Interim Report: Gait Recognition Using Deep Learning

Submitted December 6, 2021, in partial fulfillment of
the conditions for the award of the degree **BSc Hons Computer Science with Artificial Intelligence.**

Shiliang Chen
20125016

Supervised by Jianfeng Ren

School of Computer Science University of Nottingham Ningbo China

Abstract

Radar gait recognition is robust to light variations and less infringement on privacy. Previous studies often utilize either spectrograms or cadence velocity diagrams. While the former shows the time-frequency patterns, the latter encodes the repetitive frequency patterns. In this work, a dual-stream neural network with attention-based fusion is proposed to fully aggregate the discriminant information from these two representations. The both streams are designed based on the Vision Transformer, which well captures the gait characteristics embedded in these representations. The proposed method is validated on a large benchmark dataset for radar gait recognition, which shows that it significantly outperforms state-of-the-art solutions.

Contents

List of Figures

List of Tables

List of Abbreviations

CNN convolutional neural network

CVD cadence velocity diagram

FFT fast Fourier transform

KNN k-nearest neighbors

mD micro-Doppler

MLP multi-layer perceptron

SVM support vector machine

UNNC University of Nottingham Ningbo China

ViT Vision Transformer

Chapter 1

Introduction

1.1 Background and Motivation

Gait is an biological characteristic which can be measured at great distances, and its recognition has become an attractive research area with many applications such as public safety monitoring [5], health screening [6] and human-computer interaction [7, 8].

Superimposed sequences of images captured by optical sensors have been commonly used to represent walking people [9, 10]. However, this approach requires the images to be captured from the side and depends heavily on the lighting conditions. Doppler radar can capture the micro-Doppler (mD) signatures of a moving target's gait feature from the front [1], and it has been demonstrated that mD signals can reveal human dynamic features and ascertain human actions and gestures by some study [1, 11, 12]. More importantly, the radar signal is less invasive of privacy and works even in low-light conditions, making it robust in a variety of real-world situations [8]. Thus, in this project, we mainly focus on the recognition using mD signatures.

The spectrogram is a time-varying representation of mD signatures in the joint time-frequency domain [1]. As the mD effect induced by mechanical vibrating or rotating structures of a target is stable and constant, the Doppler frequency caused by that target could be well formulated by a function of dwell time [13]. Thus, the time-frequency spectrogram of mD signatures could be useful to represent a moving target and further be commonly used in gait recognition and classification [12, 14, 15, 16, 17]. However, even the same person could have slightly different gait cycles and different start location of each cycle, which would lead to degradation in classification performance. The repetition information of those frequencies is scarcely considered but is also the invariant feature of gait cycles.

The Cadence Velocity Diagram (CVD) is another form of mD signatures converted from the spectrogram by taking Fast Fourier Transform (FFT) along the time axis [18]. The CVD is less commonly used in the gait recognition [19, 20]. The CVD provides another useful measure on the frequency of different velocities of body parts repeat, which is complementary to the spectrogram.

Fusion is a common practice in order to improve the classification performance when having more than one features. It can be classified to score-level fusion and feature-level fusion. Score-level fusion only considers the probability vectors. Feature-level fusion, which aims to utilising all or part of features from different sources, can be achieved by concatenation, averaging, max pooling, min pooling, etc. Attention-based feature-

level fusion is another approach which could memorising the relationship and decide the importance and confidence of both features and therefore provide more promising results [21].

Different kinds of traditional learning paradigms, such as Naïve Bayes [22], support vector machines [6] and k-nearest-neighbour classifiers [23] have been used to classify mD signatures. Several researchers have investigated deep convolutional neural networks (CNN) to implement the classification of human gestures and human actions [2, 16]. [1] has developed temporal convolutional neural network for spectrogram, and [24] has utilised a dual-channel network to classify gait features. However, they did not exploit the latent of CVD.

In this work, an attention-based two-stream Vision Transformer is proposed to identify features that are not hand-crafted and further improve the classification performance. As time-frequency representation of mD signature only contains the time-varying nature of Doppler frequencies, CVD is introduced to provide complementary information on how often different frequencies repeat. Furthermore, an attention-based fusion is applied to fuse the most useful features. To our knowledge, this is the first investigation in fusing the two discriminating clues. We also constructed a two-stream neural network to extract features from these two representations respectively. Moreover, we first employ Vision Transformer as the backbone of our neural network, utilising the patching function of it to deeply exploit physical meanings of both representations, and achieve an accuracy of 91.2% in a 98-people dataset.

1.2 Aims and Objectives

The main objective of this project is to develop a deep learning based method in order to recognise people by their radar gait features. The objective method would be able to learn individuals' gait features through several well processed radar time-frequency diagrams and recognise their identity by the learned model using deep learning. Apart from the main objective, the method is hopefully to be presented in a conference. The main objective could be divided into several sub-aspects to achieve:

1. Study the micro-Doppler phenomenon and justify the possibility of applying it to recognise human gait.
2. Process the raw radar signal for clear, recognisable and regular diagrams. The raw micro-Doppler radar signal is obtained in a 2D representation, and this step is to investigate the signal processing techniques to obtain suitable diagrams for learning.
3. Explore and investigate possible deep learning methods existing in the industry to test and find an efficient way to classify and recognise gait features accurately.
4. Design and develop a novel method by investigating different radar signal representations and applying them to obtain better performance in gait recognition.
5. Compare and evaluate existing methods and the method we proposed.

Chapter 2

Study Design

The spectrogram is commonly used to represent mD signatures, while the CVD is often ignored though it provides additional complementary knowledge to the spectrogram. In our study, we defined a novel universal two-stream structure for extracting both features in order to improve the accuracy of gait recognition. Figure 2.1 shows the overall procedures of this study. We first focus on signal processing, since the raw Doppler radar signal is in a 2-dimension form which consists of background noise and extremely long in time scale. The produced spectrogram will be utilised and additionally is converted to CVD for another stream. A two-stream neural network is then responsible for extracting features from these two representations. The extracted features are fed into Attention-based fusion to aggregate for final classification.

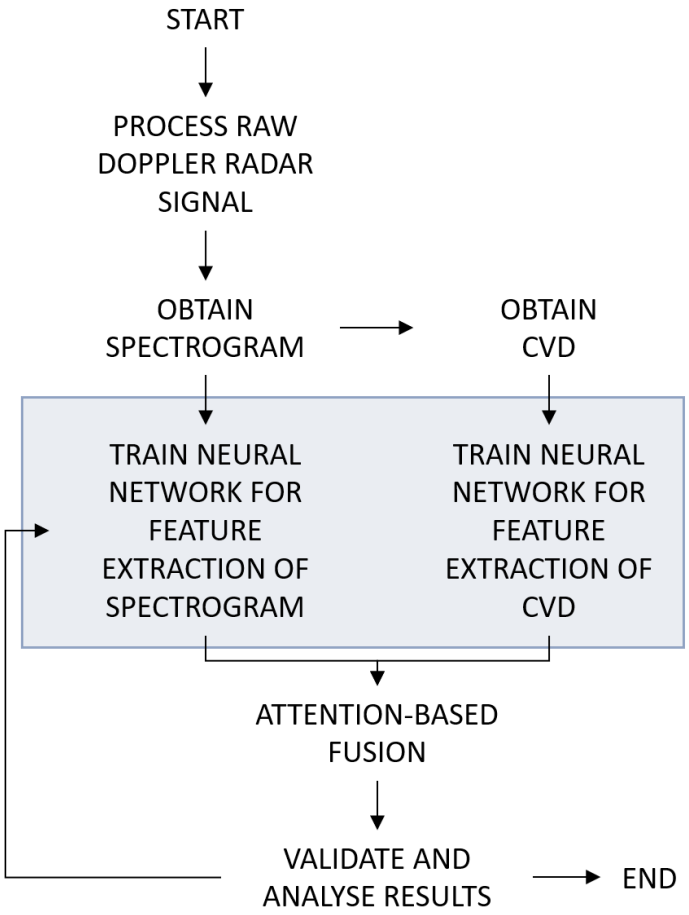

Figure 2.1: The overview of the study.

Chapter 3

Methodologies

Due to the different representations of time-frequency spectrogram and CVD, they could have complementary identification features about a walking person. In the area of gait recognition, researchers often use only one feature representation in the classification task, while there exist various fusion methods such as simple concatenation, score-level fusion, etc. to integrate features from different representations [25]. Attention mechanism has been suggested to be powerful in focusing on the important part which provides abundance information in specific feature maps. In our work, we are inspired by [25] to propose an attention-based fusion method for fusing deep features of time-frequency spectrogram and CVD in order to make the best use of them. Moreover, we propose to use Vision Transformer based two-stream neural network in extracting features from the two representations. The block diagram of this proposed method is shown in Figure 1, which consists of four main steps. Time-frequency spectrograms are fist generated from raw radar signal and CVDs are then converted from the generated time-frequency spectrograms. After the whole dataset is generated completely, spectrograms and CVDs would be fed into two separate ViT stream for extracting their own features. Attention-based fusion mechanism is then applied to find the most useful features from two and fuse them into one feature map. Finally, a fully-connected layer would transfer the feature map into the classification message.

3.1 Complementarity of Radar signal representations

Spectrogram: Tahmoush and Silvious [19] modeled Doppler of each body part of walking human as sinusoidal modulation in the spectrogram. The model reveals the velocities of each body part which is instrumental in determining the necessary human gait characteristics. The fundamental characteristics are the mean Doppler velocity and the size of the torso variation in the micro-Doppler. It has been found that the mDS of a person is relatively consistent and different walking people lead to clearly discriminative mDS [26]. Time-frequency spectrogram can well manifest such time-varying human gait characteristics. A sample spectrogram is shown in Fig. 3.1-a.

Cadence velocity diagram (CVD): Cadence velocity diagram has been proposed by [18] to analyse micro-Doppler signatures. After deriving the spectrogram, the CVD is obtained by taking the Fourier transform on the spectrogram along the time dimension for each Doppler frequency bin.

The derived CVD, as shown in Fig. 3.1-a, is a matrix with rows containing Doppler

frequency (or targets' speed) and columns containing cadence frequency, which measures how frequently different frequencies appear in the signal over time [20]. In other words, it illustrates the useful information of the repetition of velocities, which would be another key identification feature of a walking person. As a result, the CVD provides data on the shape, size, and frequency of the curves in the original spectrogram.

3.2 Two-stream Vision Transformer

Since two representations are used to provide different features, it is common to train two feature extraction networks for both of them. We therefore introduced a two-stream Vision Transformer (TSViT). An overview of this model is depicted in Figure 1. The model consists of two separate but identical networks which are both ViTs. Two sub-networks are fed with time-frequency spectrograms and CVDs respectively and output their own feature maps for fusion afterwards. Both spectrogram and CVD are obtained before training. We choose Vision Transformer (ViT) [4] as the backbone of our network.

The TSViT network can be formulated into a quadruplet $Q = (I_{TFS}, I_{CVD}, F, C)$, where I_{TFS} and I_{CVD} are two input radar representations fed into the two ViT streams, F is an attention based fusion layer and C is a classifier. Through the two ViT streams, features of two input representations would be extracted as output. Note that both input representations are resized in the same size of 224×224 with 3 channels which could be accepted by the ViTs, and since the two ViT streams are identical the resulting output would be in the same size as well. The process could be formulated as follow:

$$f_T FS = ViT_1(I_{TFS}), \tag{3.1}$$

$$f_{CVD} = ViT_2(I_{CVD}). \tag{3.2}$$

where f_{TFS} and f_{CVD} are feature maps of TFS and CVD in the size of 1×192. Here we have made modification to the original ViT Base 16×16 model, changing the last fully-connected layer to a 768×192 matrix for extracting middle-level features. The attention based fusion layer would then take the two output feature maps and generate a fused feature f_f:

$$f_f = F(f_{TFS}, f_{CVD}). \tag{3.3}$$

Classifier C finally takes the fused feature f_f and generate the classification message.

3.3 Backbone Vision Transformer

Compared to commonly used convolutional neural networks such as VGG [27], AlexNet [28], MobileNet [29] and ResNet [30], Vision Transformer (ViT) [4] can reshape the spectrogram into several local patches and use attention mechanism to obtain global knowledge of patches. Unlike scenarios of real-world images where an object could be positioned anywhere in an image but mean the same, patches of spectrogram and CVD converted from radar signal have their unique physical meaning when positioned in different locations in a diagram. This is because each patch contains knowledge about its specific area according to its time and frequency dimension for spectrogram or frequency dimensions for CVD.

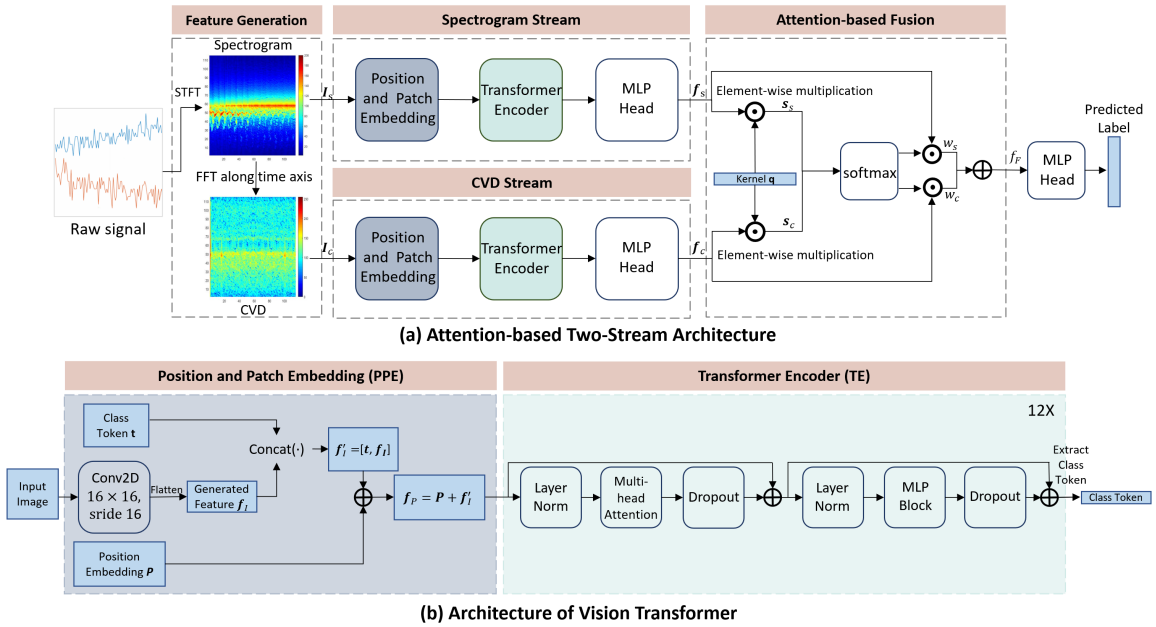

Figure 3.1: (A) shows the overview of the whole architecture; (B) details the pipeline of ViT by illustrating its position and patch embedding and transformer encoder.

Even patches similar in representation could be totally different physically based on their frequency levels and temporal position. Instead of convolutional kernels in CNN, patches are therefore ideally used to reserve the positional information of signal representations as well as the relationship between them. In our method, ViT is responsible for splitting diagrams in patches and embed their positions, and multi-head attention mechanism in transformer encoder could further extract the reserved physical information, processing physical meanings in the same way as natural language.

The architecture of ViT is depicted in Fig. 1. The input picture is first split into 196 16×16 separating patches, and each patch is then flatten to a 768-length vector. Then classification tokens are concatenated and position embedding is conducted by adding an position matrix, resulting an overall 197×768 matrix. The sequence of vectors are then fed to a standard transformer encoder, and this would be executed 12 times for the ViT base model. The transformer encoder module contains a multi-head self-attention layer, which is global and can process the relationship between patches. The output feature vector of transformer encoder after classification token extraction is in the size of 1×768. An MLP Head containing a fully-connected layer is responsible for taking the above feature vector and classification.

3.4 Attention based fusion

Inspired by Chen *et al.* [25], an attention-based fusion architecture shown in Fig. 3.1(a) is developed to fuse the complementary features extracted from both spectrogram and CVD. The target of the attention-based feature-level fusion is to find a set of weights

$\{\boldsymbol{w}_i \in \mathcal{R}^{1\times 768}\}$ for the features $\{\boldsymbol{f}_i \in \mathcal{R}^{1\times 768}\}$ to obtain an aggregated feature $\boldsymbol{f}_a \in \mathcal{R}^{1\times 768}$:

$$\boldsymbol{f}_a = \sum_{i=1}^{n} \boldsymbol{w}_i \odot \boldsymbol{f}_i, \tag{3.4}$$

where $\odot$ denotes element-wise multiplication, n is the number of features and $n = 2$ in this paper. In the attention-based fusion model, the kernel $\boldsymbol{q} \in \mathcal{R}^{1\times 768}$ is required to be trained. The process begins with the element-wise multiplication between the feature vector $\boldsymbol{f}_i$ and the kernel $\boldsymbol{q}$:

$$\boldsymbol{s}_i = \boldsymbol{q} \odot \boldsymbol{f}_i, \tag{3.5}$$

where $\boldsymbol{s}_i \in \mathcal{R}^{1\times 768}$ are the confidence scores for feature representations. A softmax function is then applied to $\boldsymbol{s}_i$ to assure that the derived weights $\sum_i \boldsymbol{w}_i = \boldsymbol{1} \in \mathcal{R}^{1\times 768}$:

$$\boldsymbol{w}_i = e^{\boldsymbol{s}_i} \oslash \sum_j e^{\boldsymbol{s}_j}, \tag{3.6}$$

where $\oslash$ denotes element-wise division. The fused feature vector is finally generated using Eq. (3.4). This attention-based fusion could well highlight the most discriminant features by training the kernel function $\boldsymbol{q}$.

Chapter 4

Implementation

4.1 Dataset overview

The evaluation is based on a gait dataset collected by authors, which consists of 1669 walking sequences from 98 volunteers. To enable the detection ability under different clothing conditions, we deliberately invite volunteers to come for the data collection twice on different days. However, the route and distance those volunteers undergo are carefully assured fixed. Volunteers walk naturally from one end to the other, towards the Doppler radar, then turn around and walk to the origin. The whole walking process mentioned above is a complete sequence which lasts about 30 seconds. Each member has walked 10 sequences for both data collection events, but some have only joined the first or the second. Since the sequence could be long in the time axis, we crop each spectrogram into multiple frames (frame size is 115×115) with overlapping 0.2 seconds approximately representing a subject's one complete gait cycle. For the consistency of all pieces of training, we deliberately generate a fixed training set randomly split from the whole dataset and let the remaining be the testing set. The ratio of data split is approximately 50% to 50%, containing 22,894 images for training and 22,874 for testing.

4.2 Experimental settings

We use the ViT with 16×16 input patch size as both the ViT model for direct training and the backbone of the proposed AT-DSViT. The parameter details of the applied ViT model is shown in Table 4.1. We also choose the original AlexNet [28], VGG16 [27], ResNet18 [30], MobileNetV2 [3] and ViT-Base [4] models as the baseline CNNs for comparison. AlexNet has been used as backbone in [2] to identify humans and achieves leading performance. [1] has considered VGG16 and ResNet18 as state-of-the-art methods in gait recognition tasks for comparison.

Table 4.1: Variants details of Vision Transformer Base

Model	Layers	Hidden size	MLP size	Heads
ViT-Base	12	768	3072	12

We train all three models using the aforementioned fixed training set and evaluate their performance with the testing set on an environment with two GeForce RTX 2080 Ti (12GB). All the signal and image processing, including the generation of CVD and concatenation of two representations, is conducted before feeding to the network. The input size is initially 115×115 and is resized to 224×224 to fit the adopted networks. SGD is

the optimiser of all these models with a momentum of 0.9 and weight decay of 5×10^{-5}. We take cross-entropy loss as the loss function. As for the learning parameters, a linear learning rate warmup and decay is used with a learning rate scheduler, and the initial learning rate is set to 0.01 for ViT and ResNet, 0.001 for AlexNet and 0.0001 for VGG. The batch size is 128 for all training. The proposed AT-DSViT loads two pre-trained ViT weights for both spectrogram and CVD streams. All the models are trained for 500 epochs each.

4.3 Realisation

4.3.1 Radar signal processing

The original raw radar signal was collected and recorded with MATLAB in a two-dimension form, where one dimension represents time and another is the magnitude of the signal. For the convenience of data reading and processing, we use MATLAB as the major tool in signal processing. As people walk towards or backwards to the radar, the magnitude of received radar signal would be recorded by the sensor with a certain frequency. Additionally, the magnitude of radar signal consists of two channels and thus presented in a complex number form. The first step is to separate the two channels for further transformation. The separated radar signal is illustrated in Figure 4.1, where the horizontal axis is time and the vertical is magnitude.

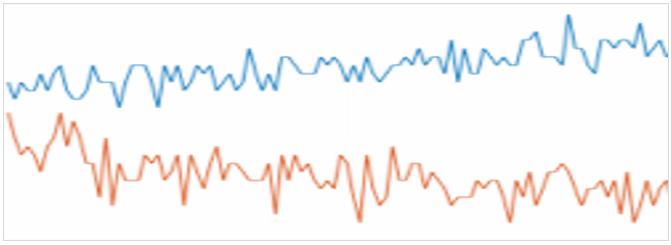

Figure 4.1: Illustration of separated raw signal

The spectrogram displays the Doppler frequency change along time and is obtained as follow. The signal $s(t)$ is segmented into M overlapping frames $\{x_0, x_1, ..., x_{M-1}\}$, where each frame $xi = \{x_i[n], n = 0, 1, ..., N-1\}$ is a column vector of length N. These M frames form a synthetic image $X = [x_1, ..., x_{M-1}]$ of size $M \times N$. Then, the discrete Fourier transform $f_i = [f_{i,0}, f_{i,1}, ..., f_{i,N-1}]$ of x_i is computed as:

$$f_{i,k} = \sum_{n=0}^{N-1} x_i[n] \exp\left\{-j2\pi \frac{kn}{N}\right\}, k = 0, 1, \ldots, N-1 \tag{4.1}$$

Cadence velocity diagram has been proposed by [18] to analyse micro-Doppler signatures. After deriving the spectrogram, the CVD is obtained by taking the Fourier transform on the spectrogram along the time dimension for each Doppler frequency bin, i.e.

$$D = F_t\{S\}, \tag{4.2}$$

where D stands for CVD and $F_t\{*\}$ means the Fourier transform along the time dimension.

For the convenience of training neural network, the obtained spectrogram and CVD are normalised and converted to gray-scale images. These images are long in time axis, so we decided to segment them into square-shape pieces with an overlapping. The sample segmented spectrogram and CVD represented in heat map are shown in Figure 4.2.

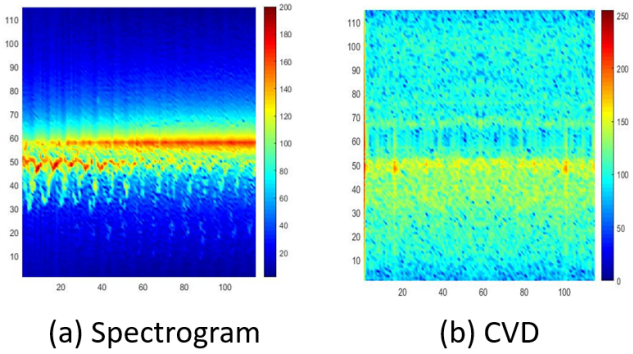

Figure 4.2: Processed spectrogram and CVD

4.3.2 Neural network training

Training a two-stream neural network means a large amount of back propagation, and since ViT does not utilise convolutional kernels for reducing parameters, this would be a fairly slow process. The solution we use is to separate the training of feature extraction for two streams and use the idea of transfer learning. In other words, the two ViT backbones for spectrogram and CVD are trained individually for their own feature extraction network. This considerably reduces the training time and provides high flexibility as two backbones can be adjusted respectively. After the two backbones are trained properly when they are able to extract most important features from two representations, the features extracted are fed into an attention-based fusion layer for aggregation. Python is one of the most popular and modern machine learning tool with profound community resources and the novel ViT model is also provided using Python with Pytorch, therefore, we choose Python for designing and training our model.

Chapter 5

Preliminary Results

5.1 Performance comparison

Table 5.1: Accuracy comparison between of the state-of-the-art methods [1, 2]. We also apply two benchmark object recognition models, MobileNet [3] and ViT [4], to the proposed radar gait recognition task. The experiments are run among single features (spectrogram or CVD) or feature fusion by the proposed AT-DSViT.

Method	Accuracy
DCNN-AlexNet on Spectrogram [2]	71.56%
TCN-VGG16 on Spectrogram [1]	69.24%
TCN-ResNet18 on Spectrogram [1]	85.56%
DCNN-AlexNet on CVD	72.44%
TCN-VGG16 on CVD	79.83%
TCN-ResNet18 on CVD	80.73%
MobileNet on Spectrogram	79.51%
MobileNet on CVD	76.84%
ViT-base 16x16 on Spectrogram	90.03%
ViT-base 16x16 on CVD	80.14%
Proposed AT-DSViT	**91.02%**

We conduct experiments to verify our approach with other state-of-the-art methods existing in the radar-signal-based recognition field. The comparison is demonstrated in Table 5.1. AlexNet [28] has been used in earlier mDS human identification [2], while VGG [27] and ResNet [30] are also commonly utilized in mDS human identification tasks more recently. [1] has used VGG and ResNet as state-of-the-art methods for comparison. In the experiments, AlexNet classifies around 72% of targets successfully on spectrogram representations, while VGG16 perform worse on spectrograms with accuracy of 69.24%. ResNet18, however, obtains the state-of-the-art with 85.56% on spectrograms. Additionally, experiments are conducted on CVDs using three aforementioned SOTA methods, where the ResNet18 reaches the highest with classification accuracy of 80.73%.

We also conduct experiments on two commonly used object detection models, MobileNet [3] and Vision Transformer (ViT) [4]. MobileNet [3] is a light-weighted convolutional neural network that performs well on natural image tasks, and ViT [4] uses patching and transformer encoder to enable a suitable network for radar signal representations. Due to the spectrogram and CVD have different physical meaning along two axes, i.e. time-frequency and frequency-frequency, extracting features simply by convoluting on the entire figure in a monolithic CNNs would be insufficient and ignore portions of the phys-

ical information. In contrast, ViT could have promising potentials to perform better on such radar signal diagrams by patching and utilising the self-attention mechanism. It is clear to see that ViT achieves 90.03% and outperforms the MobileNet in the spectrogram set by about 10%, proving that ViT could extract more useful physical information than CNNs which use convolutional kernels to extract features. As for the CVDs, performances of the two networks do not vary considerably, with ViT achieving a slightly better performance (80.14%) than the MobileNet (79.51%). Compared to the methods above, the proposed AT-DSViT with attention-based fusion mechanism performs significantly better with an accuracy of 91.02%. It can be concluded from the experimental results that the proposed AT-DSViT achieves a promising performance compared to the three state-of-the-art methods and thus is a suitable approach in human gait recognition tasks using mDS.

Chapter 6

Progress

6.1 Project Management

The timeline of the project is illustrated in a Gantt chart shown below. The chart could automatically present the bar of a corresponding length by entering the start date and duration. The bars in different colours denote different status. Purple means complete status, and yellow refers to beyond the plan. Beginning from October 10 to April 4, there are 28 weeks in total, and the plan contains seven stages in total. The project is deliberately most arranged in the first semester for submitting a conference paper.

Until the end of the first semester, the project is overall managed on schedule.As is shown on the Figure 6.1, a thorough literature review has been done, while the study of deep learning and gait recognition took one more week than planned. As for the design and experiment phase, since deep learning has been studied well, the network design began earlier than the plan, and the implementation of the network underwent successfully as scheduled. Thanks to the time saved in the previous phases, more thorough experiments could be conducted, which lasted four week and explored more possibilities that we have assumed.

1. Preliminary work
2. Literature review
3. Design & experiment
4. Interim report
5. Conference paper
6. Final report

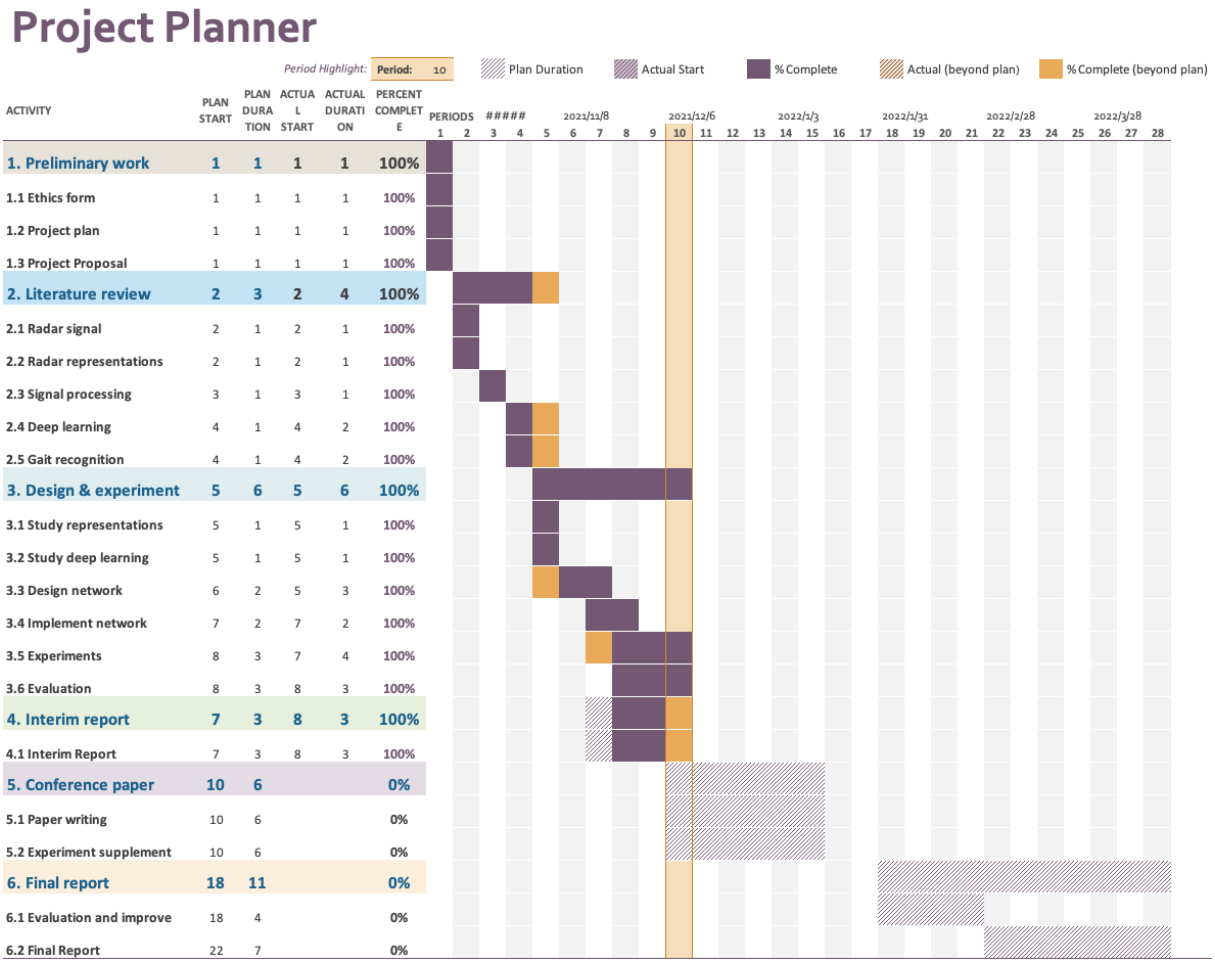

Figure 6.1: Project plan

6.2 Reflection and future work

The project is research oriented, so several research processes have been strictly included. Thanks to a previous research that I have participated, I could obtain an overview of research procedures. However, this project is much more formal and professional. Under the pressure of submitting a conference paper and the time limitation, I am required to absorb knowledge from unknown area fast and manage the whole project in a smooth way. All of these are challenging my ability of absorbing knowledge and management of time. However, with the great help of my supervisor, progress has been continuously smooth and successful.

The first necessary step is literature review. Since radar signal is a new field for me, relevant knowledge about different representations of radar signal and the concept of micro-Doppler is studied under the guidance of my supervisor who is an expert in signal. Radar signal processing techniques are necessary to master as well. After figuring out how to process and represent radar signal in a proper form, feature extraction techniques are studied. Since this project is closely linked to deep learning, this step contains exploration in deep leaning methods related or unrelated to the gait recognition as some research may already find ways to cope with radar gait features, and some methods can be transferred

to this area. What is surprising is that the self-study in deep learning has significantly helped me with my understanding of other forms of statistical learning in another module. Just like the transfer learning in machine leaning, knowledge can certainly be transferred and be of great help in other fields. Absorbing new knowledge, especially exposure to a completely new field can be terrifying, but the point here is to take the first step and have courage to consult experts. After literature review, reproduction of some valuable existing ideas are conducted to verify their availability, and new ideas are raised in this stage by evaluating the existing ones. Experiments are conducted to test some of the proposed ideas about the methods. According to the experimental results, there are several sprints to evaluate and improve the proposed method. A proper deep learning based method for radar gait recognition is proposed and evaluated at the end.

Although this is only the half of the project, I have already learned a lot, from rigorous academic writing skills, complete research cycle, ideas raising from brain storm to embracing new knowledge and technologies. As the preliminary results have been obtained, the focus of next stage would be generating and polishing the conference paper. Although many experiments have been done, they may not still be sufficient for supporting the proposed idea. Therefore, some additional experiments would be designed and completed in the conference stage.

References

[1] Pia Addabbo, Mario Luca Bernardi, Filippo Biondi, Marta Cimitile, Carmine Clemente, and Danilo Orlando. Temporal convolutional neural networks for radar micro-Doppler based gait recognition. *Sensors*, 21(2):381, 2021.

[2] Peibei Cao, Weijie Xia, Ming Ye, Jutong Zhang, and Jianjiang Zhou. Radar-ID: human identification based on radar micro-Doppler signatures using deep convolutional neural networks. *IET Radar Sonar Navig.*, 12(7):729–734, 2018.

[3] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Proc. CVPR*, pages 4510–4520, 2018.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. 8th Int. Conf. Learn. Represent.*, 2020.

[5] Arunesh Roy, Nicholas Gale, and Lang Hong. Automated traffic surveillance using fusion of Doppler radar and video information. *Mathematical and Computer Modelling*, 54(1-2):531–543, 2011.

[6] Rezaul K Begg, Marimuthu Palaniswami, and Brendan Owen. Support vector machines for automated gait classification. *IEEE Trans. Biomed. Eng.*, 52(5):828–838, 2005.

[7] Zhenyuan Zhang, Zengshan Tian, and Mu Zhou. Latern: Dynamic continuous hand gesture recognition using FMCW radar sensor. *IEEE Sensors J.*, 18(8):3278–3289, 2018.

[8] Zhen Meng, Song Fu, Jie Yan, Hongyuan Liang, Anfu Zhou, Shilin Zhu, Huadong Ma, Jianhua Liu, and Ning Yang. Gait recognition for co-existing multiple people using millimeter wave sensing. In *Proc. AAAI*, volume 34, pages 849–856, 2020.

[9] Rijun Liao, Shiqi Yu, Weizhi An, and Yongzhen Huang. A model-based gait recognition method with body pose and human prior knowledge. *Pattern Recognit.*, 98:107069, 2020.

[10] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proc. AAAI*, volume 33, pages 8126–8133, 2019.

[11] Xueru Bai, Ye Hui, Li Wang, and Feng Zhou. Radar-based human gait recognition using dual-channel deep convolutional neural network. *IEEE Trans. Geosci. Remote Sens.*, 57(12):9767–9778, 2019.

[12] Hoang Thanh Le, Son Lam Phung, Abdesselam Bouzerdoum, and Fok Hing Chi Tivive. Human motion classification with micro-Doppler radar and Bayesian-optimized convolutional neural networks. In *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, pages 2961–2965. IEEE, 2018.

[13] V.C. Chen. Analysis of radar micro-Doppler with time-frequency transform. In *Proc. IEEE Workshop on Statistical Signal and Array Processing*, pages 463–466, 2000.

[14] Dave Tahmoush and Jerry Silvious. Radar micro-Doppler for long range front-view gait recognition. In *Proc. 2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–6. IEEE, 2009.

[15] Zhaonian Zhang and Andreas G Andreou. Human identification experiments using acoustic micro-Doppler signatures. In *Proc. the 2008 Argentine School of Micro-Nanoelectronics, Technology and Applications*, pages 81–86. IEEE, 2008.

[16] Xingshuai Qiao, Tao Shan, and Ran Tao. Human identification based on radar micro-Doppler signatures separation. *Electronics Letters*, 56(4):195–196, 2020.

[17] Youngwook Kim and Taesup Moon. Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks. *IEEE Geosci. Remote Sens. Lett.*, 13(1):8–12, 2015.

[18] Svante Björklund, Tommy Johansson, and Henrik Petersson. Evaluation of a micro-Doppler classification method on mm-wave data. In *Proc. 2012 IEEE Radar Conf.*, pages 0934–0939. IEEE, 2012.

[19] Ann-Kathrin Seifert, Abdelhak M Zoubir, and Moeness G Amin. Radar-based human gait recognition in cane-assisted walks. In *Proc. 2017 IEEE Radar Conf.*, pages 1428–1433. IEEE, 2017.

[20] AW Miller, C Clemente, A Robinson, D Greig, A M Kinghorn, and J J Soraghan. Micro-Doppler based target classification using multi-feature integration. In *Proc. the IET Intelligent Signal Processing Conference 2013*, pages 1–6. IET, 2013.

[21] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proc. CVPR*, pages 4362–4371, 2017.

[22] Francesco Fioranelli, Matthew Ritchie, and Hugh Griffiths. Centroid features for classification of armed/unarmed multiple personnel using multistatic human micro-Doppler. *IET Radar Sonar Navig.*, 10(9):1702–1710, 2016.

[23] WL van Rossum, L Anitori, P van Dorp, JJM de Wit, and RIA Harmanny. Classification of human gaits using interrupted radar measurements. In *Proc. 2017 IEEE Radar Conf.*, pages 0514–0519. IEEE, 2017.

[24] Baptist Vandersmissen, Nicolas Knudde, Azarakhsh Jalalvand, Ivo Couckuyt, Andre Bourdoux, Wesley De Neve, and Tom Dhaene. Indoor person identification using a low-power FMCW radar. *IEEE Trans. Geosci. Remote Sens.*, 56(7):3941–3952, 2018.

[25] Haonan Chen, Guosheng Hu, Zhen Lei, Yaowu Chen, Neil M Robertson, and Stan Z Li. Attention-based two-stream convolutional networks for face spoofing detection. *IEEE Trans. Inf. Forensics Secur.*, 15:578–593, 2019.

[26] Dave Tahmoush and Jerry Silvious. Simplified model of dismount microDoppler and RCS. In *Proc. 2010 IEEE Radar Conf.*, pages 31–34. IEEE, 2010.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Proc. NeurIPS*, 25:1097–1105, 2012.

[29] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. CVPR*, pages 770–778, 2016.