CSI Correlation Feature Extraction (CCFE) method was developed to condense the activity-related information and mitigate activity-unrelated data, thereby improving the quality of input signals for DL networks. Meanwhile, the dimension of input signals can be reduced, which significantly decreases the computational complexity. For feature extraction, we proposed to employ the matching network to learn and extract deep and hidden features from the output of CCFE. To better explore the properties of the matching network, an innovative training strategy was designed, which can bridge a good connection between the data sets from PSEs and samples from the testing environment. The proposed strategy can realize an effective feature extraction even for the situation in which only one PSE is available. For a successful HAR, only one sample for each activity from the testing environment and the data set from one PSE are required, which however is difficult for state-of-the-art methods to achieve. We designed and performed numerous experiments from a variety of aspects to evaluate the performance of the proposed scheme. Extensive results indicated that the proposed MatNet-eCSI is notably superior in achieving higher detection accuracy and decreasing training time, compared to the existing HAR methods.

In Chapter 5, we investigated the environmental robustness of DL-based HAR methods. Towards that end, we proposed two novel environment-robust HAR schemes drawing support from advanced signal processing algorithms and properties

of DL networks. We first developed an environment-robust channel state information based HAR by leveraging the properties of a matching network (MatNet) and enhanced features (HAR-MN-EF). A CSI cleaning and enhancement method (CSI-CE) was designed to improve the quality of input signals of MatNet, by condensing behavior-related information, eliminating behavior-unrelated data, and filtering out the noise. The extracted features are put into MatNet to facilitate environment-robust HAR. To further improve HAR performance, we proposed an innovative scheme, combining an activity-related feature extraction and enhancement (AFEE) method and Matching Network (AFEE-MatNet). We designed the AFEE approach to enhance the quality of input signals by mitigating environmental noise, preserving activity-related information, and reducing the size of input signals. Then, we proposed to use MatNet to learn and extract transferable features shared among source environments. Moreover, we developed a prediction checking and correction scheme to improve recognition accuracy. Extensive experiments were conducted to verify the performance of proposed schemes, and results demonstrated that our proposed schemes greatly outperform the other relevant HAR methods, both in recognition accuracy and training time.

6.2 Future Work

In this thesis, we studied critical issues of WiFi-based HAR using properties of DL networks and proposed several novel HAR schemes to improve sensing accuracy, reduce the number of required training samples and achieve environmental robustness, respectively. Some other interesting topics in DL-HAR are yet to be investigated, which would be a natural extension to this work in the future.

Multiple-person activity recognition

For multiple-person activity recognition, DL-HAR can identify behaviors performed by multiple persons simultaneously. This would be beneficial for large-scale HAR deployment and boost its applicability in practical scenarios. Therefore, multi-person HAR is an interesting topic requiring more attention. Based on that, we

may develop multi-person HAR approaches with the help of DL architectures and innovative signal processing techniques.

Through-the-wall activity recognition

For through-the-wall WiFi-based HAR, it is possible to detect and monitor human behaviors effectively and accurately across different rooms/environments. This would benefit a wide range of applications, including elder care, emergency rescue and safety surveillance. Therefore, it is of paramount importance to facilitate the through-the-wall WiFi-based HAR. However, WiFi-based HAR in the through-the-wall scenario is still an open research problem. Given that, we may propose HAR schemes for through-the-wall scenarios drawing support from DL networks and advanced signal processing methods.

Abbreviations

ABLSTM Bi-directional long short-term memory

AE-LRCN Autoencoder Long-term Recurrent Convolutional Network framework

AF Activity filter

AFEE Activity-related feature extraction and enhancement method

AFEE-MatNet Activity-related feature extraction and enhancement method and matching network

ARIE Activity-related information extraction

BiGRU Bidirectional Gated Recurrent Units

CCE CSI compensation and enhancement

CCFE CSI Correlation Feature Extraction

CFES CSI feature enhancement scheme

CFM Correlation feature matrix

CIR Channel impulse response

CM Conjugate multiplication

CNN Convolutional neural network

COTS Commercial off-the-shelf

CSI Channel state information

CSI-CE CSI cleaning and enhancement method

CSI-based HAR CSI-based human activity recognition

DL Deep learning

DL-based HAR Deep learning for human activity recognition

DLNs Deep learning networks

DLN-eCSI Deep Learning Networks with enhanced Channel State information

DM Differential method

DRL Deep reinforcement learning

DWT Discrete wavelet transform

EWMA Exponentially weighted moving average

FFT Fast Fourier Transform

HAR Human activity recognition

HAR-AF-DLN HAR using activity filter-based deep learning network

HAR-MN-EF HAR using matching network and enhanced features

HHT Hilbert-Huang Transform

KNN k nearest neighborhood

LM Local mean

LOS Line-of-sight

LSTM Long-short term memory

LSTM-RNN Long-short term memory recurrent neural networking

MatNet Matching network

MatNet-eCSI Matching Network with enhanced channel state information

MatNet-PCC MatNet with prediction checking and correction

mmWave millimeter wave

NIC Network interface card

Non-LOS Non-line-of-sight

PCA Principal component analysis

PCC Prediction checking and correction

PSE Previously seen environment

RFID Radio-frequency identification

RNN Recurrent neural networking

RSSI Received signal strength indicator

RSSI-based HAR RSSI for WiFi-based HAR

SAE Sparse autoencoder

SDAE Stacking denoising autoencoder

SVD Singular value decomposition

WS-HAR Wireless signal-based HAR

Bibliography

[1] L. Guo, H. Zhang, C. Wang, W. Guo, G. Diao, B. Lu, C. Lin, and L. Wang, "Towards CSI-based diversity activity recognition via LSTM-CNN encoder-decoder neural network," *Neurocomputing*, vol. 444, pp. 260–273, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S092523122031777X

[2] X. Yang, D. Liu, J. Liu, F. Yan, P. Chen, and Q. Niu, "Follower: A novel self-deployable action recognition framework," *Sensors*, vol. 21, no. 3, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/3/950

[3] Y. Fang, B. Sheng, H. Wang, and F. Xiao, "WiTransfer: A cross-scene transfer activity recognition system using WiFi," in *Proceedings of the ACM Turing Celebration Conference - China*, ser. ACM TURC'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 5963. [Online]. Available: https://doi.org/10.1145/3393527.3393538

[4] Y. Zhao, P. Tu, and M.-C. Chang, "Occupancy sensing and activity recognition with cameras and wireless sensors," in *Proceedings of the 2nd Workshop on Data Acquisition To Analysis*, ser. DATA'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 16. [Online]. Available: https://doi.org/10.1145/3359427.3361911

[5] H. Lee, C. R. Ahn, and N. Choi, "Exploiting multiple receivers for CSI-based activity classification using a hybrid CNN-LSTM model," in *Proceedings of the 1st ACM International Workshop on Device-Free Human Sensing*, ser.

DFHS'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 1821. [Online]. Available: https://doi.org/10.1145/3360773.3360878

[6] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys Tutorials*, vol. 15, no. 3, pp. 1192–1209, 2013.

[7] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu, and Y. Liu, "Deep learning for sensor-based human activity recognition: Overview, challenges and opportunities," 2021.

[8] I. Nirmal, A. Khamis, M. Hassan, W. Hu, and X. Zhu, "Deep learning for radio-based human sensing: Recent advances and future directions," 2021.

[9] B. van Berlo, A. Elkelany, T. Ozcelebi, and N. Meratnia, "Millimeter wave sensing: A review of application pipelines and building blocks," *IEEE Sensors Journal*, vol. 21, no. 9, pp. 10 332–10 368, 2021.

[10] J. Wang, L. Zhang, C. Wang, X. Ma, Q. Gao, and B. Lin, "Device-free human gesture recognition with generative adversarial networks," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7678–7688, 2020.

[11] J. Liu, G. Teng, and F. Hong, "Human activity sensing with wireless signals: A survey," *Sensors*, vol. 20, no. 4, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/4/1210

[12] C. Li, M. Liu, and Z. Cao, "WiHF: Enable user identified gesture recognition with WiFi," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020, pp. 586–595.

[13] L. Guo, L. Wang, J. Liu, and W. Zhou, "A survey on motion detection using WiFi signals," in *2016 12th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)*, 2016, pp. 202–206.

[14] H. Ren, H. Jin, C. Chen, H. Ghayvat, and W. Chen, "A novel cardiac auscultation monitoring system based on wireless sensing for healthcare,"

IEEE Journal of Translational Engineering in Health and Medicine, vol. 6, pp. 1–12, 2018.

[15] J. Liu, G. Teng, and F. Hong, "Human activity sensing with wireless signals: a survey," *Sensors*, vol. 20, no. 4, p. 1210, 2020.

[16] M. Z. Ozturk, C. Wu, B. Wang, and K. J. R. Liu, "GaitCube: Deep data cube learning for human recognition with millimeter-wave radio," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[17] B. Fu, N. Damer, F. Kirchbuchner, and A. Kuijper, "Sensing technology for human activity recognition: A comprehensive survey," *IEEE Access*, vol. 8, pp. 83 791–83 820, 2020.

[18] Z. Wang, S. Chen, W. Yang, and Y. Xu, "Environment-independent Wi-Fi human activity recognition with adversarial network," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3330–3334.

[19] H. Cai, B. Korany, C. R. Karanam, and Y. Mostofi, "Teaching RF to sense without RF training measurements," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 4, Dec. 2020. [Online]. Available: https://doi.org/10.1145/3432224

[20] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3D human pose construction using WiFi," in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '20. New York, NY, USA: Association for Computing Machinery, 2020. [Online]. Available: https://doi.org/10.1145/3372224.3380900

[21] H. Kang, Q. Zhang, and Q. Huang, "Context-aware wireless based cross domain gesture recognition," *IEEE Internet of Things Journal*, pp. 1–1, 2021.

[22] Z. Hao, Y. Duan, X. Dang, Y. Liu, and D. Zhang, "Wi-SL: Contactless fine-grained gesture recognition uses channel state information," *Sensors*,

vol. 20, no. 14, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/14/4025

[23] J. M. Rocamora, I.-H. Ho, and M.-W. Mak, "Gaussian models for CSI fingerprinting in practical indoor environment identification," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[24] Y. Xie, Z. Li, and M. Li, "Precise power delay profiling with commodity Wi-Fi," *IEEE Transactions on Mobile Computing*, vol. 18, no. 6, pp. 1342–1355, 2019.

[25] 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

[26] W. Huang, Y. Liu, S. Zhu, S. Wang, and Y. Zhang, "TSCNN: A 3D convolutional activity recognition network based on RFID RSSI," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[27] J. Huang, B. Liu, P. Liu, C. Chen, N. Xiao, Y. Wu, C. Zhang, and N. Yu, "Towards anti-interference WiFi-based activity recognition system using interference-independent phase component," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications*, 2020, pp. 576–585.

[28] Z. Wang, Z. Huang, C. Zhang, W. Dou, Y. Guo, and D. Chen, "CSI-based human sensing using model-based approaches: a survey," *Journal of Computational Design and Engineering*, vol. 8, no. 2, pp. 510–523, 02 2021. [Online]. Available: https://doi.org/10.1093/jcde/qwab003

[29] X. Li, L. Chang, F. Song, J. Wang, X. Chen, Z. Tang, and Z. Wang, "CrossGR: Accurate and low-cost cross-target gesture recognition using Wi-Fi," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 1, Mar. 2021. [Online]. Available: https://doi.org/10.1145/3448100

[30] K. Sankhe, D. Jaisinghani, and K. Chowdhury, "CSIscan: Learning CSI for efficient access point discovery in dense WiFi networks," in *2020 IEEE 28th International Conference on Network Protocols (ICNP)*, 2020, pp. 1–12.

[31] Y. Ma, G. Zhou, and S. Wang, "WiFi sensing with channel state information: A survey," *ACM Comput. Surv.*, vol. 52, no. 3, Jun. 2019. [Online]. Available: https://doi.org/10.1145/3310194

[32] P. Khan, B. S. K. Reddy, A. Pandey, S. Kumar, and M. Youssef, "Differential channel-state-information-based human activity recognition in IoT networks," *IEEE Internet of Things Journal*, vol. 7, no. 11, pp. 11 290–11 302, 2020.

[33] C. Li, Z. Cao, and Y. Liu, "Deep AI enabled ubiquitous wireless sensing: A survey," *ACM Comput. Surv.*, vol. 54, no. 2, Mar. 2021. [Online]. Available: https://doi.org/10.1145/3436729

[34] Q. Zhu, Z. Chen, and Y. C. Soh, "A novel semisupervised deep learning method for human activity recognition," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 7, pp. 3821–3830, 2019.

[35] S. Aarthi and S. Juliet, "A comprehensive study on human activity recognition," in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, 2021, pp. 59–63.

[36] J. Zhang, F. Wu, B. Wei, Q. Zhang, H. Huang, S. W. Shah, and J. Cheng, "Data augmentation and dense-LSTM for human activity recognition using WiFi signal," *IEEE Internet of Things Journal*, vol. 8, no. 6, pp. 4628–4641, 2021.

[37] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019, deep Learning for Pattern Recognition. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016786551830045X

[38] R. H. Venkatnarayan, S. Mahmood, and M. Shahzad, "WiFi based multi-user gesture recognition," *IEEE Transactions on Mobile Computing*, vol. 20, no. 3, pp. 1242–1256, 2021.

[39] M. Sulaiman, S. A. Hassan, and H. Jung, "True Detect: Deep learning-based device-free activity recognition using WiFi," in *2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2020, pp. 1–5.

[40] X. Yang, R. Cao, M. Zhou, and L. Xie, "Temporal-frequency attention-based human activity recognition using commercial WiFi devices," *IEEE Access*, vol. 8, pp. 137 758–137 769, 2020.

[41] X. Huang and M. Dai, "Indoor device-free activity recognition based on radio signal," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5316–5329, June 2017.

[42] D. Zhang, Y. Hu, Y. Chen, and B. Zeng, "BreathTrack: Tracking indoor human breath status via commodity WiFi," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3899–3911, 2019.

[43] D. Zekri, T. Delot, M. Thilliez, S. Lecomte, and M. Desertot, "A framework for detecting and analyzing behavior changes of elderly people over time using learning techniques," *Sensors*, vol. 20, no. 24, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/24/7112

[44] T. Shany, S. J. Redmond, M. R. Narayanan, and N. H. Lovell, "Sensors-based wearable systems for monitoring of human movement and falls," *IEEE Sensors Journal*, vol. 12, no. 3, pp. 658–670, 2012.

[45] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *IEEE Sensors Journal*, vol. 15, no. 3, pp. 1321–1330, 2015.

[46] T. B. Moeslund, A. Hilton, and V. Krger, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006, special

Issue on Modeling People: Vision-based understanding of a persons shape, appearance, movement and behaviour. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1077314206001263

[47] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976–990, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885609002704

[48] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Comput. Surv.*, vol. 48, no. 4, Feb. 2016. [Online]. Available: https://doi.org/10.1145/2871183

[49] C. Xu, P. H. Pathak, and P. Mohapatra, "Finger-writing with smartwatch: A case for finger and hand gesture recognition using smartwatch," in *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*, ser. HotMobile '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 914. [Online]. Available: https://doi.org/10.1145/2699343.2699350

[50] J. Wang, D. Vasisht, and D. Katabi, "RF-IDraw: Virtual touch screen in the air using RF signals," *SIGCOMM Comput. Commun. Rev.*, vol. 44, no. 4, p. 235246, Aug. 2014. [Online]. Available: https://doi.org/10.1145/2740070.2626330

[51] J. Parkka, M. Ermes, P. Korpipaa, J. Mantyjarvi, J. Peltola, and I. Korhonen, "Activity classification using realistic data from wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, vol. 10, no. 1, pp. 119–128, 2006.

[52] N. Hegde, T. Zhang, G. Uswatte, E. Taub, J. Barman, S. McKay, A. Taylor, D. M. Morris, A. Griffin, and E. S. Sazonov, "The Pediatric SmartShoe: Wearable sensor system for ambulatory monitoring of physical activity and Gait," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 2, pp. 477–486, 2018.

[53] J. M. Fontana, M. Farooq, and E. Sazonov, "Automatic ingestion monitor: A novel wearable device for monitoring of ingestive behavior," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 6, pp. 1772–1779, 2014.

[54] D. Castro, W. Coral, C. Rodriguez, J. Cabra, and J. Colorado, "Wearable-based human activity recognition using an IoT approach," *Journal of Sensor and Actuator Networks*, vol. 6, no. 4, 2017. [Online]. Available: https://www.mdpi.com/2224-2708/6/4/28

[55] J. M. Rehg and T. Kanade, "Visual tracking of high DOF articulated structures: An application to human hand tracking," in *Computer Vision — ECCV '94*, J.-O. Eklundh, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 1994, pp. 35–46.

[56] S. Park, J. Park, M. Al-masni, M. Al-antari, M. Uddin, and T.-S. Kim, "A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services," *Procedia Computer Science*, vol. 100, pp. 78–84, 2016, international Conference on ENTERprise Information Systems/International Conference on Project MANagement/International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050916322943

[57] M. H. Siddiqi, R. Ali, M. S. Rana, E.-K. Hong, E. S. Kim, and S. Lee, "Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis," *Sensors*, vol. 14, no. 4, pp. 6370–6392, 2014. [Online]. Available: https://www.mdpi.com/1424-8220/14/4/6370

[58] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim, "Robust human activity recognition from depth video using spatiotemporal multi-fused features," *Pattern Recognition*, vol. 61, pp. 295–308, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320316302126

[59] Y. Gu, X. Zhang, Z. Liu, and F. Ren, "BeSense: Leveraging WiFi channel data and computational intelligence for behavior analysis," *IEEE Computational Intelligence Magazine*, vol. 14, no. 4, pp. 31–41, 2019.

[60] Y. He, Y. Chen, Y. Hu, and B. Zeng, "WiFi vision: Sensing, recognition, and detection with commodity MIMO-OFDM WiFi," *IEEE Internet of Things Journal*, vol. 7, no. 9, pp. 8296–8317, 2020.

[61] T. S. Murray, D. R. Mendat, K. A. Sanni, P. O. Pouliquen, and A. G. Andreou, "Bio-inspired human action recognition with a micro-doppler sonar system," *IEEE Access*, vol. 6, pp. 28 388–28 403, 2018.

[62] W. Mao, J. He, and L. Qiu, "CAT: High-precision acoustic motion tracking," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 6981. [Online]. Available: https://doi.org/10.1145/2973750.2973755

[63] S. Yun, Y.-C. Chen, H. Zheng, L. Qiu, and W. Mao, "Strata: Fine-grained acoustic-based device-free tracking," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, ser. MobiSys '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 1528. [Online]. Available: https://doi.org/10.1145/3081333.3081356

[64] W. Wang, A. X. Liu, and K. Sun, "Device-free gesture tracking using acoustic signals," in *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, ser. MobiCom '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 8294. [Online]. Available: https://doi.org/10.1145/2973750.2973764

[65] K. Ling, H. Dai, Y. Liu, and A. X. Liu, "UltraGesture: Fine-grained gesture sensing and recognition," in *2018 15th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2018, pp. 1–9.

[66] Z. Wang, Y. Hou, K. Jiang, W. Dou, C. Zhang, Z. Huang, and Y. Guo, "Hand gesture recognition based on active ultrasonic sensing of smartphone: A survey," *IEEE Access*, vol. 7, pp. 111 897–111 922, 2019.

[67] Y. Wang, J. Shen, and Y. Zheng, "Push the limit of acoustic gesture recognition," *IEEE Transactions on Mobile Computing*, no. 01, pp. 1–1, oct 5555.

[68] H. Yin, A. Zhou, L. Liu, N. Wang, and H. Ma, "Ubiquitous Writer: Robust text input for small mobile devices via acoustic sensing," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5285–5296, 2019.

[69] M. A. A. Al-qaness, M. Abd Elaziz, S. Kim, A. A. Ewees, A. A. Abbasi, Y. A. Alhaj, and A. Hawbani, "Channel state information from pure communication to sense and track human motion: A survey," *Sensors*, vol. 19, no. 15, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/15/3329

[70] Z. Yu, H. Du, D. Xiao, Z. Wang, Q. Han, and B. Guo, "Recognition of human computer operations based on keystroke sensing by smartphone microphone," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 1156–1168, 2018.

[71] J. Yang, H. Zou, H. Jiang, and L. Xie, "Device-free occupant activity sensing using WiFi-enabled iot devices for smart homes," *IEEE Internet of Things Journal*, vol. 5, no. 5, pp. 3991–4002, Oct 2018.

[72] X. Li, D. Zhang, J. Xiong, Y. Zhang, S. Li, Y. Wang, and H. Mei, "Training-free human vitality monitoring using commodity Wi-Fi devices," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 3, Sep. 2018. [Online]. Available: https://doi.org/10.1145/3264931

[73] M. Shahzad and S. Zhang, "Augmenting user identification with WiFi based gesture recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 3, Sep. 2018. [Online]. Available: https://doi.org/10.1145/3264944

[74] J. Wang, Q. Gao, M. Pan, and Y. Fang, "Device-free wireless sensing: Challenges, opportunities, and applications," *IEEE Network*, vol. 32, no. 2, pp. 132–137, 2018.

[75] K. Bregar, A. Hrovat, and M. Mohori, "UWB radio-based motion detection system for assisted living," *Sensors*, vol. 21, no. 11, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/11/3631

[76] E. Soltanaghaei, R. A. Sharma, Z. Wang, A. Chittilappilly, A. Luong, E. Giler, K. Hall, S. Elias, and A. Rowe, "Robust and practical WiFi human sensing using on-device learning with a domain adaptive model," in *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*, ser. BuildSys '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 150159. [Online]. Available: https://doi.org/10.1145/3408308.3427983

[77] Y. Bai and X. Wang, "CARIN: Wireless CSI-based driver activity recognition under the interference of passengers," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 1, Mar. 2020. [Online]. Available: https://doi.org/10.1145/3380992

[78] 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.

[79] 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.

[80] B. Tan, Q. Chen, K. Chetty, K. Woodbridge, W. Li, and R. Piechocki, "Exploiting WiFi channel state information for residential healthcare informatics," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 130–137, May 2018.

[81] J. Wang, X. Zhang, Q. Gao, X. Ma, X. Feng, and H. Wang, "Device-free simultaneous wireless localization and activity recognition with wavelet feature," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 2, pp. 1659–1669, Feb 2017.

[82] G. Lan, M. F. Imani, P. d. Hougne, W. Hu, D. R. Smith, and M. Gorlatova, "Wireless sensing using dynamic metasurface antennas: Challenges and opportunities," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 66–71, 2020.

[83] H. Lau, R. McConville, M. J. Bocus, R. J. Piechocki, and R. Santos-Rodríguez, "Self-supervised WiFi-based activity recognition," *CoRR*, vol. abs/2104.09072, 2021. [Online]. Available: https://arxiv.org/abs/2104.09072

[84] C. Dian, D. Wang, Q. Zhang, R. Zhao, and Y. Yu, "Towards domain-independent complex and fine-grained gesture recognition with RFID," *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. ISS, Nov. 2020. [Online]. Available: https://doi.org/10.1145/3427315

[85] B. Sheng, F. Xiao, Y. Fang, and H. Wang, "WiFi based passive action recognition with fully-connected network," in *Proceedings of the ACM Turing Celebration Conference - China*, ser. ACM TURC'20. New York, NY, USA: Association for Computing Machinery, 2020, p. 113117. [Online]. Available: https://doi.org/10.1145/3393527.3393547

[86] 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

[87] E. Cianca, M. D. Sanctis, and S. D. Domenico, "Radios as sensors," *IEEE Internet of Things Journal*, vol. 4, no. 2, pp. 363–373, April 2017.

[88] Y. Gu, F. Ren, and J. Li, "PAWS: Passive human activity recognition based on WiFi ambient signals," *IEEE Internet of Things Journal*, vol. 3, no. 5, pp. 796–805, Oct 2016.

[89] Z. Wang, B. Guo, Z. Yu, and X. Zhou, "Wi-Fi csi-based behavior recognition: From signals and actions to activities," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 109–115, May 2018.

[90] N. Lakshmanan, I. Bang, M. S. Kang, J. Han, and J. T. Lee, "SurFi: Detecting surveillance camera looping attacks with Wi-Fi channel state information," in *Proceedings of the 12th Conference on Security and Privacy in Wireless and Mobile Networks*, ser. WiSec '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 239244. [Online]. Available: https://doi.org/10.1145/3317549.3324928

[91] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "RadHAR: Human activity recognition from point clouds generated through a millimeter-wave radar," in *Proceedings of the 3rd ACM Workshop on Millimeter-Wave Networks and Sensing Systems*, ser. mmNets'19. New York, NY, USA: Association for Computing Machinery, 2019, p. 5156. [Online]. Available: https://doi.org/10.1145/3349624.3356768

[92] R. Zhang and S. Cao, "Real-time human motion behavior detection via CNN using mmWave radar," *IEEE Sensors Letters*, vol. 3, no. 2, pp. 1–4, 2019.

[93] P. Zhao, C. X. Lu, J. Wang, C. Chen, W. Wang, N. Trigoni, and A. Markham, "mID: Tracking and identifying people with millimeter wave radar," in *2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2019, pp. 33–40.

[94] H. Yan, Y. Zhang, Y. Wang, and K. Xu, "WiAct: A passive WiFi-based human activity recognition system," *IEEE Sensors Journal*, vol. 20, no. 1, pp. 296–305, 2020.

[95] D. Wu, D. Zhang, C. Xu, H. Wang, and X. Li, "Device-free wifi human sensing: From pattern-based to model-based approaches," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 91–97, OCTOBER 2017.

[96] 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.

[97] B. Sheng, Y. Fang, F. Xiao, and L. Sun, "An accurate device-free action recognition system using two-stream network," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 7, pp. 7930–7939, 2020.

[98] J. Wilson and N. Patwari, "Radio tomographic imaging with wireless networks," *IEEE Transactions on Mobile Computing*, vol. 9, no. 5, pp. 621–632, May 2010.

[99] S. Orphomma and N. Swangmuang, "Exploiting the wireless RF fading for human activity recognition," in *2013 10th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, May 2013, pp. 1–5.

[100] H. Huang and S. Lin, "WiDet: Wi-Fi based device-free passive person detection with deep convolutional neural networks," *Computer Communications*, vol. 150, pp. 357–366, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140366419301331

[101] P. Barsocchi, "Position recognition to support bedsores prevention," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 53–59, Jan 2013.

[102] T. Gong, Y. Kim, R. Choi, J. Shin, and S.-J. Lee, "Adapting to unknown conditions in learning-based mobile sensing," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2021.

[103] Y. Lu, F. Wu, S. Tang, L. Kong, and G. Chen, "Pushing the limit of CSI-based activity recognition: An enhanced approach via packet reconstruction," in *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2019, pp. 1–9.

[104] H. Abdelnasser, M. Youssef, and K. A. Harras, "WiGest: A ubiquitous WiFi-based gesture recognition system," in *2015 IEEE Conference on Computer Communications (INFOCOM)*, April 2015, pp. 1472–1480.

[105] S. Shi, S. Sigg, and Y. Ji, "Joint localization and activity recognition from ambient FM broadcast signals," in *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, ser. UbiComp '13 Adjunct. New York, NY, USA: ACM, 2013, pp. 521–530. [Online]. Available: http://doi.acm.org/10.1145/2494091.2497610

[106] S. Sigg, S. Shi, F. Buesching, Y. Ji, and L. Wolf, "Leveraging RF-channel fluctuation for activity recognition: Active and passive systems, continuous and RSSI-based signal features," ser. MoMM '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 4352. [Online]. Available: https://doi.org/10.1145/2536853.2536873

[107] S. Kianoush, S. Savazzi, F. Vicentini, V. Rampa, and M. Giussani, "Device-free rf human body fall detection and localization in industrial workplaces," *IEEE Internet of Things Journal*, vol. 4, no. 2, pp. 351–362, 2017.

[108] S. Sigg, M. Scholz, S. Shi, Y. Ji, and M. Beigl, "RF-sensing of activities from non-cooperative subjects in device-free recognition systems using ambient and local signals," *IEEE Transactions on Mobile Computing*, vol. 13, no. 4, pp. 907–920, 2014.

[109] C. Siebert, M. Leng, S. G. Razul, C. M. S. See, and G. Wang, "Human motion detection and classification using ambient WiFi signals," in *2017 Sensor Signal Processing for Defence Conference (SSPD)*, 2017, pp. 1–5.

[110] A. R. Guraliuc, P. Barsocchi, F. Potorti, and P. Nepa, "Limb movements classification using wearable wireless transceivers," *IEEE Transactions on Information Technology in Biomedicine*, vol. 15, no. 3, pp. 474–480, May 2011.

[111] C. Liu, D. Fang, Z. Yang, H. Jiang, X. Chen, W. Wang, T. Xing, and L. Cai, "RSS distribution-based passive localization and its application in sensor networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2883–2895, April 2016.

[112] X. Dang, Y. Cao, Z. Hao, and Y. Liu, "WiGId: Indoor group identification with CSI-based random forest," *Sensors*, vol. 20, no. 16, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/16/4607

[113] W. Zhang, S. Zhou, L. Yang, L. Ou, and Z. Xiao, "WiFiMap+: High-level indoor semantic inference with WiFi human activity and environment," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7890–7903, 2019.

[114] M. Muaaz, A. Chelli, and M. Ptzold, "WiHAR: From Wi-Fi channel state information to unobtrusive human activity recognition," in *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020, pp. 1–7.

[115] F. Zhang, K. Niu, J. Xiong, B. Jin, T. Gu, Y. Jiang, and D. Zhang, "Towards a diffraction-based sensing approach on human activity recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 1, Mar. 2019. [Online]. Available: https://doi.org/10.1145/3314420

[116] H. F. Thariq Ahmed, H. Ahmad, and A. C.V., "Device free human gesture recognition using wi-fi csi: A survey," *Engineering Applications of Artificial Intelligence*, vol. 87, p. 103281, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197619302441

[117] 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

[118] H. Fei, F. Xiao, B. Sheng, H. Huang, and L. Sun, "Motion path reconstruction in indoor environment using commodity Wi-Fi," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 8, pp. 7668–7678, 2019.

[119] 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.

[120] L. Wenyuan, W. Siyang, W. Lin, L. Binbin, S. Xing, and J. Nan, "From Lens to Prism: Device-free modeling and recognition of multi-part activities," *IEEE Access*, vol. 6, pp. 36 271–36 282, 2018.

[121] J. Huang, B. Liu, C. Chen, H. Jin, Z. Liu, C. Zhang, and N. Yu, "Towards anti-interference human activity recognition based on WiFi subcarrier correlation selection," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 6, pp. 6739–6754, 2020.

[122] 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.

[123] S. Zhong, Y. Huang, R. Ruby, L. Wang, Y. X. Qiu, and K. Wu, "Wi-fire: Device-free fire detection using WiFi networks," in *2017 IEEE International Conference on Communications (ICC)*, May 2017, pp. 1–6.

[124] H. Zhu, Y. Zhuo, Q. Liu, and S. Chang, "π-Splicer: Perceiving accurate CSI phases with commodity WiFi devices," *IEEE Transactions on Mobile Computing*, vol. 17, no. 9, pp. 2155–2165, Sept 2018.

[125] C. Xiao, D. Han, Y. Ma, and Z. Qin, "CsiGAN: Robust channel state information-based activity recognition with GANs," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 191–10 204, 2019.

[126] S. Fang, C. Li, W. Lu, Z. Xu, and Y. Chien, "Enhanced device-free human detection: Efficient learning from phase and amplitude of channel state information," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 3048–3051, 2019.

[127] K. Chen, L. Yao, D. Zhang, X. Wang, X. Chang, and F. Nie, "A semisupervised recurrent convolutional attention model for human activity recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 5, pp. 1747–1756, 2020.

[128] 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.

[129] 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.

[130] Q. Zhou, J. Xing, J. Li, and Q. Yang, "A device-free number gesture recognition approach based on deep learning," in *2016 12th International Conference on Computational Intelligence and Security (CIS)*, Dec 2016, pp. 57–63.

[131] M. Edel and E. K?ppe, "Binarized-BLSTM-RNN based human activity recognition," in *2016 International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Oct 2016, pp. 1–7.

[132] X. Wu, Z. Chu, P. Yang, C. Xiang, X. Zheng, and W. Huang, "TW-See: Human activity recognition through the wall with commodity Wi-Fi devices," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 1, pp. 306–319, Jan 2019.

[133] W. Li, R. J. Piechocki, K. Woodbridge, C. Tang, and K. Chetty, "Passive WiFi radar for human sensing using a stand-alone access point," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 3, pp. 1986–1998, 2021.

[134] R. Zhang, X. Jing, S. Wu, C. Jiang, J. Mu, and F. R. Yu, "Device-free wireless sensing for human detection: The deep learning perspective," *IEEE Internet of Things Journal*, vol. 8, no. 4, pp. 2517–2539, 2021.

[135] J. Wang, Q. Gao, X. Ma, Y. Zhao, and Y. Fang, "Learning to sense: Deep learning for wireless sensing with less training efforts," *IEEE Wireless Communications*, vol. 27, no. 3, pp. 156–162, 2020.

[136] 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.

[137] 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.

[138] X. Gao, H. Luo, Q. Wang, F. Zhao, L. Ye, and Y. Zhang, "A human activity recognition algorithm based on stacking denoising autoencoder and LightGBM," *Sensors*, vol. 19, no. 4, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/4/947

[139] 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.

[140] R. Alazrai, M. Hababeh, B. A. Alsaify, M. Z. Ali, and M. I. Daoud, "An end-to-end deep learning framework for recognizing human-to-human interactions using Wi-Fi signals," *IEEE Access*, vol. 8, pp. 197 695–197 710, 2020.

[141] Z. Chen, C. Cai, T. Zheng, J. Luo, J. Xiong, and X. Wang, "RF-based human activity recognition using signal adapted convolutional neural network," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2021.

[142] D. Konings, R. Grace, and F. Alam, "A stacked neural network-based machine learning framework to detect activities and falls within multiple indoor environments using Wi-Fi CSI," *IEEE Sensors Letters*, vol. 5, no. 5, pp. 1–4, 2021.

[143] B. Chikhaoui, F. Gouineau, and M. Sotir, "A CNN based transfer learning model for automatic activity recognition from accelerometer sensors," in *Machine Learning and Data Mining in Pattern Recognition*, P. Perner, Ed. Cham: Springer International Publishing, 2018, pp. 302–315.

[144] Z. Zhou, Y. Zhang, X. Yu, P. Yang, X.-Y. Li, J. Zhao, and H. Zhou, "XHAR: Deep domain adaptation for human activity recognition with smart devices," in *2020 17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, 2020, pp. 1–9.

[145] J. Ding, Y. Wang, and X. Fu, "Wihi: WiFi based human identity identification using deep learning," *IEEE Access*, vol. 8, pp. 129 246–129 262, 2020.

[146] 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, OCTOBER 2017.

[147] 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.

[148] 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.

[149] H. Wang, D. Zhang, Y. Wang, J. Ma, Y. Wang, and S. Li, "RT-Fall: A real-time and contactless fall detection system with commodity WiFi devices,"

IEEE Transactions on Mobile Computing, vol. 16, no. 2, pp. 511–526, Feb 2017.

[150] Z. Chen, L. Zhang, C. Jiang, Z. Cao, and W. Cui, "WiFi CSI based passive human activity recognition using attention based BLSTM," *IEEE Transactions on Mobile Computing*, vol. 18, no. 11, pp. 2714–2724, 2019.

[151] R. Zhao, X. Ma, X. Liu, and J. Liu, "An end-to-end network for continuous human motion recognition via radar radios," *IEEE Sensors Journal*, vol. 21, no. 5, pp. 6487–6496, 2021.

[152] T. Zheng, Z. Chen, S. Ding, and J. Luo, "Enhancing RF sensing with deep learning: A layered approach," *IEEE Communications Magazine*, vol. 59, no. 2, pp. 70–76, 2021.

[153] X. Zhang and J. Zhang, "Subject independent human activity recognition with foot IMU data," in *2019 15th International Conference on Mobile Ad-Hoc and Sensor Networks (MSN)*, 2019, pp. 240–246.

[154] J. Wang, Y. Zhao, X. Ma, Q. Gao, M. Pan, and H. Wang, "Cross-scenario device-free activity recognition based on deep adversarial networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 5, pp. 5416–5425, 2020.

[155] J. Hu, H. Zhang, K. Bian, M. D. Renzo, Z. Han, and L. Song, "Metasensing: Intelligent metasurface assisted RF 3D sensing by deep reinforcement learning," 2020.

[156] X. Ding, T. Jiang, Y. Zhong, S. Wu, J. Yang, and W. Xue, "Improving WiFi-based human activity recognition with adaptive initial state via one-shot learning," in *2021 IEEE Wireless Communications and Networking Conference (WCNC)*, 2021, pp. 1–6.

[157] X. Ding, T. Jiang, Y. Li, W. Xue, and Y. Zhong, "Device-free location-independent human activity recognition using transfer learning based on CNN," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2020, pp. 1–6.

[158] B. Sheng, F. Xiao, L. Sha, and L. Sun, "Deep spatialtemporal model based cross-scene action recognition using commodity WiFi," *IEEE Internet of Things Journal*, vol. 7, no. 4, pp. 3592–3601, 2020.

[159] H. Xue, W. Jiang, C. Miao, F. Ma, S. Wang, Y. Yuan, S. Yao, A. Zhang, and L. Su, "DeepMV: Multi-view deep learning for device-free human activity recognition," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 4, no. 1, Mar. 2020. [Online]. Available: https://doi.org/10.1145/3380980

[160] 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.

[161] 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.

[162] F. Wang, J. Liu, and W. Gong, "Multi-adversarial in-car activity recognition using RFIDs," *IEEE Transactions on Mobile Computing*, vol. 20, no. 6, pp. 2224–2237, 2021.

[163] F. Wang, J. Liu, and W. Gong, "WiCAR: WiFi-based in-car activity recognition with multi-adversarial domain adaptation," in *2019 IEEE/ACM 27th International Symposium on Quality of Service (IWQoS)*, 2019, pp. 1–10.

[164] N. Xiao, P. Yang, Y. Yan, H. Zhou, X. Li, and H. Du, "Motion-Fi$^+$+: Recognizing and counting repetitive motions with wireless backscattering," *IEEE Transactions on Mobile Computing*, vol. 20, no. 05, pp. 1862–1876, may 2021.

[165] 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.

[166] B. Guo, Y. J. Chen, N. Lane, Y. Liu, and Z. Yu, "Behavior recognition based on Wi-Fi CSI: Part 2," *IEEE Communications Magazine*, vol. 56, no. 5, pp. 108–108, 2018.

[167] S. Savazzi, S. Kianoush, V. Rampa, and U. Spagnolini, "Cellular data analytics for detection and discrimination of body movements," *IEEE Access*, vol. 6, pp. 51 484–51 499, 2018.

[168] C. Xiao, Y. Lei, Y. Ma, F. Zhou, and Z. Qin, "DeepSeg: Deep-learning-based activity segmentation framework for activity recognition using WiFi," *IEEE Internet of Things Journal*, vol. 8, no. 7, pp. 5669–5681, 2021.

[169] Y. Li, T. Jiang, X. Ding, and Y. Wang, "Location-free CSI based activity recognition with angle difference of arrival," in *2020 IEEE Wireless Communications and Networking Conference (WCNC)*, 2020, pp. 1–6.

[170] Y. Ma, S. Arshad, S. Muniraju, E. Torkildson, E. Rantala, K. Doppler, and G. Zhou, "Location- and person-independent activity recognition with WiFi, deep neural networks, and reinforcement learning," *ACM Trans. Internet Things*, vol. 2, no. 1, Jan. 2021. [Online]. Available: https://doi.org/10.1145/3424739

[171] 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

[172] 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

[173] 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

[174] 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

[175] 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.

[176] 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.

[177] 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.

[178] 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.

[179] 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.

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

[181] 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

[182] J.-Y. Chang, K.-Y. Lee, Y.-L. Wei, K. C.-J. Lin, and W. Hsu, "Location-independent WiFi action recognition via vision-based methods," in *Proceedings of the 24th ACM International Conference on Multimedia*, ser. MM '16. New York, NY, USA: ACM, 2016, pp. 162–166. [Online]. Available: http://doi.acm.org/10.1145/2964284.2967203

[183] 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.

[184] 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.

[185] 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.

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

[187] 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.

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

[189] 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

[190] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[191] X. Li, D. Zhang, Q. Lv, J. Xiong, S. Li, Y. Zhang, and H. Mei, "IndoTrack: Device-free indoor human tracking with commodity Wi-Fi," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, Sep. 2017. [Online]. Available: https://doi.org/10.1145/3130940