Angle-insensitive Human Motion and Posture Recognition Based on 2D FMCW MIMO Radar and Deep Learning Classifiers.

Master Thesis Report

Yubin Zhao

Angle-insensitive Human Motion and Posture Recognition Based on 2D FMCW MIMO Radar and Deep Learning Classifiers.

Dissertation

to obtain the degree of Master of Science
in Electrical Engineering
at Delft University of Technology
to be defended publicly on 22nd February 2022

by

Yubin Zhao

Born in Heilongjiang Province, China

This thesis has been approved by

promotor: Prof. DSc. A. Yarovoy

Thesis committee:

Prof. DSc. A. Yarovoy,	Technische Universiteit Delft
Dr. F. Fioranelli,	Technische Universiteit Delft
Dr. M. Zúñiga,	Technische Universiteit Delft

*Keywords: Deep Learning, Human Motion and Posture Recognition,. . .
Imaging Radar, Multiple Angle Classifier*

An electronic version of this dissertation is available at
`http://repository.tudelft.nl/`.

Acknowledgements

This master journey started in February 2020, and ends in February 2022. Now looking back to this fruitful two years, I had a great study and research experience. Despite the minor yet long-lasting difficulties brought by Covid, and the strong and frequent rains in Netherlands, I really enjoyed every moment of being here. Sincere thanks to everyone who put their effort in creating such a creative and comfortable environment, and to be more specific ...

I am deeply grateful to my daily supervisor Dr. Francesco Fioranelli for giving full freedom to choose the topic I am interested in, for creating a productive atmosphere and most importantly for his academic inputs throughout this period. My heartfelt appreciation also goes to my responsible supervisor Prof. DSc. Alexander Yarovoy, his prompt and efficient answers on any radar-related topics are the basis of this work. I would like to express my gratitude to many of the peer students, Xingzhuo, Liyuan, Zhongyuan, Zhuangzhuang, Yongdian, Simin and Yuxiang for participating in the measurements. Many thanks to MS3 PhD candidates for their inspiring insights as well as their kindness to contribute.

With no doubt, I would also like to thank my open-minded families who financially and mentally supported me during these years of study, my friends here in Netherlands, back in China and in other corners of the world who shared their happy moments with me, and my dearest girlfriend Shiyu who has been under the pressure of distance relationship.

Again, it has been a hard time in such a global pandemic, without anyone of the aforementioned, I may not be here. Thank you all.

ABSTRACT

Nowadays, the aging problem is shaking the root of the healthcare system in many countries, an automatic human activity recognition (HAR) is seen as a promising solution to that problem. In particular, radar-based HAR attracts people's attention thanks to its respect for privacy and functionality in poor lighting conditions. With a lot of research paying attention to this topic, there is still a lack of conclusive and practical methods. In particular, it is realized that dynamic motions at large aspect angles close to 90° or static postures have not been investigated in-depth as a part of the radar-based HAR problem. To extensively investigate this type of problem, we propose to use mm-wave FMCW MIMO radar to obtain accurate information of the human subject.

This thesis work aims to fully exploit the six dimensions of information provided by an imaging radar: range, azimuth, elevation, velocity, power and time. Two complementary data representations- point cloud and spectrogram- are utilized to represent these dimensions of information. A signal processing flow is implemented to generate the desired data representations. A hierarchical pipeline consisting of three cascaded deep learning-based classification modules is proposed to process the input data. Particularly, human orientation classification is achieved through the so-called "T-Net" network learning the geometric distribution of point clouds. The positive contribution of each module in the proposed pipeline is validated via an ablation study. The superior performances of the proposed pipeline are also established by comparing with those of the state-of-the-art baselines. The robustness of the proposed pipeline concerning a noisy environment is also discussed. It is also presented that the size of the aperture of imaging radar plays an important role in such HAR problems.

Contents

LIST OF FIGURES

List of Tables

List of Algorithms

ABBREVIATIONS

ASM Angle Sensitivity Matrix. 50, 55

ASV Angle Sensitivity Vector. 50, 55

BS Batch Size. 49

CA Cell-Averaging. 26

CFAR Constant False Alarm Rate. 26, 27, 29, 66

CNN Convolutional Neural Network. 2, 6, 7, 10, 11

CUT Cell Under Test. 26, 27, 35

DL Deep Learning. 2, 7–10, 20, 34, 66

FFT Fast Fourier Transform. 25–27

FMCW Frequency-Modulated Continuous-Wave. 13, 15, 23

GAN Generative Adversarial Network. 9, 66

GRU Gated Recurrent Unit. 7

HAR Human Activity Recognition. 1–10, 13, 17, 21, 25, 26, 30, 33, 34, 36, 40, 45, 63–65

KNN K-Nearest Neighbor. 2, 6, 7

LR Learning Rate. 49

LSTM Long-Short Time Memory. 7

MAC Multiple Angle Classifier. 20, 21, 55

MIMO Multiple-Input and Multiple-Output. 5, 13, 14, 23

ML Machine Learning. 2, 7, 8

mm-Wave millimeter wave. 12, 16, 22, 45

OS Ordered-Statistic. 26, 29

PC Point Cloud. 6, 11, 12, 20–22, 27, 29–31, 34, 40, 44–46, 54, 61, 62, 64–66

RCS Radar Cross Section. 20, 23

RNN Recurrent Neural Network. 2, 7, 11, 62

RX Receiver. 13, 17, 23–25

SAC Single Angle Classifier. 20, 21, 54, 55

STFT Short-Time Fourier Transform. 6

SVM Support Vector Machine. 2, 6, 7

TL Transfer Learning. 8–10

TX Transmitter. 13, 17, 23–25

Nomenclature

AF Array factor of antenna array. 13

$F1-score$ F1 score. 49

P_a Classification accuracy. 49

R_{max} Maximum measurement range. 15

$\bar{X}$ Mean value of the product of ASV and ASM to quantify the angle sensitivity of the classifier. 50

ΔR Range resolution. 15

$\Delta\phi$ Angular resolution in azimuth. 16

$\Delta\theta$ Angular resolution in elevation. 16

Δv Velocity resolution. 15

$\mathscr{L}$ Loss function (also known as objective function). 31

v_{max} Maximum unambiguous velocity. 15

$||v||_2$ L2 distance of the product of ASV and ASM to quantify the angle sensitivity of the classifier. 50

1 INTRODUCTION

This chapter describes the background on the radar-based Human Activity Recognition (HAR) in section 1.1, problem formulation according to the gaps of existing research studies in section 1.2, the contribution of this work and finally the structure of this report in sections 1.3 and 1.4 respectively.

1.1. BACKGROUND

There have been substantial changes in the population's age composition. Latest statistics show that in 2020, more than one fifth (20.6%) of the EU population was aged 65 and over [4], causing a severe shortage of health care professionals (e.g. 22 EU countries report a shortage of doctors, nurses and health care assistants) and shaking the roots of the health service system [5]. This sheds the importance of indoor HAR since it enables automatic monitoring systems that can improve life quality, reduce health costs, and most importantly provide timely medical help to emergencies, such as in case of a bad fall or stroke event [6].

From a technical perspective, HAR was mostly based on visual aids [7] or wearable sensors [8]. However, both types of sensors- camera and inertial measurement units- exhibit inherent limitations, such as disrespect to privacy as well as poor functionality in darkness or intense light conditions for camera, and in-life inconvenience for wearable sensors. Radar, on the other hand, is gaining attention on civil usage for the exact contrary reasons, i.e. radar collects no visual information on human subjects, provides consistent sensing quality regardless of the light conditions, and is absolutely contactless causing no uncomfortable feelings to human subjects. A more thorough comparison between radar, camera and the wearable sensor is given in Table 1.1.

Theoretical groundwork in the micro-Doppler phenomena for human movements [9] and the pioneering experimental results such as [10] have validated the possibility of using radar to achieve HAR (more details in Chapter 2). The basic philosophy of radar-based HAR is that each human activity has unique kinematic patterns by nature, and such patterns can be represented by intrinsic kinematic features, e.g. the velocity of

Table 1.1: Pros and cons of three common types of sensors for indoor HAR [2].

Sensor Type	Advantage	Limitation
Wearable sensors	a) High velocity accuracy b) Respect to privacy	a) Expensive b) Inconvenience to users
Visual sensors	a) Maintain record b) Functionality under c) changeable conditions	a) Disrespect to privacy b) No functionality without light c) High computational cost
Radar	a) Respect to privacy b) Accurate range measurements c) Functionality in darkness	a) Directional functionality b) Sensitivity to temperature and direction of arrival c) Required installation and calibration

different human body parts along with the physical extent of their movements. To be more specific, radar actively transmits and receives electromagnetic waves to determine the range to object(s) by computing the delay time and radial velocity of the object(s) by Doppler shift. Figure 1.1 is a visualization of the two main blocks of radar-based HAR, and the description is as follows:

1. Data generation (as the dashed-line boxed in Figure 1.1) includes experiments, simulations, and/or transfer learning (more in-depth explanation is given in section 2.1.4).

    - Signal processing refers to processing the generated raw radar data (e.g. I/Q data) to more informative data representations such as spectrograms. Feature extraction may or may not be included in the overall pipeline depending on the desired input format and the classifier.

2. Classification finally is achieved with the help of a classifier using the obtained data representation(s). It could be conventional Machine Learning (ML) classifier, such as K-Nearest Neighbor (KNN) and Support Vector Machine (SVM); or, Deep Learning (DL) technologies such as neural networks, such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) [11].

In conclusion, despite radar-based HAR being investigated extensively in recent years, outstanding challenges still remain to formulate and validate the best signal processing and the best classifiers to achieve satisfactory performances in realistic situations.

1.2. PROBLEM FORMULATION

By reviewing the related work on radar-based HAR, it is realized that the majority of the radar-based HAR approaches are primarily dependent on Doppler and/or micro-Doppler information being presented in various representations. That is to say, these algorithms are merely functional for kinematically dynamic human activities, e.g. walking, sitting down, standing up, and falling, amongst many others. Also, in the past literature, these activities are mostly performed in the line-of-sight direction to maximize the Doppler shift captured by radar. Nevertheless, suppose in the following two scenarios

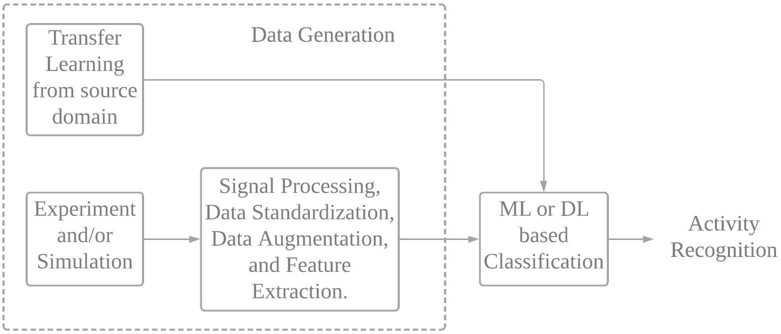

Transfer Learning from source domain
Data Generation
Experiment and/or Simulation
Signal Processing, Data Standardization, Data Augmentation, and Feature Extraction.
ML or DL based Classification
Activity Recognition

Figure 1.1: General radar-based HAR pipeline using DL or ML techniques.

1. A human subject is feeling uncomfortable, so he or she chooses to statically hold a posture to ease the pain [12];
2. A human subject sits onto a chair in the direction orthogonal to the line-of-sight of radar.

Undeniably, both scenarios induce minimal radial velocity and thus Doppler, so these activities are not clearly expressed by Doppler-dependent data representations. Guendel et al. [13] proposed to use a multistatic radar system to obtain the full Doppler information on human subjects, whereas, Yang et al. [1] proposed a so-called omnidirectional classifier. As a result of these works, preliminary results are demonstrated to tackle scenario-2. Despite the initial results in the literature, further improvements are still possible with respect to the *1)* informativeness of the radar data representations, and, *2)* the complexity and performance of the proposed classifier.

Moreover, kinematically static activities- also referred to as 'postures' or 'poses'- as in scenario-1 remain unrecognizable using the existing methods. In other words, previous radar-based HAR is confined to motions instead of postures, whereas in this thesis, the radar-based HAR problem is expected to be more comprehensive by considering human postures as well. This thus encourages us to explore novel radar data representations and DL classifiers.

As a summary, the complete problem can be formulated as:

In more realistic radar-based HAR scenarios, in which human subjects perform kinematically dynamic motions and static postures toward line-of-sight as well as non-line-of-sight directions, how to use only radar and DL technology to achieve accurate HAR?

1.3. Thesis Contribution

The contributions of this thesis are mainly in the following aspects:

- This work, to the best of my knowledge, is the first investigation of radar based HAR including motions as well as static postures together.

- This work tested the developed pipeline with an experimental dataset collected with different operational parameters, and demonstrated superior performance with respect to state of art alternatives.
- Part of this thesis is being written up as a paper to be submitted to IEEE Sensors Journal.

1.4. THESIS STRUCTURE

The following chapters of this thesis are structured as follows. Chapter 2 reviews previous literature related to radar-based classification problems with a focus on HAR. Chapter 3 gives an in-depth and in-detail description of the radar parameters and radar features. Chapter 4 firstly gives an overview of the proposed method and then describes the employed signal and data processing techniques. The laboratory measurement setup and the construction of the dataset are described in Chapter 5. Chapter 6 shows the classification results of the proposed method compared with certain baseline approaches, and its robustness under varying environments. Chapter 7 concludes this thesis and outlines potential future work.

2

Literature Review

This chapter describes the related work on the topic of radar-based classification tasks with the focus put on HAR. For section 2.1, attention is on the conventional radar that has only one transmitter-receiver pair and therefore provides only range resolution. It starts with introducing the popular choices of data representations, then moves on to the categories of human activities that have been investigated and what classifiers were employed, and finally depicts a dilemma on radar data generation. Section 2.2 discusses HAR based on the emerging Multiple-Input and Multiple-Output (MIMO) radar that provides angular information. A pioneering study of MIMO radar-based HAR is investigated in detail. Other existing techniques that may contribute to HAR are then present.

2.1. HAR using conventional radar

conventional radar with simply one transmitter and one receiver is suitable for indoor HAR thanks to its respect for human privacy and functionality in darkness. Related work can be traced back to the year 2008 [14], and radar-based HAR is still attracting attention due to the emerging needs of automatic indoor human life assistance. This section analyzes the problem of conventional radar-based HAR from four mutually dependent perspectives, namely data representation, classifier, choice of analyzed activities, and data generation.

2.1.1. Radar Data Representations

Radar data representations play a crucial role in the performance of the deep neural network-based HAR problem [15]. Radar data representations, according to their dimensionality and the way they are perceived by classifiers, can be divided into three main categories:

1. 2D radar data representations that are treated as an image-alike input. Examples of **2D radar data representations** include spectrogram [16] [17] [18], range-Doppler [19], range map [20]. In particular, range-Doppler and range map are

less popular than spectrogram for two reasons: *(a)* where human subjects perform activities is essentially arbitrary, and *(b)* range resolution for conventional pulse radar is typically not sufficient for differentiating different human parts.

2. Representations that treat radar data as a **temporal sequence**. For instance, Li et al. [21] employed a sequence of features extracted from spectrograms as the input; Li et al. [22] directly used the range profile to represent the human activities; Zheng et al. [23] utilized Kalman Filter to recursively track the subject (hand) and generated the so-called range-Doppler-angle trajectory as the input to the classifier.

3. **Handcrafted features** as a latent space-alike input, and typically these are extracted from the above two categories of data representations. For example, Principle Component Analysis is used to represent spectrograms in a high-dimensional latent space [10]; and, Jia et al. [24] extracted various handcrafted features, such as energy curve, skewness centroid and bandwidth, from spectrograms.

Apart from these three main categories of radar data representation, there is a very rare usage of other data representations. For instance, Yang et al. [25] directly used raw I/Q data as the input to the classifier; Hazra and Santra [26] adopts range-Doppler-time radar cube as a 3D radar data representation; furthermore, Du et al. [27] generate a 3D Point Cloud (PC) from the range-Doppler-time cube for classification; instead of taking the absolute value of Short-Time Fourier Transform (STFT), Wang et al. [28] directly use the complex matrix generated from STFT as the input to the classifier; He et al. [29] constructs a so-called range-Doppler surface through detecting the extended target from a sequence of range-Doppler images.

Previous work is constructed on an assumption that the kinematic characteristics of human activities can be reflected in the radar data representations. Attained experimental results [16]-[27] serve as the ground of this assumption, and therefore, build a cornerstone for further research.

The remaining challenge is to look into *new data representations that provide a new aspect of the observed motion* which is not readily discernible in existing data representations.

2.1.2. CLASSIFIERS

Both 2D or temporal sequence data representations are typically processed by neural networks adapted from the field of Computer Vision or Natural Language Processing. Conventional ML classifiers such as SVM, KNN, Random-Forest are contrarily used to achieve classification through handcrafted features rather than directly on the radar data representations.

1. HAR using 2D image-alike data representations is analog to the task of image classification in the field of computer vision. Deep **CNN** (including multiple convolution layers in cascade) is the primary architecture for classifying images and has been employed in [16]. Other popular CNN architectures, such as ResNet [30], are also adapted or directly re-used and lead to quite encouraging performance [31].

Combinations of neural network blocks are also utilized, for instance, [32] combines the structure of convolutional layer with autoencoder, thus their proposed structure inherits the benefits of both blocks.

2. Regarding temporal sequence as the input, most HAR classifiers are inspired by the work in the field of Natural Language Processing and speech recognition since the inputs in both problems are inherently sequential. **RNN** and its variant **Long-Short Time Memory (LSTM)** are a popular choice, and **Gated Recurrent Unit (GRU)** is an alternative. For example, Bidirectional LSTM [21] and LSTM [22] [33] all exhibit promising performance. Moreover, a sequence of 2D image-alike data representations can be processed by **CNN-RNN** [34], where CNN extracts out high-dimensional feature and RNN exploits the temporal relations amongst the images within the sequence.

3. Supervised ML algorithms such as **KNN** [35], **SVM** [36], **Random-Forest Bagging Tree** [37] and **Naïve Bayes combiner** [37], are compatible with handcrafted features. These algorithms are relatively easy to train in comparison with neural networks (ResNet for instance) since the latter generally have more than tens of thousands of trainable parameters. However, the curse of dimensionality constrains these supervised ML algorithms to work with very high dimensionality. Therefore, the information hidden in the radar data representations may not be fully represented by handcrafted features. This is why the current tendency is to use DL technologies to approach radar-based HAR.

As can be seen, classifier and data representation are mutually dependent. In other words, there is not an optimal data representation nor an optimal classifier, but only a superior combination of them given a specific task.

This suggests that while exploiting new radar data representation, *searching for the complementary classifier* is of great necessity. Only with a superior combination of the input data representation and classifier, the classification performances can be optimized.

2.1.3. ACTIVITIES

Choices of the human activities included in the custom datasets vary from work to work. A list of the number of training samples per literature is given in Table 2.1. In summary, for the the main works in literature [10], [13]-[17], [20]-[38],

- the number of activities ranges from 2 (fall detection) to 12 [39], with the majority of studies involving 6 to 8 classes.
- the most common activity is walking which is seen in all the datasets.
  - it is also noteworthy that walking is a general class with some variants, for example walking with a stick and walking with one hand in pocket.
- Activities, such as sitting down, standing up and falling, are frequently present in the constructed datasets.

Table 2.1: Summary of the measurement datasets from the exemplary radar-based HAR literature, where # Activities expresses the number of activities and # Sample represents the number of training and validation/test samples per activity, regardless the variable of different subjects or orientations.

Paper	# Activities	# Sample	Paper	# Activities	# Sample
[1]	6	60	[10]	7	288
[14]	7	288	[15]	2	200
[20]	7	142	[27]	8	500
[24]	6	555	[25]	8	500
[35]	4	30	[39]	12	72

Activities such as sitting down, bending over, and standing up (from a chair or bending), are selected since they are widespread in indoor daily life and could be applied for human life assistance. Therefore, the choice of activities essentially depends on the designated application. This thesis aims to contribute to daily life assistance, and therefore common daily activities are of interest.

There are overall two realized limitations. First, the existing human activities are not comprehensive enough to express kinematically static activities. To be more specific, motionless activities are rarely recognized as output classes except for the activity of sitting still [10] [14] and standing still [27], and even in these studies, only one of such kinematically static activities are considered so that this 'Doppler-poor' class is made recognizable. That is to say, previous work has not investigated the problem of radar-based human posture recognition, e.g., sitting still versus standing still. Second, since most of the related work only adopts a monostatic radar, with an exception like Guendel et al.'s [13], and is dependent on the Doppler feature, human activities are often artificially limited to the line-of-sight direction, i.e. walking back and forth in front of the radar, which is an unrealistic constraint for true daily life assistance purpose.

2.1.4. DATA GENERATION

The emerging ML or DL technologies essentially are data-driven. That is to say, a classifier must be trained on a comprehensive and well-labeled dataset (only considering applicable for supervised learning). From some of the aforementioned DL-assisted HAR literature, the size of the datasets used to evaluate the classifier performance are listed in Table 2.1 in terms of the number of samples in the recorded dataset (training plus validation/test), showing the significance of promising data generation methods [1].

In the meantime, data generation is recognized as one of the most challenging tasks for radar-based HAR [40]. Generally speaking, there are three methods to generate radar data: experiment, simulation, and Transfer Learning (TL). These methods exhibit their own pros and cons which will be discussed in this section.

Experimental results are regarded as the ground truth of data since realistic data col-

[1] Some studies in the literature did not explicitly show the number of training samples, so they are not included in Table 2.1. The demonstrated numbers are experimental data-only, i.e., otherwise generated data is not counted.

lection due to hardware imperfections, clutter, multipath and noise are implicitly considered. All the radar-based HAR literature [10], [13]-[38] use experimental data to validate their classification performance. However, radar experiments are typically *(a)* time-consuming: tens of hours are spent in a laboratory to complete one dataset as in [13], *(b)* expensive: experiments in realistic conditions imply the involvement of many human subjects for a long time and investing the required hardware resources, *(c)* inflexible: experiment setup cannot be freely adjusted due to the laboratory constraints.

Simulation is the second option for data generation, and can be further divided into two categories according to the application of parametric or non-parametric computations:

1. **Parametric computations**: Kinematic models can be used to record the skeleton human motions, such as Thalmann Model, motion capture and Kinect. These skeleton human motions are transformed into radar data representations through analysis of human body **scattering behaviour** and **radar signal modelling** [41] [42] [43]. Thanks to the existence of open-source datasets of the kinematic motions, such as the CMU MOCAP dataset, synthetic radar data of human activities can be generated to improve the classification performance [44] [42].

2. **Non-parametric computations**: **Generative Adversarial Network (GAN)** is a special type of neural network-based simulator that consists of a generator and a discriminator. Through learning how to generate new data (for generator) and what is the criteria of true data (for discriminator), GAN is able to generate unseen data given sufficient training data. Examples of this type of work include [45] and [46], which use GAN to produce spectrograms to improve classification accuracy. Additionally, Vishwakarma et al. [47] used GAN to learn the noise distributions in the echos of electromagnetic waves to learn to generate realistic data [2].

Once a model that depicts radar echoes from human movement is completed, such simulation is an effective approach to generate a huge amount of data since it has very high flexibility. However, three reasons are limiting the general application of synthetic data: *(a)* there is no generally applicable criteria on the quality of the synthetic data, as a result, the robustness of classifiers purely trained with synthetic data is in doubt; *(b)* the parametric model is not sufficiently specified to describe most of the aforementioned activities; *(c)* the simulators fail to learn the variations in signal-to-noise ratio, clutter and other artifacts. This is why simulated data is mostly employed to augment the experimental dataset, rather than as the direct source for learning on its own.

TL is a concept raised from DL referring to an approach that transfers data from a source domain to a target domain. For radar-based HAR, the target domain must be one of the radar data representations, while the source domain is different, such as optical image [31], speech signal [40] and spectrogram images of human subjects not included in the target domain [39]. TL is easy to implement since technically it is ’re-using the pre-trained weights of the neural network in the training with new data, and more importantly, TL leads to an improvement in classification accuracy as shown in [31] and

[2] They employed passive WiFi radar instead of the conventional active radar sensor. However, they are the same in the principle of range, Doppler and cross-range information.

[40]. However, the drawback of TL is so-called negative learning, which is typically triggered by the dissimilarity between source and object domains and actually could be the case for transferring from optical images to heatmap-alike spectrograms [48].

2.2. HAR Using Imaging radar

Thanks to the rapid development in automotive radar, imaging radar [3] is seen as an alternative solution of LiDAR to provide reliable perception schemes, e.g., segmentation [49] or classification [50]. Most imaging radar-related classification studies now pay attention to the field of automotive driving by fusing with other optical sensors for instance. However, there are many significant differences between the principles of indoor HAR and automotive driving. For example, the requirement of update rate for automotive driving is much higher than that for HAR; the targets on road could be static rather than in motion and/or of an arbitrary aspect angle, making the Doppler/micro-Doppler pattern not particularly representative; and the fact that maximal detection range, as well as Doppler ambiguity, must be higher than indoor conditions. These facts make the automotive driving literature, such as [51] [52], [53], [50] and [54], of minimal reference value.

Therefore, this section still focuses on pioneering imaging radar-based HAR literature. Enabled by the additional angular information provided by imaging radar, static posture recognition or large aspect angle motion recognition, which was impossible using conventional radar as discussed in section 2.1, now become possible.

To start with, Cui and Dahnoun's work [55] proved that imaging radar is capable to acquire sufficient data to represent the postures of human subjects despite lower body parts such as feet are typically not identified. One step further, using the optical images as the ground truth of human skeletons for CNN-based supervised learning, Zheng et al. [56] successfully reconstructed human skeletons with the help of an imaging radar in through-the-wall conditions, where imaging radar effectively learns from the ground truth optical images. Sengupta et al. [57] also presented a DL-based approach for skeletal estimation, where the results are remarkably comparable with the ground-truth. These are encouraging work that suggests through DL technologies radar can learn from the data representations of other sensors, and subsequently remedies the radar's inherent deficiencies, such as low angular resolutions. These works thus built a concrete basis for high-accuracy posture recognition.

Kılıç et al. [58] achieved through-wall imaging of human subjects, and attained very high accuracy in posture classification between standing and sitting using CNNs. The input data representation to their proposed CNN architecture is a flattened range profile, so the angular resolutions provided by imaging radar are not fully exploited. He et al. [59] achieved over 80% of accuracy amongst four classes through background noise elimination technique and parallel CNNs. Tiwari, Bajaj and Gupta [60] utilized an 8-channel imaging radar to classify seven fitness-related classes through range-Doppler image as the input to CNNs, giving an over 90% of accuracy. Nickalls, Wu and Dahnoun

[3] Imaging radar is referred to as a type of radars that is used to generate images of the illuminated field of view. 3D imaging radar provides three dimensions of data: range (depth), azimuth and Doppler, while 4D imaging radar additionally provides elevation information.

[61] mounted the radar on the ceiling to collect the PC representation of human subjects. For the constructed PCs, both decision tree model and CNN lead to very high (more than 98%) accuracy amongst three kinematically static postures. Further, Sasakawa et al. [62] proposed a simple but efficient posture recognition scheme using the estimated height and radar cross-section of human subjects, and similarly, Honma et al. [63] used the estimated height and the intensity of the Doppler component as the handcrafted features attaining extraordinary results. The final prediction accuracy achieved by KNN is outstanding, averagely 95% for the former and 76% for the latter. Kim, Alnujaim and Oh's work [64] utilized 2D PC images generated by 4D imaging radar as the input to two classifiers, CNN and CNN-RNN. The recognition accuracy amongst the 7 motion classes is extraordinarily high (over 94%).

Table 2.2: Summary of imaging radar-based classification works that considers human posture recognition.

paper	**posture**	**radar features**	**classifier**
[58]	standing, sitting	SFCW	CNNs
[59]	punching, walking, standing, siting down	FMCW 8 azimuthal channels	parallel CNNs
[60]	dumbbell, shoulder press, squat, lateral raise, boxing, right triceps and left triceps	FMCW 8 azimuthal channels	CNNs
[63]	standing, sitting on chair, sitting on ground, laying on ground	Bi-static MIMO	threshold comparison
[62]	standing, sitting on chair, sitting on ground, laying on ground	Bi-static MIMO	KNN
[61]	standing, sitting, laying on ground	FMCW 12 azimuthal channels	decision tree
[65]	bowing, kicking, marching, punching, sitting, standing	FMCW 192 elevational and azimuthal channels	CNN/ CNN+RNN

A summary of the aforementioned posture recognition work is listed in Table 2.2. All these works establish a solid ground for imaging radar-based human posture recognition. Generally speaking, this attributes to the spatially representative modality of PC. Also, the works on PC construction furthermore proved imaging radar's capability to capture necessary representations on human subjects. It is also learned that for high-dimensional data representations (PCs or snapshots of PC), DL techniques exhibit promising performance. However, the difficulty of data collection remains a non-trivial problem.

2.3. Limitations and Challenges

By reviewing the previous literature, several limitations of the existing work are found, together with the thus raised challenges are listed as follows.

- It is not yet established what classifiers are most suitable to learn the important information from the new format of data representations enabled from millimeter wave (mm-Wave) imaging radar.

  - This encourages us to consider what new **data representations** are likely to be representative for the human activities thanks to the additional angular resolutions? Furthermore, what **DL classifiers** are compatible?

- Kinematically static postures and kinematically dynamic motions have not been jointly examined. Moreover, human activities are deliberately constrained to be performed in the line-of-sight orientation due to the excessive dependency on the Doppler feature.
  - How can motions and **postures** performed on the line-of-sight as well as **non-line-of-sight** direction be recognized, i.e. omnidirectional classification (as in [1]) of motions together with postures?
- Radar experiments are seen as time-consuming and expensive, while there is still a lack of conclusive criteria on the quality of synthetic data for simulation.
  - Given underlying new data representations, would it be possible to perform certain **augmentations** of the dataset to be economical on time?

In this work of MSc thesis, the focus is put around the first two points. To be more specific, two data representations- PC and spectrogram- are utilized to exploit the six intrinsic features enabled by mm-Wave radar. Correspondingly, modified T-Net and PointNet [66] and AlexNet [3] are the realized compatible classifiers. According to the spatial information contained by two data representations, the final prediction is made in a successive manner.

3

Radar Parameters

In order to extend radar-based HAR to include kinematically static postures, the use of a 2D Frequency-Modulated Continuous-Wave (FMCW) MIMO radar is proposed. The used radar is designed and manufactured by Texas Instrument. It provides considerable freedom to configure the radar parameters such as chiro ramp slope and ramp time interval, and accordingly, radar features such as bandwidth. This chapter firstly discusses the employed 2D FMCW MIMO radar and presents its beam pattern (section 3.1), then explains the chosen waveform parameters and shows the derived radar features (section 3.2).

3.1. 2D MIMO Radar Description

The used radar board consists of four cascaded AWR2243 chips, where each AWR2243 chip has 4 Transmitter (TX)s and 3 Receiver (RX)s. With all RXs used simultaneously but TXs enabled successively, in total 16 TXs and 12 RXs are obtained.

A picture of the board is given in Figure 3.1, and a sketch of the (not-to-scale) relative antenna positions is given in 3.2. The corresponding virtual antenna positions are shown in Figure 3.3. This virtual array is obtained by making spatial convolution of the TX array and the RX array over the X and Z axis. As shown at the orange positions in Figure 3.3, some of the antennas are virtually overlapped in space. The signal received by one of the two overlapped antennas is discarded, so the resulting 'effective' number of channels (transmitter-receiver pair) is 134 rather than 192=16×12, and the 'effective' virtual array has a MIMO aperture of $86 \times 6\lambda_{antenna}$. It is convinced that such imaging radar could provide sufficient additional cross-range information for radar-based HAR as proved in [64], therefore enable static human posture recognition or HAR with a large aspect angle.

The array factor of the total virtual array [1] is given in Figure 3.4, where AF is the computed result of equation 3.1 with progressive phase shift β_n set to 0 and a constant antenna excitation $A_n = 1$. It is easily realized that array factor is maximized for

[1] The array factor discussed in this thesis are in far-field such that the transmitted electromagnetic waves can be treated as plane waves and the virtual array is deterministic.

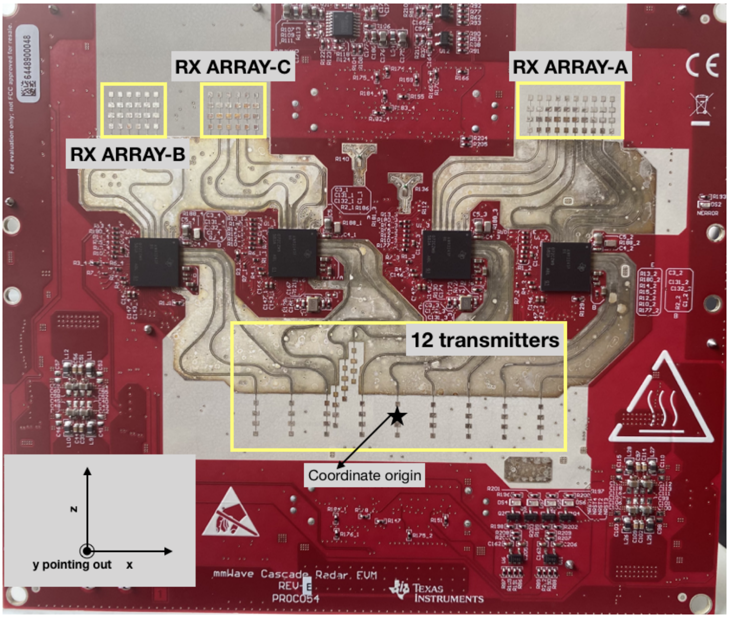


Figure 3.1: Picture of four-device cascaded AWR2243 radar board.

$exp(\cdot) = 1$. That is to say, the progressive phase shift (β_n) and the phase shift due to scattering ($k_{chirp}\vec{d}_n \cdot \hat{e}_r$) should sum up to 0, i.e., the progressive phase shift is expected to compensate the phase shift due to aspect angle.

$$AF(\hat{e}_r) = \sum_n A_n exp(jk_{chirp}\vec{d}_n \cdot \hat{e}_r + j\beta_n)[68], \tag{3.1}$$

where $\hat{e}_r$ is the unit vector in spherical coordinate, $exp(\cdot)$ is exponential operator, j is the imaginary unit, $\vec{d}_n$ is the vector pointing from coordinate origin to the geometric position of the antenna, and k_{chirp} is the wavenumber of the center frequency of the employed chirp (refer to Table 3.1).

The total beam pattern is the linear multiplication of array factor and antenna pattern. The individual antenna pattern simulated by Texas Instrumentation is shown in Figure 3.5, as can be seen, the field of view in azimuth is approximately -60 to +60 degree and -20 to +20 degrees in elevation, for attenuation of less than 10dB. Outside this field of view, the energy scattered by objects could be easily covered by clutters or multi-path from the field of view.

In summary, the principles of imaging radar are presented, and it is also shown that the used 2D MIMO imaging radar enables angular resolutions on both azimuth and elevation. This additional spatial information will play a key role in this project to recognize human postures and motions on multiple aspect angles.

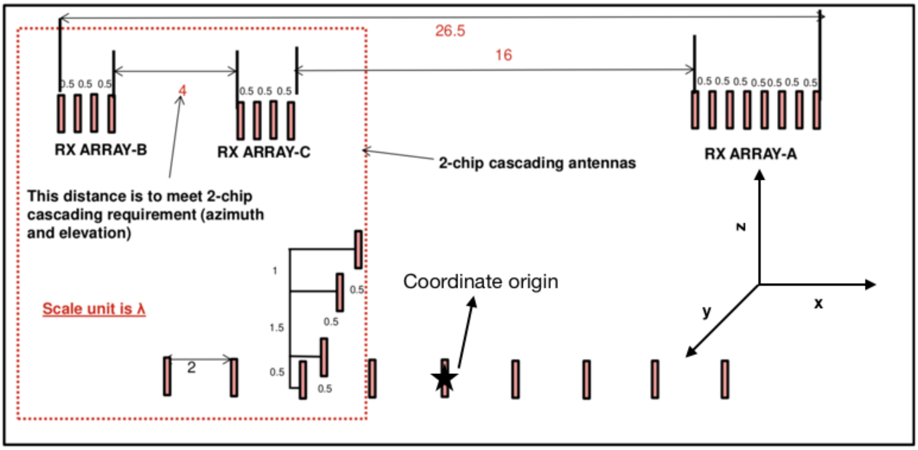


Figure 3.2: Antenna array positions (adapted from [67]).

3.2. DERIVED RADAR FEATURES AND WAVEFORM

For an FMCW radar, the constraint on the so-called BT (bandwidth time) product is no longer valid. The maximum measurement range (R_{max}) [2], maximum velocity ambiguity interval (v_{max}), range resolution (ΔR) and velocity resolution (Δv), respectively, can be expressed in equations 3.2a-3.2d.

$$R_{max} = \frac{c f_{ADC}}{2 r_{chirp}}, \tag{3.2a}$$

$$v_{max} = \pm \frac{\lambda_{chirp}}{4 N_{TX} T_{chirp,total}}, \tag{3.2b}$$

$$\Delta R = \frac{c}{2B} = \frac{c}{2 T_{chirp} r_{chirp}}, \tag{3.2c}$$

$$\Delta v = \frac{\lambda_{chirp}}{2 T_{CPI}}, \tag{3.2d}$$

where c is the speed of light, B is the valid sweep bandwidth (see red dashed line in Figure 4.2) [3], $T_{chirp} = N_{ADC}/f_{ADC}$ is the ADC-sampled chirp ramp interval, r_{chirp} is the ramp slope of the chirp, T_{CPI} is the coherent processing interval.

Unlike range resolution or Doppler resolution that only depends on the bandwidth of the waveform or the coherent processing interval, angle resolution is approximately equal to the proportion between the chirp center wavelength and the MIMO aperture size, as in equation 3.3, where the impact of aspect angle is safely neglected since the human subjects are genuinely on the broadside direction.

[2] Maximum measurement range differs from the maximum range ambiguity since it is essentially dependent on ADC sampling

[3] The total sweep bandwidth is not necessarily equal to the valid sweep bandwidth because the ramp time may not be completely sampled by the ADC.

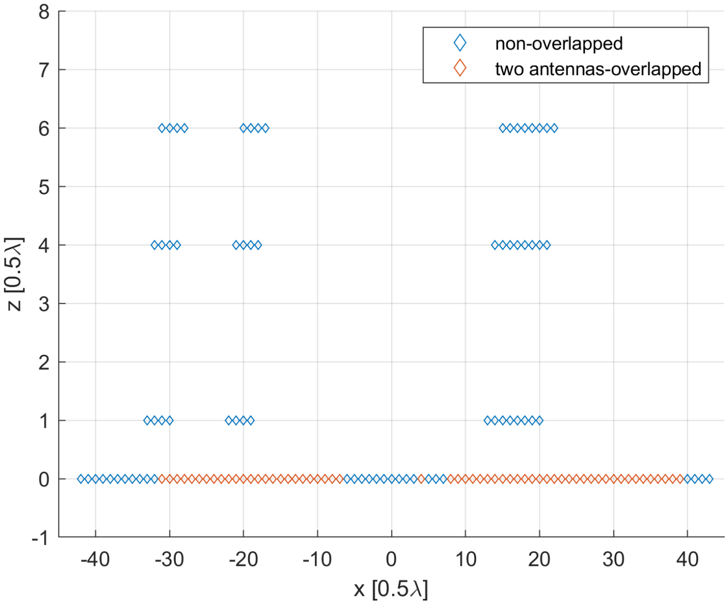


Figure 3.3: Synthetic 2D virtual array.

$$\Delta\phi = \frac{\lambda_{chirp}}{L_{\phi}}, \tag{3.3a}$$

$$\Delta\theta = \frac{\lambda_{chirp}}{L_{\theta}}, \tag{3.3b}$$

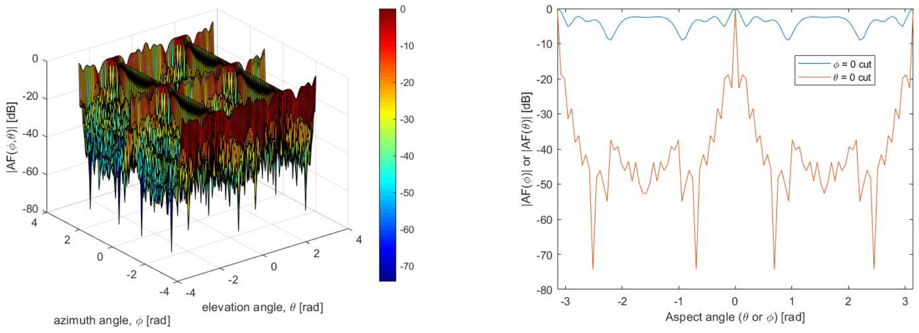
where $\Delta\phi$ and $\Delta\theta$ express the azimuth resolution and elevation resolution, respectively, and L_{ϕ} and L_{ϕ} express the MIMO apertures in azimuth and elevation, respectively [4].

For a mm-Wave radar operating at 77GHz to 81GHz, the center wavelength is, as the name suggests, of millimeters. This limits the maximum velocity ambiguity, whereas enables better velocity resolutions. Other trade-offs are also quite self-explanatory, e.g. the chirp interval simultaneously determines the maximum detection range and the maximum unambiguous velocity, the ADC sampling rate (f_{ADC}) should be set to the largest configuration value.

The ultimate goal of waveform design is to configure the radar parameters to the most fitting combination that provides as good resolutions as possible. There are also a few of rules of thumb to be considered:

- The typical indoor human activity velocity is less than 2.6m/s [69];

[4] The angular resolutions discussed here actually vary with respect to the aspect angle such that $\Delta\phi \propto \cos phi$. However, since the human activities are performed in-place and in front of the radar, such variations are assumed to be negligible and thus not considered in this thesis.



(a) (b)

Figure 3.4: Array Factor ($|AF(\hat{e_r})|$) of the total virtual array presented (a) in a 3D surface, and (b) on the azimuth and elevation cut.

- The maximum measurement range should be sufficient to cover the possible positions of the human subject, i.e., fully cover the experiment environment.
- The coherent processing interval of human activities should be reasonably small such that Fourier transform can be applied on the ground of coherent processing, i.e., one body part remains in one range and cross-range bin within the coherent processing interval;
- Hardware limitations, such as ADC sampling rate, ramp rate and data storage space, should also be considered;
- The designed waveform should have similar performances as in [64] so that comparative study is enabled.

With all the aforementioned factors considered, the waveform for HAR is empirically configured to have the performances as in Table 3.1.

Table 3.1: Waveform parameters and derived features of the radar. The definition of parameter and feature is subject to whether it is directly configurable, the directly configurable term is referred to as parameter, the other as derived feature.

Parameter	**Symbol**	**Value**
Antenna design wavelength	$\lambda_{antenna}$	$3.90mm$
Number of TXs	N_{TX}	12
Number of RXs	N_{RX}	16
Total number of virtual channels	$N_{channel}$	192
MIMO aperture on azimuth	L_ϕ	$42.5\lambda_{antenna}$
MIMO aperture on elevation	L_θ	$3\lambda_{antenna}$
ADC Sampling Rate	f_{ADC}	$2.7MHz$
Chirp Ramp Interval	T_{chirp}	$60\mu s$

Total Chirp Interval	$T_{chirp,total}$	$63\mu s$
Number of chirps per sub-frame	N_{chirp}	1536
Start Frequency	f_{start}	$77GHz$
Chirp Ramp Slop	r_{chirp}	$60MHz/\mu s$
Sub-frame Periodicity	$T_{sub-frame}$	$100ms$
Field of view on azimuth	FOV_{ϕ}	$[-40deg, 40deg]$
Field of view on elevation	FOV_{θ}	$[-20deg, 20deg]$
Coherent processing interval	T_{CPI}	$0.1s$
Derived Features	**Symbol**	**Value**
Equivalent number of channels on x-axis	N_{ϕ}	86
Equivalent number of channels on z-axis	N_{θ}	7
Transmitted Chirp Bandwidth	B_{Tx}	$3.6GHz$
Received Chirp Bandwidth	B_{Rx}	$2.84GHz$
Valid chirp center wavelength	λ_{chirp}	$3.82mm$
Maximum Measurement Range	R_{max}	$6.75m$
Maximum Unambiguous Velocity	v_{max}	$\pm 1.26m/s$
Range Resolution	ΔR	$5.28cm$
Velocity Resolution	Δv	$0.0286m/s$
Azimuth Angle Resolution (broadside)	$\Delta\phi$	$1.4deg$
Elevation Angle Resolution (broadside)	$\Delta\theta$	$18deg$

The generated raw data per sub-frame is a 3D cube in the format of fast-time, slow-time, channel. Quantitatively, each 3D is made of $N_{ADC} - (N_{chirp}/N_{TX}) - (N_{TX} \cdot N_{RX})$, where the slow-time has the number chirps equal to N_{chirp}/N_{TX} is the result of time-division multiplexing. By rearranging the structure of the cube and interpolating zeros to the empty positions in the virtual array, the raw data becomes a 4D hypercube with the dimensions of fast-time, slow-time, virtual elevation channel, and virtual azimuth channel. Signal processing algorithms can be applied to this hypercube for further analysis.

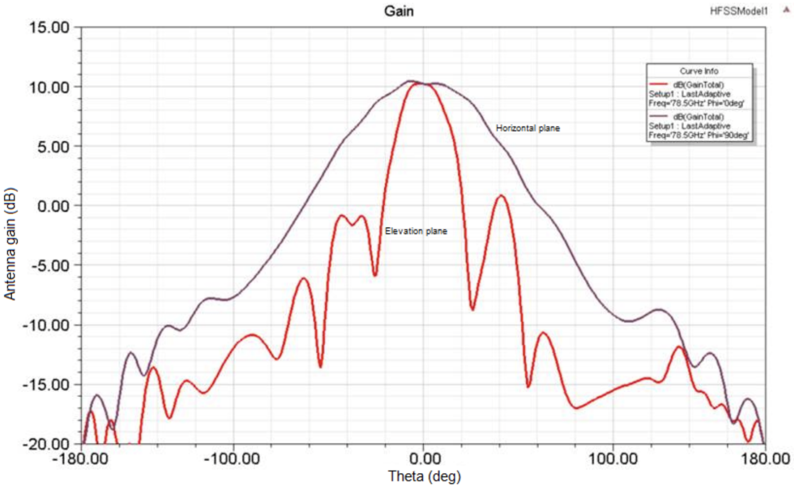


Figure 3.5: Radiation pattern of a single micro-strip antenna (simulated and provided by Texas Instrument [67]).

4

Proposed HAR Pipeline

The structure of this chapter is arranged as follows. First, an overview of the proposed pipeline is given in section 4.1. The signal processing algorithms applied to obtain the desired data representations are explained in section 4.2. Then, the DL classifiers and implementation details are given in section 4.3. Last, section 4.4 presents two baselines that could be used to comparatively evaluate the performance of the proposed method.

4.1. Pipeline Overview

Conventional monostatic radar, by continuously transmitting and receiving electromagnetic waves, generates four intrinsic features of the object: range, Doppler, received power proportional to the Radar Cross Section (RCS), and temporal relations. The usage of these dimensions has been thoroughly analyzed (chapter 2), and it appears that the state-of-art works mostly depend on the intrinsic feature- Doppler. To achieve human static activity (posture) recognition, additional dimensions of information must be introduced assuming the micro-Doppler features introduced by respiration and heartbeat are not sufficiently representative and often difficult to be reliably estimated in realistic scenarios. The imaging radar which provides additional azimuth and elevation information will be introduced in this interest.

The overview of the proposed method is given in Figure 4.1, definitions of Single Angle Classifier (SAC) and Multiple Angle Classifier (MAC) will be addressed in section 4.4. To summarize, the proposed pipeline fully explores all six intrinsic features provided by imaging radar and uses both PCs and spectrograms as the input data representations. The hierarchical structure of the classification pipeline includes a module (namely orientation classification) to classify to which orientation the human subject is facing toward, e.g., a 0 degree aspect angle. Based on the prediction made by this module, PC classification module predicts which posture or motion pair the input belongs to. For those predicted as motion pairs (sitting/standing pair or bending/standing pair), the spectrogram classification module is then utilized to classify to which specific motion the input sample belongse.g. bending over or standing up from bending, or, sitting down or standing up from sitting.

Furthermore, the descriptions of the four main modules are as follows:

1. The *data generation module* starts with measuring experimental data via the imaging radar. Given the measured raw radar data, the signal processing flow involves PC generation on the 4D radar cube, and spectrogram generation using the slow-time axis from particular range bins and one channel. The outputs of the signal processing flow are two data representations- PC and spectrogram. The former expresses three intrinsic features: range, azimuth, and elevation, i.e., (X,Y,Z) as in the cartesian coordinate system. The latter, on the other hand, expresses the remaining intrinsic features- received energy, Doppler, and time.

2. The *orientation (ori.) classification module* includes the so-called T-Net (introduced by Qi et al. in [66]). Through learning the geometric characteristics of the PCs, T-Net predicts the human orientation of the subject.

3. For each predicted orientation of the human subject, *PC classification module* makes a classification (P_{PC}) according to the spatial similarity of the activity within the segment interval (2 seconds, for more details refer to chapter 5.2), e.g., *(a)* sitting still, *(b)* standing still, *(c)* sitting down or standing up from chair (yellow-shaded in Table 5.1), and *(d)* bending over or standing up from bending (red-shaded in Table 5.1).

4. The last module is the so-called *spectrogram (spc.) classification module*. This module treats the predicted class from the previous module as prior knowledge and further uses the spectrogram as the input to the AlexNet [3] to recognize the activities within *(c)* or *(d)*, outputting the probability vector P_{spc}. Therefore, the final prediction result is jointly subject to the decisions made by $\arg\max(P_{PC})$ and $\arg\max(P_{spc})$. The choice of AlexNet is in accordance with its relatively simple architecture and presumably easy-to-converge.

Furthermore, inspired by Yang et al. [1], a good angle-insensitive HAR pipeline should be robust to make predictions given data on multiple orientations for training, or, even more ideally given data on one orientation only. Thus, two definitions are given for the way to use a pipeline in terms of different training data combinations with respect to aspect angle, e.g. training with data from one human orientation and testing with multiple human orientations is termed SAC, while namely MAC if training and testing with multiple angles. Naturally, the orientation classification module is insignificant for SAC and thus bypassed.

The main innovations of the proposed pipeline are as follows.

- Unlike in the past work such as [64] and [58], the proposed modular pipeline is designed with a goal to exploit all intrinsic features obtained by imaging radar, i.e. range, azimuth, elevation, Doppler, received power and time, and not just one specific data representation.

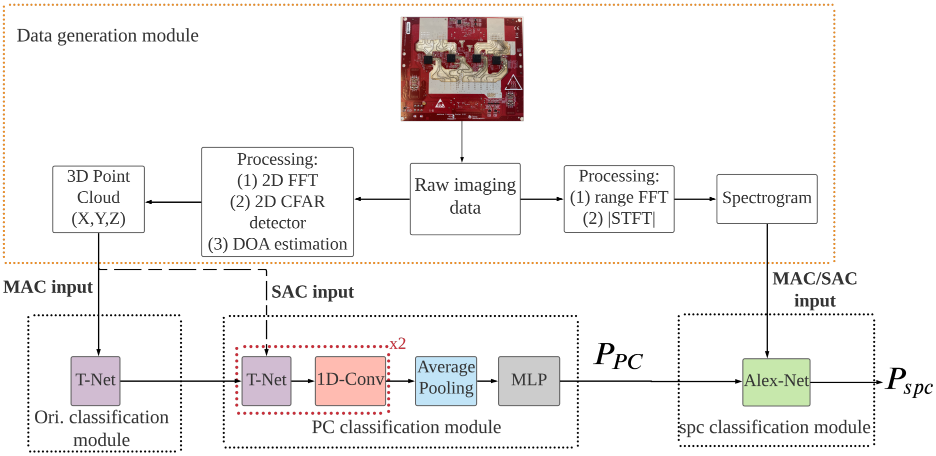


Figure 4.1: Overview of the proposed pipeline, where the main contributions are the parallel processing and fusion of PCs and spectrograms, and the usage of T-Net for angular orientation insensitivity.

- The cascaded architecture of the subsequent modules simplifies the task of multiple-angle human activity (posture and motion) classification with the help of the designated neural networks.
  - In particular, the usage of T-Net to learn geometric features of PCs and therefore achieve orientation classification is of an original contribution of this thesis.
- The pipeline is designed to be compatible with noisy and limited amount of radar data by replacing the symmetric function- Max Pooling- with Average Pooling and using light-weight neural networks, respectively. More detailed explanation on this is given in Section 4.3.

4.2. SIGNAL PROCESSING ALGORITHMS

Resolutions of range, azimuth and elevation enable the construction of 3D PCs. An implicit assumption here is that human bodies are perceived by radar as an ensemble of multiple scatters since a human body fits the definition of extended target, i.e., the body size exceeds the range and angular resolutions provided by mm-Wave imaging radar and thus could originate multiple detected points. Therefore the PCs supposedly represent the shape of the human body seen from the line-of-sight direction of the radar. Several signal processing algorithms applied for the generation of PCs are described in this section.

4.2.1. SIGNAL MODEL

The concept of MIMO is based on the orthogonality of the transmitted signals. In this thesis, time-division multiplexing is utilized to ensure this requirement. A overview of the time-multiplexing FMCW signal model is given Figure 4.2.

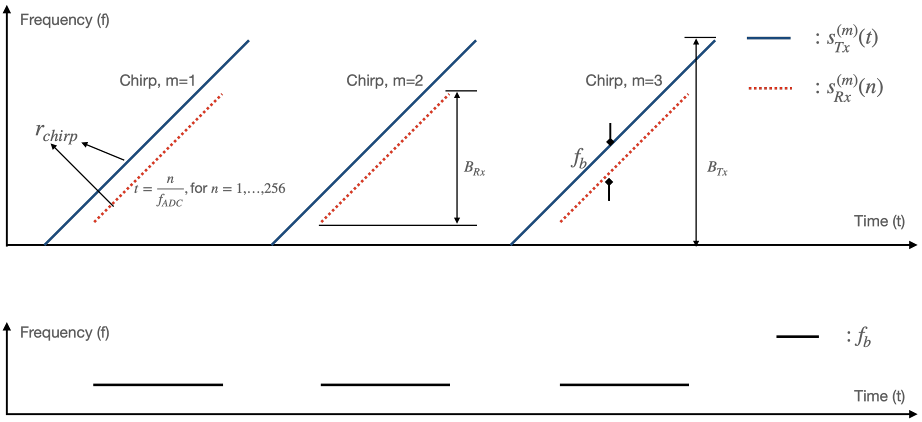


Figure 4.2: Visualization of time-multiplexing FMCW signal and beat signal.

To be more specific with the mathematical expressions of the FMCW signal model, a few assumptions can be made for people as "objects of interest" within an indoor environment within a sub-frame interval of time (96.7ms) [1]:

- The movement rate of human body parts is small enough so that these parts stay in a particular range and angular voxel and the body shape remains approximately the same within a sub-frame. As a result, the following parameters can be assumed to be constant within a sub-frame interval:
  - The distance between the i'th TX and object, R_{ti}, and the distance between object and the j'th RX, R_{rj};
  - the TX gain G_{Tx} and RX gain G_{Rx};
  - the RCS of the body, σ; the polarization of the receiver $\vec{p}_{rec}$ and that of the scattered EM waves $\vec{p}_{scatter}$;
- The antennas transmit stable power levels within a chirp interval, i.e. P_{Tx} is a constant.

It is logical to start with expressing frequency modulation. Frequency modulation involves frequency ramping within a chirp, and the instantaneous frequency of a linear chirp signal (as used in this work) can be described by as follows,

[1] One frame expresses the time duration of one complete segment of human activity (2 seconds), and sub-frame expresses the coherent processing interval for Doppler. 20 sub-frames constitute one frame that can be used for classification)

$$f_{Tx}(t) = f_{start} + \frac{B_{Tx}}{T_{chirp}} t = f_{start} + r_{chirp} t \tag{4.1}$$

The instantaneous phase of a cosine signal can be found by the integration of its instantaneous frequency as follows,

$$\phi_{Tx}(t) = 2\pi \int f(t)dt = 2\pi(f_{start} t + r_{chirp} t^2), \tag{4.2}$$

while assuming the initial phase of the signal is 0, the transmitted signal can be simplified and written as follows,

$$s_{Tx}(t) = A_{Tx} \cos(\phi(t)) = A_{Tx} \cos[2\pi(f_{start} t + r_{chirp} t^2)], \tag{4.3}$$

where A_{Tx} is the transmitted signal amplitude.

Taking the phase, noise and clutter into consideration, the complete signal transmitted by i'th TX and received by j'th RX can be modelled as follows,

$$s_{Rx,i,j}(t) = A_{media} \cdot s_{Tx}(t-\tau) \cdot e^{-jk(R_{ti}+R_{rj})-j\phi_{target}} \cdot [\vec{p}_{rec} \cdot \vec{p}_{scatter}(t)] + C(t) + n(t) \tag{4.4}$$

$$P_{Rx,i,j} == \frac{P_{Tx} G_{Tx} G_{Rx} \lambda^2_{chirp} \sigma}{(4\pi)^3 R^2_{ti} R^2_{rj}}, \tag{4.5a}$$

$$A_{media,i,j} = \frac{P_{Rx}}{P_{Tx}}, \tag{4.5b}$$

where A_{media} is the amplitude attenuation of the signal due to the propagation path and antenna gains as expressed in equation 4.5; $\tau = \frac{R_{tj}+R_{rj}}{c}$ is the delay caused by EM wave's round trip (Tx-to-object and object-to-Rx); phase of the received signal is also subject to object scattering behavior (ϕ_{target}), as well as the phase delay caused by the total propagation distance ($k(R_{ti} + R_{rj})$); the polarization of the receiver ($\vec{p}_{rec}$) can be seen as constant, while that of object's scattering behavior should be time-dependent ($\vec{p}_{scatter}$); and the clutter-originated signals and noise are expressed in C and n. It is noteworthy that one of the disadvantages of $mm-wave$ radar can be observed that media attenuation of $mm-wave$ is much larger than the counterpart of radar operating with at MHz.

Suppose in an ideal indoor environment, where clutter and noise do not exist, scattering phase delay and polarization mismatch are negligible, the received signal model can be simplified to be:

$$\begin{aligned} s_{Rx,i,j}(t) &= A_{Tx} \cdot A_{media} \cdot s_{Tx}(t-\tau) \\ &= A_{Tx} \cdot A_{media}(t) \cdot cos[2\pi(f_{start}(t-\tau) + r_{chirp}(t-\tau)^2)] \cdot e^{-jk(R_{ti}+R_{rj})} \end{aligned} \tag{4.6}$$

After mixing the received signal with the local oscillator (equivalent to the transmitted instantaneous frequency, $f(t)$, from equation 4.1), the frequency component is divided into high and low parts. The low-pass filter is supposed to effectively mitigate the high-frequency component, leaving out the base-band signal, also known as beat signal (as visually explained in Figure 4.2). The beat signal can be expressed as follows:

$$s_{Rx,i,j}(t) = A_{Tx} \cdot A_{media} e^{j2\pi(r_{chirp}\tau t + f_{start}\tau)} = A_{Tx} \cdot A_{media} e^{j\phi_{Rx,i,j}(t)} \tag{4.7}$$

The beat frequency, which equals the derivative of the instantaneous phase, then can be expressed as follows:

$$f_{b,i,j}(t) = \frac{1}{2\pi}\frac{\partial\phi_{Rx,i,j}(t)}{\partial t} = \frac{2r_{chirp}((R_{ti} + R_{rj}) + 2v_r t)}{c} + \frac{2f_{start}v_r}{c}, \tag{4.8}$$

where, the small difference between two paths of i'th TX to object and object to j'th RX is negligible since they are divided by the velocity of light ($3 \cdot 10^8 m/s$), so the approximate distance between the radar and object $R = \frac{R_{ti}+R_{rj}}{2}$ is used; and the range variation due to (human) target movement within the sub-frame interval is far smaller than R and thus the component, $2v_r t$, can be safely neglected. The final beat frequency model is given as follows:

$$f_b = \frac{4r_{chirp}R}{c} + \frac{2f_{start}v_r}{c}, \tag{4.9}$$

Writing the beat signal being sampled at frequency, f_{ADC}, in discrete-time domain gives us as follows:

$$s_{Rx}(n) = A_{Tx} \cdot A_{media} \cdot e^{j2\pi\left[\frac{2f_{start}R}{c} + \left(\frac{2f_{start}v_r}{c} + \frac{2r_{chirp}R}{c}\right)\frac{n}{f_{ADC}}\right]} \tag{4.10}$$

Furthermore, for a time-multiplexing FMCW signal, each frame contains multiple chirps, constituting the slow-time domain (also known as chirp domain). Denoting the received sequence of chirps with $m = 1, 2, \dots, M$, we can get the expression of the received fast-time and slow-time FMCW signal per channel as follows:

$$s^{(m)}_{Rx}(n) = A_{Tx} \cdot A^{(m)}_{media} \cdot e^{j2\pi\left[\frac{2f_{start}R}{c} + \left(\frac{2f_{start}v_r}{c} + \frac{2r_{chirp}R}{c}\right)\frac{n}{f_{ADC}}\right]} \tag{4.11}$$

4.2.2. Range-Doppler FFT

As shown in equation 4.10, the beat signal can be interpreted as a discrete 2D sinusoidal signal. Therefore, it is crucial to estimate the beat signal frequency in order to accurately obtain the range and Doppler information on the object. In this particular HAR use case, we have no prior knowledge on the number of sources, i.e., how many range or Doppler bins the object would occupy; or, multiple realizations of the 2D sinusoidal signal. The best estimation method is periodogram (Fast Fourier Transform (FFT) for discrete signal), assuming the beat signal is stationary (i.e., R is constant within a chirp of $756\mu s$ and v_r is constant with a frame of $0.1s$), which is fairly reasonable for human activities.

Expressions of 1D range FFT and 2D range-Doppler FFT on the received signal are given in equations 4.12 and 4.13, respectively.

$$S_{Rx}^{(m)}(k) = \sum_{n=0}^{N-1} s_{Rx}^{(m)}(n) e^{-j2\pi kn/N} \text{ for } k = 0, \ldots, N-1 \tag{4.12}$$

$$\begin{aligned} S_{Rx}(k,l) &= \sum_{m=0}^{M-1} S_{Rx}^{(m)}(k) e^{-j2\pi lm/M} \\ &= \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} s_{Rx}^{(m)}(n) e^{-j2\pi kn/N} e^{-j2\pi lm/M} \end{aligned} \tag{4.13}$$

for $l = 0, \ldots, M-1$, and $k = 0, \ldots, N-1$

4.2.3. CFAR

2D FFT is firstly applied on the fast time and slow time dimensions, outputting range-Doppler matrix. Detecting the presence of human body part(s) at a specific range and Doppler is the next step, and for that Ordered-Statistic (OS) Constant False Alarm Rate (CFAR) Detection is used.

CFAR is a common technique used for object detection in clutter and multiple target situations. CFAR is renowned for its capability to adaptively suppress homogeneous background clutter and therefore prevent the presence of undesired targets. The principle is that for a given range-Doppler cell, namely Cell Under Test (CUT), the probability of detection is positively dependent on the probability of false alarm. Unlike in an ideal situation where the noise can be modeled as linear and Gaussian so that a mathematical model can be used to compute the optimal threshold for detection, such assumptions seldom hold in the real-world environment. Estimation of noise level is required, and this raises the OS technique, which selects a certain pre-defined value (median in this case) from the ordered sequence of reference cells. OS CFAR is chosen for its superior performance over Cell-Averaging (CA) CFAR given the presence of non-uniform clutter [70], such as the edges of tables and chair in our experimental environment which may be included in the reference cells.

A visualization of the implementation of OS CFAR is given in Figure 4.3. In this method, the reference cells (N_{ref} = 8 for range detection, N_{ref} = 4 for Doppler detection) are assumed to contain the background and clutter as in the CUT, whereas the most adjacent cells, namely guard cells (N_{guard} = 16 for range detection, N_{guard} = 0 for Doppler detection), are used to prevent object being treated as noise or clutter since human subject is treated as extended target for indoor HAR.

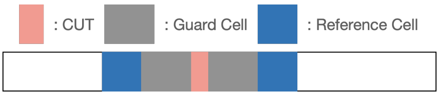


Figure 4.3: Visualization of OS CFAR for range and Doppler detection.

Quantitatively speaking, a scaling factor is applied on the estimated noise to find the optimal threshold as in equation 4.14,

$$T = \alpha P_{noise}, \tag{4.14a}$$

$$\mathrm{P}_{FA} = \left(1 + \frac{\alpha}{\mathrm{M}}\right)^{-\mathrm{M}} \tag{4.14b}$$

where, T is the detection threshold, P_{noise} is the estimated noise from adjacent cells, and α is a scaling factor that can be computed via the desired false alarm rate P_{FA}, and M is the number of reference cells. If the CUT is of an intensity larger than T, the detection result is positive, i.e., a scatter is present at the range-Doppler CUT.

4.2.4. Angle Spectrum Estimation and Detection

With a CUT detected as a target, the particular 2D range-Doppler voxel from all MIMO channels can be rearranged to the virtual array matrix form, e.g. 86 × 7 matrix. Next, 2D FFT can be applied to estimate the azimuth spectrum and elevation spectrum. The 86 × 7 virtual array matrix is padded with zeros to a 256 × 256. With an azimuth angle and elevation angle map, multiple local maximum values on azimuth and on elevation are chosen as the final detection results. The peak value detection algorithm is given in Algorithm 1. Use of the peak value selection algorithm is subject to the fact that human body tend to occupy a varying number of azimuth or elevation cells in terms of the present activity and the distance from body to the radar, causing CFAR with a constant number of reference and guard cells less preferable. The final detected scatter(s) after the transformation from spherical coordinate to Cartesian coordinate to construct a PC in XYZ space.

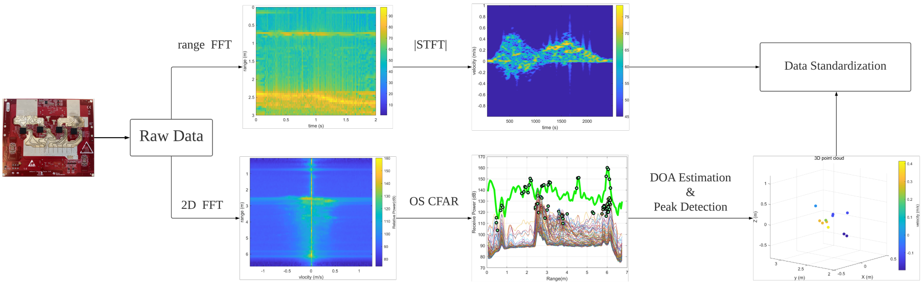


Figure 4.4: A zoomed-in visualization of the data generation flow.

4.2.5. STFT

A human activity typically lasts for over 1 second. During this time, human subject unnecessarily stays in designated range or Doppler bins. In other words, it is unreasonable to assume the received signal is stationary for the whole coherent processing interval. Thus, STFT becomes a reasonable solution that better exhibits the details of the micro-motions, under the same assumption as in section 4.2.2.

```
Data:
$M$ cross-range (azimuth or elevation) spectrum estimated by FFT with length $N$;
$s$, cross-range spectrum axis with length $N$;
Result:
$P$ power of the detection peaks ;
$\Theta$, angle of the detection peaks;
$\gamma \leftarrow 1.6$;
$P_{min} \leftarrow \infty$;
$P_{max} \leftarrow 0$;
$i \leftarrow 0$;
Peak Detection ← false;
while $i < N$ do
    $i \leftarrow i + 1$;
    $CUT \leftarrow M(i)$;
    if $CUT > P_{max}$ then
        $P_{max} \leftarrow CUT$;
        $\Theta_{max} \leftarrow \theta(i)$;
    end
    if $CUT < P_{min}$ then
        $P_{min} \leftarrow CUT$;
    end
    if Peak Detection then
        if $CUT < \frac{P_{max}}{\gamma}$ then
            $P \leftarrow P_{max}$;
            $\Theta \leftarrow \theta_{max}$;
            Peak Detection ← false;
        end
    else
        if $CUT > P_{min} \cdot \gamma$ then
            Peak Detection ← true;
        end
    end
end
```

Algorithm 1: Peak value detection algorithm for cross-range detection

First, a slow-time vector is generated by accumulating particular range bins from an arbitrary channel within n_{frame}th interval can be used and expressed as $s(n_t + m, n_{frame})$.

The formula to generate Doppler-time matrix and spectrogram are given in equation 4.15 and 4.16, respectively.

$$STFT(n_t, k, n_{frame}) = \sum_{m=0}^{L_{win}-1} s(n_t + m, n_{frame}) h(m) e^{-j2\pi \frac{mk}{L_{win}}}, \tag{4.15}$$

where $n_t = 1, \ldots, N_t$ is the time index and $k = 1, \ldots, K$ is the discrete frequency index of a 2D Doppler-time matrix [2] $S(n_t, k)$; $s(\cdot)$ expresses a slow-time vector for the given frame; and $L_{win} = 128$ is the length of the employed Hann window function $h(\cdot)$. Absolute square should be applied on the Doppler-time matrix to obtain the real-valued spectrogram, which expresses the energy intensities.

$$S(n_t, k, n_{frame}) = ||STFT(n_t, k, n_{frame})||_2^2 \tag{4.16}$$

A zoomed-in visualization of the data generation pipeline is given in Figure 4.4. Although some of the blocks (OS CFAR and peak detection) are implemented by Texas Instrument, the full pipeline is implemented as a contribution of this thesis. Intermediate outputs on the other hand show how the raw radar data is step-by-step transformed into informative data representations that can be processed by neural networks.

4.3. DL CLASSIFIERS

Referring to Figure 4.1, using the generated data representations three DL modules are implemented and combined in the proposed pipeline. This section explains these modules in detail.

4.3.1. ORIENTATION CLASSIFICATION MODULE

The orientation classification module is the same as the T-Net in [66] except the last fully-connected layer outputs the number of orientations instead of the 3 × 3 regularization matrix. The implementation details are given in Table 4.1. The structure of Conv1D are (C_{in}, C_{output}, K), where C_{in} and C_{out} are number of input and output channels, and K is the kernel size of 1D convolution.

To understand the design philosophy of T-Net, the characteristic of PC compared with other data representations should be firstly explained:

- Unlike images, PC is a set of **unordered** points, i.e., these points have no physical significance in their order. Neighboring points must be defined by the metric space distance instead of the position in the input set.

Considering this characteristic of PC, it is understandable why the 1D convolution layers (layer-1, 3 and 5 in Table 4.1) do not learn the orderly relations amongst points, but only extracts the geometric feature of each point to a high-dimensional latent space.

[2] Doppler-time and spectrogram are sometimes interchangeably used, whereas in this thesis Doppler-time is a complex-valued matrix after STFT, and spectrogram is the absolute matrix of Doppler-time

Table 4.1: Implementation details of the T-Net, where k is the number of outputs which could express the probability of classes or the regularization matrix in PointNet. The gray row expresses the modified symmetric function- average pooling layer- with the aim of achieving robustness against noise.

Structure
1. Conv1D(3,64,1)
2. Batch Norm & ReLU
3. Conv1D(64,128,1)
4. Batch Norm & ReLU
5. Conv1D(128,1024,1)
6. Batch Norm & Flattening
7. Average Pooling(L)
8. Fully Connected(1024,512)
9. Batch Norm & ReLU
10. Fully Connected(512,216)
11. Batch Norm & ReLU
12. Fully Connected(216,$k=6$)

Moreover, it is also understood why a global pooling layer is employed to find the relations amongst all points since pooling can be viewed as a symmetric function such that $F(a,b) = F(b,a)$, and through such a function the order of PCs is non-influential. Average pooling layer is used instead of the max pooling layer because average pooling is better suited to learning the global features for point cloud registration problem on noisy data [71].

The loss for the orientation classification module is the cross-entropy loss as expressed in equation 4.17.

$$\mathscr{L}_{T-Net} = \mathscr{L}_{C.-E.} = -\frac{1}{N}\left(\sum_{i=1}^{N} \mathbf{y}_i \cdot \log\left(\hat{\mathbf{y}}_i\right)\right), \tag{4.17}$$

where $\mathbf{y}_i$ and $\mathbf{y}_i$ express the ground truth and the predictions of the orientation, respectively.

In summary, the T-Net is modified to fit our application of radar-based angle-insensitive HAR by predicting the orientation of the human activities through corresponding PCs.

4.3.2. PC CLASSIFICATION MODULE

The PC classification module essentially is a modified PointNet [66] with a visualization shown in Figure 4.5, and the implementation details are given in Table 4.2. To explain the functionality of this module, it is logical to start with explaining PointNet, and another characteristic of PCs is noteworthy:

- As a geometric description of the object, PC should be **invariant** to rotations. In other words, if the PC is rotated to an angle, the physical meaning of the input should be maintained.

This is the reason for a transformation applied on the input PC and on the latent space. Such transformation is achieved by learning the geometric pattern (orientation) of the subject first, then generates a so-called regularization matrix to regularize the PC or latent space.

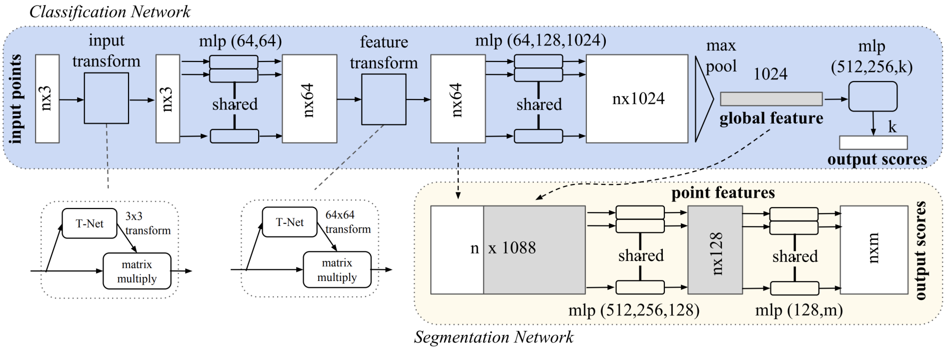


Figure 4.5: PointNet architecture overview (image taken from [66] and only the classification network is used).

As the T-Net is exactly meant to find the oriental information on the subject, it is pointless to re-train the 1D convolution layers of the first T-Net. This constitutes the first modification: the weights of 1D convolution layers in T-Net are pre-trained in the orientation classification module, as a result the loss function ($\mathscr{L}$) of the modified PointNet excludes this regularization loss as in equation 4.18,

$$\mathscr{L}_{PointNet} = \mathscr{L}_{C.-E.} + \mathscr{L}_{reg} = -\frac{1}{N}\left(\sum_{i=1}^{N} \mathbf{y}_i \cdot \log\left(\hat{\mathbf{y}}_i\right)\right) + ||I - A \cdot A^T||_2^2, \qquad (4.18)$$

where I expresses a 64×64 identity matrix, and A is the transform matrix generated through T-Net. The first part of the loss function ($\mathscr{L}$) is the cross-entropy loss between the predicted vector ($\boldsymbol{y}_i$) and the ground truth vector ($\hat{\boldsymbol{y}}_i$); while loss term is a regularization loss of the latent transformation matrix, forcing the regularization matrix to approximate an unitary matrix.

The second modification is replacing the max pooling layer as in [66] with average pooling layer to be robust to noise. This is consistent with the modification of the T-Net to achieve robustness against noise.

4.3.3. SPECTROGRAM CLASSIFICATION MODULE

For kinematically dynamic human motions, it is still convinced that Doppler-dependent representations, such as spectrogram, will be most efficient and representative according to the findings in the literature review (section 2). It would be unreasonable to abandon this representation. Also, since the PCs already contain the full spatial information on the target, spectrogram complements the other intrinsic features.

As for the architecture for the spectrogram (image) classification, a conventional CNN model- AlexNet [3] is chosen in this project, considering the difficulty in radar data

Table 4.2: Implementation details of the modified PointNet, where the gray row expresses the modified symmetric function- average pooling layer- in the interest of robustness against noise.

Layer	**Structure**
T-Net(3x3)	Table 4.1, with k = 9

1D Convolution(64,64)	Conv1D(3, 32, 1)
	BatchNormalization(32)
	ReLU(slope=0)
	Conv1D(32,64,1)
	BatchNormalization(64)
	ReLU(slope=0)

T-Net(64,64)	Table 4.1, with k = 4096

1D Convolution(64,128,1024)	Conv1D(64, 128, 1)
	BatchNormalization(128)
	ReLU(slope=0)
	Conv1D(128,1024,1)
	BatchNormalization(1024)
	ReLU(slope=0) & Flattening

Global Average Pooling	Average Pool

MLP(1024,256,k=4)	Fully Connected(1024, 512, 1)
	BatchNormalization(512)
	ReLU(slope=0)
	Fully Connected(512, 256, 1)
	BatchNormalization(256)
	ReLU(slope=0)
	Fully Connected(256, k=4, 1)

collection and the potential consequent insufficiency of training samples. To be more specific, AlexNet [3] is of a simple architecture, consisting overall 61,100,840 trainable parameters, which is smaller than other exemplary CNN models such as VGG [72] and ResNet-50 [73], with 138,357,544 and 68,883,240 trainable parameters, respectively. This hints that AlexNet [3] is a relatively a "easy-to-converge" model (or at least easier), and thus used. Moreover, AlexNet is pre-trained on ImageNet in order to achieve transfer learning from optical images to spectrogram images.

The implementation details of AlexNet [3] are given in Table 4.3, where Drop Out is a function that randomly sets a portion of elements in the input tensor to be zero.

Table 4.3: Architecture of the AlexNet [3]. The gray row expresses the last layer of feature extraction, which is followed by perception layers such as fully connected layer.

Input size: (channel,height,width) = (3x224x224)
Conv2D(3,64, kernel size=(11,11),stride=(4,4),padding=(2,2))
ReLU(slope=0) & Max Pooling(kernel size=(3,3),stride=(2,2),padding=(2,2))
Conv2D(64,192, kernel size=(5,5),stride=(1,1),padding=(2,2))
ReLU(slope=0) & Max Pooling(kernel size=(3,3),stride=(2,2),padding=(2,2))
Conv2D(192, 384, kernel size=(3, 3), stride=(1, 1), padding=(1, 1))
ReLU(slope=0)
Conv2D(384, 256, kernel size=(3, 3), stride=(1, 1), padding=(1, 1))
ReLU(slope=0)
Conv2D(256, 256, kernel size=(3, 3), stride=(1, 1), padding=(1, 1))
ReLU(slope=0) & Max Pooling(kernel size=(3,3),stride=(2,2),padding=(0,0))
Drop Out(probability=0.5) & Flattening
Fully connected(9126, 4096) & Additive bias
ReLU(slope=0) & Drop Out (probability=0.5)
Fully connected(4096, 4096) & Additive bias
ReLU(slope=0)
Fully connected(4096, $N_{class} = 6$)

4.4. Baselines for Performance Comparison

This work is novel in terms of the included activities, specifically the kinematically static postures combined with dynamic activities, as well as the classification pipeline exploiting 6 intrinsic features of imaging radar data, so there is a very limited amount of precedent work available for direct comparison. Therefore, only two baselines are defined as follows:

1. **Baseline-1 of state-of-the-art radar-based HAR**: The motion classification part of this pipeline is essentially consistent with the definition of angle-insensitive classifier proposed by [1]. Hence, in section 6.3 the performance of the proposed pipeline on 4 motions is to be compared with the counterpart of the method proposed by Yang et al. [1] [3]

[3] Even if the code of their model is not provided, the model is re-implemented using the identical architecture

2. **Baseline-2 of state-of-the-art radar-based HAR**: It is also valuable to compare if the 3D PC gives better results than simply viewing PCs as snapshots/image. This approach of treating PC as an image is implemented in [64], where 3D PCs corresponding to 2 seconds are stored as 2D images (see for examples, Figure 5.5). These images are then proceeded by Deep CNN for recognition of activities. Therefore, their proposed Deep CNN [64] is trained from scratch to compare the performance with my proposed pipeline [4]

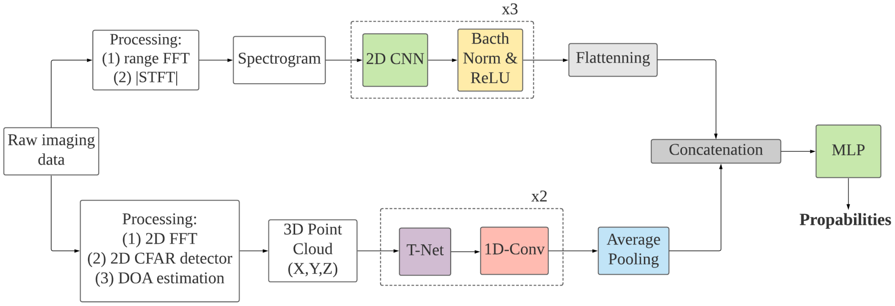


Figure 4.6: Baseline-3 that fuses (concatenates) the latent space extracted from spectrogram and PC.

3. **Baseline-3 to 4 for other data representations and pipeline structure**: 2 other-proposed baselines that exploit all intrinsic features of imaging radar data are discussed to comparatively study the performance of the final proposed pipeline. These two alternative baselines were developed in the initial part of this MSc project as precursors of the final proposed architecture.

   - **Baseline-3**: Fusing two data representations at the feature (latent space) level is a common approach in DL, for instance, [74] fuses the latent spaces extracted from spectrogram and cadence velocity diagrams. Baseline-3 is consequently proposed by fusing the latent space extracted from PCs and spectrograms to exploit all six intrinsic features of imaging radar data. A visualization of the baseline architecture is shown in Figure 4.6. As can be seen, two neural networks are in parallel instead of cascaded as in the final proposed pipeline, also T-Net is not employed in this baseline.

   - **Baseline-4**: A baseline that firstly uses T-Net for orientation classification, secondly uses a point set that consists of 6 intrinsic features as the input for activity classification. The 6 intrinsic features are attainable from an imaging radar, including x, y, z, estimated SNR expressed in logarithmic scale (power

as given in the paper, while hyper-parameters are fine-tuned empirically since they are not specified in the literature.

[4] Even if the code of their model is not provided, the model is re-implemented using the identical architecture as given in the paper, while hyper-parameters are fine-tuned empirically since they are not specified in the literature.

of CUT divided by averaged power of reference cells), Doppler and time (expressed by the frame index of the detected point, ranging from 1 to 20). The architecture is the same as described in Table 4.2 except the input size is $N \times 6$ and the output of the first T-Net changes to 6×6.

5

Measurement Setup and Dataset Construction

With all the radar parameters determined, signal processing and classification chain presented, proper data generation is thus desired. This thesis uses real-life imaging radar data to comparatively analyze the performance of the proposed pipeline. This chapter describes the laboratory set up in section 5.1, and the construction of the dataset in section 5.2.

5.1. Measurement Setup

The measurements were taken in the Lage Hallen of EWI in TU Delft. The specific dimensions of the measurement setup are as follows:

- The distance from the ground to the radar is $0.75m$ such that the antennas could cover the human subjects in the elevational field-of-view;
- The distance from the radar to the left, right, and front walls are $2.1m$, $1.8m$ and $5m$ respectively;
- Shown in Figure 5.1 is a sketch of the room layout indicating the positions of radar and human subject, human orientation directions, and potential sources of clutter (tables, walls and closet).

5.2. Dataset Description

To the best of my knowledge, there is no existing open-source imaging radar-based human activity dataset that combines static postures and dynamic activities for HAR. to examine the performance of the proposed HAR pipeline, collecting custom data sets appears to be the only option that also gives a degree of freedom to include both kinematically dynamic motions and static postures. The list of the included motions and

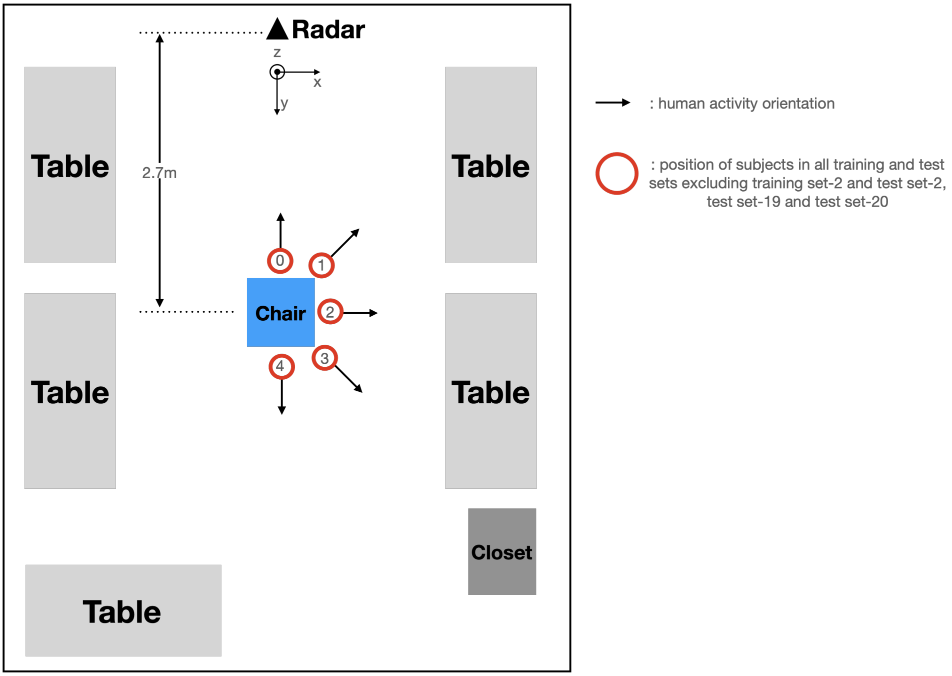

Radar
z
x
y
2.7m
Table
Table
Table
Table
Table
Chair
0
1
2
3
4
Closet
: human activity orientation
: position of subjects in all training and test sets excluding training set-2 and test set-2, test set-19 and test set-20

Figure 5.1: Data measurement setup in an office-like room at TU Delft. The 12 positions for the measurements refer to different subsets of data used for training and testing and are described in section 5.2.1

postures is shown in Table 5.1. The chosen motions are the most common ones according to the previous literature analyzed in Chapter 2, and those postures can be seen as a kinematically static state between the motions. As a result, the radar-based HAR problem is extended beyond ’motion’ to ’motion plus postures’.

Table 5.1: List of motions and postures to simulate daily human activities, where the four different colors represent the output classes of PC classification module.

Dynamic motion	**Static posture**
1. Sitting down to a chair	5. Standing still
2. Standing up from a chair	6. Sitting still on chair
3. Bending over	
4. Standing up after bending	

The body characteristics of the eight human subjects who participated in the measurements are described in Table 5.2. All the subjects are male and aged between 20 and 30 years old.

Table 5.2: Body characteristics of the participated human subjects.

Subject index	1	2	3	4	5	6	7	8	Mean±Std
Height (cm)	180	168	170	180	185	178	177	177	176.9±5.5
Weight (kg)	75	70	70	70	95	82	72	72	75.7± 8.8

5.2.1. Training and Test Sets

Overall **20 training sets** and **23 test sets** are generated via experimental measurement and their manipulations such as adding white noise to the measured data. Additionally, two pairs of training and test sets are generated for baseline approaches ([1] and [64]). The main objective behind these multiple subsets of data is to have *1)* a large dataset (*training/test set-1*) for basic performance evaluation of the proposed pipeline, *2)* several evaluations on smaller sets (*training set-2 to -20 and test set-2 to -20*) designed to explore the impact of specific variables such as the aspect angle, the kinematic diversity from subject to subject, amongst others. The descriptions of these training and test sets are as follows (with reference to the positions shown in Figure 5.1 as well):

- For *training set-1* and *test set-1*, postures and motion are recorded separately at positions 0 to 4. Specifically, one complete time interval for measurement is 2 minutes. During this time, human subjects were asked to perform either one static posture, e.g. sitting still on the chair, or a pair of complementary motions, e.g., sitting down and standing up at a period of approximately 2 seconds for each motion. Labels are then generated manually for these data by visual segmentation. This particular way of data collection is an attempt to ease the process of labeling and calibration by conditioning the execution time of each movement, at the price of less realistic kinematics. The distributions of 6 activities in dataset-1 are given in Figure 5.2, with overall 13,433 samples. Dataset-1 then is divided by 80% and 20%, forming *training set-1* and *test set-1* respectively.

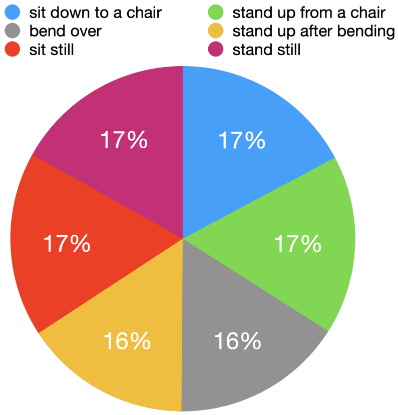


Figure 5.2: Distributions of dataset-1 (splitted into *training set-1* and *test set-1*).

- Furthermore, as explained in the literature review part, radar data collection and labeling are genuinely time-consuming. This additional test set attempts to exploit the symmetry of human bodys to generate PCs and spectrograms. According to Figure 5.1, PCs on position 7, for instance, can be treated as the PCs mirrored from position 1 with respect to the line-of-sight direction (Y axis in the given coordinate system). Spectrograms can be exactly duplicated because the aspect angle of movement orientation for positions 1 and 7 are the same. The same method also applies for positions 6 (from position 2) and 5 (from 3). The simulated *training set-2* therefore includes overall 3 orientations and 6 activities [1]. To test whether human subjects perform in-place activities in a strictly symmetric manner, *test set-2* is experimentally created and contains the data of one human subject sequentially performing all 6 activities on positions 5 to 7.

- Another dataset is obtained through adding additive white Gaussian noise to the measured raw data while assuming the measured raw data as noise-free. The attained SNR levels are, 20dB, 18dB, 15dB, 13dB, 10dB and 8dB. Also, it is noteworthy that the change in SNR from 20dB to 10dB corresponds to a synthetic range variation of 1.78 times further range (e.g., from 2.7m to 4.8m), according to the relation $P_{Rx} \propto \frac{1}{R^4}$ (extracted from 4.5a). The noisy datasets are divided by 80% and 20%, compositing training and *test set-3* (20dB), *test set-7* (10dB), and *test set-8* (8dB).

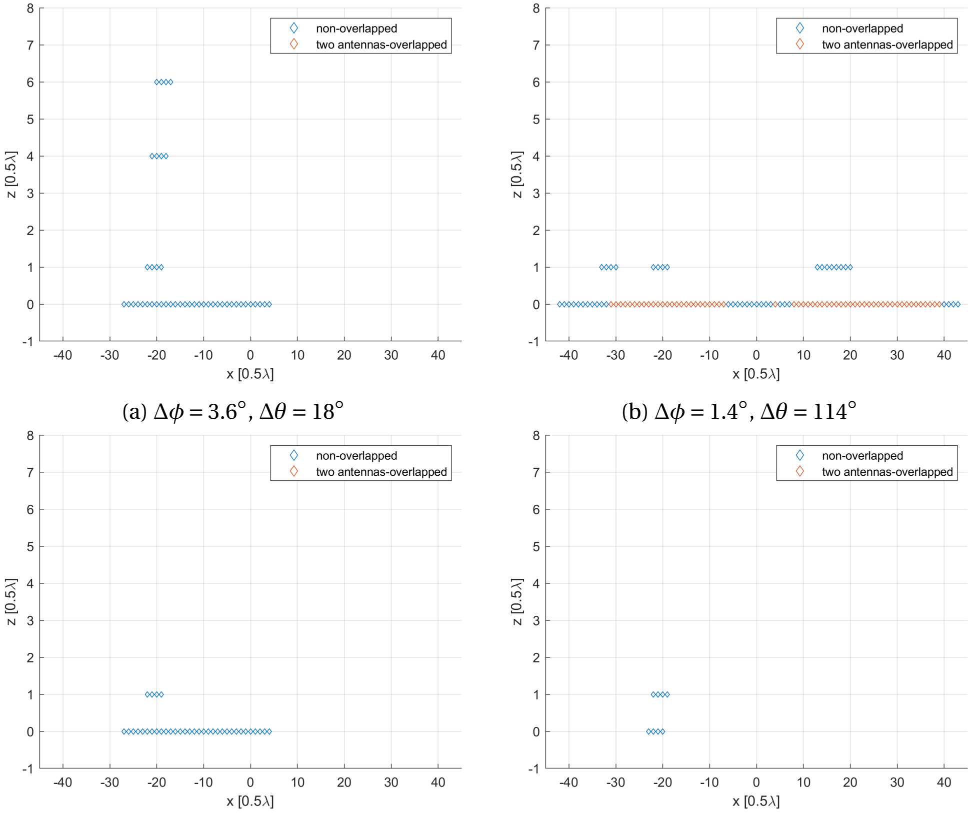
- To examine to what extent of performance variation the diversity of the human subjects may lead, the so-called leave-one-subject-out test is a common approach as in [13]. The leave-one-subject-out test is dividing the total dataset in terms of the subject, for instance in our dataset, data of one of the eight subjects is left-out for testing, data of all other seven subjects are used for training instead. Therefore, overall eight training and test sets are generated (*training and test set-9 to -16*), where the left-out subject is 1 to 8 (see Table 5.2) in order.

- Having a MIMO aperture of 43λ on the horizontal channel domain and 3λ on the vertical channel domain provides fairly good cross-range resolutions. However, it may be unpractical or too expensive to use such a high-resolution imaging radar for HAR. *Training set and test set-17 to -20* are created by processing only a subset of the original virtual array. The corresponding MIMO apertures are shown in Figure 5.3 together with their corresponding derived angular resolutions.

- Last but not least, two datasets are created for the baselines of Yang's work [1] and Kim's work [65], separately. *Baseline training set-1* and *baseline test set-1* are respectively the spectrogram part of *training set-1* and *test set-1*. Similarly, *baseline training set-2* and *baseline test set-2* only include those of PCs.

A summary of all training sets and test sets is given in Table 5.3. Exp. stands for data collected directly from experiments; Sim. stands for simulation of data by adding white

[1] It should be noted that the word 'simulated' here refers to the fact that these data were not experimentally measured, but generated synthetically from symmetry and/or duplication of the experimentally measured data.



(a) $\Delta\phi = 3.6°$, $\Delta\theta = 18°$

(b) $\Delta\phi = 1.4°$, $\Delta\theta = 114°$

(c) $\Delta\phi = 3.6°$, $\Delta\theta = 114°$

(d) $\Delta\phi = 22.9°$, $\Delta\theta = 114°$

Figure 5.3: Virtual arrays and corresponding angular resolutions of training and test sets (a) 17, (b) 18, (c) 19, and (d) 20.

noise or by synthetic manipulation (e.g., rotation) of PCs; Orien. stands for orientation; V.A. stands for virtual array; # Sample expresses the average number of samples per activity in the corresponding dataset, where one sample includes one for data representation as in the column of input.

To evaluate if the constructed dataset is sufficient to train a data-driven model for radar-based HAR, the numbers of samples are listed in Table 2.1 from the relevant literature and those in Table 5.3 used in this thesis. For instance, 288, 500 and 1791 samples per activity are included in the datasets of [10], [27] and in this work, respectively. It is established that the generated dataset is larger than the typical radar-based HAR datasets, whereas no data augmentation or simulation methods are used in this work. Moreover, taking the fact that the employed neural networks are of simple architecture into consideration, it is reasonable to conclude that even though the obtained training samples are far smaller than the typical number of samples in other fields, e.g., 14,197,122 in ImageNet [75] or more than 8,120,000 in natural language dataset-Yelp, the datasets in Table 5.3 are sufficient to see the convergence of the neural networks processing radar data representations.

Table 5.3: Summary of training and test sets described in section 5.2.1.

Dataset	**Exp./ Sim.**	**# Subjects**	**# Orien.**	**SNR**	**V.A.**	**Position (5.1)**	**Input**	**# Samples**
Training set-1	Exp.	8	5	meas.	3.3	0-4	PC + spc.	1791
Test set-1	Exp.	8	5	meas.	3.3	0-4	PC + spc.	448
Training set-2	Sim.	8	3	meas.	3.3	5-10	PC + spc.	8070
Test set-2	Exp.	1	3	meas.	3.3	5-10	PC + spc.	357
Training set-3	Sim.	8	5	$\approx 20dB$	3.3	0-4	PC + spc.	1791
Test set-3	Sim.	8	5	$\approx 20dB$	3.3	0-4	PC + spc.	448
Training set-4	Sim.	8	5	$\approx 18dB$	3.3	0-4	PC + spc.	1791
Test set-4	Sim.	8	5	$\approx 18dB$	3.3	0-4	PC + spc.	448
Training set-5	Sim.	8	5	$\approx 15dB$	3.3	0-4	PC + spc.	1791
Test set-5	Sim.	8	5	$\approx 15dB$	3.3	0-4	PC + spc.	448
Training set-6	Sim.	8	5	$\approx 13dB$	3.3	0-4	PC + spc.	1791
Test set-6	Sim.	8	5	$\approx 13dB$	3.3	0-4	PC + spc.	448
Training set-7	Sim.	8	5	$\approx 10dB$	3.3	0-4	PC + spc.	1791
Test set-7	Sim.	8	5	$\approx 10dB$	3.3	0-4	PC + spc.	448
Training set-8	Sim.	8	5	$\approx 8dB$	3.3	0-4	PC + spc.	1791
Test set-8	Sim.	8	5	$\approx 8dB$	3.3	0-4	PC + spc.	448
Training set-9	Exp.	7	5	meas.	3.3	0-4	PC + spc.	1567
Test set-9	Exp.	1	5	meas.	3.3	0-4	PC + spc.	51
Training set-10	Exp.	7	5	meas.	3.3	0-4	PC + spc.	1567
Test set-10	Exp.	1	5	meas.	3.3	0-4	PC + spc.	51
Training set-11	Exp.	7	5	meas.	3.3	0-4	PC + spc.	1567
Test set-11	Exp.	1	5	meas.	3.3	0-4	PC + spc.	51
Training set-12	Exp.	7	5	meas.	3.3	0-4	PC + spc.	1567
Test set-12	Exp.	1	5	meas.	3.3	0-4	PC + spc.	51
Training set-13	Exp.	7	5	meas.	3.3	0-4	PC + spc.	1567
Test set-13	Exp.	1	5	meas.	3.3	0-4	PC + spc.	51
Training set-14	Exp.	7	5	meas.	3.3	0-4	PC + spc.	1567

Dataset	**Exp./ Sim.**	**# Subjects**	**# Orien.**	**SNR**	**V.A.**	**Position (5.1)**	**Input**	**# Sample**
Test set-14	Exp.	1	5	meas.	3.3	0-4	PC + spc.	51
Training set-15	Exp.	7	5	meas.	3.3	0-4	PC + spc.	1567
Test set-15	Exp.	1	5	meas.	3.3	0-4	PC + spc.	51
Training set-16	Exp.	7	5	meas.	3.3	0-4	PC + spc.	1567
Test set-16	Exp.	1	5	meas.	3.3	0-4	PC + spc.	51
Training set-17	Sim.	8	5	meas.	5.3a	0-4	PC + spc.	1791
Test set-17	Sim.	8	5	meas.	5.3a	0-4	PC + spc.	408
Training set-18	Sim.	8	5	meas.	5.3b	0-4	PC + spc.	1791
Test set-18	Sim.	8	5	meas.	5.3b	0-4	PC + spc.	408
Training set-19	Sim.	8	5	meas.	5.3c	0-4	PC + spc.	1791
Test set-19	Sim.	8	5	meas.	5.3c	0-4	PC + spc.	408
Training set-20	Sim.	8	5	meas.	5.3d	0-4	PC + spc.	1791
Test set-20	Sim.	8	5	meas.	5.3d	0-4	PC + spc.	408
Baseline training set-1	Exp.	8	5	meas.	3.3	0-4	spc.	1791
Baseline test set-1	Exp.	8	5	meas.	3.3	0-4	spc.	448
Baseline training set-2	Exp.	8	5	meas.	3.3	0-4	PC(image)	1791
Baseline test set-2	Exp.	8	5	meas.	3.3	0-4	PC(image)	448

5.3. DATA PREPARATION

After the signal processing algorithms, some preparation is still needed such that input samples can be better processed by neural networks. This section introduces the data standardization methods employed for the thesis.

5.3.1. DATA STANDARDIZATION

Data standardization is a common technique to prepare the pre-processed data for the DL classifiers by removing the non-relevant differences amongst data samples.

PCs are standardized by *(1)* subtracting the mean value of each dimension from points, then *(2)* dividing each of the X,Y,Z coordinate in the point set with its Euclidean distance. For *baseline training/test set-2*, such standardization is also applied over the dimension of Doppler, power and time index. Mathematically, standardization of PCs can be expressed as:

$$c_{standardized} = \frac{c - mean(c)}{||p||_2} \tag{5.1}$$

where X and $X_{standardized}$ is x, y or z coordinate of a point before and after standardization, respectively; and $p = (c_x, c_y, c_z)$ expresses the position of the point.

On the other hand, spectrogram images are standardized by *(1)* subtracting the mean value of each Red, Green or Blue (RGB) channels, then *(2)* dividing each of the RGB channel in the image with its standard deviation. The mathematical expression is the same as the counterpart of PC standardization (equation 5.1) except replacing coordinate with an RGB channel, and Euclidean distance of a point with the standard deviation of an RGB channel of the overall dataset.

In this way of standardization, the input samples are not within a particular range (e.g. $[0, 1]$). However, the standardized samples are robust against outliers [76]. With such standardization applied, the results would be conclusive to show that orientation classification is not based on the variation in positions but the actual geometric patterns. For instance, the trivial difference caused by the position variation (position 0 to position 8 for instance) could be eliminated. Moreover, data standardization is a common technique to help increase the convergence rate of neural networks [76].

5.4. VISUALIZATION OF DATA REPRESENTATIONS

This section shows examples of the generated data representations in terms of different activities and human orientations from randomly selected human subjects (Figure 5.4 and 5.5). Specific implementation details to generate these input samples are as follows:

- Spectrograms are generated with a time interval of 2 seconds since the designated activity period is 2 seconds for all data collections. The length of one window is 128 chirps (equivalent to $80.64ms$), while the length of overlapped window is 127 chirps for a smoother appearance of the resulting spectrograms.
- In this thesis, the number of reference cells, M, is equal to 8 in range domain and 4 in Doppler domain; while the scaling factor, α, is chosen as $6.3 = 8dB$ in , giving the probability of false alarm rate $P_{FA} = 0.0096$ in range domain and 0.0227 in Doppler domain according to equation 4.14.

- Theoretically PC can be generated from an arbitrary CPI, while 20 sub-PCs with each sub-PC generated from 0.1 seconds of data are aggregated as a complete PC for two reasons,
  - the physical time interval of one PC is therefore consistent with the spectrograms so that final classification is accomplished on a coherent dimension,
  - the sub-PC generated from one sub-frame is generally too sparse to represent a human subject [64], however, aggregation of PCs over multiple intervals leads to dense and representative samples,
- Spectrograms are expressed in decibel scale:

$$Image(n_{frame}) = 10 \cdot \log_{10} S(n_t, k, n_{frame}) \tag{5.2}$$

- To suppress the presence of noise, a threshold is applied to set the minimum energy level to be 45dB.
- Using the prior knowledge of where the human subjects are, only the points inside the designated areas are kept as the target-originated when generating PCs. These areas are:
  - $x \in [-0.75, 0.75]m$, $y \in [2, 3.1]m$, and $z \in [-0.75, 1.25]m$ for position indices 0, 1 and 2;
  - $x \in [-0.75, 0.75]m$, $y \in [2.4, 3.5]m$, and $z \in [-0.75, 1.25]m$ for position indices 3 and 4;
  - $x \in [-0.75, 0.75]m$, $y \in [, 3.5]m$, and $z \in [-0.75, 1.25]m$ for position indices 8, 9 and 10;
  - $x \in [-0.75, 0.75]m$, $y \in [, 3.5]m$, and $z \in [-0.75, 1.25]m$ for position indices 11 and 12;
- As in Table 4.2, the input size of PC is fixed (1024×3). However, this is not ensured in practice due to the statistical nature of detection algorithms and the unpredictable scattering behavior from the human body at mm-Wave frequencies. Therefore, random up-sampling or down-sampling is applied on the detected PCs. While this is a crude approach to counteract this problem, it is the quickest way to get the pipeline to work without extensive modifications to the network to accept PCs of variable lengths.

Based on the spectrogram images in Figure 5.4, the movements (top 4 rows) and postures (bottom 2 rows) are visually distinctive. Yet, two postures as expected originate almost zero Doppler regardless of the human orientation. This shows that the spectrogram-based HAR methods cannot consider static postures as classes. It is also clearly observable the compression of Doppler level with respect to aspect angle, suggesting the difficulty in spectrogram-based classification of movements at large aspect angle close to 90° (e.g. 3rd column of Figure 5.4).

On the other hand, it is more difficult to get visual cues from PCs, though the differences caused by variation in aspect angle could be seen for instance by comparing the third column with the first column in Figure 5.5. Moreover, the detected PCs rarely have points close to the floor ($z = -0.75m$). This could be the result of the relatively small reflective area of the lower legs and feet, and the attenuated radiation power of individual transmit antennas (as shown in Figure 3.5). However, it can be safely assumed that such absence of data in lower human body parts is less important in this work since they are not as representative as the torso for human activities [55].

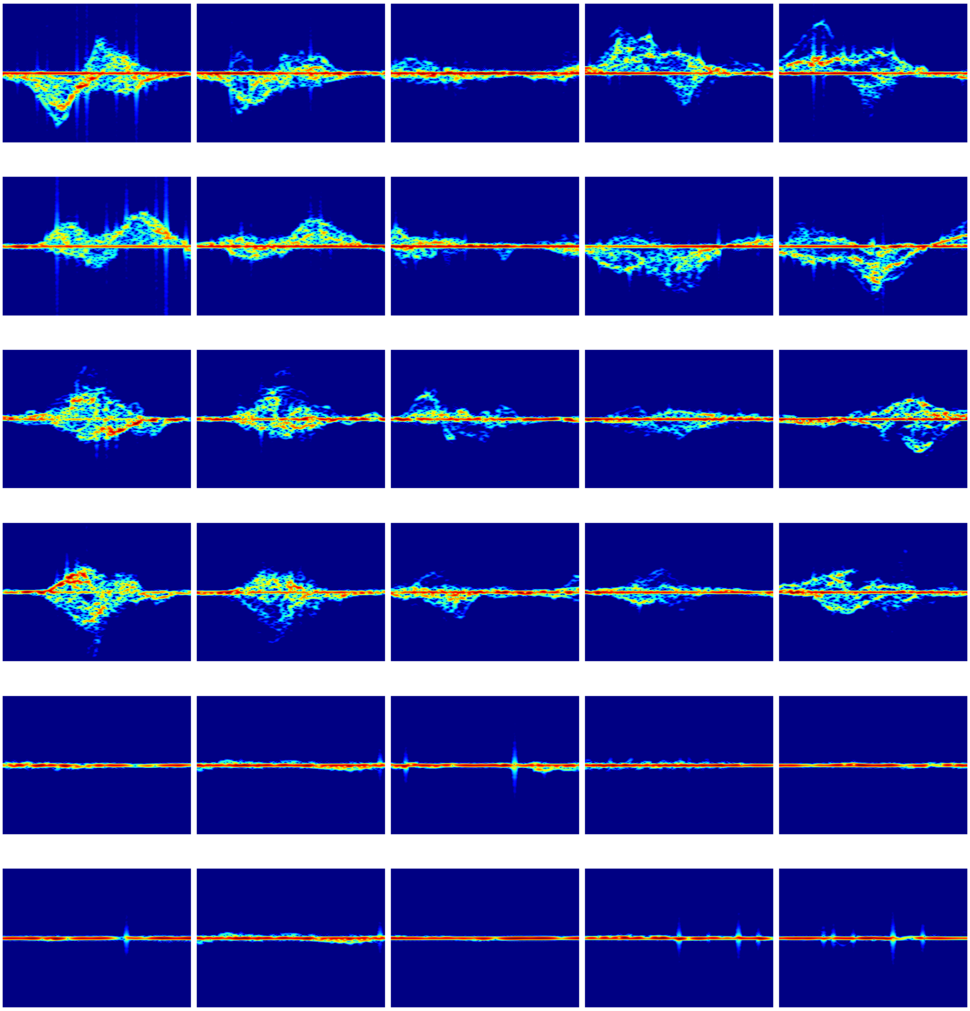


Figure 5.4: Examples of spectrogram images from the measured dataset (*training set-1*). From top to bottom are motions of sitting down, standing up from sitting, bending over and standing up from bending, and postures of sitting still and standing still, respectively. From left to right are of different aspect angles, at positions 0 to 4 (Figure 5.1), respectively. The horizontal axis is of approximately 2 seconds, and vertical axis expresses velocity in ±1.26m/s.

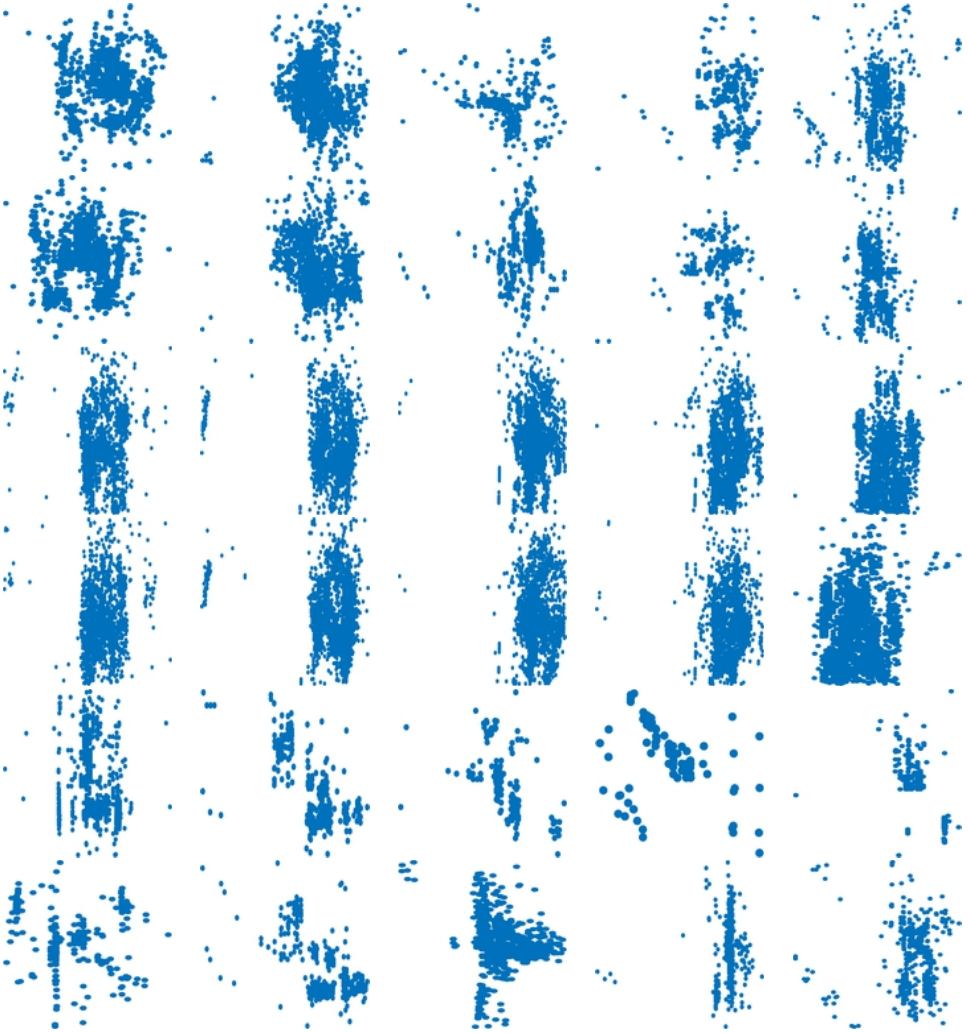


Figure 5.5: Examples of the front view of PCs from the measured dataset (training set-1). From top to bottom are motions of sitting down, standing up from sitting, bending over and standing up from bending, and postures of sitting still and standing still, respectively. From left to right are of different aspect angles, at positions 0 to 4 (Figure 5.1), respectively.

6 Results

This chapter analyzes the performance of the proposed method through different experiments and metrics. The gain due to each module in the proposed pipeline is examined as an ablation study in Section 6.2. Section 6.3 shows the performance of the proposed pipeline in comparison with the state-of-art baseline architectures. Sections 6.4, 6.5 and 6.6 investigate the performance of the proposed pipeline in terms of different aspects of robustness. Last but not least, section 6.7 analyzes the error sources and suggests what can be improved.

Before going into the details, some common techniques are used in the experiment(s) to improve the reliability of the results as follows:

- The hyper-parameters, e.g., Learning Rate (LR) and Batch Size (BS), are empirically fine-tuned for different training sets and classifiers, including the proposed pipeline and the baselines, to optimize the classifier performances;

- Typically, data-driven methods approach the global minima using gradient descent. Hence, it normally takes more than 1 epoch/iteration for a neural network to reach its optimal state. Choosing a good number of training epochs could provide optimal performance. In this work, the number of training epochs is selected based on the empirical observation of the loss/accuracy history (e.g., Figures 6.2 and 6.3). During the training process, the model that provides the highest classification accuracy for the particular training set is stored and then used for independent testing.

- To compare the classifier performance, accuracy P_a and $(F1 - score)$ (Equation 6.1) are seen as promising metrics for a well-balanced dataset as the one collected for this work (see Figure 5.2).

$$P_a = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6.1a}$$

$$P_p = \frac{TP}{TP+FP}, \tag{6.1b}$$

$$P_r = \frac{TP}{TP+FN}, \tag{6.1c}$$

$$F1-score = 2\frac{(P_r \cdot P_p)}{(P_r + P_p)} = 2 \cdot \frac{TP}{2TP+FP+FN}, \tag{6.1d}$$

where TP, TN, FP, and FN denote the true positive, true negative, false positive and false negative, respectively. These metrics are used to quantitatively reflect the classification performance of a pipeline.

- To examine how robust the proposed pipeline is against the variations in human orientations, the concept of angle sensitivity is inherited from [1] and explained in detail in section 4.1. Correspondingly, Angle Sensitivity Matrix (ASM) and Angle Sensitivity Vector (ASV) are used to show the angle sensitivity. The former has two dimensions of training and test orientation, and the latter squeezes from a matrix into a vector since data of all orientations are used for training together. Last, to give quantitative metrics on the angle sensitivity of a classifier, mean value $\bar{X}$ and L2 distance $||v||_2$ of ASM and ASV are employed. Suppose there are M individual training sets, each corresponds to the data of one particular aspect angle, and similarly, there are N test sets. As a result, we get the metrics as follows:

  - ASM is an $M \times N$ matrix, where the cell on ith row and jth column expresses the classification accuracy with the training data from ith angle and test data from jth angle;
  - Similarly, ASV is a vector of length N, where each element again expresses the accuracy given the particular data for testing;
  - *Mean*: $\bar{X} = \frac{1}{M}\sum_{i,j} x_{i,j}$, where X expresses ASM or ASV and x expresses the entry of the element. Ideally, ASM or ASV should be filled with classification accuracy of 100%. Therefore, this metric expresses the classification accuracy of an angle-insensitive classifier/pipeline.
  - L2-distance: $||v||_2 = \frac{1}{M}\sqrt{||\sum_{i,j} x_{i,j} - 1||}$. The result reflects the sensitivity of the classifier/pipeline and should be as close to 0 as possible.

6.1. Results of the Proposed Pipeline

This section presents the results of the proposed pipeline classifying 6 activities (i.e., trained with *train set-1* and tested with *test set-1*).

Figure 6.1 uses t-SNE (t-distributed stochastic neighbor embedding) [77] to visualize the examples of the flattened feature vectors to help explain the working mechanism of neural networks. As can be seen, the clusters are merged but distinguishable in the flattened latent space in the orientation classification module, where clusters are grouped according to the human subject's aspect angle. Meanwhile, the clusters are well separated and even the existence of different orientations (e.g., five separate red clusters) can

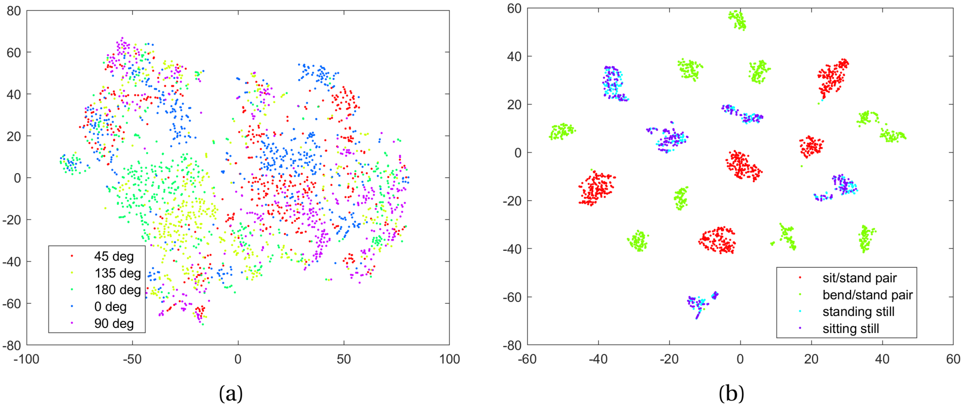


Figure 6.1: t-SNE visualizations of (a) the flattened latent space in orientation classification module, and (b) the flattened latent space extracted in PC classification module.

be clearly observed. This discloses the functionality of the feature-extracting layers, i.e., the layers before the flattening operation can be viewed as a complete feature extraction function, analog to handcrafted feature extraction The fully-connected layers following the flattened vector can be viewed as an approximation of nonlinear functions to achieve classification, e.g., drawing a line between the clusters in Figure 6.1 to separate them.

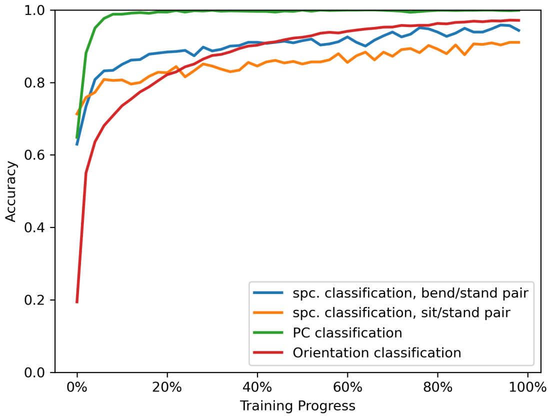


Figure 6.2: Examples of the training accuracy history of each module in the proposed pipeline, where the training process includes 500 epochs for orientation classification module and 50 epochs for the others.

Figures 6.2 and 6.3 present the examples of the accuracy and loss history of each module in the training process, respectively, where the training process includes 500 epochs for orientation classification module and 50 epochs for the others.. As can be

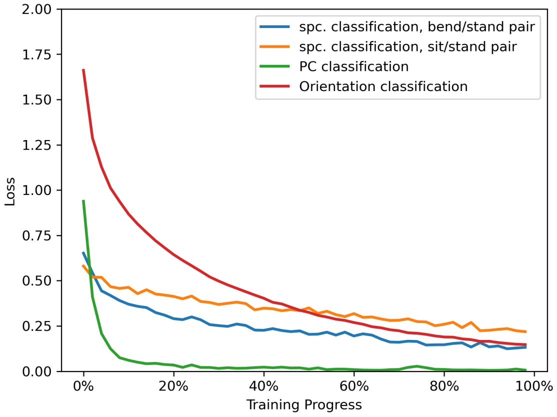


Figure 6.3: Examples of the training loss history of each module in the proposed pipeline, where the training process includes 500 epochs for orientation classification module and 50 epochs for the others.

seen, the loss and accuracy of the validation curve reach a stable state, i.e. no longer decreases or increases with respect to epochs. Therefore, it is reasonable to conclude that the trained models have reached stability and converged, and therefore, represent the optimal performance of the proposed pipeline.

Quantitative metrics of three modules are as follows:

- The orientation classification module makes very accurate predictions of the subject orientation ($P_a = 0.975$), which presumably will simplify the task for PC and spectrogram classification module;

- PC classification module, using the spatial information of the subject, also makes promising classification with the average accuracy, $Pa = 0.969$;

- Spectrogram classification, however, has less favorable performance:

  - The average binary classification accuracy for sitting down/standing up from sitting shows $P_{a,sit/stand} = 0.838$; and

  - The average binary classification accuracy for bending down/standing up from bending shows $P_{a,bend/stand} = 0.855$.

In conclusion, the working mechanism of the proposed pipeline is unveiled through visualization, and the promising classification performance of the proposed pipeline classifying 6 activities is validated with an accuracy of 87.0% and an F1-score of 0.867.

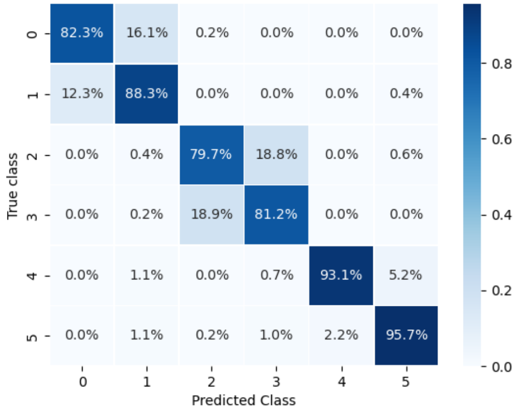


Figure 6.4: Normalized confusion matrix for the proposed pipeline to classify 6 activities, where class-0 to 5 expresses sitting down, standing up from sitting, bending over, standing up from bending, sitting still and standing still, respectively.

6.2. RESULTS OF ABLATION STUDY

Through the removal of certain classification modules in the proposed pipeline, this section is expected to study the contribution of each module. Table 6.1 includes the results obtained from the ablation study. As suggested by the columns of **Difference** in Table 6.1, it is reasonable to conclude that each module makes a crucial contribution to the advantageous performance of the proposed pipeline. Meanwhile, comparing first two rows with third to fifth rows reveals the importance of the two complementary data representations as the results attained through one data representation is significantly lower.

Table 6.1: Summary of the results of ablation study

Used Module	**Accuracy (%)**	**Difference (%)**	**F1-score (%)**	**Difference (%)**
PC classification	74.7	-12.3	74.9	-11.8
spec. classification	74.0	-13	74.2	-12.5
Ori classification & spec. classification	77.1	-9.9	77.2	-9.5
PC classification & spec. classification	81.8	-5.2	80.7	-6
Ori classification & PC. classification	81.8	-5.2	81.4	-5.3
Full Pipeline	87.0	0	86.7	0

6.3. Results of the Baseline Approaches

Comparisons with state-of-the-art algorithms are a common and important step to critically evaluate the proposed method. This section demonstrates the results of four baselines. It should be noted that neural networks in Yang et al.'s work [1] and in Kim's work [64] are implemented by the author and trained from scratch as the pre-trained models or training data was not made publicly available by the authors, while the AlexNet (Table 4.3) in baseline-3 is pre-trained on ImageNet.

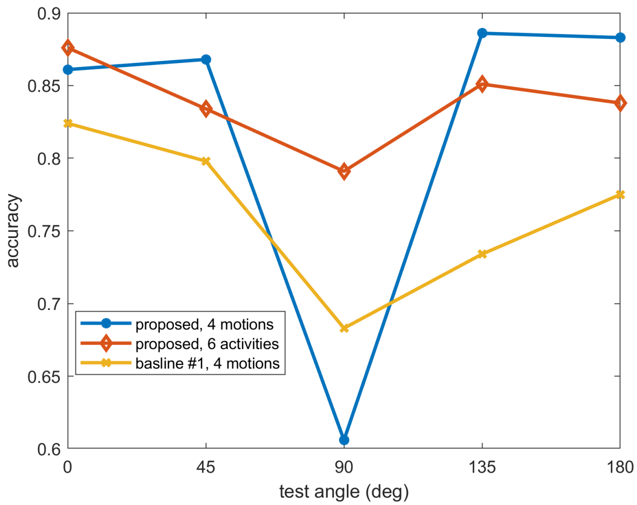


Figure 6.5: ASVs of the proposed pipeline and the baseline-1 [1] as a function of aspect angle.

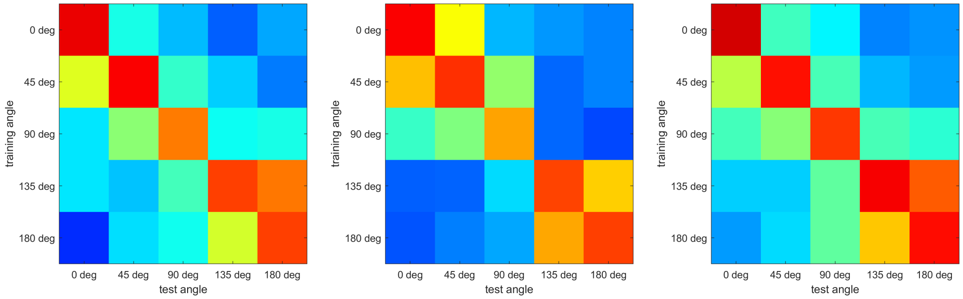


Figure 6.6: ASMs of (a) the proposed pipeline for 4 motions, (b) the baseline [1] for 4 motions, and (c) the proposed pipeline for 6 activities, where classification accuracy of each cell ranges from 0 (blue) to 1 (red).

Regarding the proposed pipeline as a SAC, the orientation classification module becomes pointless since only the data of one orientation can be used for training. As a result, only PC and spectrogram classification modules are utilized, the corresponding

ASM and ASV in comparison with those of the baseline ([1]) are given in Figures 6.5 and 6.6 and Table 6.2. It should be noted that, since the conventional spectrogram-based classifiers lack the necessary spatial information to classify postures, only four motions are examined in baseline-1. Four conclusive points are drawn as follows:

- Figure 6.5 shows that the closer to 90 degree human orientation is, the worse results are attained, considering all methods. This is presumably subject to the fact that the movement of a torso with an aspect angle of 90 deg do not generate representative Doppler features for classification;
- Figure 6.6 shows that the proposed pipeline outperforms Yang's method [1] for the seen test data as shown by the cells on the diagonal. Yet, the proposed pipeline appears to be more sensitive to angle variations than Yang's method [1] given test data with an orientation close to the training data (e.g., the cells on the nondiagonal upper left or bottom right parts). Meanwhile, for those results obtained from a test angle far from the angle of the training data (e.g., the cells on the upper-right or bottom left parts), the proposed pipeline again shows superiority thanks to the use of additional spatial information.
- According to Table 6.2, the proposed pipeline's mean accuracy is closer to 100% than the baseline's, and the L2-distance is smaller. Therefore, quantitatively speaking, the proposed pipeline is better at classifying motions and/or postures than the baseline in terms of MAC or SAC.
- Moreover, the classification results of 6 activities are given in Table 6.3, where the advantageous performance of the proposed pipeline over baselines in terms of classification accuracy and F1-score is presented in the columns of **Difference**.
  - It is noticeable that the F1-score of baseline-3 has an F1-score of NaN (Not a Number). This is caused by the wrong prediction of Doppler-related motions, i.e. none of the samples is predicted as sitting down to chair nor bending over.
  - The advantage of the proposed pipeline over the baseline-2 and 3 is obvious, considering a difference of at least 12.1% in accuracy and 13.1% in F1-score.

Table 6.2: Quantitative metrics of the baseline [1] in comparison with the proposed pipeline.

Method	**Activity**	**MAC-based results**		**SAC-based results**	
		$\bar{X}$	$\|\|v\|\|_2$	$\bar{X}$	$\|\|v\|\|_2$
proposed	4 motions	0.8208	0.0935	0.4767	0.1050
proposed	6 activities	0.8380	0.0735	0.5077	0.0988
baseline-1 [1]	4 motions	0.7628	0.1084	0.4558	0.1091

As discussed in the literature review, radar data measurement is expensive and time-consuming, limiting the typical dataset to be much smaller than the computer vision

Table 6.3: Classification results of the baselines in comparison with the proposed pipeline (*training/test sets-1 and baseline training/test set-2*).

Method	Accuracy (%)	Difference (%)	F1-score (%)	Difference (%)
proposed	87.0	0	86.7	0
baseline-2 [64]	74.9	-12.1	73.6	-13.1
baseline-3	51.4	-35.6	NaN	NaN
baseline-4	29.4	-57.6	27.7	-59

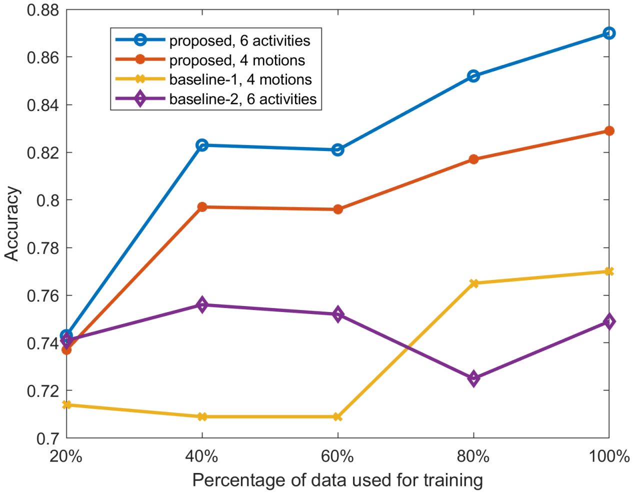


Figure 6.7: Classification accuracy of the proposed pipeline and two baselines with respect to varying number of training samples.

datasets. Therefore, the architecture in Yang's work [1] is deliberately designed to be lightweight, and the neural networks in the proposed pipeline are also expected to achieve convergence with as few samples as possible. Therefore, an important comparison is around the classifier/pipeline performance with respect to variation in the number of training samples. Figure 6.7 shows the test accuracy (*test set-1*) of the baselines and the proposed pipeline trained with only a randomly selected subset of *training set-1*. The results in Figure 6.7 show the following findings:

- The performance of the proposed pipeline shrinks sharply from 87% to 74% for 6 activities and from 83% to 73.5% for 4 motions, along with the decrease in the number of training samples.
- Nevertheless, the test accuracy and F1-score remain higher than or approximately equal to the state-of-the-art baselines given 20% of *training set-1* (equivalent to 358 samples per activity).

In conclusion, in terms of various quantitative metrics, the proposed pipeline generally has superior performances over the baselines in *1)* angle insensitivity, *2)* classification accuracy, and *3)* robustness against a limited number of training samples.

6.4. RESULTS OF LEAVE-ONE-SUBJECT-OUT TEST

It is expected that different people have corresponding specific kinematic patterns when performing certain activities. In the interest of training a generally applicable pipeline, it is crucial to learn how such diversity in kinematic patterns could influence pipeline performance. As can be seen from Figure 6.8, in the case of classifying an unseen human subject's activities, the average accuracy and F1-score are 63.5% and 62.5% respectively, which are approximately 20% lower than the original results.

- Comparing the test results of subject-1 with subject-7, or subject-5 with subject-6, shows that the similarity of kinematic patterns unnecessarily exists for people who have very close body characteristics.
- The drop in classification performance in the leave-one-subject-out test however fits the expectation since such pattern is also presented in [78].
- These results establish the importance of data generation and/or data augmentation from seen to unseen people, even if completely addressing this problem goes beyond the scope of this MSc thesis.

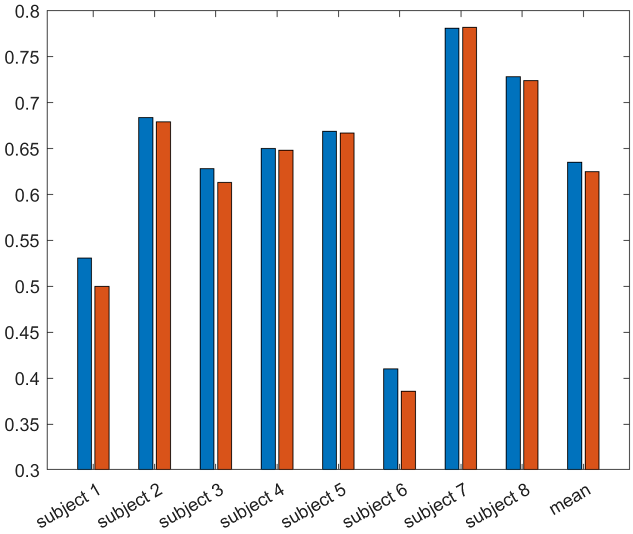


Figure 6.8: Classification accuracy in the leave-one-subject-out test, where the horizontal axis represents the index of the left-out subject and blue bar expresses the classification accuracy and red for F1-score.

6.5. RESULTS OF NOISY DATA

This section focuses on one of the most influential parameter in radar system-SNR, for example training with *train set-3* and testing with *test set-3*. Because of the randomness nature of additive noise, data generation, training and testing are independently repeated for five times. Figure 6.9 shows the average test accuracy and the standard deviation of different realizations in terms of varying SNR levels. As can be seen, the classification performances almost linearly decrease along with lowering SNR levels, and the accuracy could drop to nearly 50% for an SNR of 8dB. These suggest that a noisy environment could significantly undermine the performance of the proposed pipeline. Last and most importantly, the pipeline gain due to replacing max pooling with average pooling is clearly shown by comparing the blue curve with the red (see Figure 6.9), indicating that average pooling as a symmetric function better fits the task of processing noisy radar data. Therefore, the modification of replacing max pooling with average pooling as in the proposed pipeline indeed leads to superior performances and thus is reasonable.

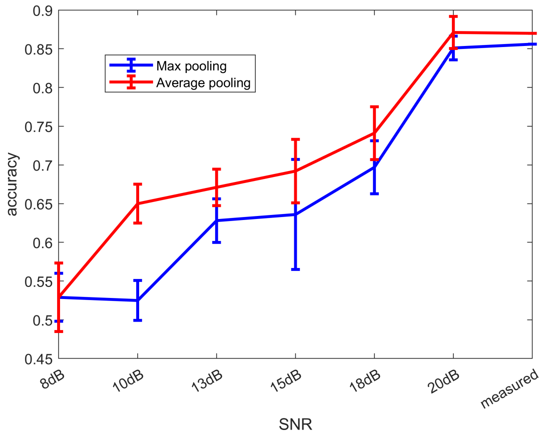


Figure 6.9: Classification accuracy and F1-score with respect to varying SNR levels, where average pooling refers to the proposed pipeline, whereas max pooling refers to the original PointNet and its transformation network T-Net [66].

For data-driven methods, it is also interesting to evaluate if certain controllable variables between training and test data could influence the performance. In this case, we can cross-validate the results of training-with-measured-data and testing-with-noisy-data, and vice versa. The results are listed in Table 6.4. These results establish that the consistency between the SNR levels of training and test data is also important. Therefore, estimating the SNR level in practice and accordingly selecting the trained model could be of merit.

Table 6.4: Cross-examination of the robustness of the trained pipeline given data of unseen SNR levels, where *training and test set-1* are from experimental measurement and *training and test set-3* to *8* are with SNR levels of $20dB$, $18dB$, $15dB$, $13dB$, $10dB$ and $8dB$, respectively.

Train set	**Test set**	**Accuracy (%)**	**Difference (%)**	**F1-score (%)**	**Difference (%)**
1	*1*	87.0	0	86.7	0
1	*3*	67.8	-19.2	67.4	-19.3
1	*4*	65.3	-21.7	64.6	-22.1
1	*5*	60.0	-21.0	58.7	-28.0
1	*6*	55.3	-31.7	52.8	-33.9
1	*7*	52.8	-34.2	49.8	-36.9
1	*8*	47.8	-39.2	44.4	-42.3
3	*1*	69.0	-18.0	68.7	-18
4	*1*	50.1	-36.9	50.4	-36.3
5	*1*	49.6	-37.4	50.8	-35.9
6	*1*	41.8	-45.2	43.2	-43.5
7	*1*	44.9	-42.1	45.1	-43.6
8	*1*	31.4	-55.6	30.8	-55.9

6.6. Results of Variation in MIMO Aperture

Using an imaging radar with relatively smaller MIMO aperture than the that used to generate *training/test set-1* is economic and power-saving, thus the results obtained from four pairs of training/test sets are given in Table 6.5. Meanwhile, the cross-examination of training on the large-aperture dataset and testing on the small-aperture dataset and vice versa is also listed.

These results evidence the findings as follows:

- since the largest decrease in accuracy is as small as 9.2%, it is reasonable to conclude that the proposed pipeline has promising robustness to be applied with different imaging radars (varying MIMO aperture); and,
- the significance of the consistency between training and test data is again highlighted according to the huge accuracy drops as in the cross-source evaluation parts. On the contrary, since some features are learned (i.e. accuracy still is better than flipping a dice), such trained models could be used as the pre-trained model in transfer learning.

Table 6.5: Performance of the proposed pipeline given data from varying MIMO aperture.

Train set	**Test set**	**Accuracy (%)**	**Difference (%)**	**F1-score (%)**	**Difference (%)**
1	*1*	87.0	0	86.7	0
17	*17*	83.6	-3.4	83.4	-3.3
18	*18*	83.3	-3.7	83.0	-3.7
19	*19*	79.4	-7.6	79.2	-7.5
20	*20*	77.8	-9.2	77.5	-9.1
1	*17*	26.4	-60.6	25.7	-61
1	*18*	58.7	-29.3	58.8	-27.9
1	*19*	27.4	-50.6	26.8	-59.9
1	*20*	31.6	-55.4	31.8	-54.9
17	*1*	28.1	-58.9	26.3	-60.4
18	*1*	72.8	-14.2	67.8	-18.9
19	*1*	26.8	-60.2	25.9	-60.8
20	*1*	29.0	-58.0	23.6	-63.1

6.7. ERROR ANALYSIS

This section critically evaluates the sources of mistakes for the proposed pipeline.

Error Source-1 To begin with, the performances of the proposed pipeline per se, as shown in Figure 6.4, sitting down is often mixed with standing up from sitting, and bending over is mixed with standing up from bending. This is due to the poor performances of the spectrogram classification module. To be more specific, Figure 6.4 shows that classification accuracy of spectrogram based module (first to fourth cell on the diagonal) is worse than the counterpart of PC classification module (fifth and sixth cell on the diagonal). To understand the cause of such results, data representations should be again discussed. The used PCs involve directly represent three intrinsic features of imaging radar: range, azimuth and elevation (equivalently: X,Y, and Z); spectrogram exploits Doppler, power and time. Therefore, spectrogram and PC contain distinct information. The error of spectrogram classification module is caused by the dimension of time. The data collection is such that each subject repeatedly performs one pair of motions for approximately same period of 2 seconds. However, human subjects often fail to comply this rule, whereas the segments of samples are divided by exactly 2 seconds. Hence, some of the spectrogram samples are inherently incorrect in time. Figure 6.10 presents two examples of the wrongly predicted sample in *test set-1*, where it is seen that spectrograms do not include the full movement but starts from the half-way. Naturally, spectrogram images cannot represent any real movement and then leads to the poor classification performance.

In summary, the poor performances of spectrogram classification module is due the crude way of segmentation, which causes that certain spectrogram images only contain a part of the movement information, which is presumed to be the cause of wrong classification results. This suggests that an adaptive and automatic segmentation method is crucial to attain more superior performances.

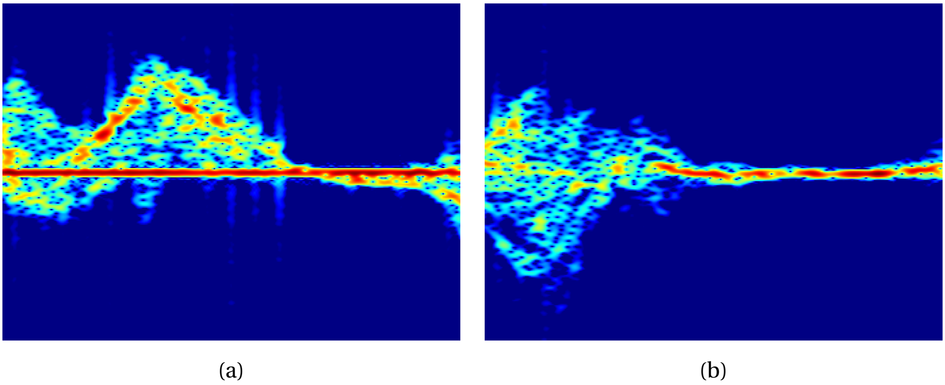


Figure 6.10: Examples of wrongly predicted spectrogram images of motion samples (a) bending over, and (b) standing up. The horizontal axis is of approximately 2 seconds, and vertical axis expresses velocity in ±1.26m/s.

Error Source-2 Figure 6.5 in section 6.3 confirms the superiority of the proposed pipeline in general. However, the proposed pipeline classifying 4 motions at test test angle of 90 deg exhibits poorer results than the baseline-1 [1]. It is interesting to understand the reasons of such dramatical drop in classification performance from adjacent aspect angles to 90 deg. To elaborate on that, the confusion matrix of spectrogram classification module processing the data of sitting down/standing up after sitting pair at the aspect angle of 90 deg are presented in Figure 6.11. As demonstrated, CNN model finds it hard to extract informative features from the spectrogram images of motions at large aspect angle. This is due to the fact the macro-motion (of torso) and micro-motion (of arms for instance) are not in radial direction, and therefore do not generate Doppler shift. For example, Figure 6.12 shows an example of the spectrogram image of human subject sitting down and standing up from sitting at 90 deg where almost no Doppler variation is exhibited.

Therefore, with the Doppler information on certain motions missing at large aspect angle, it is evident that further exploitation of the spatial information to make motion classifications is important. For instance, RNN could be introduced to learn the temporal variations in the PCs within one segment of movement.

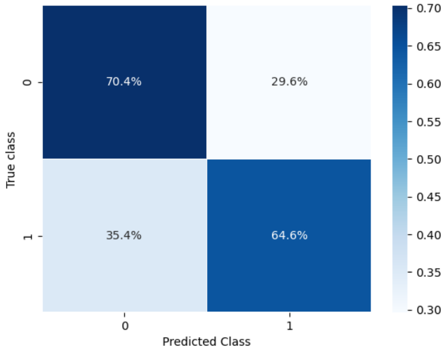
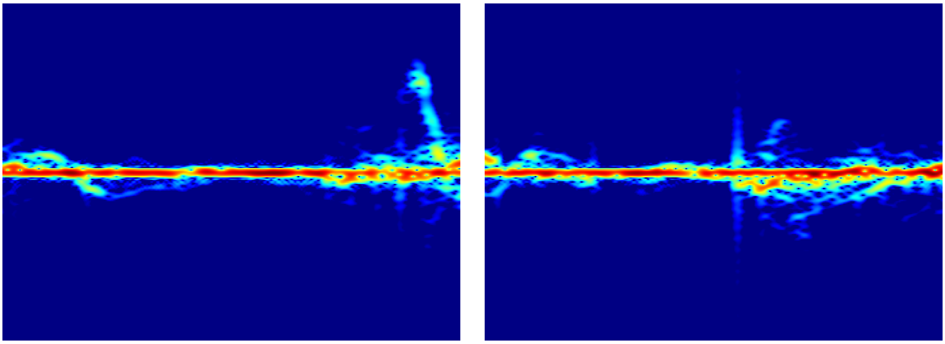


Figure 6.11: Confusion matrix of spectrogram classification module processing the data of sitting/standing pair at the aspect angle of 90 deg, where class-0 expresses sitting down and class-1 expresses standing up from sitting.

Error Source-3 Furthermore in section 6.3, it is realized that the proposed pipeline is sensitive to the amount of training samples (see Figure 6.7). Figure 6.13 shows the confusion matrix of training with 20% of training set-1 and testing with *test set-1*. By comparing Figure 6.13 with Figure 6.4, it is clear that more samples are mixed for the motion pair of bending over/standing up from bending, and two postures. Therefore, the poor classifier performance is subject to both PC and spectrogram classification modules. It is reasonable to suspect that for the proposed pipeline, such drop in classification performance is due to its hierarchical structure. For instance, when 20% of *training set-1* are utilized for training, spectrogram classification module has only 72 samples per class for



(a)

(b)

Figure 6.12: Example of the spectrogram image of human subject (a) sitting down, and (b) standing up from sitting at 90 deg.

training. This number not only is far smaller than the number in ImageNet (more than 1000 samples per class), but also smaller than the typical radar-based HAR datasets (e.g., 142 in [20], 500 in [27] and more examples in Table 2.1).

The disadvantages of the proposed hierarchical structure is realized that the number of training samples for the second or the third stage may be too small to see neural network's convergence. Therefore, certain data augmentation methods could be adopted to further enhance the robustness of radar-based HAR methods against limited number of training samples.

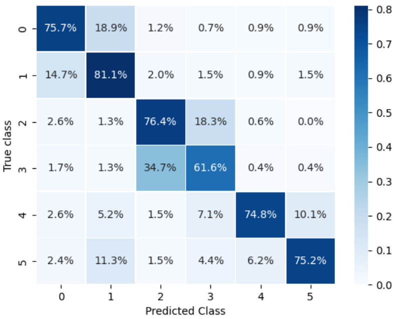


Figure 6.13: Confusion matrix of the proposed pipeline training with 20% of *training set-1* and testing with test set-1, where class-0 to 5 expresses sitting down, standing up from sitting, bending over, standing up from bending, sitting still and standing still, respectively.

7 Conclusion and Future Work

Automatic indoor HAR is seen as a key technology to solve the severe shortage of health care professionals brought by aging problem. Radar as a sensor is attracting people's attention due to its inherent advantageous characteristics such as respect to privacy and promising functionality in poor lightening conditions. Through literature review, the limitation of the past work on the topic of radar-based HAR is realized as the **Excessive Dependency on Doppler** such that kinematically static postures and kinematically dynamic motions have not been jointly examined. Moreover, human activities are often constrained to be performed in the line-of-sight orientation due to the excessive dependency on the Doppler feature.

With this MSc thesis work focusing on the second point of the aforementioned limitation, imaging radar is seen as a feasible solution. The main advantage of imaging radar compared to conventional radar is that it provides additional spatial information on the object. Such information enables more informative depiction of the shape of a human body, and therefore, could be the key to recognize kinematically static postures and movements at large aspect angles. Correspondingly, a few challenges are raised as follows:

- It is challenging to find the informative data representations given additional spatial information.

- It is not yet established what classifiers are most suitable to learn the important information from the new format of data representations.

To respond to these challenges, two complementary data representations- PC and spectrogram- are utilized. The former essentially represents three dimensions of information- Doppler, power and time; the latter however expresses range, azimuth and elevation. A hierarchical pipeline consisting of three modules is proposed to process these data. The main innovations of this pipeline include the usage of all six intrinsic features provided by imaging radar as well as the modifications of neural networks, enabling HAR with respect to multiple aspect angles and in noisy environments.

To validate the performances of the proposed pipeline, a custom experimental dataset is generated. Furthermore, some ’simulated’ datasets are also generated based on the measured data, for instance by adding white Gaussian noise to the measured data. Ablation study is performed to learn the contribution of each module. Moreover, a few baselines are implemented to comparatively study the performances of the proposed pipeline. The main findings in the attained results are as follows:

- The proposed pipeline extracts high-dimensional latent space from the input representations, and achieves promising accuracy (87.0%) for a classification problem involving 4 motions and 2 postures;

- Each of the modules in the proposed pipeline is proved to be of crucial contribution via ablation studies;

- The proposed pipeline attains advantageous performances compared with the baselines, i.e. the accuracy of the proposed pipeline is 6% higher than [1] for classifying four motions , and 12.1% higher than [64] for classifying all six activities.

- Leave-one-subject-out test shows that different human subjects inherently have own kinematic patterns, which are unnecessarily dependent on the body characteristics.

- Through adding white Gaussian noise to the measured data, it is valid that SNR level directly affects the detection results and quality of spectrogram images, and thus indirectly affect the classification performances.

- It is also established that large synthetic aperture of the imaging radar is effective in attaining superior performances through accurate generation of PCs.

Last but not least, the error sources are analyzed to critically study the limitations of the proposed method. These limitations and thus suggested future work can be summarized as follows:

- Current segments are obtained by a fixed duration of 2 seconds. An adaptive and automatic segmentation method is crucial to attain more superior performances.

- Spectrogram representation is not informative enough to represent the kinematic characteristics of movements at large aspect angles (90 degree for instance). Further exploitation of the spatial information to make motion classifications is important.

- Hierarchical structure limits the amount of training data for second or third stage. Certain data augmentation methods could be applied to improve the sensitivity of the proposed pipeline against number of training samples.

Future work: Apart from resolving the limitations of the proposed pipeline per se, there are numerous aspects of future work on radar-based HAR worthy following up. Referring back to the three main blocks of radar-based HAR (see Figure 1.1), future work can also be divided into such three categories. Some rough ideas are given as follows:

1. More realistic (less artificial) datasets should be collected. In these datasets, human subjects should have the freedom to causally perform daily activities, in-place and transitional. In the meantime, simulations of imaging radar data could be a crucial step to generate/ augment a large-scale, comprehensive and reliable radar dataset. DL techniques will also play an important role in this aspect, since GAN has been proved capable of generating spectrograms [46] even with noise and clutter factors [47]. In the field of computer vision, GAN-based point cloud generation has already been explored [79], future work for radar community is to adapt such works to radar PCs which inherently have non-equidistant sampling density. Transfer learning and many other data augmentation methods could also contribute to this goal. And to my opinion, an open-source radar dataset, such as [80], [81] and [82], is the most important step leading to the further scientific advances of our community.

2. Two data representations- spectrogram and PC- are used in this project. There still exist numerous other combinations of imaging radar's six intrinsic features. I would also like to stress a fact that the usage of time information in this project is close to a 'micro-temporal relation' within each segment of movement instead of a 'macro-temporal relation' that considers the dependencies between previous, present, and even future activities. Sequences of spectrogram images is a good example how of such macro-temporal relation can be shown in data representations [34]. Naturally, sequences of frames of PCs will also be applicable.

3. Triggered by novel data representations, advanced classifiers (pipelines) must be proposed to reasonably consume the input.

   - Given a large-scale dataset, the architecture of neural network(s) can be more complex. For instance, instead of PointNet [66], more advanced PointNet++ [83] which jointly considers the global feature and local semantic features of PCs can be applied without concerning the problem of hard-to-converge; meanwhile, those more advanced image classification models, such as ResNet [73] or Vision Transformer [84] which were initially proposed with large-scale image dataset such as ImageNet [75] would have the potential to more accurately classify spectrogram images.
   - Suppose a more realistic dataset is created and the macro-temporal relation is to be introduced. As discussed in the chapter of literature review (2), RNN or its variant [21], Transformers [85], etc., could be applied to capture such macro-temporal relation. This supposedly will further improve the classification accuracy in the temporally adjacent samples.
   - The pipeline should not only includes the feedforward lines (from input samples to classification results as in my pipeline), but also feedback lines (from classifications results to input samples). For instance, how the orientation classification and classification results can help with detection and tracking of the human subject, e.g. by choosing the adaptive number of guard and reference cells in CFAR, or, by choosing the correct motion model in a extended target tracking algorithm [86].

REFERENCES

[1] Y. Yang, C. Hou, Y. Lang, T. Sakamoto, Y. He, and W. Xiang, “Omnidirectional motion classification with monostatic radar system using micro-Doppler signatures,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 5, pp. 3574–3587, 2019.

[2] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, “Monitoring activities of daily living in smart homes: Understanding human behavior,” *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[4] “Population structure and ageing statistics explained.” [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/SEPDF/cache/1271.pdf

[5] “Analysis of shortage and surplus occupations 2020.” [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/22189434-395d-11eb-b27b-01aa75ed71a1

[6] S. A. Shah and F. Fioranelli, “RF sensing technologies for assisted daily living in healthcare: A comprehensive review,” *IEEE Aerospace and Electronic Systems Magazine*, vol. 34, no. 11, pp. 26–44, 2019.

[7] N. Lu, Y. Wu, L. Feng, and J. Song, “Deep Learning for fall detection: Three-Dimensional CNN combined with LSTM on video kinematic data,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 314–323, 2019.

[8] S. C. Mukhopadhyay, “Wearable sensors for human activity monitoring: A review,” *IEEE sensors journal*, vol. 15, no. 3, pp. 1321–1330, 2014.

[9] V. C. Chen, *The micro-Doppler effect in radar.* Artech House, 2019.

[10] Y. Kim and H. Ling, “Human activity classification based on micro-Doppler signatures using a Support Vector Machine,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 5, pp. 1328–1337, 2009.

[11] S. Z. Gurbuz and M. G. Amin, “Radar-based human-motion recognition with deep learning: Promising applications for indoor monitoring,” *IEEE Signal Processing Magazine*, vol. 36, no. 4, pp. 16–28, 2019.

[12] T.-L. Le, M.-Q. Nguyen, and T.-T.-M. Nguyen, “Human posture recognition using human skeleton provided by Kinect,” in *2013 International Conference on Computing, Management and Telecommunications (ComManTel)*, 2013, pp. 340–345.

[13] R. G. Guendel, M. Unterhorst, E. Gambi, F. Fioranelli, and A. Yarovoy, “Continuous human activity recognition for arbitrary directions with distributed radars,” in *2021 IEEE Radar Conference (RadarConf21)*, 2021, pp. 1–6.

[14] Y. Kim and H. Ling, "Human activity classification based on micro-Doppler signatures using an artificial neural network," in *2008 IEEE Antennas and Propagation Society International Symposium*, 2008, pp. 1–4.

[15] B. Jokanovic, M. G. Amin, and F. Ahmad, "Effect of data representations on deep learning in fall detection," in *2016 IEEE Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 2016, pp. 1–5.

[16] 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.

[17] B. Erol and M. G. Amin, "Fall motion detection using combined range and Doppler features," in *2016 24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 2075–2080.

[18] A. Angelov, A. Robertson, R. Murray-Smith, and F. Fioranelli, "Practical classification of different moving targets using automotive radar and deep neural networks," *IET Radar, Sonar & Navigation*, vol. 12, no. 10, pp. 1082–1089, 2018.

[19] P. Molchanov, S. Gupta, K. Kim, and K. Pulli, "Multi-sensor system for driver's hand-gesture recognition," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, 2015, pp. 1–8.

[20] Y. Shao, S. Guo, L. Sun, and W. Chen, "Human motion classification based on range information with deep Convolutional Neural Network," in *2017 4th International Conference on Information Science and Control Engineering (ICISCE)*, 2017, pp. 1519–1523.

[21] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Bi-LSTM network for multimodal continuous human activity recognition and fall detection," *IEEE Sensors Journal*, vol. 20, no. 3, pp. 1191–1201, 2020.

[22] X. Li, Y. He, Y. Yang, Y. Hong, and X. Jing, "LSTM based human activity classification on radar range profile," in *2019 IEEE International Conference on Computational Electromagnetics (ICCEM)*, 2019, pp. 1–2.

[23] X. Zheng, Z. Yang, K. He, and H. Liu, "Hand gesture recognition based on range Doppler-angle trajectory and LSTM network using an MIMO radar," in *Eleventh International Conference on Signal Processing Systems*, vol. 11384. International Society for Optics and Photonics, 2019, p. 113840P.

[24] M. Jia, S. Li, J. L. Kernec, S. Yang, F. Fioranelli, and O. Romain, "Human activity classification with radar signal processing and machine learning," in *2020 International Conference on UK-China Emerging Technologies (UCET)*, 2020, pp. 1–5.

[25] S. Yang, J. L. Kernec, F. Fioranelli, and O. Romain, "Human activities classification in a complex space using raw radar data," in *2019 International Radar Conference (RADAR)*, 2019, pp. 1–4.

[26] S. Hazra and A. Santra, "Short-range radar-based gesture recognition system using 3D CNN with triplet loss," *IEEE Access*, vol. 7, pp. 125 623–125 633, 2019.

[27] H. Du, T. Jin, Y. Song, Y. Dai, and M. Li, "A three-dimensional deep learning framework for human behavior analysis using range-Doppler time points," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 4, pp. 611–615, 2020.

[28] X. Wang, P. Chen, H. Xie, and G. Cui, "Through-Wall human activity classification using complex-valued convolutional neural network," in *2021 IEEE Radar Conference (RadarConf21)*, 2021, pp. 1–4.

[29] Y. He, P. Molchanov, T. Sakamoto, P. Aubry, F. L. Chevalier, and A. Yarovoy, "Range-Doppler surface: a tool to analyse human target in ultra-wideband radar," *Radar Sonar Navigation Iet*, vol. 9, no. 9, pp. 1240–1250, 2015.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[31] H. Du, Y. He, and T. Jin, "Transfer learning for human activities classification using micro-Doppler spectrograms," in *2018 IEEE International Conference on Computational Electromagnetics (ICCEM)*. IEEE, 2018, pp. 1–3.

[32] M. S. Seyfioğlu, A. M. Özbayoğlu, and S. Z. Gürbüz, "Deep convolutional autoencoder for radar-based classification of similar aided and unaided human activities," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 4, pp. 1709–1723, 2018.

[33] G. Klarenbeek, R. I. A. Harmanny, and L. Cifola, "Multi-target human gait classification using LSTM recurrent neural networks applied to micro-Doppler," in *2017 European Radar Conference (EURAD)*, 2017, pp. 167–170.

[34] S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges, "Interacting with soli: Exploring fine-grained dynamic gesture recognition in the radio-frequency spectrum," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, 2016, pp. 851–860.

[35] F. Fioranelli, M. Ritchie, and H. Griffiths, "Bistatic human micro-Doppler signatures for classification of indoor activities," in *2017 IEEE Radar Conference (RadarConf)*, 2017, pp. 0610–0615.

[36] M. G. Amin, Y. D. Zhang, F. Ahmad, and K. D. Ho, "Radar signal processing for elderly fall detection: The future for in-home monitoring," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 71–80, 2016.

[37] H. Li, J. le Kernec, A. Mehul, S. Z. Gurbuz, and F. Fioranelli, "Distributed radar information fusion for gait recognition and fall detection," in *2020 IEEE Radar Conference (RadarConf20)*, 2020, pp. 1–6.

[38] M. Li, T. Chen, and H. Du, “Human behavior recognition using range-velocity-time points,” *IEEE Access*, vol. 8, pp. 37 914–37 925, 2020.

[39] X. Li, Y. He, F. Fioranelli, X. Jing, A. Yarovoy, and Y. Yang, “Human motion recognition with limited radar micro-Doppler signatures,” *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[40] Y. Li, K. He, D. Xu, and D. Luo, “A transfer learning method using speech data as the source domain for micro-Doppler classification tasks,” *Knowledge-Based Systems*, vol. 209, p. 106449, 2020.

[41] B. Erol and S. Z. Gurbuz, “A Kinect-based human micro-Doppler simulator,” *IEEE Aerospace and Electronic Systems Magazine*, vol. 30, no. 5, pp. 6–17, 2015.

[42] B. Eroi, C. Karabacak, S. Z. Gürbüz, and A. C. Gürbüz, “Radar simulation of different human activities via Kinect,” in *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, 2014, pp. 1015–1018.

[43] B. Erol, C. Karabacak, S. Z. Gürbüz, and A. C. Gürbüz, “Simulation of human micro-Doppler signatures with Kinect sensor,” in *2014 IEEE Radar Conference*, 2014, pp. 0863–0868.

[44] S. Vishwakarma, W. Li, C. Tang, K. Woodbridge, R. Adve, and K. Chetty, “Simhumalator: An open source wifi based passive radar human simulator for activity recognition,” *arXiv preprint arXiv:2103.01677*, 2021.

[45] I. Alnujaim, D. Oh, and Y. Kim, “Generative adversarial networks for classification of micro-Doppler signatures of human activity,” *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 3, pp. 396–400, 2020.

[46] B. Erol, S. Z. Gurbuz, and M. G. Amin, “GAN-based synthetic radar micro-Doppler augmentations for improved human activity recognition,” in *2019 IEEE Radar Conference (RadarConf)*, 2019, pp. 1–5.

[47] S. Vishwakarma, C. Tang, W. Li, K. Woodbridge, R. Adve, and K. Chetty, “GAN based noise generation to aid activity recognition when augmenting measured wifi radar data with simulations,” in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2021, pp. 1–6.

[48] Z. Wang, Z. Dai, B. Poczos, and J. Carbonell, “Characterizing and avoiding negative transfer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[49] I. Orr, M. Cohen, and Z. Zalevsky, “High-resolution radar road segmentation using weakly supervised learning,” *Nature Machine Intelligence*, vol. 3, no. 3, pp. 239–246, 2021.

[50] H. Sim, S. Lee, B.-h. Lee, and S.-C. Kim, “Road structure classification through artificial neural network for automotive radar systems,” *IET Radar, Sonar & Navigation*, vol. 13, no. 6, pp. 1010–1017, 2019.

[51] H. Jha, V. Lodhi, and D. Chakravarty, “Object detection and identification using vision and radar data fusion system for ground-based navigation,” in *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE, 2019, pp. 590–593.

[52] K. Patel, K. Rambach, T. Visentin, D. Rusev, M. Pfeiffer, and B. Yang, “Deep learning-based object classification on automotive radar spectra,” in *2019 IEEE Radar Conference (RadarConf)*, 2019, pp. 1–6.

[53] I. Roldan, C. R. del Blanco, Á. Duque de Quevedo, F. Ibañez Urzaiz, J. Gismero Menoyo, A. Asensio López, D. Berjón, F. Jaureguizar, and N. García, “DopplerNet: a convolutional neural network for recognising targets in real scenarios using a persistent range-Doppler radar,” *IET Radar, Sonar & Navigation*, vol. 14, no. 4, pp. 593–600, 2020.

[54] S. Abdulatif, Q. Wei, F. Aziz, B. Kleiner, and U. Schneider, “Micro-Doppler based human-robot classification using ensemble and deep learning approaches,” in *2018 IEEE Radar Conference (RadarConf18)*, 2018, pp. 1043–1048.

[55] H. Cui and N. Dahnoun, “Human posture capturing with millimetre wave radars,” in *2020 9th Mediterranean Conference on Embedded Computing (MECO)*. IEEE, 2020, pp. 1–4.

[56] Z. Zheng, J. Pan, Z. Ni, C. Shi, S. Ye, and G. Fang, “Human posture reconstruction for through-the-wall radar imaging using convolutional neural networks,” *IEEE Geoscience and Remote Sensing Letters*, 2021.

[57] A. Sengupta, F. Jin, R. Zhang, and S. Cao, “mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns,” *IEEE Sensors Journal*, vol. 20, no. 17, pp. 10 032–10 044, 2020.

[58] A. Kılıç, İ. Babaoğlu, A. Babalık, and A. Arslan, “Through-wall radar classification of human posture using convolutional neural networks,” *International Journal of Antennas and Propagation*, vol. 2019, 2019.

[59] Z. He, X. Feng, H. Zheng, and W. Li, “Posture recognition with background noise elimination using FMCW radar,” in *2020 Cross Strait Radio Science & Wireless Technology Conference (CSRSWTC)*. IEEE, 2020, pp. 1–3.

[60] G. Tiwari, P. Bajaj, and S. Gupta, “mmFiT: Contactless fitness tracker using mmWave radar and edge computing enabled deep learning,” 2021.

[61] D. Nickalls, J. Wu, and N. Dahnoun, “A real-time and high performance posture estimation system based on millimeter-wave radar,” in *2021 10th Mediterranean Conference on Embedded Computing (MECO)*. IEEE, 2021, pp. 1–4.

[62] D. Sasakawa, N. Honma, T. Nakayama, and S. Iizuka, “Human posture identification using a MIMO array,” *Electronics*, vol. 7, no. 3, p. 37, 2018.

[63] N. Honma, D. Sasakawa, N. Shiraki, T. Nakayama, and S. Iizuka, “Human monitoring using MIMO radar,” in *2018 IEEE International Workshop on Electromagnetics:Applications and Student Innovation Competition (iWEM)*, 2018, pp. 1–2.

[64] Y. Kim, I. Alnujaim, and D. Oh, “Human activity classification based on point clouds measured by millimeter wave MIMO radar with deep Recurrent Neural Networks,” *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13 522–13 529, 2021.

[65] Y. Kim and T. Moon, “Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks,” *IEEE geoscience and remote sensing letters*, vol. 13, no. 1, pp. 8–12, 2015.

[66] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “Pointnet: Deep learning on point sets for 3D classification and segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[67] “Design guide: Tidep-01012 imaging radar using cascaded mmwave sensor reference design,” Texas Instrumentation, online; accessed 19 August 2021.

[68] C. A. Balanis, *Antenna theory: analysis and design*. John wiley & sons, 2015.

[69] B. Çağlıyan and S. Z. Gürbüz, “Micro-Doppler-based human activity classification using the mote-scale Bumblebee radar,” *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 10, pp. 2135–2139, 2015.

[70] H. Rohling, “Radar cfar thresholding in clutter and multiple target situations,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. AES-19, no. 4, pp. 608–621, 1983.

[71] Y. Aoki, H. Goforth, R. A. Srivatsan, and S. Lucey, “Pointnetlk: Robust & efficient point cloud registration using PointNet,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7163–7172.

[72] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015.

[73] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.

[74] S. Chen, W. He, J. Ren, and X. Jiang, “Attention-based dual-stream vision transformer for radar gait recognition,” 2021.

[75] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[76] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O’Reilly Media, 2019.

[77] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html

[78] Y. Zhao, R. G. Guendel, Y. Alexander, and F. Francesco, “Distributed radar-based huamn activity recognition using vision tranformer and CNNs,” in *European Radar Conference 2022*, April 2022.

[79] D. W. Shu, S. W. Park, and J. Kwon, “3D point cloud generative adversarial network based on tree structured graph convolutions,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[80] “Radar signatures of human activities,” 10.5525/gla.researchdata.848, accessed: 2022-01-30.

[81] “Asl-sequential-dataset (77 ghz fmcw mimo),” https://github.com/ci4r/ASL-Sequential-Dataset, accessed: 2022-01-16.

[82] “Dataset of continuous human activities performed in arbitrary directions collected with a distributed radar network of five nodes,” https://data.4tu.nl/articles/dataset/Dataset_of_continuous_human_activities_performed_in_arbitrary_directions_collected_with_a_distributed_radar_network_of_five_nodes/16691500, accessed: 2022-01-30.

[83] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *arXiv preprint arXiv:1706.02413*, 2017.

[84] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021.

[85] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017.

[86] J. Pegoraro and M. Rossi, “Real-time people tracking and identification from sparse mm-wave radar point-clouds,” *IEEE Access*, vol. 9, pp. 78 504–78 520, 2021.