View-invariant Feature Selector and its Application on Gait Recognition

by

Ning Jia

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Computer Science

December 2016

THE UNIVERSITY OF
WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

First and foremost I would like to express my sincere gratitude to my supervisor Professor Chang-Tsun Li and Associate Professor Victor Sanchez, for their motivation, enthusiasm, patience and immense knowledge. They have been constantly supportive and inspiring to my study as well as my life during the whole PhD career.

I would like to express my cordial gratitude to my parents, Mr. Laien Jia and Ms. Lifen Ma for their love, understanding and encouragement throughout my life, which give me the strength to chase whatever I want. To my wife Ms Xiaofan Zhang, and all the inspirations and love she brings to my life.

My sincere thankfulness to my annual progress panel members Associate Professor Nathan Griffiths for his guidance and valuable suggestions on my research. My gratitude to Professor Liang Wang and Associate Professor Yongzhen Huang for their kindness and favour during my visit to Chinese Academy of Science.

I would also like to thank the colleagues at the University of Warwick, Dr. Yu Guan, Dr. Xin Lu, Dr. Xinjie Wei, Dr. Yi Yao, Mr. Xufeng Lin, Mr. Xin Guan, Mr. Alaa Khadidos, Mr. Qiang Zhang, Mr. Roberto Fernandez, Mr Shan Lin, and Mr. Ching-Chun Chang for their kindness not just inside the lab. To my friends, Dr. Liang Sun, Dr. Yunlian Sun, Dr. Houwen Peng, Mr. Chunfeng Song, Mr. Chunshui Cao, Mr. Jingyu Liu, Mr. Yuqi Zhang, and Mr. Jianyong Xue for making my graduate life full of fun.

Declarations

This thesis is submitted to the University of Warwick in partial fulfilment of the requirements for admission to the degree of Doctor of Philosophy. The work presented here is my own, except where specifically stated otherwise, and was performed in the Department of Computer Science at the University of Warwick under the supervision of Professor Chang-Tsun Li and Associate Professor Victor Sanchez during the period Oct 2012 to Aug 2016. The research materials have not been submitted, either in the same or different form, to this or any other university for a degree. All sources of information are specifically acknowledged.

Abstract

The proliferation of the national-wide deployment of surveillance cameras and identity management systems has promoted the development of biometric systems. Gait as a behavioural biometric trait can be measured unobtrusively at a moderate distance, thus it is predominant in remote human tracking and identification tasks. The past two decades have witnessed a considerable development of gait recognition systems. Yet there are challenges that confine the practical application of gait analysis. The motivation of our work is to identify the problems and find corresponding solutions to explore the potentials of gait recognition and promote its applicability in open-world scenarios.

Gait recognition systems use human profile as features, while the appearance of human profile, also known as silhouette, can be affected in various manners. For example, clothing changes the shape of torso (coat) or legs (skirt); carrying bag attaches extra region to the silhouette; walking surface or speed variation changes the appearance of legs. On the other hand, camera viewpoint variation changes the shape of both the upper and lower body, while segmentation errors may cause massive corruption of the gait features. We summarise them into two categories: partial interference and holistic deformation. The former has been well addressed by existing literatures. The holistic deformation on gait silhouette results in large intra-class variation, and we notice that the performance of conventional approaches decreases under such circumstance. Thus our work focus mostly on the latter challenge.

Accordingly, we propose ViFS, an automatic feature selection approach that seeks for the optimal representation features from gallery set, and evaluate its perfor-

mance under various conditions. We find that ViFS minimises the intra-class variation between gallery and probe data, and by introducing proper feature enhancers, we can further reduce the number of holistic deformation modalities required in the gallery set. We test the proposed method on public dataset that contains viewpoint variations, and the matching accuracy has achieved 99.1% on CASIA Dataset B and 97.7% on OU-ISIR Large Population Dataset. The formulation and discussion are presented in Chapter 3.

The success of Convolutional Neural Network (CNN) based methods in image classification field has drawn attention from researchers. Recently a large number of literatures have covered the application of CNN in computer vision tasks, including face and gait recognition in the biometrics field. CNN has much greater discriminant learning ability in the highly non-linear space. Thus we merge CNN feature maps with the proposed ViFS approach, which achieves the state-of-the-art performance on view-invariant gait recognition problem. The methodology and results are presented in Chapter 4.

Among the holistic deformation challenge, the silhouette quality issue is seldom addressed, while no published dataset concerns with the influence of segmentation quality on gait recognition algorithms. We create a dataset that contains silhouettes with six different segmentation qualities in both gallery and probe set, and evaluated the conventional methods as well as the proposed ViFS approach on this dataset. It is proved that ViFS based framework and its extension outperforms the conventional methods by 8%-10%, which further indicates the effectiveness of ViFS based framework on gait holistic deformation challenge. This work is presented in Chapter 5.

This thesis aims at tackling the gait silhouette holistic deformation challenge, and ViFS based frameworks are proposed to achieve robust recognition performance. We evaluate the effect of different feature enhancers for ViFS, and find out that the discriminant power of CNN feature maps is much more powerful than subspace learning methods (3% higher accuracy under same conditions), thus it requires less gallery data to achieve deformation-invariant recognition.

Table 1: Thesis chapter and the corresponding publications.

Thesis Chapters	Publications	Contribution
Chapter 3	Publication 4	ViFS and its application on view-invariant gait recognition
Chapter 4	Publication 3&5	Combining CNN feature maps with ViFS
Chapter 5	Publication 1&2	Explore quality disparity problem

We provided the publication list for my PhD research on gait recognition as follows. Chapter 3-5 are highly related to the listed papers.

1 **N. Jia**, V. Sanchez, C.-T. Li, H. Mansour, "On reducing the effect of silhouette quality on Iindividual gait recognition : a feature fusion approach", *Proceedings of the 14th International Conference of the Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, Sept. 2015.

2 **N. Jia**, V. Sanchez, C.-T. Li, H. Mansour, "The influence of segmentation on individual gait recognition", *Proceedings of the 7th IEEE International Workshop on Information Forensics and Security (WIFS)*, Roma Tre University, Italy, Nov. 2015.

3 M. Rauf, **N. Jia**, C. Song, Y. Huang and L. Wang, "Knowledge Transfer Between Networks and Its Application on Gait Recognition", *Proceedings of the 21st International conference on Digital Signal Processing (DSP)*, Beijing, China, Oct. 2016.

4 **N. Jia**, C.-T. Li, V. Sanchez, "Fast and Robust Framework for View-invariant Gait Recognition", *Proceedings of the 5th International Workshop on Biometrics and Forensics (IWBF)*, the University of Warwick, UK, April 2017.

Chapter 1

Introduction

1.1 Human Identification Using Biometrics

Biometrics is interpreted as the way to recognise humans based on their distinctive biological characteristics [4]. The characteristics can be measured from either *physiological traits*, such as face, fingerprint, iris, or *behavioural traits*, e.g. handwriting, gait, voice. It is argued that biometrics is the most reliable way to identify people, as we carry them with us [5]. Biometrics provides solutions to a wide range of issues, with the overarching intention of denying imposters from the protected resources. Electronic products such as cellphones and personal computers are installed with fingerprint or face recognition applications to secure personal content. Organisations and corporations adopt face or iris recognition based check-in systems or access control devices in the restricted areas. Fingerprint and face recognition have been implemented for border control. Applying for identity card or visa requires biometric information registration. Security agencies use biometrics to identify and track suspects. The ever-growing demand of fast and precise automatic human identification systems has promoted the rapid development of computer vision based biometric authentication techniques. A considerable amount of innovative sensors, computational hardware and machine learning methods are developed over the past two decades to meet the demand of storing and utilising the large mounts of unstructured yet connected data.

A biometric authentication system is designed based on the presence of distinguishing attributes between individuals, and that the measurable attributes can be retrieved by sensors and represented in a digital format to proceed decision making via machine. Therefore it can be modelled with pattern recognition frameworks, which consist of three basic modules:

- *Source information acquisition.* The source information including biometric attributes is acquired from appropriately designed sensors. For instance, in the case of fingerprint, the friction ridges on fingertips are captured by the reflection of LED light from an optical sensor [6]. Figure 1.1 (a) demonstrates a typical sensor for fingerprint acquisition, while the captured fingerprint image is shown in Figure 1.1 (b).

- *Feature extraction.* Every biometric modality has its unique features that distinguish one subject from others. Feature extraction is aimed at extracting these features and store them in numerical form as the input for identification systems. Generally the features should be aligned and normalised following a specific criterion to avoid unexpected noise caused by misalignment in metric learning [7]. A demonstration of feature extraction on fingerprint image is shown in Figure 1.1 (c-f). (c) The original fingerprint image, (d) the local-region ridge orientation fields, (e) the fingerprint with sharpened ridge and noise elimination, (f) the extracted ridge bifurcations and terminations, which are also known as minutia (marked in red).

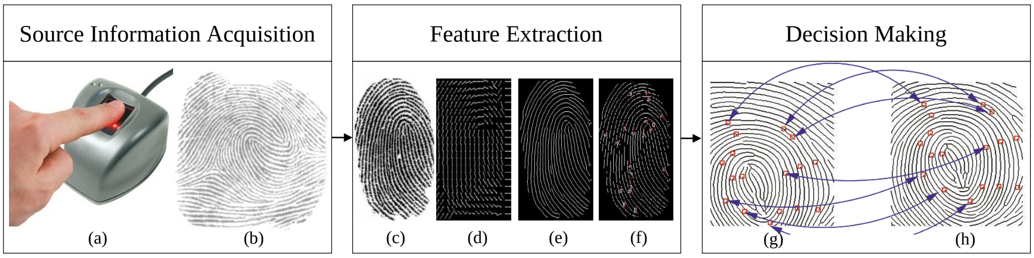


Figure 1.1: The process of fingerprint-based biometric authentication system.

- *Decision making.* Metric learning techniques are applied to compute the similarity scores between the extracted features from the input data and the reference templates stored in the database [4]. Figure 1.1 (g) demonstrates the minutia extracted from the input fingerprint image, Figure 1.1 (h) shows the minutia on the reference image, and the blue arrows refer to the corresponding matchings. Biometric systems may make decisions in verification or identification manner, based on the application context. *Verification* is to assess the validity of the claimed identity by comparing the scores of the input features with the pre-enrolled reference templates. For example, in the case of handwritten verification, the scores are compared with varying decision threshold to validate the claimed identity [8]. *Identification* requires the input features to be compared with all the reference templates, and the top matching score often decide the identity of the input data. In addition, identification can be open-set or closed-set, based on whether the identity of the input sample exists in the reference database. *Recognition* can be used to imply either verification or identification [9], thus we prefer to use it in this thesis for a general reference.

Apart from these basic components, *feature selection* is an indispensable procedure in modern biometric systems as well as pattern recognition approaches. As is often the case, the source information comprises a fair amount of redundant contents. Therefore the dimensionality of the numerical representation for biometric features could be very high. For the sake of recognition accuracy and computational efficiency, a conventional solution is to apply dimension reduction tools to simplify the problem and increase the decision-making accuracy. Furthermore, the recent development and application of convolutional neural networks (CNN) to computer vision tasks has boosted the performance of biometric identification systems. With the ability to automatically learn problem-specific features directly from input image, the integrated deep architecture of CNN merges the basic modules into one framework, providing an end-to-end solution to modern biometric problems [10–12].

Humans recognise one another not only using body characteristics (e.g. face, voice, gait), but also rely on other contextual information (e.g. hairstyle, height, accessories). Recent studies suggest that a well designed fusion framework is capable of combining multiple primary biometric traits and ancillary information, for example, age, height, and gender, which contributes to the improvement of identification accuracy. These attributes are named soft biometrics or light biometrics, since they provide information concerning human identity, while these attributes cannot be used to verify people independently [13]. They can be deduced from the prime biometrics for the same subject, or acquired from other approaches to support the identification. Recently soft biometrics have been introduced to monitoring and indexing databases [14].

Considering the enrolment rate, time cost and matching accuracy, it is inadequate to employ a single biometric trait in large scale deployment of biometric systems [15]. By merging the advantages of multiple biometric source, a multi-biometric system can significantly improve the population coverage and the indexing speed, as well as the recognition rate. As mentioned in [16], the current deployed systems integrate at least two biometrics: fingerprint with face, or face with iris. Apart from face, iris and fingerprint, which have been testified as efficient traits for biometric systems, researchers started to explore more subtle biometrics, such as ear shape, vein pattern, and gait, in order to amplify excessive modalities. As one of the behavioural biometrics, gait has gained growing attention in the past two decades and is regarded as a promising biometric trait for human identification. At present, there is an increasing number of closed-circuit televisions installed in public places for surveillance and security purposes. As reported by the British Security Industry Authority (BSIA), there are up to 5.9 million CCTV cameras installed in UK [17], and the demand of supervising such a large number of CCTV cameras is acute. Biometric modalities, which require subject cooperation, like fingerprint or palm, are infeasible for these purposes. It is also known that face recognition is strongly affected by illumination and view angles in the open world. Also, when the subject keeps a moderate distance from the camera, neither face nor iris is recognisable due to low resolution. Furthermore, it is known that face, fingerprint and iris may not be perceivable due to the occlusions (sunglasses, scarf, etc.), or the viewpoint of cameras (side-view and back-view). Gait on the other hand seems to be a better choice in dealing with such problems. It is suggested by Nixon et al. [5] and a profusion of literatures that the walking pattern of human is unique and highly repeatable, and people can identify one another based on their gait. As a behavioural biometric trait, gait can be used to identify human from a moderate distance. In addition, gait can be an expedient technique in establishing fast and robust multi-biometric system by narrowing down the searching space. This thesis explores the potential of gait analysis to be an independent tool in human recognition tasks.

1.2 Gait Recognition

Gait recognition focuses on both the human body shape in spatial domain and movement in temporal domain. A conventional gait recognition system consists of 3 steps:

- *Acquire gait signature (source information acquisition).* A spectrum of images containing the walking subject are captured using a video camera (or cameras). The binary silhouettes (or shape) of the subject, regarded as the gait signatures, are subtracted from the background.
- *Construct gait template (feature extraction).* Gait template refers to the normalised binary silhouette [18] and its extensions, e.g. the average of all silhouette within a gait cycle [19]. Compared with the original silhouette, a good template uses less storage space and achieves a higher recognition accuracy.
- *Similarity measurement (decision making).* After modelling input and reference data using the same template, we measure the distance between them and find the matching identity. Simple approaches like nearest-neighbour methods can be implemented for this purpose [20]. Furthermore, we may introduce other distance functions to reduce the dimensionality of feature space and find projective subspace where useful information is preserved and the redundant features is discarded.

A gait cycle starts with the *stance* state, followed by two *strides*, and stop with another stance state. An intuitive example is presented in Figure 1.2.

In a cooperative manner, the close-distance biometrics such as fingerprint, face and iris can achieve very high identification rate [16]. However, when the subject

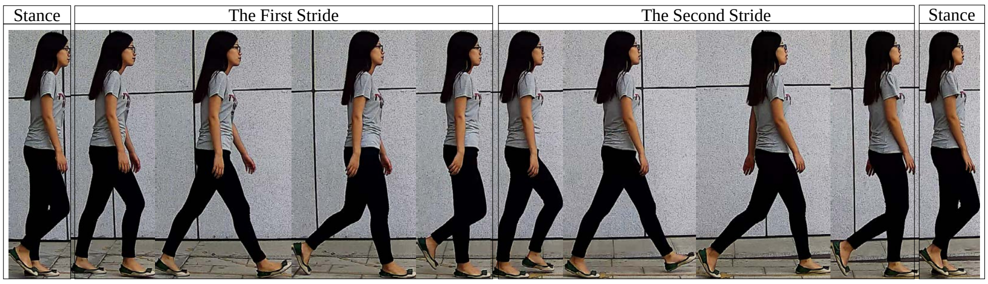


Figure 1.2: The demonstration of one gait cycle.

does not cooperate with certain protocols, e.g. reluctant to contact with the optical fingerprint sensor or look straightforward to the camera, those biometric modalities are invalid or severely degraded. Instead, gait recognition can achieve high performance under harsh conditions, such as long distance, non-uniform, illumination variance and low resolution. Moreover, gait is robust to disguise and occlusions, which makes it an important tool in tracking and identifying subjects through public surveillance cameras. Figure 1.3 presents two scenarios where subjects are captured by the camera in a non-invasive manner. It is nearly impossible to perform face or iris recognition, since these features are too degraded to be perceived, or completely occluded. Under such circumstances, gait appears to be the best choice to identify people. The main advantages of gait recognition are summarised as follows:

- *Identification at a distance.* As a behavioural biometrics trait, gait does not acquire the texture details of human body. Instead, it focuses on the body shape as well as spatial-temporal movement, which hardly degrades within a moderate distance. On the contrary, face or iris features are not perceivable at a distance or under low resolution. Hence computer vision based gait recognition is deemed as the optimal choice for human identification at a distance.

- *Robustness to ambient interference.* The presence of pose, illumination and ex-

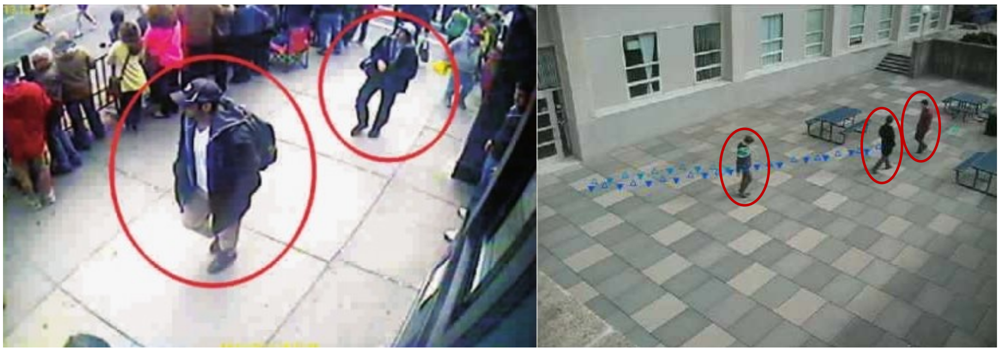


Figure 1.3: Two scenarios where main biometrics (e.g. face, fingerprint, iris) are not perceivable. Gait seems to be the only available trait under such circumstances.

pressions/accessories variations lead to low recognition accuracy in face recognition, and these variations are ubiquitous in non-cooperative scenarios. Gait recognition has better tolerance to these interference factors [2] [21].

- *Non-obtrusiveness.* Since it is performed at a distance, a gait recognition system does not require the subject to interact with the camera. Also, gait does not need the subject to be cooperative, on account of its better tolerance against view angle variations.

- *Non-invasiveness.* The camera could be installed at a concealed place, which does not arouse the awareness or alarm of the subject. Hence the subject is less likely to disguise purposefully.

Apart from being used as the prime biometric trait in human identification, gait recognition is also a good auxiliary to other biometrics in multi-biometric systems. It is proved that the fusion of face and gait has achieved promising performance in human identification and tracking [22] [23].

In [24] [25], pioneering experiments have been carried out on translating gait biometrics to forensic task with real crime scene videos, and gait based evidences has been introduced by court to increase the confidence of identity. Larsen et al.

identified a bank robber in Denmark using evidences from forensic gait analysis [24]. Bouchrika et al. manually marked human segments from raw video sequences and analysed the locomotion of anatomical annotations, which helps identifying a burglar in UK [25]. However, the performance of computer vision based gait recognition is limited by a variety of factors, such as walking speed, clothing and carry condition, camera view point, silhouette quality. Still, there is much work to do for automatic gait recognition in practical application. In this thesis, we aim to explore the potential of gait recognition, and propose robust algorithms that are less sensitive to these factors.

1.3 Contributions and Thesis Outline

As gait recognition is still in its early stage, researchers have discovered a few factors that affect the identification accuracy, as mentioned above. However, there are more to be explored, like silhouette quality disparity. In order to bring gait recognition from laboratory experiment to real world applications, the problems need to be addressed. The purpose of this research is to study and explore the potential and limitation of gait recognition, and to propose robust algorithms that are less sensitive to two problems: recognition under arbitrary view, which is caused by camera viewpoint changes, and recognition with silhouette quality disparity, which is caused by inconformity of gait signature acquisition. In this thesis, we view these problems as the holistic feature deformation challenge, and propose view-invariant feature selector (ViFS) based frameworks to tackle this challenge.

1.3.1 Contributions

The three contributions of this thesis are listed as follows. The corresponding chapters and publications are demonstrated in Table 1.

1. For gait recognition under arbitrary views (also known as view-invariant gait recognition problem), we propose a view-invariant feature selector (ViFS) based framework to automate feature selection process and perform fast view-invariant gait recognition. Specially, we assemble gallery templates from different view angles into one set, and regard the probe template from a specific view angle as the reference set. By minimising the cross-view distance between gallery and probe set, we realise the ViFS. We use subspace learning methods as feature enhancers to increase the discriminative power of gait features. Benefiting from parameter-free linear computations, our framework has very low-computational cost, making it suitable for real time applications. In addition, we notice that the prosed ViFS is amenable to be used with various subspace learning methods as feature enhancers. The formulation of ViFS and its evaluation are detailed in Chapter 3.

2 We further enhance the performance of ViFS based framework using CNN feature maps. Specifically, we train a conventional CNN with sufficient multi-view data. Then we use CNN as the feature enhancer to obtain the gait feature maps from a certain layer of the network. We use ViFS to automatically select view-invariant features from the gallery CNN feature maps, and use them as reconstructed feature templates to match with the probe CNN feature maps. In addition, we analysed the feature maps extracted from the well trained CNN model, in order to understand the functional mechanisms of CNN as feature

enhancer. The joint force of CNN feature maps and ViFS achieved higher matching accuracy than the previous work on CASIA Dataset B, which further promoted the potential of ViFS in real world gait application. The details of this work is presented in Chapter 4.

3 Apart from the widely known factors, there are others hindering the performance of gait recognition algorithms, including silhouette quality disparity. We generate a dataset to statistically analyse this problem and evaluate the performance of conventional algorithms. We discovered that when gallery and probe templates are generated using different approaches, there is a strong possibility that the recognition rate would be very low, due to the holistic feature deformation on the silhouettes. Meanwhile, if the segmentation results is very close to the ground truth, even different approaches might lead to similar results. Considering that the segmentation errors result in holistic feature deformation, we apply ViFS to reduce the effect of silhouette quality disparity. Specifically, we evaluate the performance of conventional subspace learning algorithms on the generated dataset, including a weak classifier fusion approach that make use of the fused features from different quality pairs, and make classification decisions through majority voting. Alternatively, we adopt ViFS to select the optimal features from multi-quality gallery set, and perform classification based on the Euclidean distances between the reconstructed gallery features and the probe features. Experimental results show that the ViFS based proposed framework outperforms other algorithms, suggesting that ViFS is efficient in tackling gait holistic deformation challenge.

1.3.2 Thesis Outline

The rest of this thesis is organised as follows.

- Chapter 2 provides a review of the prior works related to the research themes of the thesis to prepare the readers for the more specific works developed by the author in the past four years. This chapter starts with the development history of gait representation, and then points out the challenges facing gait recognition techniques. Many prior works aiming at addressing various issues are also discussed. A number of underlying techniques, such as dimension reduction, discriminant learning, and CNN, to be used in our own work are presented.

- Chapter 3 deals with the challenges due to view variation, which is often encountered in gait recognition. We will first discuss the impact of view variation on gait recognition and then propose a framework, which extracts a set of view-invariant features and then use subspace learning methods to enhance the selected features to facilitate recognition.

- In Chapter 4, we investigate the combination of CNN features and feature ensemble approaches on tackling view-invariant gait recognition problem on gait recognition. We firstly discuss the effectiveness of CNN feature maps and analyse the reasons behind the incompatible performance of CNN based approaches on computer vision tasks. Then we evaluate its compatibility to traditional approaches on main-stream database.

- Chapter 5 is concerned with the observed fact that gait recognition accuracy is dependent on the disparity between the quality of gait silhouette of the probe and gallery images. We firstly propose a weak classier fusion strategy to deal with the identified problem. The classifier fusion strategy makes use of the information from the weak classifiers and majority voting to increase the probability of the right guess. Later we adopt ViFS to further improve the performance of the methodology.

- Chapter 6 concludes this thesis and points out a few lines of investigation in the future.

Chapter 2

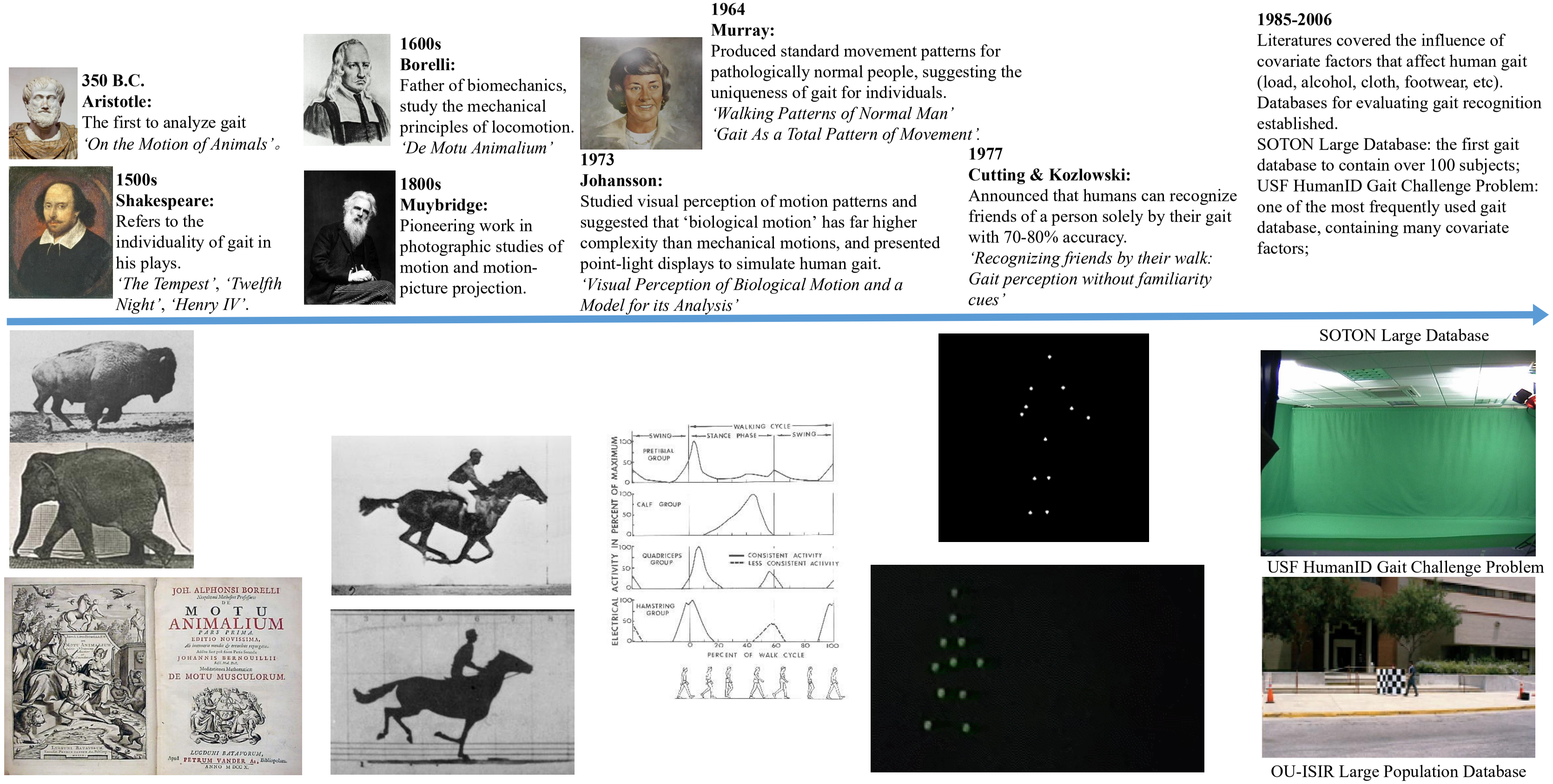
Literature Review

2.1 The History of Gait Representation

It is mentioned by Nixon et al. [5] that the foremost study on gait is by Aristotle in his book *On the Gait of Animals* [26]. With the development of physics and mechanics, the early pioneers such as Leonardo da Vinci, Galileo and Borelli were able to promote the development of biomechanics on animal locomotion. Muybridge (1830-1894) conducted his famous experiment of recording racing horse using 12 cameras, which is regarded as the pioneer work on photographic study of motion. Murray et al. [27] measured the kinematics of various human body components during walking, based on a sequence of photos recording human gait. According to their experiments on 60 healthy men of different ages and height, Murray et al. reported the normal interval of parameters on human gait, such as the duration of gait cycle, the length of stride and step, the rotation of body joints (pelvis, knee, ankle, etc.). It is a pioneer work on using a simple and low cost photographic method to record and statistically analyse normal human gait, while other contemporary literatures mainly focused on studying the pathology of human gait. Another work by Murray indicated that human gait comprises the coordinated motion of 20 body components [28]. Johansson [29] proved that people can recognise human biological motions (walk, dance, etc.) which are represented by dynamic bright points on the main joints. Later Cutting et al. [30] pointed out that this dot representation model of biological locomotion is sufficient for human identification[1], which substantiated the discriminatory power of gait. Those works revealed that the complexity of the interactions between body segments typifies the distinctive motion characteristics for each individual, thus gait can be used for human identification. Moreover, the representation of human gait can be very abstract. Those landmarks are summarised in Figure 2.1. In the early stage of gait research, several literatures proposed gait representation models describing gait spatial and temporal patterns [31, 32]. Since the 21st century a number of literatures have been seeking for an optimal way of representing human gait, such as [33–40]. Recently a few templates are proposed, such as [41,42]. Despite that a spectrum of feature representation were proposed, e.g. contour, structural skeletons, and dots representation, it is proved that gait binary silhouette is the most effective representation that encodes gait biometrics [18]. Therefore, as mentioned in Chapter 1, we refer gait signature to the binary silhouette extracted from gait video sequences.

There are several terms frequently occurred in this thesis, as well as gait recognition literatures, with regard to gait nomenclature. The term *gallery* refers to labelled gait sequences or templates stored in the database, corresponding to the *reference* data as introduced in Chapter 1. *Probe* means the sequences to be identified or verified, corresponding to the *input* data. By performing metric learning approaches, we may find the minimum distance between the probe data and the

[1] A similar demonstration of dot representation on human gait can be found on `https://www.biomotionlab.ca/Demos/BMLwalker.html`.



350 B.C.
Aristotle:
The first to analyze gait
'On the Motion of Animals'。

1500s
Shakespeare:
Refers to the individuality of gait in his plays.
'The Tempest', 'Twelfth Night', 'Henry IV'.

1600s
Borelli:
Father of biomechanics, study the mechanical principles of locomotion.
'De Motu Animalium'

1800s
Muybridge:
Pioneering work in photographic studies of motion and motion-picture projection.

1964
Murray:
Produced standard movement patterns for pathologically normal people, suggesting the uniqueness of gait for individuals.
'Walking Patterns of Normal Man'
'Gait As a Total Pattern of Movement'.

1973
Johansson:
Studied visual perception of motion patterns and suggested that 'biological motion' has far higher complexity than mechanical motions, and presented point-light displays to simulate human gait.
'Visual Perception of Biological Motion and a Model for its Analysis'

1977
Cutting & Kozlowski:
Announced that humans can recognize friends of a person solely by their gait with 70-80% accuracy.
'Recognizing friends by their walk: Gait perception without familiarity cues'

1985-2006
Literatures covered the influence of covariate factors that affect human gait (load, alcohol, cloth, footwear, etc).
Databases for evaluating gait recognition established.
SOTON Large Database: the first gait database to contain over 100 subjects;
USF HumanID Gait Challenge Problem: one of the most frequently used gait database, containing many covariate factors;

Figure 2.1: The milestones in the development history of gait representation.

gallery dataset, and label the probe data with the corresponding gallery data. *Silhouette* is defined as the region of pixels from a person [18]. Within one gait cycle, the silhouette goes through spatial-temporal changes, which models the pattern of human locomotion over time. In addition, it removes the interference information[2], which mainly refers to the texture details and colour regions on human body.

Gait recognition approaches can be broadly classified into two categories: *model-based* and *appearance-based*, in terms of different feature types. Model-based gait recognition refers to identifying people by modelling their distinctive gait characteristics with underlying mathematical structures [43]. One or more physical models of the human body are established and integrated with a set of parameters that are extracted from human silhouettes, for example, the size of body parts (height, length of torso and thighs), the length of the stride and cadence, the speed of the stride, and the variation of joint angles. These parameters, which are coded with the discriminant information, are logically and quantitatively connected with each other. These human physical models can be 2-dimensional (2D) or 3-dimensional (3D), depending on the number of digital cameras recording gait sequences from different views. Niyogi et al. proposed to extract contour of gait silhouette, and they built a five-stick model to fit the bounding contour, representing human body structure [44]. The periodic angle variation recovered from the joints of the stick model were used for classification. Cunado et al. modelled human gait as an articulated pendulum and recorded the parameters of joint angles, thigh hight and frequency during the walking cycle [45]. Their work also mainly focused on the bottom half of human body, i.e. the movement of human legs. Inspired by these two pioneering

[2] In computer-vision based gait recognition system, the interference information could misguide the distance measurement between vectorised gait signature. Nevertheless, these information, including clothes, shoes, gender, and accessories, may be regarded as soft biometrics and contribute to identify people in a fusion framework [14].

works, many researchers proposed optimised structural and motion models, as well as feature extraction methods, to improve identification accuracy [46] [47]. Wang et al. proposed to extract features from both shape and movement features, then fuse the static and dynamic information of the shape and movement features, respectively [48]. In [49] [50] the authors constructed a 3D model from raw video sequences to represent human gait, which is robust to factors such as view angle, clothing, illumination. However, these methods require high definition video source and are computationally expensive. Also they did not test their methods on large scale databases. When using gait recognition for surveillance and access control, depth cameras can be used to achieve higher accuracy due to the richer information than 2D-based gait features [51–54]. Figure 2.2 illustrates the extraction of normal and depth-based gait silhouette and templates. It is clear that depth-based silhouette retains complete information, while the GMM based methods fails to preserve the features on the top and bottom part of the silhouette. This situation is constantly encountered on outdoor scenario due to the illumination changes on the subject and the interference of walking surface or objects on the background. Most model-based methods rely on high quality gait sequences captured under controlled environment (e.g. indoor environment, close-distance between subject and camera, multi-view cameras, in-depth camera or kinetic camera), thus they are better at handling occlusions and changes in scale, as well as camera view point changes. However, the restrictions to sensors and the low tolerance to video quality makes model-based methods less applicable for outdoor gait recognition.

Appearance-based methods adopt gait silhouette as the feature source to build effective gait templates. The classification is performed by measuring the pixel-to-pixel distance between gallery and probe templates. A commonly used

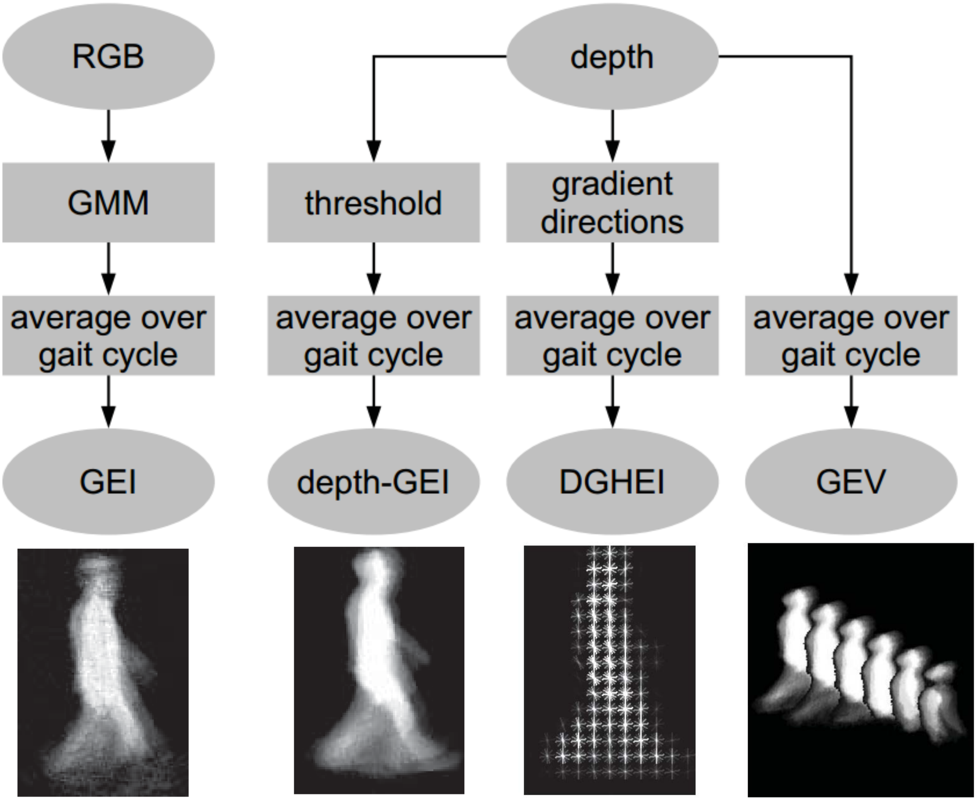


Figure 2.2: Comparison between normal and depth-based methods on building gait templates [54]. The GEI on the leftmost side seems blurred and distorted, comparing with the depth-GEI on its right side. The other two images illustrate other approaches based on depth-based gait silhouette.

appearance-based template is the Gait Energy Image (GEI), which averages all the binary silhouettes from a gait cycle and generate a single gait template [19]. Since GEI encrypts spatial and temporal information of one gait cycle into a single template, methods based on GEIs usually have low computational costs and low storage requirements. Another similar approach named Motion Silhouette Image (MSI) is proposed in [55]. The pixel value of MSI at a certain position depends on its temporal history of motion over a gait cycle. Both GEI and MSI embed spatial information over a gait cycle into one template, thus they are vulnerable to shape variation caused by rotation, clothes or carry condition. Moreover, since GEI converts the spatial-temporal information during one walking cycle into a single 2D gait template, which avoids matching features in temporal sequences, it loses the dynamical variation between successive frames. Wang et al. proposed Chrono-Gait Image (CGI) method to cover temporal information which is absent in GEI [56]. They extract the outer contour from gait silhouette image, and encode with different colour according to the time stamp in a gait sequence. Then the coloured contour images from a gait cycle are encoded into one multichannel template, i.e. a CGI. Bashir et al. proposed to calculate the Shannon Entropy of gait silhouette image and gather them into gait entropy image (GEnI) [57]. Experiments on several large gait dataset (over 4000 subjects) suggest that GEIs, among other templates that are mentioned above, are the most statistically stable and efficient template for gait recognition [58] [59], while other templates are proved to be useful for a special challenge or on a certain database. Therefore for the sake of simplicity and universality, most researchers still apply GEI as gait representation template to evaluate their algorithms. The advantages of appearance-based approaches are the low requirements of silhouette quality and computational ability. However, these 2D

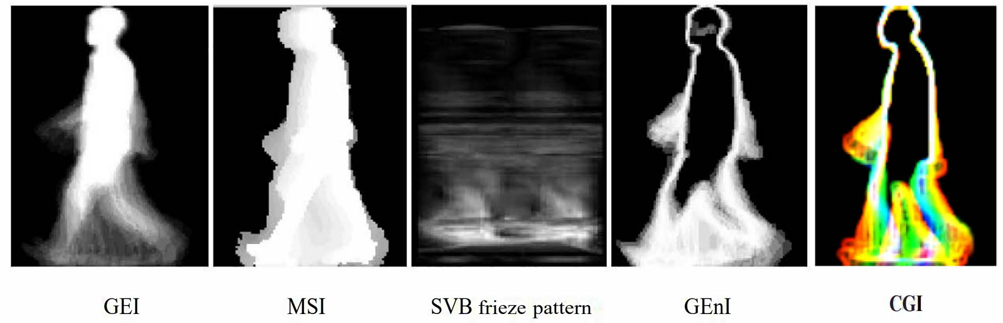


Figure 2.3: Examples of GEI, MSI, SVB frieze pattern, GEnI and CGI.

based features are not robust to viewpoints and scale changes. Figure 2.3 displays examples of the templates mentioned above.

In [60], the gait template containing both spatial shape information and temporal gait variation are summarised as spatial temporal-based method. Comparing with the appearance-based templates mentioned above, the spatial temporal-based method preserves the temporal variation between successive frames. Inspired by the success of crystallographic group theory on recognising periodic patterns, Liu et al. proposed a mathematical model called frieze patterns [61]. Their algorithm automatically recognises the underlying lattice pattern from gait silhouette sequences and extracts the representative motifs of features. Followed by Liu et al.'s work, Lee et al. ameliorated the frieze patterns based method that jointly used the intra and inter-shape variations [62]. Instead of extracting frieze patterns from gait silhouette, they perform their algorithm on difference frame between a series of key frames. The frieze pattern based templates is specially designed for tackling the local region shape variations caused by clothing or carrying bags, or walking speed. However, if no appearance changes are present, these methods is inferior comparing with classical appearance-based templates such as GEI [57]. In addition, frieze pattern methods requires higher demand of computational power, and preserve less meaningful 2D graphic information due to the vectorised computation [63].

Our intention is to push the process of gait recognition in real world application. Therefore in this thesis we adopt appearance-based methods, due to its effectiveness on outdoor gait recognition. Since the contributions of this do not include proposing new gait templates, we only employ GEI for the sake of simplicity and efficiency. The reviews of machine learning techniques on gait recognition are presented in Section 2.3.

2.2 Challenges and Public Databases

In this section, we introduce the main challenging factors for gait recognition, as well as the commonly used gait databases comprising these factors for evaluating algorithms.

2.2.1 Main Challenges

Computer vision based gait recognition approaches automatically extract a set of features comprising identity characteristics from the raw video sequences. Often these features comprise noises (corrupted features) that are unrelated to personal identity. These noises arise from a number of interior or exterior factors, which cause transient or permanent changes to gait patterns and consequently degrade the discriminative power of gait features. In order to ensure the recognition accuracy, it is critical (and challenging) to eliminate the influence of these factors from the extracted features set. To summarise, we classify these factors into three types in terms of their causes:

- *Personal factors.* Human gait can be affected by walking speed, healthy condition, time elapse, mood, clothes, carry conditions, etc. Variations come from subject himself/herself.

- *Environmental factors.* The extracted features can be contaminated by illumination changes, occlusion, walking surface, etc. Furthermore, when performing background subtraction (in order to obtain the foreground subject silhouette), moving objects on the background can easily generate strong noise.
- *Sensor factors.* Computer vision based approaches require a sensor to capture raw data (sound, images, video sequences, or other forms of digital signals). For gait recognition, the resolution or frame rate of camera, the distance and view point between camera and subjects can generate interference to the extracted features.

Alternatively, we also summarise two main challenges for appearance-based gait recognition, in terms of the active region of the factors:

- *Partial interference challenge.* The interference comes from the variation of carry condition, walking speed, shoes and walking surface, etc. The feature corruption is mainly on partial region of human gait silhouette, and each of the factor affects a relatively constant position. Most of these factors has been introduced in the published databases, and a number of approaches have been proposed to tackle this challenge.
- *Holistic deformation challenge.* The deformation could be engendered by the change of camera view point, degradation of video quality, etc., causing the holistic shape variation on human silhouette. The effect of segmentation quality has not been focused in the published databases, while the large view variation issue remains an open problem.

Among the existing approaches, Guan et al. proposed a robust framework based on Random Subspace Method (RSM) and its extensions, which achieves the state-of-the-art accuracy on local interference challenge on the mainstream databases. In the summary work of Guan et al. [21], they systematically analysed the effectiveness of RSM as a random feature selection technique. Based on the assumption that the noisy region is less than 50% of the whole feature area, the number of correct classification decisions is in excess of the wrong decisions from weak classifiers. Thus by performing majority voting scheme, RSM based method significantly improves the matching accuracy on local interference challenge (by more than 10%). Similar approaches include gait silhouette partition [64], in which the noisy regions are isolated from the clean regions, or patch segmentation [65], where the noisy patches are outnumbered by clean patches in majority voting. These are effective solutions for local interference challenge. However, for global deformation challenge, the clean feature regions are overwhelmed by noisy regions. Figure 2.4 shows the GEI samples of one subject captured from 11 different views. When the subject walks parallel to the camera image plane, i.e. 90° with the camera viewpoint, it is commonly referred to as side-view or *lateral-view*. Besides, 0° is illustrated as *front-view* and 180° as *back-view*, while other views are referred to as oblique views [66]. When view variation is larger than 18°, the shape of GEI has explicit global deformation. Figure 2.5 presents the examples of gait sequences captured under three different scenes in the first row with row label *Frame*, and the gait signatures obtained using three different foreground detection approaches in the rest three rows labelled as Seg (a) to Seg (c). Scene 1 has static background, Scene 2 contains slight interference from moving leaves on the background, and Scene 3 has dynamic objects (cars and pedestrians) on the background. Seg (a) refers to the

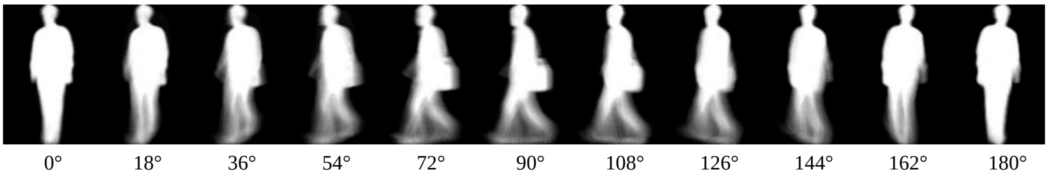


Figure 2.4: The demonstration of gait recognition under arbitrary view problem. There are 11 GEIs from view 0° to 180°, with an interval of 18°.

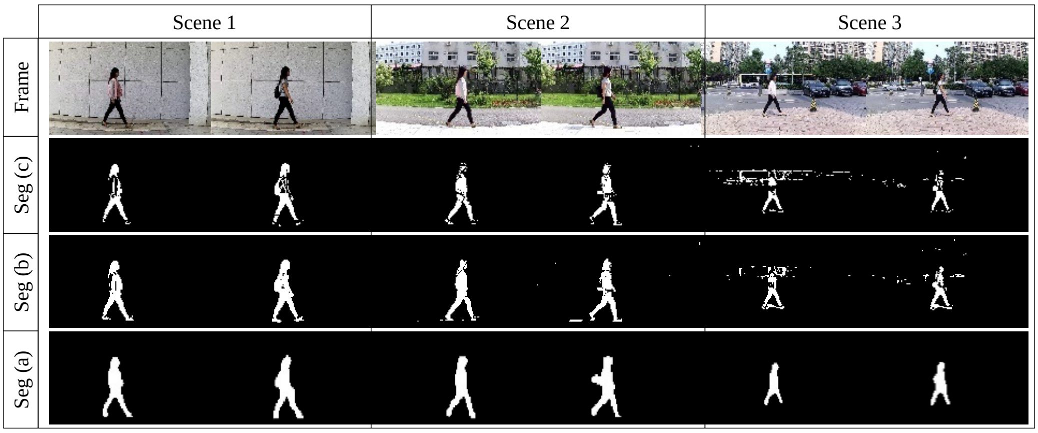


Figure 2.5: The demonstration of gait recognition with silhouette quality disparity problem. Gait signatures obtained from various scenes are segmented using different approaches.

Gaussian Mixture Model, Seg (b) refers to the background subtraction method, and Seg (c) is a well-trained CNN model for human segmentation. Those three methods provides very different segmentation results, while the feature corruption occupies more than 50% of the silhouette area. Thus the partial feature corruption based approaches as mentioned above cannot deal with these problems. Instead, we need to develop new solutions for global deformation challenge.

2.2.2 Databases

With the popularisation of web search engine and on-line social network applications, it is very easy to access large numbers of human face images. As one of the most publicised biometrics, fingerprint has been used for forensics and identification

Table 2.1: The most widely used gait databases and their attributes.

Name	Covariates	Subjects	Sequences	Views	In/Outdoor
CASIA B	4	124	1240	11	In
SOTON Large	3	115	2128	2	Both
USF HumanID	6	122	1870	2	Out
OU-ISIR, B	1	68	1350	1	In
OU-ISIR, LP	1	4007	7842	4	In

for over a century. Comparing with gait data, face and fingerprint data are more easily acquired. On the contrary, gait recognition databases mainly focus on factors and application potentials. The published gait databases only contain a small number of subjects, comparing with other biometric databases such as face and fingerprint. Table 2.1 listed several gait databases that are mostly used for gait system evaluation. In this section, we review the main challenges at the present stage, and detail two mainstream databases that are widely used for evaluating gait recognition algorithms, namely CASIA Dataset B (CASIA B) and OU-ISIR Large Population Dataset (OU-ISIR, LP).

CASIA Gait Dataset B

CASIA Gait Dataset B has its distinctive advantage in the number of view directions for each recorded sequence. Dataset B is a large multiview gait database created in January 2005. There are 124 participants in this dataset, each recorded with three variations: *normal walking (nm)*, *wearing coat (cl)*, and *carrying bag (bg)*. Every time a subject walks through the designated path, 11 videos from different view angles are captured simultaneously. The camera set-up for capturing gait sequences is illustrated in Figure 2.6. As is shown, 11 cameras are placed on the left side of the walking path for subjects. The angle difference between two adjacent cameras are 18°. Hence gait sequences with 11 different angles are captured whenever a subject

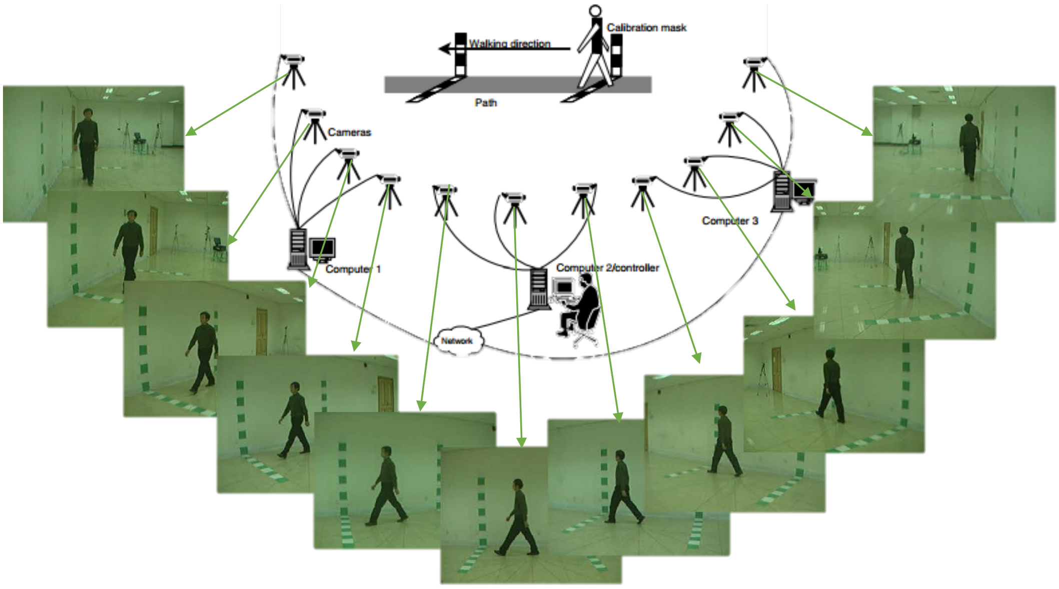


Figure 2.6: The camera set-up and demonstration video sequences of CASIA Gait Dataset B [67]. The angle difference between two adjacent cameras are 18°. With this set-up, 11 gait sequences will be captured each time a subject walks through the path.

walks through the path. Each time a subject walked naturally along a straight line 6 times, and 11×6=66 normal walking video sequences were captured. Followed by normal walking, the subject walked twice along the straight line with his (or her) coat. Likewise, the subject then carried a bag[1] and walked twice again. The frame size of the recorded video is 320×240, and the frame rate is 25 fps. Each video sequence contains two or three gait cycles. Among the participants there were 93 males and 31 females, 123 Asians and 1 European, ageing from 20 to 30. 10 video files are recorded for each participant (6 nm+2 cl+2 bg). There are a total of 10×11×124=13640 video sequences in the database.

[1] The bag could be a knapsack, a satchel, or a handbag, depending on the preference of the subject.

OU-ISIR Gait Database

The OU-ISIR Gait Database is published by the Institute of Scientific and Industrial Research (ISIR), Osaka University (OU). It comprises two dataset widely adopted for evaluating gait recognition algorithms: Treadmill Dataset and Large Population Dataset. The main considerations are: 1) to justify whether the proposed gait recognition framework is robust to common variations, and 2) to ensure the experiment result is statistically reliable.

The Treadmill Datasets consist of four subsets denoted from A to D, each introducing a specific variation [68]. Dataset A, C (under preparation, has not been published yet) and D comprises speed variation, view variation and gait fluctuation, respectively. Collected on 2007, Treadmill Dataset B contains 68 participants, each with 24 clothes variation. When recording gait sequences, the participants walked on a treadmill with up to 32 combinations of clothes variations[3], and their gait sequences are captured by a side view camera at 60 fps, 640 by 480 pixels. A set of binary silhouettes is extracted from the sequences and stored into the dataset, with standard size 128×88. This dataset is specialised in studying the effect of clothes on gait recognition. Figure 2.7 shows three examples from Treadmill Dataset B. The three images are captured from the same subject with different clothes.

The Large Population Dataset is collected on 2009 in Japan [58]. It has over 4000 participants. The data set consists of persons walking on the ground surrounded by the 2 cameras at 30 fps, 640 by 480 pixels. The datasets are basically distributed in a form of silhouette sequences registered and size-normalized to 88 by 128 pixels size. Figure 2.8 shows three examples from Large Population Dataset. The participants contains 2135 males and 1872 females, aged from 1 to 94. The gait

[3] They have 32 different types of clothes altogether, from which 24 are selected for every subject.

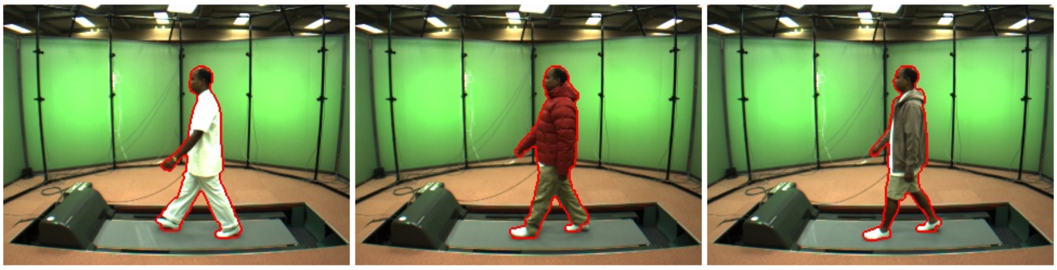


Figure 2.7: Samples of OU-ISIR Treadmill Dataset B [68]. This database focuses on clothes variation.

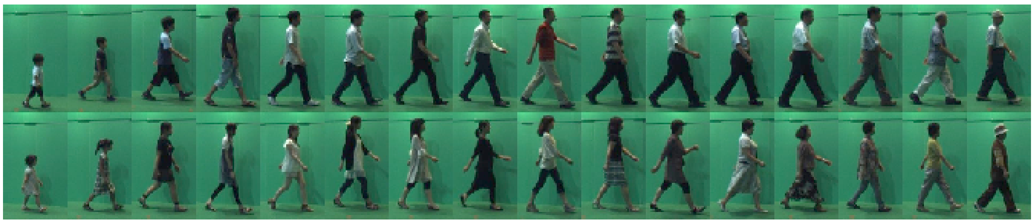


Figure 2.8: Samples of OU-ISIR Large Population Dataset [58]. This database has the largest number of subject (over 4000) of all the published gait databases.

sequences of each subject are recorded simultaneously from 4 observation angles using 4 cameras, namely 55°, 65°, 75° and 85°.

2.3 Related Works

This section reviews the state-of-the-art approaches related to the global deformation challenge, including recognition under arbitrary view and recognition with silhouette quality disparity, as mentioned in Section 1.3.

2.3.1 Gait Recognition Under Arbitrary View

Current view-invariant gait recognition algorithms consists of three categories: 1) methods based on constructing 3D human model, 2) methods based on view-invariant feature, and 3) methods based on unitary projection.

The first category is to build a 3D model to represent human body structures using multi-view gait sequences. As summarised by [69], it is more preferable to use 3D model reconstruction approaches when providing forensic evident in court, since the distinctive gait parameters can be intuitively spotted. This argument is also supported by [24]. A geometric concept named visual hull is proposed in an early work by Laurentini [70], using 2D silhouettes of an object to reconstruct a 3D approximation model. Later, this idea was introduced to an integrated face and gait model on multi-view recognition task, where a 3D visual hull based gait model is constructed using the silhouettes captured by 4 cameras around the subject, and the gallery and probe silhouettes are obtained by projecting the 3D model to the canonical lateral-view [71]. Seely et al. presented the University of Southampton Multi-Biometrics Tunnel, a specially designed facility for non-contact biometrics recognition task [72]. Following the same idea as [71], they construct a 3D volumetric gait model using the sequences captured by 8 cameras around the tunnel, thus the silhouettes from arbitrary view can be synthesised and fed into a standard 2D gait analysis system. Benefited from the development of appearance-based gait representation, they use gait templates such as GEI to improve the recognition efficiency. Ariyanto et al. improved this model by fitting the jointed 3D points cylinders to fit the visual hull shape, in order to accurately model the movement of thighs and shins during a gait period [73]. They argued that the analysis on the trajectory of leg joints helps improve the discriminatory capability of the silhouette based method. Kwolek et al. proposed to model key human components, e.g. head, spine, pelvis, and the four limbs, with 11 rigid segments [74]. They use calibrated and synchronise cameras to track the subject in video sequences and fit this 3D model with human gait, and measure the similarities between body component parameters to perform reconnection.

Ahmed et al. use kinect skeleton data to reconstruct human skeleton model [75], and use the joint distance variation and joint angle variation as features to measure the similarity between different models. Despite that most of the works mentioned above claimed that their approach achieved high recognition accuracy, above 96% or so, these 3D model construction based literatures has main draw-backs: their methods have not been tested on a collective database. Seely et al. [72] tested their method on a dataset containing 103 subjects, and all the sequences are recorded in the Multi-Biometrics Tunnel. Ariyanto et al. [73] use a dataset containing 46 subjects recorded in the Multi-Biometrics Tunnel, but they did not mention whether these subjects have intersect with [72]. Kwolek et al. [74] tested their method on their own indoor dataset comprising 22 subjects. Ahmed et al. [75] used the kinect skeletal gait database provided by the SMART Technologie ULC Calgary, Canada, containing 20 subjects. Furthermore, all these datasets mentioned above are collected in highly constrained indoor environment. Their practicability for real world applications need to be testified. Sandau et al. adopt 8 high definition cameras to create an explicit 3D human model, while the static and dynamic parameters are measured and recorded manually by expert observers [69]. The authors reported that when manually measuring the gait parameters on the 3D model, the joint centre annotation varies between different observers, which causes inter-observer classification variability. Figure 2.9 illustrated the procedures to establish the 3D model for a subject in [69], where (a) is the subject with leopard spandex dress[4], (b) the generated 3D model and surface smoothing, (c) the reconstructed 3D human model and (d) its rotation screen shot. Despite that the 3D model is very explicit and accurate, observers provided very different annotations and measurements on gait

[4] According to their paper, it is used for enhancing the body curvatures and textures.

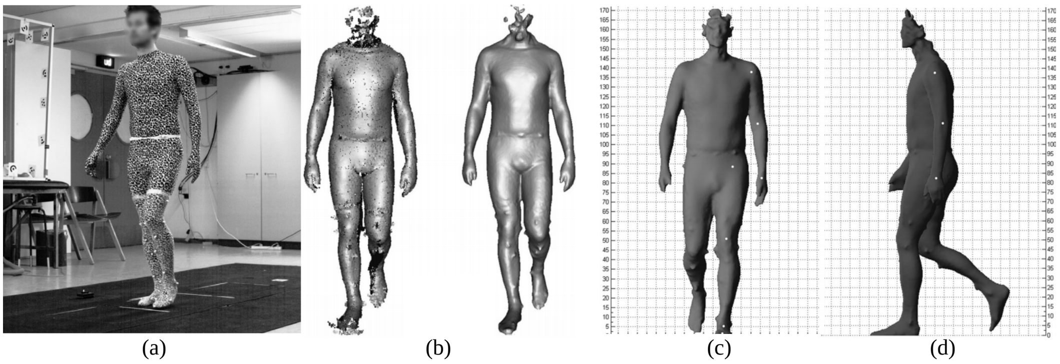


Figure 2.9: An example of 3D human model construction by Sandau et al. [69]. (a) the subject with leopard spandex dress (b) the generated 3D model and surface smoothing (c) the reconstructed 3D human model (d) the rotation screen shot.

parameters. Without a standard protocol and a number of published databases, it is very difficult to evaluate the effectiveness of the proposed approaches. In addition, the expensive 3D acquisition equipments and the high computational cost of building 3D models restricted the 3D model-based approaches from extensive use [69] [59].

In the second category, researchers seek for view-invariant features from single-view gait silhouette sequences, and perform recognition under lateral view. Kusakunniran et al. [76] and Goffredo et al. [77] provided view-invariant gait features for cross-view recognition. In [76], the authors proposed Gait Texture Image (GTI) and applied Transform Invariant Low-rank Textures to obtain common canonical view (side view) gait features from other view angles. However, their method is limited in view-invariant gait recognition, and it is difficult to transfer from front or back view to the side view. [77] proposed model based view-invariant gait feature, which use lower limbs' poses estimation to perform viewpoint rectification. It is also limited to view-invariant gait recognition, plus it is restrained from other factors of model-based methods.

Since gait sequences are normally acquired from a distance with the natural occlusion of body components and low resolution, it is difficult to extract model-based parameters (height, length of limbs, joint angle, etc.) from captured gait sequences. Therefore most researchers adopt appearance-based features, which refers to the whole binary silhouettes of the subject. Because there is no view-invariant features on 2D gait silhouette, cross-view gait matching is normally performed by means of subspace learning [78–81] or view transformation model techniques [82–84]. Subspace learning based methods are proved to be efficient on tackling gait challenges. Wu et al. proved that Local Fisher Discriminant Analysis (LFDA) is efficient in cross-view gait recognition, due to its discriminative feature learning ability in local geometry subspace, and proposed an iterative learning approach to optimise the construction of local affinity matrix, thus further promoted the LFDA based learning approaches [79]. Huang et al. suggests that the extension of Locality Preserving Projection (LPP) performs well for Cross-Speed Gait Recognition [85]. Han et al. applied linear discriminant analysis (LDA) for multi-class discriminant feature learning in his famous GEI paper [19]. Lu et al. proposed a framework that combined LDA with multi-linear tensor principle component analysis (PCA) [86], and later Lu et al. used boosting regularised LDA along with multi-linear PCA to further enhance the performance [87]. The boosting LDA randomly selects a subset from the original feature set each time, and this procedure is repeated thousands of times to produce weak classifiers. A recent work by Fan et al. applied LDA to the discrete cosine transform of gait templates [88]. Most subspace learning methods perform linear computations on training dataset to obtain an unitary subspace projection matrix. By combining the function of discriminant learning and dimensionality reduction, their advantages lies in the low computational cost and high efficiency. Makihara et al. proposed a method to deal with camera view point changes, named view transformation model (VTM) [89]. There are also several literatures [82–84]. In [90], CCA is implemented for fusion between two types of features and in [91] CCA is proved to be efficient in tackling clothes and carry conditions for multi view gait recognition.

The breakthrough work by Krizhevsky et al. [92] is regarded as a great success of CNN on large scale image classification, and since then CNN is widely known as a powerful tool on hard computer vision tasks. Firstly CNN is used for image classification, object localisation and detection. A multi-scale deep CNN proposed by Sermanet et al. largely outperforms other approaches on object localisation and detection tasks [93]. Farabet et al. applied a two layer CNN for scene labelling task, which also achieved record beating performance [94]. In biometrics identification field, Taigman et al. proposed a 9 layer deep CNN based framework to perform face recognition, and achieves human-level matching accuracy. Sun et al. obtains comparable results using a framework integrated with a 6 layer CNN and the Joint-Bayesian technique [95]. Recently CNN has been introduced to tackle gait recognition challenges. Alotaibi et al. applied a full convolutional network with 4 convolutional layers and a softmax layer for simple gait recognition tasks, i.e. matching gallery and probe under same conditions [96]. Yan et al. used a 5 layer network with 3 convolutional layers and 2 full connected layers for cross-condition gait recognition. They also introduced multi-task learning approach, which perform gait recognition, view prediction and scene prediction simultaneously. According to their report, the multi-task learning accelerated the convergence of CNN in training process. However, the cross-condition performance of their network has no significant improvement comparing with the traditional approaches, for example PCA + LDA. Shiraga et al. proposed a 4 layer network consisting of 2 convolutional layers and 2 full connected layers, and used it for large-scale gait recognition on OU-ISIR Large Population Dataset [97]. Their network has great advantage over other approaches on the large-scale dataset, while the view variation is small (at most 30°). They also addressed that CNN based method can significantly reduce the equal error rates (EER) and thus improve the gait verification accuracy. A very thorough study on CNN based gait recognition is provided by Wu et al., where they extensively evaluated the effect of training procedure and network architecture on the performance of CNN [2]. Instead of training the network with image-label pair, as is done by most literatures mentioned above, they train their network with pair-image and binary label. Specifically, they exhaustively pair all the GEIs in the dataset, if the GEI pair belongs to the same subject, the corresponding label is set to 1, while GEIs from different subject is labelled to 0. This pair-image training process is similar to the work by Sun et al. [95] for face verification, which simulates the process of linear unitary subspace projection. The experiment results proved that the pair-image training network outperforms other approaches by a large margin. The feature maps learned from the CNN has strong discriminant power that it has great robustness on cross-condition gait recognition. However, as is seen from the details, the cross-view recognition on large view variation (54°and above) is still not idea.

2.3.2 Gait Recognition with Silhouette Quality Disparity

Like in other biometric modalities, source information acquisition are the fundamental procedure of gait recognition system. As mentioned in Section 2.2.1, many factors can introduce noisy regions on the numerical representation of biometric traits [98], thereby affect the accuracy in distance measurement. Thus the quality of gait signature has direct impact on the recognition accuracy [99]. The recognition algorithms based on human physical models with spatial and temporal parameters, i.e. model-based approaches, are not yet very suitable for real world applications, due to the high requirement for video resolution and computational power. In another word, model-based approaches are more easily affected by the degradation on gait signature. On the contrary, appearance-based gait recognition shows higher tolerance to noises. The canonical representation of gait features is the binary silhouette extracted from a static background [18] [100] and normalised to a specific size (128×88, according to the baseline framework proposed by [18]). Several efficient templates have been proposed to encrypt spatial-temporal information of a gait cycle into one image, among which GEI [19] is regarded as one of the most easily implemented yet highly efficient templates [59].

The exhibit of factors, such as cloth and carry condition, walk speed, view angle, walking surface, and time elapse (ageing), caused the fundamental changes of human gait, while another question arises as to how would the low-level feature representation affect the recognition algorithms. Hence silhouette quality as a new factor has gained increasing attention in recent years [59]. The quality of gait silhouettes can be influenced, for example, by the background environment when capturing gait sequences and the accuracy of the segmentation method used to detect the gait silhouette. As is mentioned in [101], Chellappa et al. emphasised the importance of time alignment in gait template, since gait features are normally modelled from a temporal sequence. The traditional gait template, for example GEI, consider two strides as one gait cycle, and using time series signal of lower half silhouette to detect a whole cycle. Specially, when the subject stands still, the lower half silhouette contains less pixel than taking a stride, and the variances of pixel numbers (or entropy spectrum) can generate a sinusoid time series signal. Automatic systems may not be able to detect an accurate gait cycle from a set of low quality sequences since the noise generated by strong shadow or illumination variations may cause severe distortion to silhouette, thus cannot perform the sinusoid-based analysis.

Due to the fact that the acquired distinctive features are easily contaminated, gait has not been considered as a dependable biometric trait in practical human recognition, and accordingly gait recognition systems are rarely deployed in law enforcement departments or commercial organizations. Therefore it is suggested that the main task for gait recognition is to explore the limitations of traditional gait templates as well as the recognition frameworks. Sarkar et al. proposed the gait baseline algorithm, and discussed silhouette segmentation errors in the HumanID Gait Challenge Problem database caused by shadows under participants, varying lighting conditions and moving objects in the background [18]. It is noticed that the appearance-based gait recognition baseline algorithm is resistant to minor segmentation errors. Liu et al. carried out an in-depth study of the factors that had negative effect on the baseline algorithm, and indicated that the quality degradation caused by segmentation errors that correlated with the certain background, e.g. illumination and shadow, could contribute to higher recognition accuracy when gallery and probe are recorded under same conditions [102]. Thus they argued that it is ineffective to acquire better silhouette segmentation results; instead, it would be a better choice to seek for condition-invariant features from different components of human silhouette. Zhang et al. [103] investigated the low resolution gait recognition problem, and proposed a framework combining super resolution and multi-linear tensor manifold learning to perform high resolution back projection. Their approach suggested that subspace learning approaches is resistant to scale variation of data samples. Guan et al. [104] further addressed the issue of poor recognition accuracy when the resolution of gait silhouettes are very low, and proposed to use enhanced RSM method to reduce the over-fitting and improve recognition accuracy. AKae et al. [105] and Guan et al. [106] proposed similar solutions to gait recognition with extreme low frame-rate sequences respectively. Matovski et al [99] introduced quality metrics to improve the quality of silhouette, and by testing their method on a new indoor database, it is proved that the improvement of silhouette quality generates better recognition performance. Still, the field of quality-invariant gait recognition, which is compulsory for practical application of human gait recognition, remains unexplored.

Most literatures focus on improving silhouette quality, while few of them consider it as a factor and study the case when silhouette quality disparity exists. Thus the quality disparity challenge remains unsolved, which prevent gait recognition from being deployed in real world applications. Above all, the quality of gait signatures is vital, and it is easily contaminated when obtained from the source video sequences. Recognition accuracy may be hindered if the associated gait gallery and probe silhouettes are acquired under different conditions.

2.4 Fundamental Knowledge for This Thesis

In this section we formulated the methods that have been used in our proposed framework. Subspace learning methods are well known for its efficiency in dimensionality reduction and discriminant feature learning ability. When learning features from images, CNN shows its overwhelming power, benefiting from its high non-linearity feature learning ability. In this thesis, we adopt GEI as gait template.

A GEI $G(x, y)$ is defined as:

$$G(x, y) = \frac{1}{N} \sum_{k=1}^{N} I_k(x, y), \tag{2.1}$$

where $I_k(x, y)$ is the kth binary image, and (x, y) denotes the pixel coordinates. Consider n GEI samples that are stored as d-dimensional column vectors in a matrix $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, $\mathbf{x}_i \in \Re^d$, $i \in \{1, 2, ..., n\}$. Let W be the transformation matrix that projects the original space onto an r-dimensional subspace, where $d \gg r$. The new feature matrix in the subspace is denoted as $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$, where $\mathbf{y}_i \in \Re^r$. The transformation matrix for each element is given by $\mathbf{y}_i = W^T \mathbf{x}_i, i \in \{1, ..., n\}$. Matrix W varies according to the subspace learning method used.

2.4.1 Subspace Learning

As is mentioned in Chapter 1, a common problem occurs when evaluating biometrics identification system - the high dimensionality of exacted features. The linear subspace learning methods are proved to be fast and efficient in dimensionality reduction and reducing information redundancy.

PCA

As one of the most used dimensionality reduction techniques, PCA is widely applied as a preprocessing tool to reduce the dimensionality and solve the singularity problem of feature matrices [107]. PCA seeks a compact representation of patterns in a feature subspace. The columns of the PCA transformation matrix W_{PCA} are calculated by solving the eigen-decomposition problem $\lambda_i \mathbf{e_i} = S \mathbf{e_i}$, where λ_i and $\mathbf{e_i}$ are the corresponding eigenvalues and eigenvectors, respectively, and $S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^\top$ is the covariance matrix of the original sample matrix X, where μ is the sample mean, $\mu = \frac{1}{n}\sum x_i, i \in \{1, 2, ..., n\}$. Matrix W_{PCA} is then composed by column eigenvectors corresponding the rth highest eigenvalues; $W_{PCA} = \{\mathbf{e_1}', \mathbf{e_2}', ..., \mathbf{e_r}'\}$, where $\mathbf{e_j}', j \in [1, r]$ is the jth eigenvector.

LDA

Compared to PCA, LDA embeds discriminant power between different classes in the feature subspace, which makes it a supervised subspace learning method suitable for multi-class learning problems. Assuming there are c classes in X, with n_l samples in subset $\mathcal{X}_l$, $l \in \{1, 2, ..., c\}$, so that $n = \sum_{l=1}^{c} n_l$; the within-class scatter matrix S_W is then defined as: $S_W = \sum_{l=1}^{c}\sum_{x \in \mathcal{X}_l}(x - \mu_l)(x - \mu_l)^\top$, and the between-class scatter matrix S_B is defined as: $S_B = \sum_{l=1}^{c}(\mu_l - \mu)(\mu_l - \mu)^\top$, where μ_l is the mean of the samples in class l, and μ is the mean of all samples. In order to maximize between-class scatter while minimizing within-class scatter after projection, the following criterion is used:

$$W_{LDA} = \arg\max_{W} \frac{|W^\top S_B W|}{|W^\top S_W W|}, \tag{2.2}$$

where W_{LDA} is the transformation matrix, whose columns are the generalised eigenvectors $\{\mathbf{e_1}', \mathbf{e_2}', ..., \mathbf{e_r}'\}$ that correspond to the largest eigenvalues in $S_B W_{LDA} = \lambda_i S_W W_{LDA}$.

LPP

LPP tends to preserve the local data structure after projecting the data onto a subspace [108]. It first constructs an adjacency graph G to model the local structure of the samples. The adjacency graph has n nodes, with node i corresponding to x_i in X. A pair of nodes i and j are connected if x_i and x_j are close in the space. The elements of the weighted similarity matrix A, which specifies the similarities among nodes in G, are formulated as follows:

$$A_{ij} = \begin{cases} exp\left(\frac{-\|x_i-x_j\|^2}{t}\right), & \text{if nodes } i \text{ and } j \text{ are connected,} \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

The heat kernel parameter $t \in \mathfrak{R}$ can be determined empirically; if t is very large, $exp\left(-\|x_i - x_j\|^2/t\right) = 1$ and matrix A comprises binary weights. Two possible ways exist to determine if nodes are *close*:

1. K nearest neighbours: if x_i is among the K nearest neighbours of x_j, or vice versa;

2. ϵ-nearest neighbours: if $\|x_i - x_j\|^2 < \epsilon, \epsilon \in \mathfrak{R}$.

The eigen-decomposition problem of LPP is generalized as:

$$XLX^\top W_{LPP} = \lambda XDX^\top W_{LPP}, \tag{2.4}$$

where D is a diagonal matrix with $D_{ii} = \sum_j A_{ij}$, and L is the Laplacian matrix $L = D-A$. The Laplacian of the graph is an approximation of the Laplace-Beltrami operator. The transformation matrix $W_{LPP} = \{\mathbf{e_1}', \mathbf{e_2}', ..., \mathbf{e_r}'\}$, and $\{\lambda_1', \lambda_2', ..., \lambda_r'\}$ are the corresponding r smallest eigenvalues. The feature subspace created by W_{LPP} can preserve an intrinsic geometric structure of the manifold samples [109] [110]. LPP performs supervised learning ability by assigning a weight equal to 0 to all between-class similarity matrix values. The total similarity matrix A is then given as follows:

$$A = \begin{bmatrix} A_1 & \cdots & 0 \\ \vdots & \ddots & 0 \\ 0 & 0 & A_c \end{bmatrix} \tag{2.5}$$

In our experiment, supervised LPP (SLPP) is implemented, for the unsupervised learning method (for example PCA) cannot improve the recognition performance.

Local Fisher Discriminant Analysis

Sugiyama [111] propose a novel subspace learning method, called Local Fisher Discriminant Analysis (LFDA), which embeds within-class similarity matrices into local within-class scatter matrices and local between-class scatter matrices, denoted as $\tilde{S}^{(w)}$ and $\tilde{S}^{(b)}$, respectively. These matrices are formulated as follows:

$$\begin{aligned} \tilde{S}^{(w)} &= \frac{1}{2}\sum_{i,j=1}^{n} \tilde{W}_{i,j}^{(w)}(x_i - x_j)(x_i - x_j)^\top, \\ \tilde{S}^{(b)} &= \frac{1}{2}\sum_{i,j=1}^{n} \tilde{W}_{i,j}^{(b)}(x_i - x_j)(x_i - x_j)^\top, \end{aligned} \tag{2.6}$$

with

$$\begin{aligned} \tilde{W}_{i,j}^{(w)} &= \begin{cases} 1/n_\ell & if y_i = y_j = \ell, \\ 0 & if y_i \neq y_j, \end{cases} \\ \tilde{W}_{i,j}^{(b)} &= \begin{cases} 1/n - 1/n_\ell & if y_i = y_j = \ell, \\ 1/n & if y_i \neq y_j. \end{cases} \end{aligned} \tag{2.7}$$

where n_ℓ is the number of samples in class ℓ, with $\sum_{\ell=1}^{c} n_\ell = n$. The transformation matrix of LFDA is then defined as:

$$W_{LFDA} = \underset{W \in \Re^{d \times r}}{\arg\max} \left[tr(\frac{W^\top \tilde{S}^{(w)} W}{W^\top \tilde{S}^{(b)} W}) \right]. \tag{2.8}$$

LDFA searches for the transformation matrix W_{LFDA} that separates data from different classes while clustering data from the same class as close as possible. The solution follows a similar approach as the one followed in LDA.

2D PCA

Yang et al. [112] propose the 2D extension of PCA. Consider the training set $\{\mathbf{I}_i|i = 1, ..., n\}$, where $\mathbf{I}_i$ refers to a single sample (e.g. , a GEI) in 2D form (with size $d_r \times d_c$), and n refers to the total number of samples. The *image covariance matrix* $\mathbf{C}$ is then calculated by:

$$\mathbf{C} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{I}_i - \bar{\mathbf{I}})^\intercal(\mathbf{I}_i - \bar{\mathbf{I}}), \tag{2.9}$$

where $\bar{\mathbf{I}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{I}_i$ is the mean value of all training samples. By performing eigen-decomposition on $\mathbf{C}$, we can obtain the 2DPCA projection basis $V_{pca} = \{\mathbf{v}_i|i = 1, ..., p\}$, as the p orthonormal eigenvectors corresponding to the p largest eigenvalues. Compared with the canonical PCA, 2DPCA is much more computationally efficient. For example, for GEIs of size 128×88, the covariance matrix of vectorised samples using canonical PCA has a complexity $O(2^d), d = d_r \times d_c = 11264$; while the complexity of calculating the image covariance matrix, $\mathbf{C}$, is only $O(2^{d_r}), d_r = 128$.

Note that here we use $\mathbf{I}_i$ to represent sample i in 2D form. In the following, all samples are assumed to be vectorised into features vectors instead of being in 2D form. Therefore, we denote feature vector i by $\mathbf{x}_i$ or $\mathbf{y}_i$.

2.4.2 Convolutional Neural Network (CNN)

LeCun et al. offered a very detailed explanation on the effectiveness of convolutional layers in their work [113], which can be summarised as 1) efficiency in learning shift invariance features and 2) preserving the neighbourhood structure of features in each local region. Benefiting from the availability of massive labeled training data and powerful GPU computing implementation, CNN based approaches maintain a

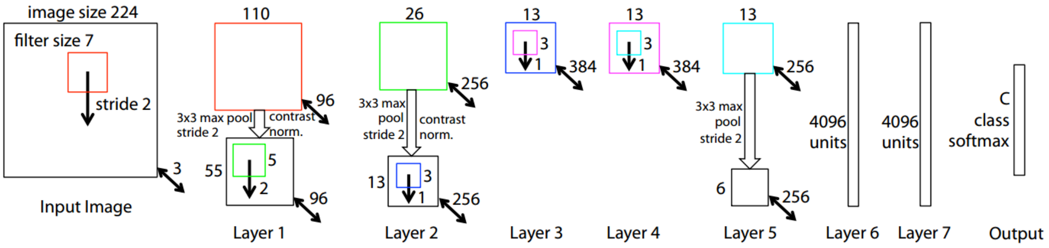


Figure 2.10: The hierarchical structure of AlexNet.

leading position in many areas, such as image classification and face recognition. A standard CNN approach builds a deep neural network model, and uses a training image set and the corresponding labels to train the intermediate layers. After completing the training, the model is able to map any input image x_i to a probability vector $\hat{y}_i$ over a number of classes. Figure 2.10 displays an example of the most popular CNN model - AlexNet. The input image size is 224×224. The input image is convolved with 7×7 filter using stride 2 in both vertical and horizontal direction. The total number of filters are 96. For layer 1, the generated feature maps are then went through ReLU function and pooled with 3×3 regions (also using stride 2). By contrasting normalised across feature maps, 96 different element feature maps of size 55×55 are obtained. The operations in the following layers (layer 2, 3, 4 and 5) are similar to layer 1. The output of the previous layer are taken as input for convolutional layers (for layer 1 the input is images), and processed with a set of filters learned from the training process. The resulting features from the 5 convolutional layers are passed to fully connected layer 6 and 7. The final softmax layer consists of C-way softmax function for classification.

A conventional CNN structure (such as AlexNet) starts with convolutional layers, followed by pooling and normalization layers, and ends with fully connected layers. Between every two layers there is an activation function. The basic activation functions are sigmoid function $f(x) = 1/(1+e^{-x})$ and $tanh(x)$, both can be distinguished with their upper and lower limit. Sigmoid ranges from [0,1] and tanh from [-1,1]. Rectified linear unit (ReLU) function is another kind of activation function that is favorable for efficiency of the network. It is defined as $f(x) = max(0, x)$.

Every convolutional layers evolved into next layer on the basis of filters, which are the basic element of convolution process they are in different sizes and different shapes. These filters move all over the data with different strides (distance between two steps) and produces feature map and with the help of pooling and normalization they try make sense of the data. Pooling is used to down-sample the data by preserving one activity of every reign of feature map. This information is either be average or maximum in the local region. The simple pooling operation in CNN can be formulized as:

$$L_{(i-1)} = pool(max(0, W_i * L_i + b_i)), \tag{2.10}$$

where * refers to the convolutional operation between data X_0 or (information X_i produced by i^{th} layer), i^{th} layer filters W_i and bias b_i. For every specific task CNN contains millions of parameters. It is built by the above mentioned layers but every CNN have different structure according to the task it going to perform. Filter size, number of layers, activation function, learning rate etc. every parameter is adjustable according to the take and requirement.

Another related work is CNN-based segmentation. As mentioned in 2.3.2, the silhouette quality has great impact on the recognition accuracy. In order to obtain high-quality human silhouette, a lot of segmentation methods are proposed, among which CNN-based one achieves the best performance in terms of accuracy. In [114], Wu et al., proposed a multi-scale segmentation method which segments images at pixel level, with high accuracy but very low speed. In [115], Song et al., proposed a fast image level CNN-based segmentation method which achieves 1000 frames per second with slight drop on accuracy compared with [114]. To achieve pixel-wise segmentation at image level, Fully Convolutional Networks (fullconvnets) is developed [116], achieving a good balance between accuracy and speed. We will further analysis these CNN-based human segmentation algorithms in the following sections.

2.5 Summary

In this section we review the development of gait related research, and argue that gait is capable of being regarded as a reliable biometric trait for human identification tasks, while the representation of human gait can be very abstract. We discuss the existing challenges as well as present the published datasets that specifically designed for evaluating corresponding approaches. We summarise the existing challenges into two categories: partial interference and holistic deformation challenge, according to the distribution of the corrupted features. Then we go through the previous works that related to our research, i.e. gait recognition under arbitrary view, and gait recognition with silhouette quality disparity, which all belongs to the gait holistic deformation challenge. At last we provide the formulation of the fundamental algorithms that are used in our research. In the next chapter we present the formulation of the proposed ViFS and its application in gait recognition frameworks. Experiments and analysis are provided to support the efficiency of the proposed method.

Chapter 3

Gait Recognition Under Arbitrary View

View-invariant gait recognition is one of the major challenges to identify people through their gait. Many researchers have evaluated view angle transformation techniques, discriminant analysis and manifold learning approaches for cross-view recognition, and their proposals are usually based on a common factor, i.e., to establish a cross-view mapping between gallery and probe templates. However, their effectiveness is restricted to small view angle variances. A promising approach to perform view-invariant gait recognition is through multi-view feature learning. In this chapter, we propose ViFS and integrate it in a framework for view-invariant gait recognition. ViFS technique select features from multi-view gait templates and reconstructs gallery templates that accurately match the data for a specific view angle. ViFS is thus able to reconstruct gallery templates from arbitrary view angles, and thus help to transfer the cross-view problem to identical-view gait recognition. We also apply linear subspace learning methods such as LDA and SLPP as feature enhancers for ViFS, which substantially reduces the computational cost while improving the recognition speed. We test the proposed framework on the CASIA Dataset B and OU-ISIR Large Population (OU-ISIR LP) Dataset. The average recognition accuracy of the proposed framework for 11 different views exceed 98%.

3.1 Problem Statement and Motivation

View-invariant gait recognition has recently gained increasing interest, and a number of efficient approaches have been proposed to tackle this challenge. As mentioned in Chapter 1, gait features are extracted from binary gait silhouettes, which comprise the shape of human profile from one direction, thus it is practically impossible to extract view-invariant features from single-view gait silhouette templates. In general, there are types of view-invariant gait recognition: *cross-view recognition* where only a single view angle is available in both, the gallery and probe sets (the view angles can be different), and *multi-view recognition*, where templates from multiple view angles are available in the gallery set[1]. For the case of cross-view gait recognition, the state-of-the-art is best represented by Wu et al.'s [2] work, which uses CNN based approaches. When multi-view gait templates are obtained, or depth information is available, it is possible to reconstruct 3D or 2.5D models representing the human body, from which arbitrary views of gait sequences can be obtained by projection, and parameters associated with body parts can be easily measured. Tang et al. [1] propose to construct parametric 3D gait models from three cameras and use partial similarity matching to improve recognition rates. Their method achieves promising results on several major gait datasets. Similarly, Luo et al. [117] propose to use 3D gait models and sparse representation-based classification

[1] In fact, this case can be reversed, i.e. multi-view templates available in the probe set, while gallery templates are from a single view angle.

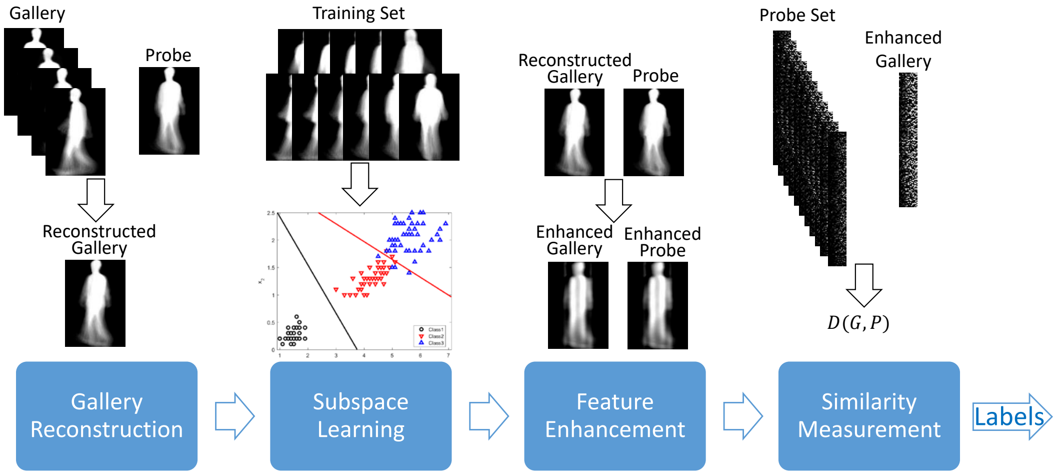


Figure 3.1: The pipeline of the proposed framework.

to perform view-invariant classification. As mentioned in Chapter 2, 3D model based approaches including [1] [117], require a specifically designed multi-view database for model construction and training purposes.

In this section, we propose a ViFS based framework to automate feature selection and perform fast view-invariant gait recognition. First, we gather gallery templates from different view angles into one matrix, and regard the probe template from a specific view angle as the reference. By minimising the distance between the gallery templates and the probe template, we realize the ViFS. We use subspace learning methods as feature enhancers to increase the discriminative power of gait features. LDA has been proven to be efficient in finding the most discriminant subspace for gait recognition [21]. SLPP is a manifold based mapping method for feature projection and dimensionality reduction [108]. It is designed to preserve the local geometric structure of feature sets in the projection subspace, and has been shown to outperform LDA [85]. Finally, we measure the Euclidean distance between enhanced gallery and probe features, and compare the returned labels with the ground-truth for accuracy.

3.2 View-invariant Feature Selector and the Gait Recognition Framework

The proposed framework uses GEIs as gait features. Figure 3.1 illustrates the flow-process diagram of the proposed framework. It consists of four stages:

- *Gallery reconstruction.* We use multi-view gallery set and the probe set to train the feature selector ViFS. By minimising the distance between reconstructed gallery and probe set, ViFS reduces the view variance between gallery and probe set.
- *Subspace learning.* We use the training set to generate the subspace projection matrix for feature enhancement. The generated unitary projection subspace can increase the discriminant ability of features, while reducing feature dimensionality.
- *Feature enhancement.* The enhanced gallery set and probe set are projected into the generated subspace. The generated gallery and probe features have better discriminative power, and thus increase the identification accuracy.
- *Similarity measurement.* Here we use Euclidean distance to calculate the distance between gallery and probe feature sets, and label the probe data according to distance scores.

The formulation of the proposed framework is presented below.

3.2.1 View-invariant Feature Selector (ViFS)

For the simplicity of formulation, assume we have h sample from h different view angles in gallery set $\mathcal{G} = \{\mathbf{x}_i\}_{i=1}^{h}$, and one probe samples $\mathbf{y}$ from a specific view in

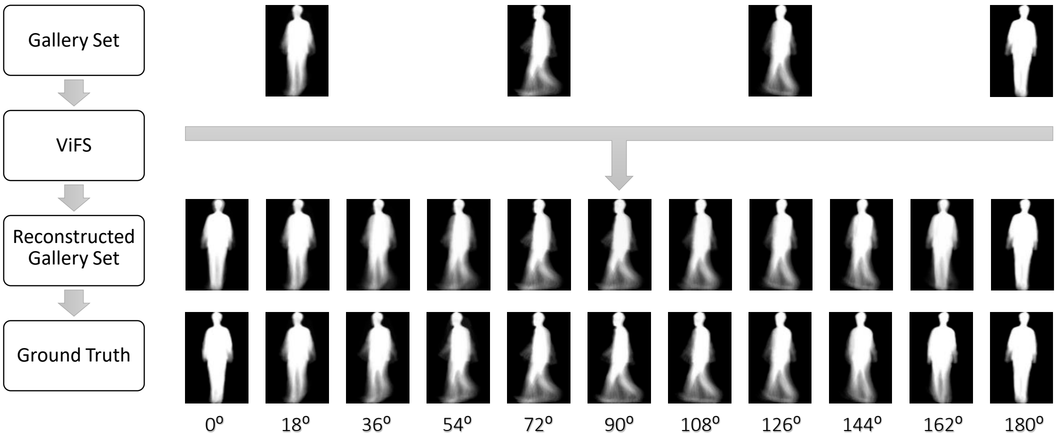


Figure 3.2: Illustration of the reconstructed gallery templates for missing views by ViFS. The ground truth shows the gallery templates from all views provided by the CASIA gait dataset B.

probe set $\mathcal{P}$. Due to the view difference between gallery and probe samples, the intra-class distance can be larger than the inter-class distance for the same subject, leading to misclassifications. In Figure 3.2, we give an intuitive view of the intra-class variation in the *cross-view* case. Take the GEIs from the first row of images as examples. The last two GEIs are from the same subject, with the left from 126°and the right from 180°. The variations mainly exist in the lower limbs with high brightness on the grayscale image. To reduce the negative effects of view angle differences on the classification results, one can minimise the cross-view distance between gallery and probe samples. Under the scenario that the view angles of the gallery and probe samples are unknown, one would like to find a feature vector $\mathbf{w} = \{w_i\}_{i=1}^{h}$ that minimises the objective function:

$$f(\mathbf{w}) = \|\mathcal{G}\mathbf{w}^{\intercal} - \mathbf{y}\|^2 = \|\sum_{i=1}^{h} w_i \mathbf{x}_i - \mathbf{y}\|^2. \tag{3.1}$$

The minimiser $\hat{\mathbf{w}}$ of $f(\mathbf{w})$ satisfies $\nabla f(\hat{\mathbf{w}}) = 0$, leading to $\nabla f(\hat{\mathbf{w}}) = 2\mathcal{G}^{\intercal}(\mathcal{G}\hat{\mathbf{w}}^{\intercal} - \mathbf{y})$ [118]. Then $\hat{\mathbf{w}}$ can be calculated by:

$$\hat{\mathbf{w}} = (\mathcal{G}^{\intercal}\mathcal{G})^{-1}\mathcal{G}^{\intercal}\mathbf{y}. \tag{3.2}$$

Since the gallery set $\mathcal{G}$ and its covariate matrix $\mathcal{G}^{\intercal}\mathcal{G}$ are practically impossible to be upper-triangular, we cannot solve Eq. (3.2) directly. Instead, we use QR-factorisation, such that $\mathcal{G} = QR$, to generate an orthogonal matrix Q and upper-triangular matrix R from $\mathcal{G}$. Thus Eq. (3.2) can be formulated as:

$$\begin{aligned} \hat{\mathbf{w}} &= (\mathcal{G}^{\intercal}\mathcal{G})^{-1}\mathcal{G}^{\intercal}\mathbf{y} \\ &= ((QR)^{\intercal}(QR))^{-1}(QR)^{\intercal}\mathbf{y} \\ &= R^{-1}Q^{\intercal}\mathbf{y}. \end{aligned} \tag{3.3}$$

We can obtain $\hat{\mathbf{w}}$ by solving $R\hat{\mathbf{w}} = Q^{\intercal}\mathbf{y}$ with back substitution. We call minimiser $\hat{\mathbf{w}}$ the ViFS, as it selects features from the multi-view gallery samples to reconstruct an optimal template $\hat{\mathcal{G}} = \mathcal{G}\hat{\mathbf{w}}^{\intercal}$ that accurately matches probe samples $\mathbf{y}$. The way of training ViFS can be very flexible, i.e. it can be trained not only from gallery and probe samples; in special cases it can also be obtained from an additional sample set with different subjects (like samples in training set). In Figure 3.2 we presents a set of examples to demonstrate the effectiveness of ViFS in feature reconstruction. We take four samples of the same subject from gallery set, each from a certain view (18°, 72°, 126°, and 180° respectively), denoted as $\mathcal{G}$. We train ViFS with probe samples from 11 angles. Specifically, for the reconstruction of 0° gallery sample, we use the four gallery samples and one probe sample to generate ViFS for 0°, denoted as $\hat{\mathbf{w}}_0$, and obtain the reconstructed template $\hat{\mathcal{G}}_0 = \mathcal{G}\hat{\mathbf{w}}_0$. The reconstructed gallery samples on the third row of Figure 3.2 is intuitively comparable to the ground truth samples on the fourth row of Figure 3.2, suggesting that ViFS achieves view-transformation (to a closer view with probe sample) on gallery samples. The differentiation between reconstructed gallery sample (from the 4 view samples mentioned above to 144°) and probe sample (144°) is also shown in Figure 3.1, on the second row of the rightmost region, denoted as *Reconstructed Features*. It is clear that the reconstructed gallery sample has less variation with the probe sample.

3.2.2 Feature Enhancement

ViFS is designed to minimise the cross-view variance between gallery and probe features, and in multi-view gait recognition problem, this equals to reducing the intra-class variance. In order to further enhance the extracted features, we apply subspace learning methods to increase the inter-class variance. Since subspace learning methods are designed to project the input features into another space with lower dimensionality, the redundant information are removed and the discriminant features are preserved. Furthermore, since they are linear transformation, the computational cost and time consumption are both very low. Although classic discriminant analysis methods fit this purpose, manifold learning methods have drawn attention from computer-vision researchers in recent years. We thus employ two representative algorithms, namely linear discriminant analysis (LDA) and supervised locality preserving projection (SLPP), in this work. The performance comparison of the feature enhancers on the proposed method are displayed in Section 3.3.1.

Before applying LDA or SLPP, a common pre-processing step is to reduce the dimensionality of the original dataset and make sure the data matrices are non-singular. To this end, we use 2D PCA as the first step of subspace learning in order to reduce the computational cost of the proposed framework. We denote the matrix containing all training samples as $\mathcal{T} = \{\mathcal{T}_i\}_{i=1}^{n_{\mathcal{T}}}$, a 3D matrix with size $d_r \times d_c \times n$ and $n_{\mathcal{T}}$ samples. Following the formulations in Section 2.4.1, the eigenvectors $V_{pca} = \{\mathbf{v}_i\}_{i=1}^{n_{\mathcal{T}}}$, as well as the corresponding eigenvalues $\lambda_{pca} = \{\lambda_i\}_{i=1}^{n_{\mathcal{T}}}$ are obtained by eigen-decomposition of the image covariance matrix. We select the first p eigenvectors following according to $\frac{\sum_{i=1}^{p}\lambda_i}{\sum_{i=1}^{n_{\mathcal{T}}}\lambda_i} > 99\%$. Thus we obtain $V_{pca} = \{\mathbf{v}_i\}_{i=1}^{p}$, a $d_c \times p$ subspace projection matrix. The subspace projection is then $\mathfrak{T}_i = \mathcal{T}_i V_{pca}$, which results in matrix $\mathfrak{T} = \{\mathfrak{T}_i\}_{i=1}^{p}$. We reshape the 3D matrix $\mathfrak{T}$ to 2D form with dimensions $d_{pca} = d_r \times p$ and n samples. We then use $\mathfrak{T}$ and the corresponding class labels to train the LDA and SLPP project matrice V_{lda} and V_{lpp}, respectively.

Assuming there are h views in gallery sample set $\mathcal{G}$, and $n_{\mathcal{G}}$ samples in total. After obtaining the ViFS projection basis, $\hat{\mathbf{w}}$, and the subspace projection matrices, V_{pca}, V_{lda} and V_{lpp}, from the training procedure, the gallery set after feature extraction is obtained by $\hat{\mathcal{G}} = \mathcal{G}\hat{\mathbf{w}}^{\intercal}$. We project $\hat{\mathcal{G}}$ onto the subspace matrices to obtain an enhanced gallery feature set, as follows:

$$\begin{aligned}
&\mathfrak{G}_{pca} = \{\hat{\mathcal{G}}_i V_{pca}\}_{i=1}^{n_{\mathcal{G}}}, \quad \textit{reshape } \mathfrak{G}_{pca} \textit{ to } 2D,\\
&\mathfrak{G}_{lda} = V_{lda}^{\intercal}\mathfrak{G}_{pca},\\
&\mathfrak{G}_{lpp} = V_{lpp}^{\intercal}\mathfrak{G}_{pca},
\end{aligned} \tag{3.4}$$

Following the same procedure, we also obtain the enhanced probe sets $\mathfrak{P}_{lda}$ and $\mathfrak{P}_{lpp}$. For simplicity, we use $\mathfrak{G}$ and $\mathfrak{P}$ to represent the enhanced feature sets for gallery and probe, respectively, in the formulation of the similarity measurement.

3.2.3 Similarity Measurement

We use the most common metric learning method, i.e. Euclidean distance, to obtain matching scores between gallery and probe. The Euclidean distance between gallery feature set $\mathfrak{G}$ and probe feature set $\mathfrak{P}$ is calculated as:

$$D(\mathfrak{G}_i, \mathfrak{P}_l) = \|\mathfrak{G}_i - \mathfrak{P}_l\|, \quad i = 1, ..., c. \tag{3.5}$$

If $D(\mathfrak{G}_k, \mathfrak{P}_l) = \min_{i=1}^{c} D(\mathfrak{G}_i, \mathfrak{P}_l)$, the probe feature vector is assigned to the same class label k of the gallery feature.

3.3 Experiments and Analysis

3.3.1 CASIA Dataset B

We validate the effectiveness of ViFS using the two feature enhancers: LDA and SLPP. We use the CASIA Dataset B [67], which is a multi-view gait dataset containing 124 subjects in total. The size of each silhouette image is normalised to 128 × 88; one video sequence produces a single GEI. In this thesis, we focus on studying the performance of our framework across different view angles. Thus we only choose *normal* sequences from all subjects for evaluation, i.e. those sequences that are not affected by changes in clothing or carrying objects. The sequences of the first 74 subjects are used for training, and the other 50 subjects are used for testing. In the testing set, each subject has six sequences; the first four sequences are regarded as gallery sequences, and the remaining two sequences as probe sequences. Because ViFS is a multi-view based feature extractor, we require gait sequences in the gallery set to have been captured from more than one view angle in order to train the minimiser $\hat{\mathbf{w}}$. We compare our results with other works that have the same or similar multi-view settings.

3.3.2 Problem Analysis

Table 3.1: The Cross-view matching result with raw GEIs (%), without feature selection or enhancement. G: Gallery Data; P: Probe Data.

P G	0	18	36	54	72	90	108	126	144	162	180
0	98.4	30	13.3	3.3	3.3	5	5	6.7	10	18.3	50
18	28.3	99.2	36.7	1.7	6.7	1.7	6.7	5	11.7	43.3	18.3
36	10	20	97.6	28.3	20	13.3	16.7	16.7	41.7	16.7	11.7
54	11.7	8.3	28.3	97.6	25	35	41.7	33.3	21.7	8.3	6.7
72	6.7	5	10	31.7	97.6	76.7	61.7	61.7	18.3	6.7	5
90	6.7	6.7	13.3	26.7	80	96.7	93.3	25	16.7	5	5
108	3.3	5	13.3	45	65	95	97.6	53.3	20	6.7	5
126	5	5	18.3	35	43.3	33.3	41.7	97.6	41.7	8.3	6.7
144	3.3	15	38.3	20	15	11.7	23.3	33.3	97.6	20	10
162	18.3	36.7	13.3	1.7	3.3	3.3	3.3	5	23.3	98.4	36.7
180	40	20	10	1.7	3.3	3.3	3.3	5	6.7	38.3	100

Table 3.1 shows the cross-view match between gallery and probe data of 11 different view angles (0° to 180°). The diagonal data refers to the matching result under identical view, which should be the maximum value on each columns. For raw GEI templates, the features have very limited discriminant power cross-view. Even the smallest view variance, i.e. 18° between gallery and probe data could cause severe misclassification. In fact, most of the cross-view match has very low matching rate, except the result between 90° and 108°, which achieves similar score as the identical view, as is marked in gray in the table. It suggests that pixel values of gait templates captured under 90° and 108° are highly correlated, and the inter-class variation is much larger than the intra-class variation caused by view angle rotation. In another word, the templates under lateral view, or close to lateral view (72° to 108°), has better tolerance to small view variance than templates under frontal or

back view.

3.3.3 Cross-view Evaluation

We first compare the effect of feature enhancers on raw GEI templates using LDA and SLPP without using ViFS. The gallery and probe GEIs are projected into the generated subspaces learned from the training dataset, and then the Euclidean distance is calculated between gallery and probe data to find the closest match. We tested the performance on cross-view cases first, therefore the accuracy scores are obtained by single-gallery single-probe match. The gallery view angles are from 0° to 180°, and the tested view angles for the probe data are 0°and 90°. As seen in Figure 3.3, SLPP attains slightly better performance than LDA when the difference between the gallery and probe view angles is large, which can be explained by the ability of SLPP to keep the local geometric structure of feature sets. Recall that we formulate both LDA and LPP in Section 2.4, where the objective function of LDA is to minimise the distance between samples and class centroid, while maximising the distance between class centroids. It might result in a case that the class centroids are well split, but the samples from different classes are overlapped. However, LPP formulates its objective function by calculating the distance between samples within one class (intra-class variation) or from different classes (inter-class variation), which ensures that the class boundary is well separated, thus it is less likely to result in misclassification. Based on this observation, we can assert that when only a few views are available in the gallery set, SLPP is expected to better enhance the features extracted by ViFS than LDA, especially if there is a large difference between the view angles of probe and gallery data.

Table 3.2 and 3.3 shows the result of cross-view matching using LDA and

Table 3.2: The Cross-view matching result with LDA (%). G: Gallery; P: Probe

P G	0	18	36	54	72	90	108	126	144	162	180
0	99.2	90	63.3	25	16.7	13.3	11.7	13.3	28.3	66.7	70
18	76.7	100	100	53.3	43.3	26.7	38.3	48.3	61.7	73.3	53.3
36	43.3	96.7	99.2	98.3	91.7	66.7	80	88.3	80	58.3	31.7
54	30	60	96.7	99.2	100	95	91.7	88.3	85	43.3	16.7
72	21.7	61.7	100	98.3	100	98.3	96.7	91.7	71.7	28.3	13.3
90	15	43.3	76.7	95	98.3	98.4	98.3	95	73.3	21.7	11.7
108	10	43.3	83.3	93.3	96.7	96.7	99.2	98.3	95	36.7	8.3
126	11.7	46.7	86.7	86.7	91.7	90	96.7	99.2	100	65	11.7
144	18.3	53.3	83.3	78.3	70	55	85	100	100	80	31.7
162	43.3	63.3	65	38.3	38.3	21.7	36.7	60	90	99.2	80
180	63.3	65	40	26.7	13.3	16.7	20	33.3	45	83.3	100

Table 3.3: The Cross-view matching result with LPP (%). G: Gallery; P: Probe

P G	0	18	36	54	72	90	108	126	144	162	180
0	99.2	95	58.3	33.3	23.3	25	20	23.3	30	68.3	75
18	85	100	93.3	71.7	48.3	33.3	36.7	41.7	60	73.3	63.3
36	56.7	98.3	99.2	93.3	91.7	76.7	76.7	83.3	80	65	43.3
54	33.3	85	98.3	99.2	100	98.3	93.3	90	90	53.3	28.3
72	28.3	73.3	95	98.3	99.2	98.3	98.3	96.7	78.3	43.3	16.7
90	25	61.7	86.7	95	95	99.2	98.3	95	85	41.7	13.3
108	18.3	51.7	90	93.3	91.7	96.7	99.2	98.3	93.3	53.3	18.3
126	21.7	45	86.7	91.7	91.7	93.3	98.3	98.4	96.7	70	23.3
144	33.3	66.7	86.7	80	71.7	66.7	86.7	98.3	98.4	90	41.7
162	51.7	78.3	65	45	35	36.7	43.3	68.3	88.3	98.4	78.3
180	80	70	46.7	33.3	26.7	26.7	33.3	30	51.7	88.3	99.2

SLPP respectively. The gallery and probe GEIs are projected into the generated subspace learned from the training dataset, and then calculate the Euclidean distance to find the closest matching. SLPP shows better performance on large angle variance matching than LDA (over 36°), which benefits from the manifold learning ability that keeps neighbourhood data structure. This conclusion is drawn from the comparison between the two tables on the grey region (and a similar pattern is shown on the symmetric regions on the lower-triangle area). On the contrary, LDA has greater performance on matching results across small view difference, which is mainly on the diagonals of the tables. Same observation comes from Figure 3.3, where SLPP shows greater power in matching across large view difference.

3.3.4 Multi-view Evaluation

Table 3.4 tabulates the matching rates, in percentage, for ViFS + LDA(ViFS1) and ViFS + SLPP(ViFS2) for all 11 view angles. We compare the performance of ViFS1 and ViFS2 with maximum performance attained by LDA or SLPP (i.e., same view angle matching, denoted by Max in the table), respectively, and their average matching performance across view angles (denoted by Avg in the table). Without ViFS, subspace learning methods achieve significantly low result on the frontal (0°) and back views (180°) - see results marked with grey. This is mainly due to the large intra-class variance caused by these two special views. However, with the power of ViFS, the enhanced features from ViFS1 and ViFS2 show great robustness to view-angle changes, achieving the highest accuracy (close or even higher than the maximum accuracy) across the majority of view angles. The relatively low performance of ViFS1 and ViFS2 on the 180°view angle can be explained by the fact that the discriminative boundary for 0° and 180° view angles is hard to learn, and features from other view angles cannot provide information that is correlated to the 180°view angle. ViFS2 has slightly better performance than ViFS1, due to a

Table 3.4: Matching rates (%) for ViFS1 and ViFS2 for different view angles.

Feature enhancer	Method	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Average
LDA	Max	99.2	100	100	99.2	100	98.4	99.2	100	100	99.2	100	99.6
	Avg	36.2	66.1	87.7	80.8	78.7	71.1	77.5	80.3	77.2	52.6	27.6	66.9
	ViFS1	98.3	95	98.3	98.3	100	98.3	98.3	98.3	100	100	81.7	97
SLPP	Max	99.2	100	99.2	99.2	100	99.2	99.2	98.4	98.4	98.4	99.2	99.1
	Avg	44.5	75.2	88.2	84	79.2	76.4	78.6	80.6	79.1	62	35.9	71.2
	ViFS2	98.3	96.7	98.3	98.3	98.3	98.3	98.3	98.3	98.3	100	86.7	97.3

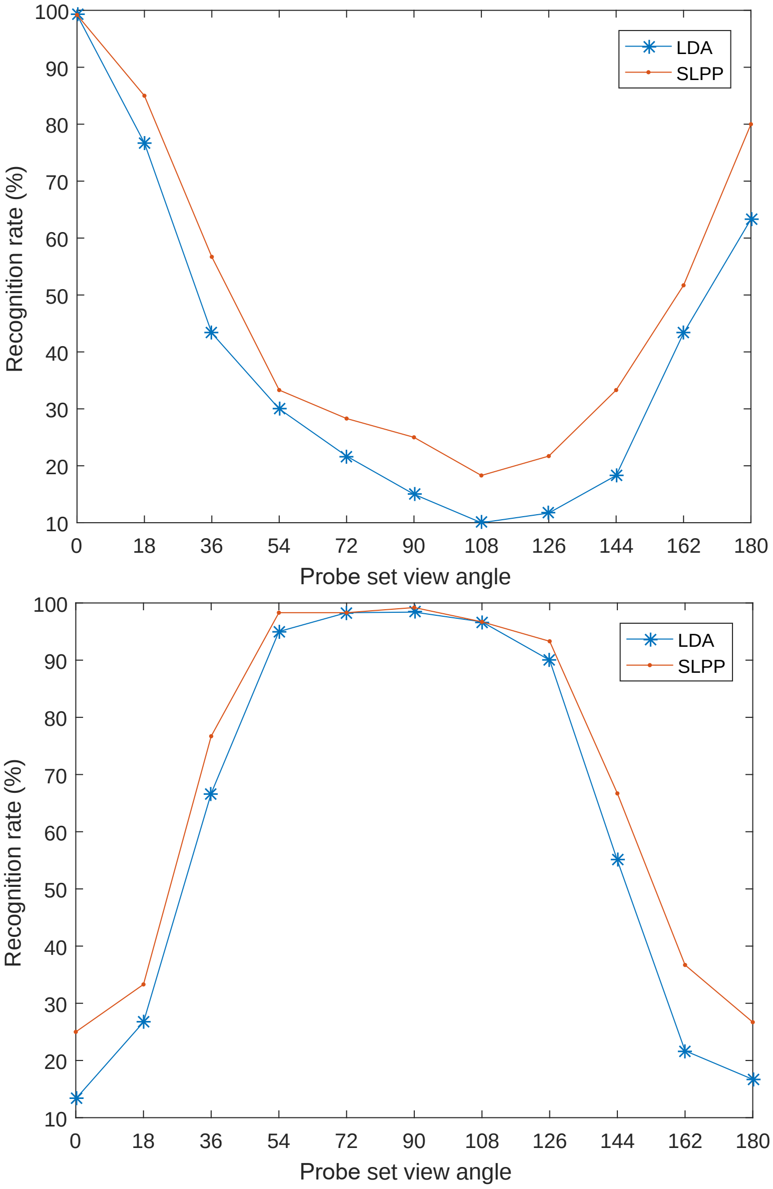


Figure 3.3: Cross-view comparison between the discriminant analysis and manifold learning approaches.

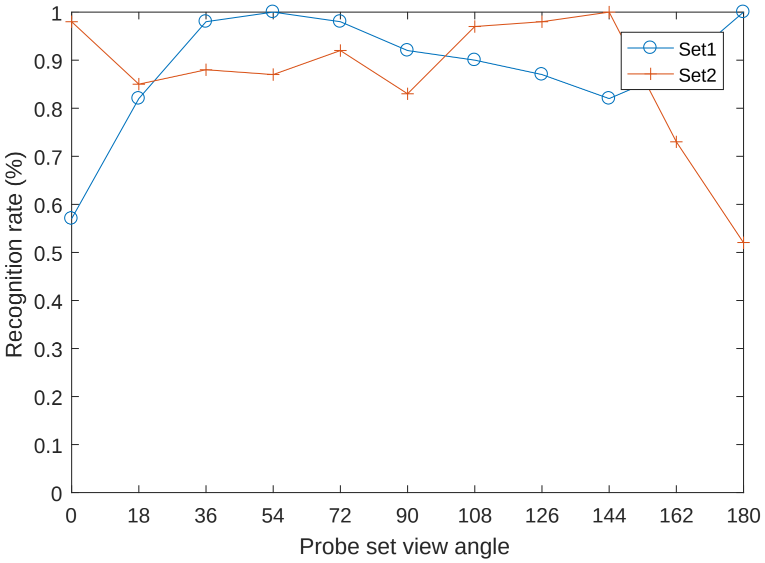


Figure 3.4: Recognition accuracy (%) of ViFS2 when templates from two views are available in the gallery set.

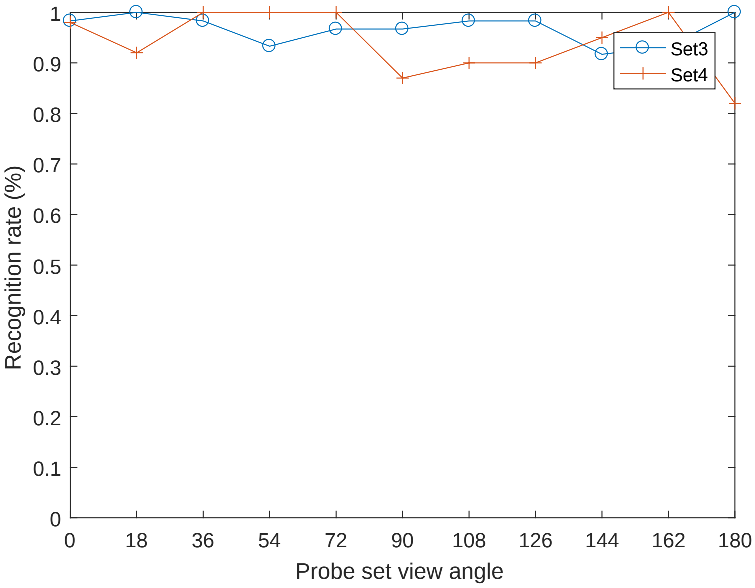


Figure 3.5: Recognition accuracy (%) of ViFS2 when templates from 3 and 4 views are available in the gallery set.

better discriminant power provided by SLPP. The results in Table 3.4 show that the proposed ViFS is amenable to be used with different subspace learning methods, and may thus be integrated with other feature enhancement approaches.

We also test ViFS2 on the scenario when the gallery set contains templates from only two view angles. We use two different gallery view sets: Set1 is $\{54°, 180°\}$ plotted with blue curve and Set2 is $\{0°, 126°\}$ with red. From Figure 3.4, we draw the conclusion that when training view angles are widely separated, we achieve fairly good results. The average accuracy of Set1 is 88.6%, and the Set2 is 86.6%. The frontal and back views of the probe set have the lowest matching rates, when they do not appear in gallery set. For example, Set1 does not include the 0° view angle, thus the performance of recognising 0°probe templates is below 60%. On the contrary, Set2 does not have the 180°view angle, and therefore the accuracy of the 180°probe templates is only slightly above 50%. We also evaluate the performance of ViFS on the 3+ view scenario. Set1 is trained with $\{0°, 18°, 108°, 180°\}$, and plotted with blue curve. Set2 is trained with $\{0°, 54°, 126°\}$ and plotted with red. Figure 3.5 shows that the increase of training views improves the overall performance of our framework, and make the matching accuracy more stable across 11 views.

In Table 3.5, we compare ViFS2 with two recently proposed methods by Tang et al. [1]. Tang(9) refers to the experiment result using 9 training views

Table 3.5: Comparison with Tang el al.'s work [1].

Method	18°	36°	54°	72°	90°	108°	126°	144°	162°	Average
Tang(9)	94	98	99	98	99	98	98	98	93	97.3
ViFS2(5)	100	100	96.7	100	98.3	98.3	98.3	100	100	**99.1**
Tang(4)	91	98	92	98	94	98	93	98	90	94.7
ViFS2(4)	100	100	98.3	100	98.3	98.3	98.3	100	91.7	**98.2**
ViFS2(3)	85	91.7	90	96.7	96.7	98.3	98.3	98.3	100	**95**

from 18° to 162°. Tang(4) refers to the experiment result using 4 training views $\{36°, 72°, 108°, 144°\}$. ViFS2(5) refers to the proposed method ViFS+SLPP using 5 training views $\{0°, 36°, 72°, 108°, 144°\}$. Likewise, ViFS2(4) and ViFS2(3) refers to the proposed method ViFS+SLPP using 4 ($\{36°, 72°, 108°, 144°\}$) and 3 ($\{0°, 108°, 162°\}$) training views. In their paper, Tang et al. only compare the results on probe data from 18° and 162°, thus we follows the same setting. The *Average* on the right most column of Table 3.5 is the average of all the accuracy numbers across the row. ViFS2 attains better performance than other state-of-the-art methods, on average. We also proved that with less training view (ViFS2(3)), the proposed framework still outperforms 3D model based methods proposed by Tang et al.

In Table 3.6 we compare the performance of ViFS2 with one of the experiment result in Wu et al.'s method [2]. For the specific experiment, they assume that gallery templates from all views are available, which fits the assumption of ViFS based framework. Their results are the average rates excluding identical view. Following the same experiment setting, we use multiple gallery data to train ViFS and match with probe excluding identical view. We i.e. training using 4 views (ViFS2(4)) and 2 views (ViFS2(2)) respectively, and matching with probe excluding identical views. ViFS2(2) refers to the proposed method ViFS+SLPP using 2 views for training, and our rates also excludes identical view. ViFS2(4) attains better result than the CNN

Table 3.6: Comparison with Wu et al.'s work [2].

Gallery	0°- 180°				Average
Probe	0°	54°	90°	126	
Wu et al.	82.2	**94.8**	88.9	93.6	89.9
ViFS2(4)	**88**	90	**94**	**94**	**91.5**
ViFS2(2)	62	92	**94**	78	81.5

based method by Wu el al. However, when training views are equal or less than 2, our method cannot produce satisfying result mainly due to the lack of view-invariant features.

We also provide an intuitive comparison between different cases where varying number of gallery view data are available. Here we use ViFS1 to generate the matching accuracy, since the ViFS+LDA method has larger accuracy margin between different cases. In Case1 we would like to evaluate the scenario when gallery templates are mainly from frontal view, i.e. 0°-54°. We denote the scenario that gallery data contains gait templates from two frontal views ($\{0°, 18°\}$) as Case1(2). Similarly, we denote the scenario that gallery data contains gait templates from four frontal views as Case1(4), and the gallery set from 0° to 108°is denoted as (Case1(7)) . As shown in Figure 3.6, the margins between the three curves representing these three cases are very large, and as the number of available gallery views grows larger, the overall performance improves. The LDA enhancer cannot provide cross-view feature learning ability when the view variation is large than 36°, thus Case1(2) fails to generate satisfying results on probe templates between 54° and 144°. Case1(4) is less effected, since it contains 54° template which is close to lateral view. Case1(7) covers all the views from 0° to 108°, thus it is able to provide good view-invariant recognition results. We also observe similar (in an opposite way) results in Figure 3.7, where gallery templates are mainly from back view. In 2 view case (Case2(2)), the gallery set contains $\{162°, 180°\}$, in 4 view case (Case2(4)), the gallery set has $\{126°, 144°, 162°, 180°\}$, in 7 view case (Case2(7)) the gallery set is from 72° to 180°. We already observe from Table 3.2 that LDA enhanced features perform well in opposite view matching cases such as 0°againt 180°, 72 °against 108°. It could explain the opposite results shown in Figure 3.6 and 3.7.

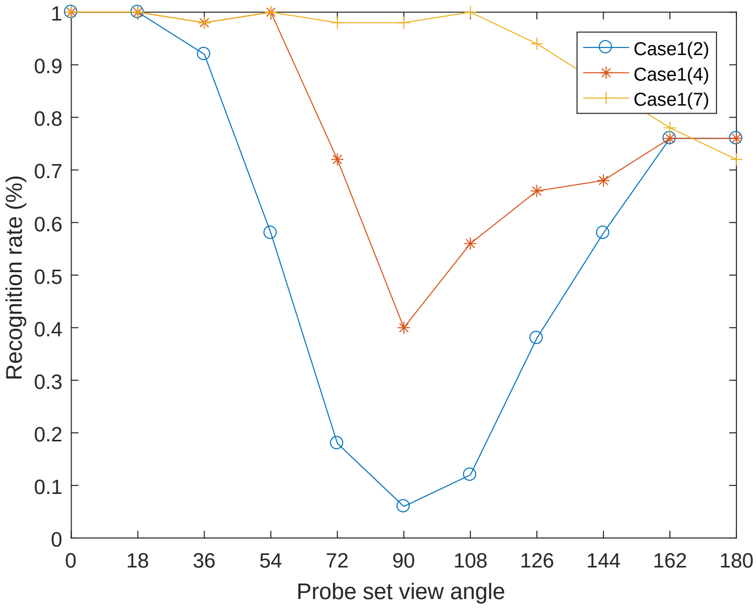


Figure 3.6: Recognition accuracy (%) of ViFS1 when gallery templates from different views are available in the gallery set. Gallery views are mainly from frontal views.

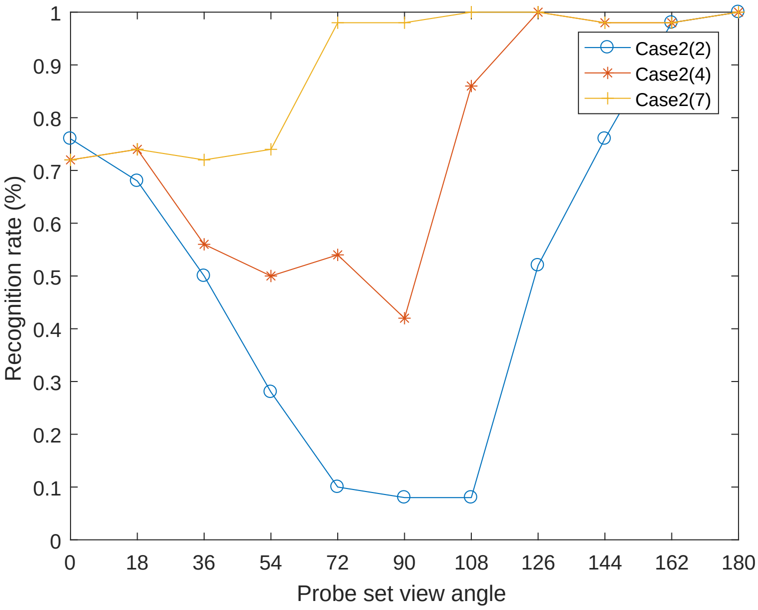


Figure 3.7: Recognition accuracy (%) of ViFS1 when gallery templates from different views are available in the gallery set. Gallery views are mainly from back views.

In Case3 and Case4 we evaluate the case where gallery templates are from widely spread views. In Case3 we evaluate the scenario when gallery templates are widely spread, but mainly from frontal view. In 2 view case (Case3(2)), the gallery set contains $\{0°, 90°\}$, in 4 view case (Case3(4)), the gallery set has $\{0°, 18°, 54°, 90°\}$, in 7 view case (Case3(7)) the gallery set is from 0° to 108°. As shown in Figure 3.6, the margins between different number of gallery views are decreased, since the enhanced gallery feature sets covers more view angles. Still, we observe that when view variation is large than 36°, Case3(2) fails to generate satisfying results on probe templates from 36° and 54°. Case2(4) has better results on these two views, benefiting from the multi-view gallery feature sets on this side, but has similar results with Case3(2) between 90° and 180°. Case3(7) covers all the views from 0° to 108°, thus it is able to provide good view-invariant recognition results. We also observe similar (in an opposite way) results in Figure 3.7, where gallery templates are mainly from back view. In 2 view case (Case4(2)), the gallery set contains $\{90°, 180°\}$, in 4 view case (Case4(4)), the gallery set has $\{90°, 126°, 162°, 180°\}$, in 7 view case (Case4(7)) the gallery set is from 72° to 180°.

Based on the observations discussed above, we discovered the weakness of ViFS, that it cannot perfectly merge the advantage of two views when they are concentrated For example, in Case1(4) when gallery is from $\{0°, 18°, 36°, 54°\}$, we would expect a better performance matching with 90° probe data, since we observe in Table 3.2 that 54° gallery data has good cross-view performance when matching with 90 ° probe data. We discovered that the ViFS minimiser for this case is $w = \{0.006, -0.042, -0.541, 1.52\}$, corresponding to the gallery samples from the 4 views listed above. The 54°gallery data is offered with a very high weight 1.52. We suggest that it should be too high that it causes feature contamination, instead of good feature selection. The over weight of a certain template causes opposite effect, since it sabotages the standardisation of feature sets, and according to our discussion in Chapter 1 and 2, it results in misclassification.

3.4 OU-ISIR Large Population Dataset

The OU-ISIR Large Population Dataset includes more than 4000 subjects, each recorded using cameras from 4 views: 55°, 65°, 75°, and 85°. Among all the published dataset for gait recognition evaluation, it has extensively bigger size in the number of subject size, thus it is more reliable in statistically evaluating the performance of the proposed framework. According to the existing protocols of benchmarks [80] [97] [119], a common experiment setting is to use a subset of 1912 subject, which is divided into two groups, where 956 subjects are used for training purpose and the rest for testing. We refer to this subset as OU-ISIR LP dataset in the following discussion. Identical to the previous section, we evaluate the framework in cross-view and multi-view manner separately.

3.4.1 Cross-view Evaluation

Table 3.7 and Table 3.8 tabulates the matching results using LDA and SLPP respectively. Following the same settings as in Section 3.3.3, we match the euclidean distance between enhanced gallery and probe feature. It is clear that SLPP has much better performance than LDA in cross-view case, which is identical to the previous observation on CASIA Dataset B. Specifically, when view angle variance between gallery and probe data is small (10°), both LDA and SLPP achieve high accuracy, close to their identical-view matching result. However, when the view angle variance is large than 20 °, the accuracy of LDA drops faster than SLPP.

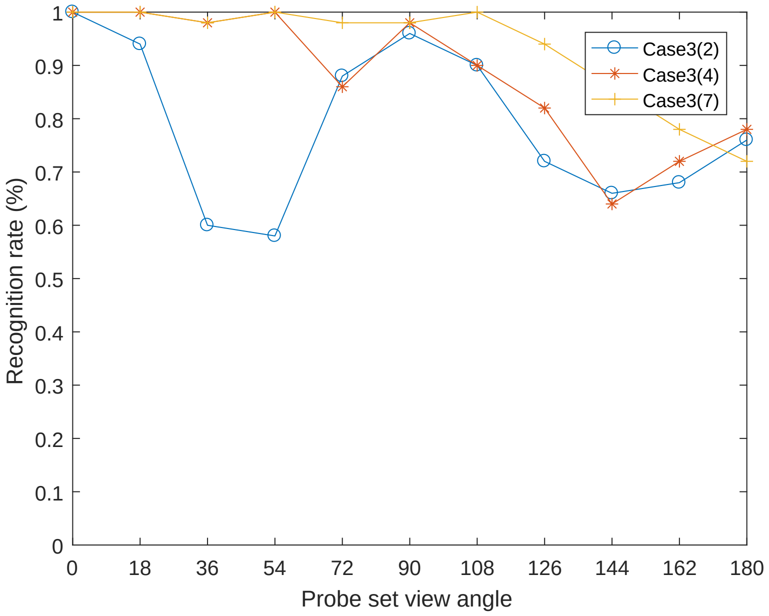


Figure 3.8: Recognition accuracy (%) of ViFS1 when gallery templates from different views are available in the gallery set. Gallery views are from wide spread views (concentrating on frontal view).

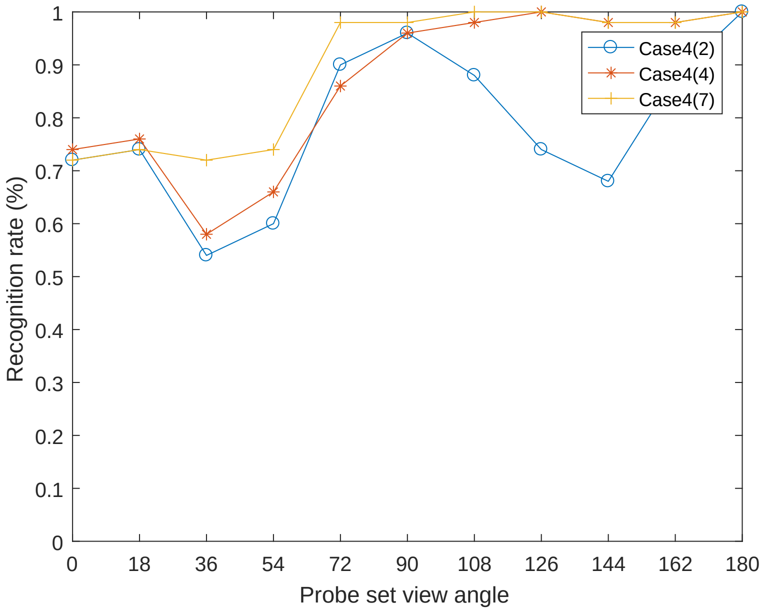


Figure 3.9: Recognition accuracy (%) of ViFS1 when gallery templates from different views are available in the gallery set. Gallery views are mainly from wide spread views (concentrating on back view).

Table 3.7: Cross-view matching result between gallery and probe set using LDA.

Probe	Gallery				Average
	55°	65°	75°	85°	
55°	90.4	89.9	73.8	53.8	77
65°	72.2	94.8	93.1	79.6	84.9
75°	44.1	92.1	95.5	93.7	81.4
85°	35.3	76.4	94.5	96.9	75.7

Table 3.8: Cross-view matching result between gallery and probe set using SLPP.

Probe	Gallery				Average
	55°	65°	75°	85°	
55°	89.7	94.8	91.7	81.3	89.4
65°	77.9	96.5	96.8	92.6	91.0
75°	59.2	95.5	97.1	96.4	87.1
85°	47.7	89.1	96.7	97.2	82.7

3.4.2 Multi-view Evaluation

We introduce ViFS2 (ViFS+SLPP) to perform multi-view matching on OULP dataset. We use all 4 views to train ViFS, and match the reconstructed gallery set with probe set from different views. Here we compare the performance of our experiment with two CNN based approaches due to Wu et al. [2] and Shiraga et al. [97]. As mentioned in both [2] and [97], five-fold cross-validations are deployed to reduce the effect of randomness. Specifically, the training and testing set (each contains 956 subjects) are randomly selected 5 times, each time we record the recognition accuracy, and the final accuracy is the average of the 5 experiments.

Table 3.9 presented the matching result of ViFS2 and the comparison between different approaches. Wu et al. (i) refers to the identical-view matching rates, while Wu et al. (a) is the average accuracy of four gallery view to a certain probe view. The same notations apply to Shiraga et al. We notice that Wu et al. achieve the highest accuracy in identical-view matching, which strongly suggest the effectiveness

Table 3.9: Multi-view matching result between gallery and probe set using ViFS2.

Gallery	55°- 85°				Average
Probe	55°	65°	75°	85	
Wu et al. (i)	**98.8**	**98.8**	**98.8**	**98.8**	**98.8**
Wu et al. (a)	93.5	94	94	95.8	89.9
Shiraga et al. (i)	94.7	95.1	95.2	94.7	94.9
Shiraga et al. (a)	89.2	93.3	93.3	90.5	91.6
ViFS2	**97.8**	**97.9**	**98.4**	**97.5**	**97.7**

of CNN in extracting discriminative features from images, especially with sufficient number of training samples. Meanwhile, Shiraga et al. uses a network shallower than Wu et al., and they do not use pair-image as input to train the network, thus their performance is lower than Wu et al. However, the proposed method ViFS2 shows better performance in view-invariant gait recognition, when 4 view gallery data are available. Besides, the overall performance is very close to Wu et al.'s method in identical-view scenario, while the training and testing time of our linear framework should be significantly smaller than the CNN based approach.

3.5 Summary

In this chapter, we proposed a multi-view feature selector ViFS, and developed a robust framework based on ViFS and subspace feature enhancers to tackle the multi-view gait recognition problem. We reconstruct the gallery data on the image level, and use linear feature enhancers to increase the discriminant power of features and reduce the number of required gallery views. We tested the proposed framework on CASIA Dataset B and OU-ISIR Large Population Dataset, and evaluated the effect of LDA and SLPP as feature enhancer on our framework. The result indicated that if gallery set covers 4 or more views, the matching accuracy of the proposed framework with the probe set from all views can exceed 98% on average on CASIA

Dataset B, and 97% on OULP. Furthermore, the proposed framework is parameter-free and has low computational cost, which indicates the great potential for real world application.

In the next chapter we use CNN as feature enhancer for ViFS based framework. Since CNN is widely recognised as the most powerful image classification tool, its discriminative feature learning ability outperforms any conventional approaches on computer vision tasks. The feature maps learned by CNN are fed into the ViFS, obtaining the reconstructed gallery feature maps, which are used to compare with probe feature maps from arbitrary views.

Chapter 4

A Joint Framework of ViFS and CNN

In the previous chapter we compared the proposed framework based on View-invariant Feature Selector (ViFS) with the state-of-the-art approaches, including CNN based methods by Wu et al. [2] and Shiraga et al. [97]. The cross-view matching results reveals the fact that features learned by CNN has greater discriminative power than traditional learning methods. Common CNN is regarded as an end-to-end solution to computer vision problem, while the feature extraction, feature selection and decision making process are integrated and fine-tuned simultaneously. It is often regarded as one of the greatest advantage for CNN, comparing with the traditional step-by-step approaches. However, as proved by [120] [121], a pre-trained CNN can also be served as a powerful generic feature extractor for general purposes. According to their report, the extracted features (also known as feature maps [122]) can be combined with traditional approaches (e.g. SVM, random forest) to solve a specific classification problem. The existing CNN based approaches to gait recognition only introduce CNN as an end-to-end solution, while the hand-crafted features and classifier ensemble methods are proved to be efficient in improving recognition accuracy [21] [123]. Therefore, in order to pursue higher identification accuracy, we would like to explore the potential of CNN feature maps on gait recognition, and the possibility of combining ViFS with CNN feature maps. Based on the conclusion from previous chapter, the accuracy of ViFS framework heavily relies on the effectiveness of feature enhancers[1]. With the conventional linear subspace enhancer (like LDA and SLPP), our method is able to compete with CNN based approaches on multi-view gait recognition problem. We would assume that the combination of CNN feature maps and ViFS could achieve the state-of-the-art performance on both cross-view and multi-view problem.

4.1 CNN Feature Maps

4.1.1 The Hierarchical Feature Learning of CNN

To understand the success of CNN, Zeiler and Fergus [3] proposed to map the feature activities in the intermediate layers back to the pixel space with a Deconvolutional Network (decovnet). Hence, the cause of activities in the feature maps can be traced back to the specific patterns on input images. Since it performs filtering and pooling reversely, decovnet is also capable of mapping features to pixels, which demonstrates the hierarchical learning process of features within the CNN model intuitively. As is shown in Figure 4.1, 9 reconstructed feature maps in pixel level is displayed, referring to the top 9 activations in a random subset. We can see that the first layer consists of colour blocks and colour edges at different frequency and

[1] As is noticed in Chapter 3, SLPP has better performance than LDA on cross-view matching, and it has better feature enhancement performance on ViFS. Therefore we draw the conclusion that the discriminative power of feature enhancers is highly related to its effectiveness on ViFS.

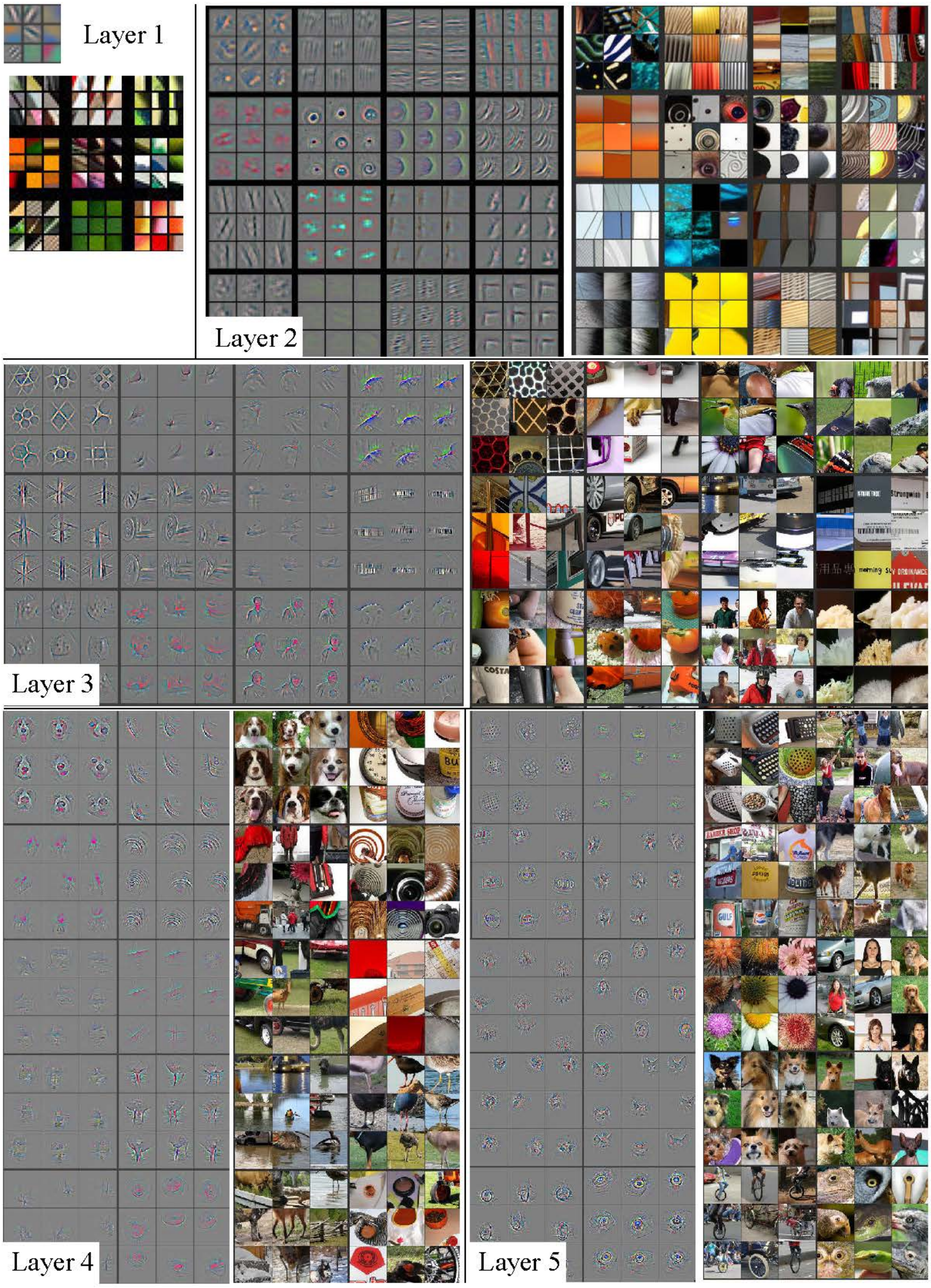
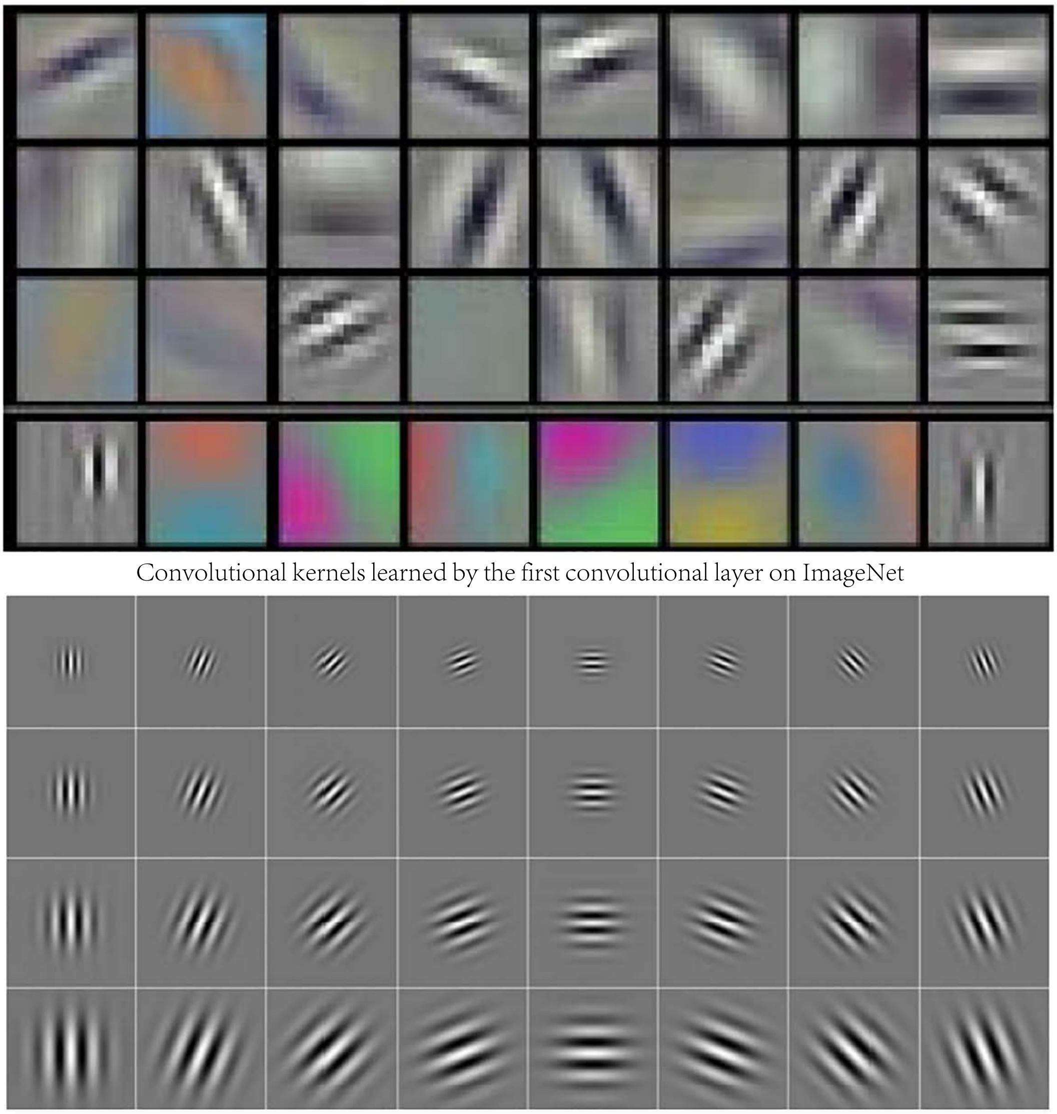


Figure 4.1: Visualised features from the 5 convolutional layers presented in [3]. The network is well trained with images from ImageNet database.

orientations. However, the corresponding input images (more specifically, the local regions of input images) may not be from the same class. In layer 2 the feature maps is mainly describing corners, edges or other conjunctions. Layer 3 captures larger region of textures with similar patterns. The visualised feature maps is able to describe a general class, e.g. the tires in row 2 column 2 and humans in row 3 column 3. As the features pass to further layers, it describes more discriminative textures, specified in different classes. In layer 5 the feature maps are able to locate the whole subject with large pose variations, e.g. keyboard in row 1 column 1 and logos in row 2 column 1.

4.1.2 Motivation

It is mentioned in Chapter 2 that Gabor wavelet representation has been proved to be effective for GEI based gait recognition [21] [78] [124] [125], since it is able to capture the salient properties from input images. For example, in [78] the authors compare two approaches on USF HumanID Gait Challenge Dataset. Approach 1 is called general tensor discriminant analysis (GTDA), which is an improved metric learning approach based on 2D LDA. Approach 2 is the combination of Gabor representation and GTDA. The recognition accuracy has roughly 10% improvement in general. They all proved that Gabor representation, i.e. the convolution of input image with Gabor filters, is able to extract discriminative features from raw images. The process of obtaining Gabor representation is similar to generating feature maps from input image in layer 1. Figure 4.2 shows the visualisation of Gabor filter and convolutional kernels in the first layer of a CNN. Both are edge detectors of various orientations. Therefore, it might be argued that the high level features learned by CNN share similar characteristics with Gabor based representations, which has



Convolutional kernels learned by the first convolutional layer on ImageNet

Gabor kernels of 8 orientations and 4 scales

Figure 4.2: Visualised Gabor kernels and convolution kernels .

greater discriminative power. The convolutional kernels in CNN has similar effect as Gabor filters. Furthermore, the multi-layer convolutional computation is proved to be more effective in learning high-level features. Hence we regard CNN as a feature mapping technique, and combine its hierarchical learning ability with the feature selection method to achieve better performance. Firstly we train the network with labelled gallery data, where the parametres and weights in the network are automatically adjusted. This step creates a mapping from pixel-level features to high-level features with better orientation-invariance and discrimination between classes. The learned gallery feature set can be extracted from the softmax layer. Then we feed the network with probe data and obtain the learned probe feature vectors from the softmax layer. We put these features into our least square regression model, from which an optimal set of gait representation is generated accordingly. Finally we match the probe features with the gallery features after regression. According to the previous research, the least square regression model is an efficient tool in tackling gait deformation problem, especially when gallery data is sufficient. We believe that it will also work for multi-view gait recognition problem. A simple demonstration of this framework is shown in Figure 4.3.

4.2 Network Structure

It is mentioned in Chapter 2 that CNN can also be used for segmentation tasks. In order to further improve the recognition accuracy, we propose to use CNN based segmentation network to improve the silhouette quality of the dataset. In this section, we firstly introduce a well trained fully convolutional network (*fullconvnet*) to obtain high quality silhouette. Then we use this set of silhouettes to generate GEIs for training the feature extraction CNN and evaluating the proposed framework.

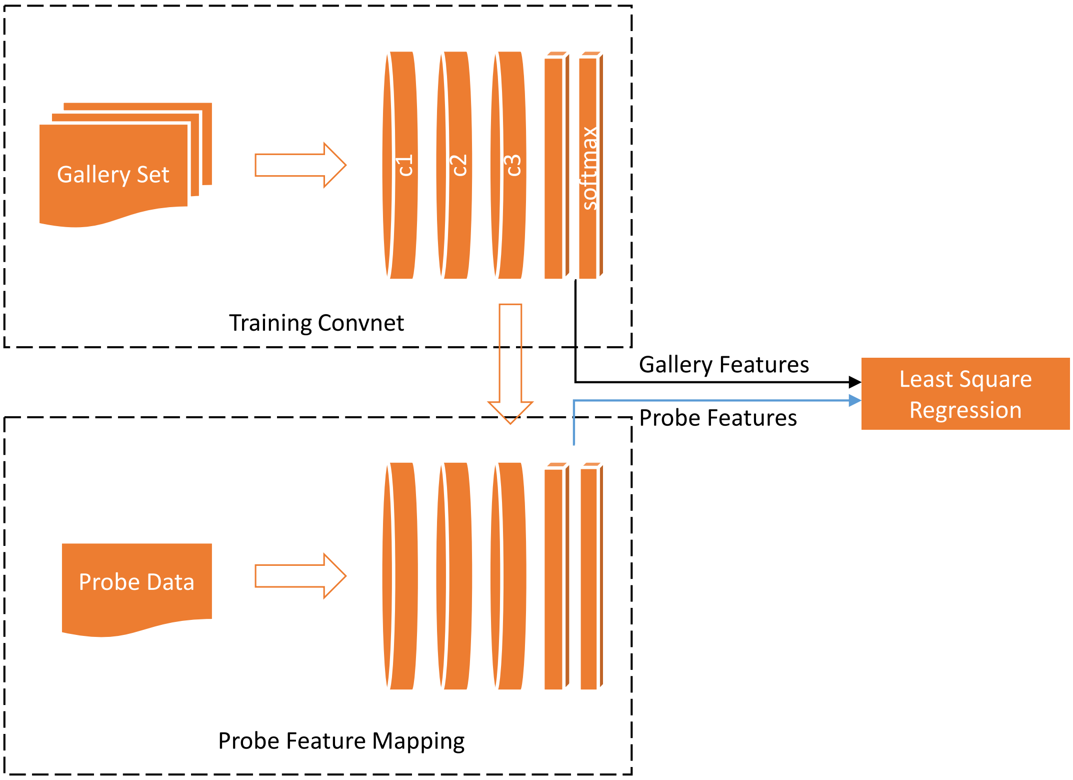


Figure 4.3: The proposed framework to combine CNN and least square regression model. The details of CNN is introduced in Chapter 2, while the details of least square regression model is introduced in Chapter 3.

4.2.1 Segmentation Network

We design a 3-channel (RGB) fullconvnet to segment human profiles from three randomly sampled frames of an input video. Each channel consists of 7 convolutional layers for feature extraction and 1 de-convolutional layer for recognition. Specifically, for the first convolution stage, an input image of size 64×64×3 goes through a single stride convolution with 48 kernels of size 5×5, followed with a two-stride 3×3 max pooling. The pooling products are normalized with a local response normalization layer. The second convolutional stage follows the same procedures as the first one. The 3rd to the 5th stage consists of convolutional operations only, while dropout is introduced in the 6th and 7th stage along with the convolutional stage. The de-convolutional layer reflects the output features from the first convolutional stage, with the size of 1×1 feature map, to produce a 64×64 prediction. To fit the range scope of a binary mask, we put a sigmoid neuron unit[2] after the de-convolutional layer. After normalizing the size of the de-convolutional layer and the mask image, we train the segmentation network with the L2 norm of the prediction and the mask. For a given image x_i, the L2 norm of channel-j is formulated as:

$$L_{seg-ch-j} = \|p_i - m_i\|_2, \tag{4.1}$$

where m_i refers to the i-th binary mask of x_i, and p_i is the i-th segmentation prediction at the top layer. The i-th prediction is formulated as

$$p_i = g(f(x_i)), \tag{4.2}$$

[2] All neurons are Rectified Linear Unit (ReLU) except for the de-convolutional layer, which is sigmoid. Besides, the max pooling layers are uniformly two-stride with size 3×3.

where $f(x_i)$ projects x_i into a feature space via one or more convolutional stages, and $g(\cdot)$ predicts the final segmentation results though the de-convolutional layers. Hinton et al. introduced a similar auto-encoder approach, where $f(\cdot)$ and $g(\cdot)$ can be regarded as encoder and decoder respectively [126]. The segmentation loss function for all three channels is formulated as:

$$L_{seg} = \sum_{ch=1:3} \|p_i - m_i\|_2, \tag{4.3}$$

An iterative learning is performed to further improve the segmentation accuracy of human profiles. During the first iteration, we use the rough segmentation results to locate human profiles, where the cropped profiles are sent to the second iteration as references for a more precise segmentation. The optimal results are obtained within several iterations. It is proved that the iterative learning obtains more precise silhouettes that improves gait recognition accuracy. In our experiments, we make a trade-off between computational cost and accuracy by setting iteration times to two. More details about fullconvnets can be found in [116].

4.2.2 Feature Mapping Network

We adopt a 5-layer CNN with 3 input channels as classifier. The inception layer transfers 3 single channel profiles simultaneously, and generate a 3 channel blob (can be regarded as a 3 channel RGB image). The first and second convolutional component of the classifier follow the same settings as the first step of fullconnets, while the third component perform convolutional operations only. The output feature map after three convolutional stages is with the size of 11×11. Based on the feature outcomes,he subsequent two fully-connected layers perform recognition with a soft-max unit.

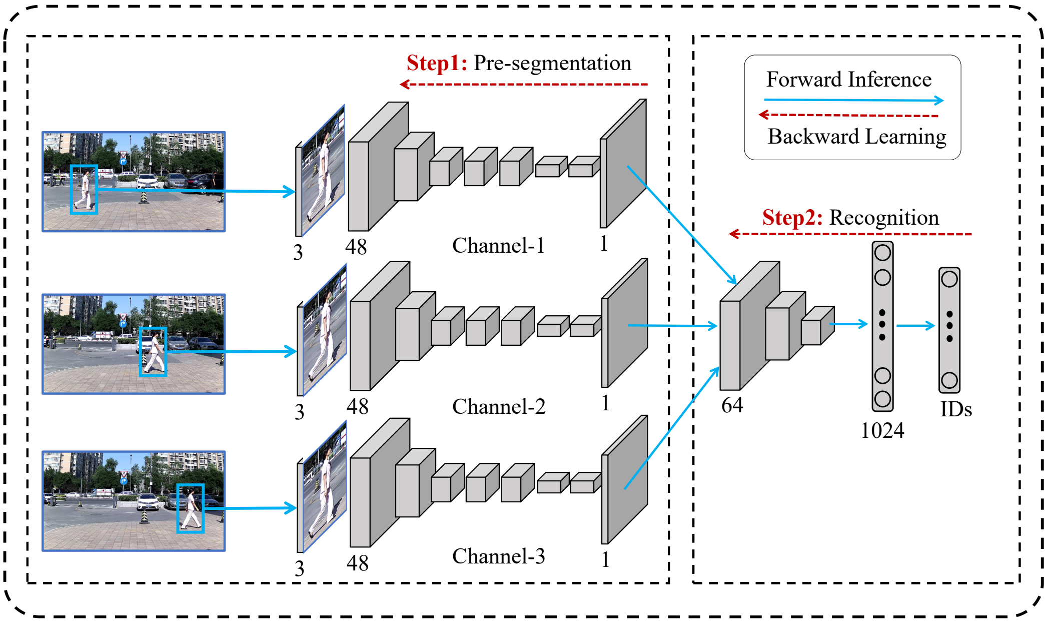


Figure 4.4: The architecture of the segmentation network and feature mapping network.

As shown in Figure 4.4, we explain the two steps for training our framework:

- *Pre-segmentation.* First, we train a 3-channel (RGB) CNN [116] model with public human segmentation databases [127]. We call this process as gait pre-segmentation, which is used as the initial model of our gait segmentation modular. Afterwards, the fullconvnet model is iteratively used. In detail, at the beginning, the input of the fullconvnet model are three randomly sampled frames from a gait video. At each iteration, we obtain a more accurate cropped human box containing the output segmented pixels of human from the last iteration of segmentation, and this cropped box is used as the input of the next iteration. After several iterations (e.g., two or three), relatively accurate human silhouettes will be produced and used as the pseudo-label of gait segmentation.

Table 4.1: The details of the CNN structure (13 layers with 3 channels).

Layer	Channel-1	Channel-2	Channel-3
C1	conv5-48 max-pool LRN	conv5-48 max-pool LRN	conv5-48 max-pool LRN
C2	conv5-128 max-pool LRN	conv5-128 max-pool LRN	conv5-128 max-pool LRN
C3	conv192-3	conv192-3	conv192-3
C4	conv192-3	conv192-3	conv192-3
C5	conv192-3	conv192-3	conv192-3
C6	conv1024-7 dropout	conv1024-7 dropout	conv1024-7 dropout
C7	conv1024-1 dropout	conv1024-1 dropout	conv1024-1 dropout
DC8	deconv1-64 sigmoid	deconv1-64 sigmoid	deconv1-64 sigmoid
Inception layer			
C9	conv5-64 max-pool		
C10	conv5-128 max-pool		
C11	conv3-192		
FC12	FC1024 dropout		
FC13	FC69 soft-max		

- *Recognition.* We train the CNN-based gait recognition in this step. The inputs are the gait segmentation results, and the supervised labels are the human identities corresponding to the input gait videos. More details will be introduced in Table 4.1 with corresponding explanation.

4.3 Experiment Design and Results

We now testify the cross-view performance of the prosed framework. Firstly we present the baseline result of CNN feature maps on CASIA Dataset B. Then we

Table 4.2: Cross-view matching using CNN feature maps setting 1. G: Gallery; P: Probe

P G	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°
0°	100	100	92	92	68	64	68	84	88	100	100
18°	100	100	100	100	84	76	76	92	92	100	92
36°	92	100	100	100	96	84	92	92	92	92	80
54°	80	100	100	100	100	100	96	100	92	84	72
72°	56	80	100	100	100	100	100	100	100	72	60
90°	52	80	88	96	100	100	100	100	100	72	48
108°	52	76	88	96	100	100	100	100	100	72	56
126°	68	84	84	100	100	100	100	100	100	92	72
144°	84	96	96	96	96	88	100	100	100	100	76
162°	96	100	96	84	80	68	80	100	100	100	100
180°	96	96	84	80	72	64	64	96	88	100	100

introduce the combined framework and shows the improved results.

4.3.1 Cross-view Matching Using CNN Feature Maps

We firstly presents the cross-view matching result using CNN Feature Maps. Specifically, we put gallery and probe GEI templates into the well-trained network, and extract the feature map vectors from the penultimate layer of CNN. We measure the euclidean distance between gallery and probe feature map vectors, and the cross-view matching accuracy is tabulated in Table 4.2. CNN feature maps show significant improvement comparing with the result using traditional subspace learning methods (recall Table 3.2 and 3.3 in Chapter 3). Apart from the large view disparity cases, which are marked in grey colour, other cross-view matching results are all above 80%, suggesting that CNN feature maps have great discriminant power.

4.3.2 The Combination of ViFS and CNN Feature Maps

In order to perform robust view-invariant recognition, we apply ViFS to multi-view gallery CNN feature maps. Assume that gallery set has 2 views available, the number of all views are 11, thus there could be $\binom{11}{2} = \frac{11!}{2!(11-2)!} = 55$ different combinations. We select 3 representative sets for comparison. Set1 contains $\{0°, 90°\}$, where the gallery views are widely spread. Set2 contains $\{0°, 54°\}$, where 0° has good performance on the frontal/back view and 54° has good results on other views (18° to 144°). Set3 contains $\{18°, 108°\}$. As is shown in Table 4.3, Set3 achieves very high accuracy in average, while Set1 and Set2 has only slight improvement comparing with the average results of Table 4.2, which is denoted as Avg. Wu el al. refers to the state-of-the-art CNN method presented by [2]. Their experiment setting assumes that gallery from 0° to 180° are available. Set3 outperforms Wu et al.'s method by 3% on average, but the results on 72° and 90° probe data is lower than theirs.

Table 4.3: The matching results using the combination of CNN feature maps and ViFS.

Set	Probe											
	0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Average
Set1	100	100	92	92	80	80	80	96	92	100	100	92
Set2	76	100	100	100	100	96	100	100	92	80	64	91.6
Set3	100	100	100	100	92	88	96	100	100	100	92	**97.1**
Avg.	86.9	92	92.7	93.1	88	85.1	85.5	90.9	93.8	91.3	85.5	89.5
Wu et al.	88.7	95.1	98.2	96.4	94.1	91.5	93.9	97.5	98.4	95.8	85.6	94.1

Table 4.4: Comparison with Tang el al.'s work [1].

Method	18°	36°	54°	72°	90°	108°	126°	144°	162°	Average
Tang(9)	94	98	99	98	99	98	98	98	93	97.3
Tang(4)	91	98	92	98	94	98	93	98	90	94.7
ViFS3(2)	100	100	100	92	88	96	100	100	100	**97.3**

Following the same criterion, we provide an intuitive comparison between different cases where varying number of gallery view data are available, in order to explore the potential of CNN+ViFS framework. In Case1 we would like to evaluate the case when gallery templates are mainly from frontal view. In 2 view case

(Case1(2)), the gallery set contains $\{0°, 18°\}$, in 4 view case (Case1(4)), the gallery set has $\{0°, 18°, 36°, 54°\}$, in 7 view case (Case1(7)) the gallery set is from 0° to 108°. As is shown in Figure 4.5, the margins between different number of gallery views is much closer comparing with the result in Figure 3.6, since the CNN feature maps has much greater discriminant power across views, while the LDA enhancer cannot provide cross-view feature learning ability when the view variation is large than 36°. However, as view variation grows larger, ViFS loses its effectiveness, and the good results are all came from CNN feature maps as is tabulated in Table 4.2. Case1(4) and Case1(7) are overlapped in this case, which is unexpected, since from the previous experiments the more view should brings better performance. We look deep into the generated ViFS descriptors, and find out that the weights are almost evenly distributed, while the sum of the weights are way above 1, which indicates the same issue as is discussed in Section 3.3.4 that it sabotages the standardisation of feature sets. We also observe similar (in an opposite way) results in Figure 3.7, where gallery templates are mainly from back view. In 2 view case (Case2(2)), the gallery set contains $\{162°, 180°\}$, in 4 view case (Case2(4)), the gallery set has $\{126°, 144°, 162°, 180°\}$, in 7 view case (Case2(7)) the gallery set is from 72° to 180°.

In Case3 and Case4 we evaluate the case where gallery templates are from widely spread views. In Case3 we evaluate the case when gallery templates are widely spread, but mainly from frontal view. In 2 view case (Case3(2)), the gallery set contains $\{0°, 90°\}$, in 4 view case (Case3(4)), the gallery set has $\{0°, 18°, 54°, 90°\}$, in 7 view case (Case3(7)) the gallery set is from 0° to 108°. The three curves are even more closer to each other, indicating that the great discriminant power of CNN 'confuses' ViFS, making it difficult to select the most representative features from the multi-view feature set.

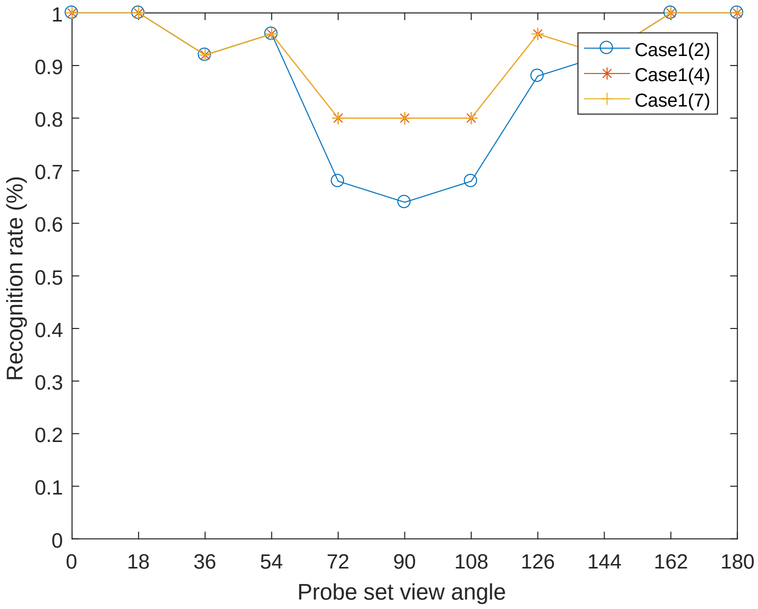


Figure 4.5: Recognition accuracy (%) of ViFS+CNN when gallery templates from different views are available in the gallery set. Gallery views are mainly from frontal views.

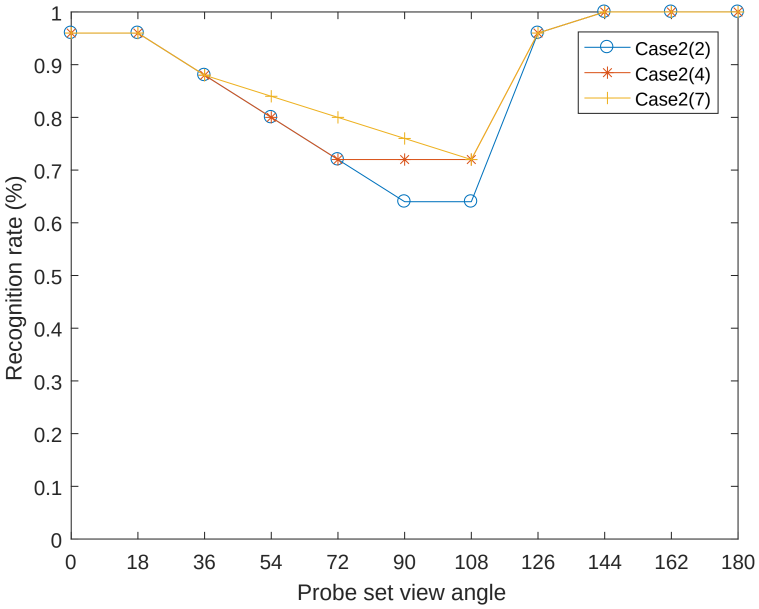


Figure 4.6: Recognition accuracy (%) of ViFS+CNN when gallery templates from different views are available in the gallery set. Gallery views are mainly from back views.

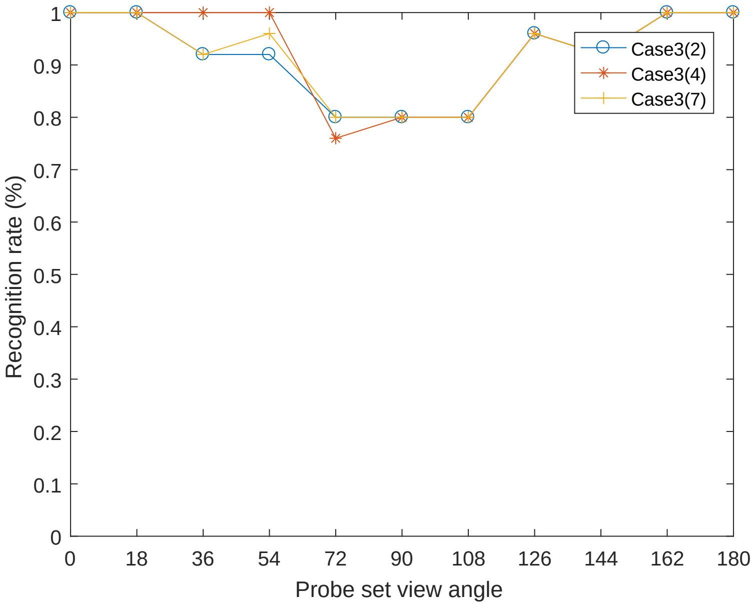


Figure 4.7: Recognition accuracy (%) of ViFS+CNN when gallery templates from different views are available in the gallery set. Gallery views are from wide spread views (concentrating on frontal view).

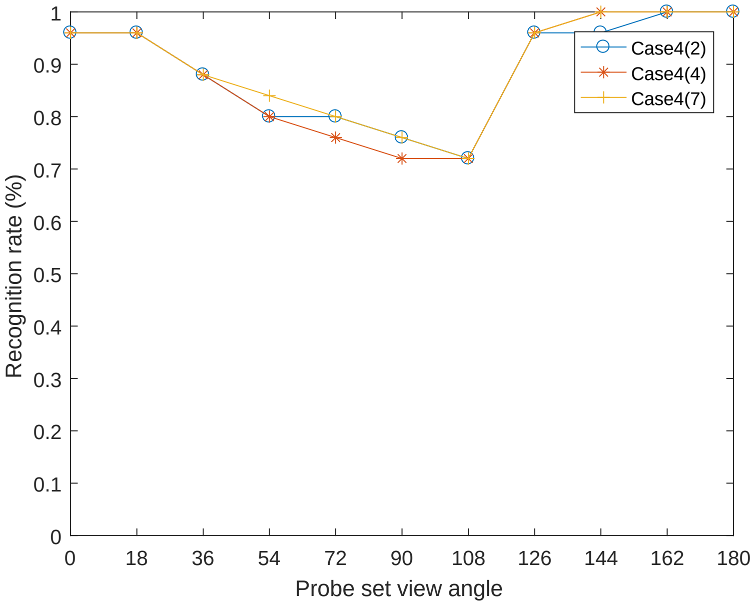


Figure 4.8: Recognition accuracy (%) of ViFS+CNN when gallery templates from different views are available in the gallery set. Gallery views are mainly from wide spread views (concentrating on back view).

4.4 Summary

In this chapter we successfully combined ViFS with CNN feature maps and achieved view-invariant gait recognition with less gallery data. By taking advantage of projecting the assembled gallery vectors into the same hyperplane with probe vector, the inter-class variation is enlarged while the intra-class variation is minimised. The promising result from our experiments indicated that when strong correlation exists between gallery and probe set, the ViFS model is able to automatically find this correlation among a set of gallery data with various deformations, and then returns with the best recognition result. Furthermore, we use CNN to extract highly efficient features from the input GEIs, and use ViFS to maximise its recognition accuracy. The result from our experiments convincingly proved that as the best feature extraction tool, CNN is compatible with traditional feature selection models.

The effect of silhouette quality disparity on gait recognition has not been well addressed by literatures. Besides, the former studies claimed that instead of improving the quality of silhouette, it is more productive to study noise-invariant features. However, an intuitive thought is that the noise should interference the recognition accuracy, instead of enhancing it. In the next chapter we generate a dataset to statistically analyse this problem and evaluate the performance of conventional algorithms, as well as the proposed ViFS based frameworks.

Chapter 5

Gait Recognition with Silhouette Quality Disparity

5.1 Problem Statement

As is mentioned in Chapter 2, the field of quality-invariant gait recognition remains unexplored. Most literatures focus on improving silhouette extraction quality, while few of them consider it as a covariate factor and study the case when silhouette quality disparity exists. There are generally two aspects to take into account when studying quality-invariant gait recognition. Firstly, it is inevitable that gallery and probe gait sources are obtained from different scenarios. Assume that gallery and probe data are acquired from different cameras, one with high resolution surveillance camera monitoring a shopping mall with sufficient illumination; another with normal low resolution camera monitoring an open field. A problem is easily occurred that the silhouettes from indoor high resolution camera is of better quality than those from outdoor low resolution camera. It is commonly seen since the public surveillance cameras differs from place to place. In practice, we might find that a subject commits crime in an outdoor environment which is recorded by the surveillance camera, and when a suspect is captured, the gait data is obtained in another controlled indoor environment; or a crime is hold and the gait data is retained in the system, while he escapes from the prison and is needed to be located from the cameras around city. Another aspect is that background subtraction algorithms may generate silhouettes of different quality levels under different scenes. For example, if the subject appears in the first frame of the video sequence, and stays in a relatively fixed position until the end of recording, it is very difficult for the Gaussian Mixture Model (GMM) or background subtraction (BS) based method to perform segmentation, since GMM initialises the models in the first frame, and all the constant regions throughout the whole video sequence including the subject (or part of the subject) will be regarded as background. Figure 5.1 illustrates this situation using video sequences from CASIA Dataset B. Images from the first row is recorded from a 0° camera and the second row are from 180°, all sequences are from the same subject. The first RGB image of each row is the raw video image, and the 13 continuous binary silhouettes are the segmentation results (normalised to standard size) using GMM method. The rightmost grey scale image of each row is the constructed GEI accordingly. There is an obvious human-shape hollow in both the silhouettes and GEIs, indicating that GMM fails to model this constant region as a moving foreground object. In both case, there exits silhouette quality disparities due to the illumination, shadow variance, or other forms of contamination from indoor to outdoor environment. The illumination variations can cause holes and missing parts of body on the silhouette, while post-processing approaches, for instance morphological dilation and erosion, can cause further quality disparities.

By looking into the factors that may cause silhouette quality disparity, it

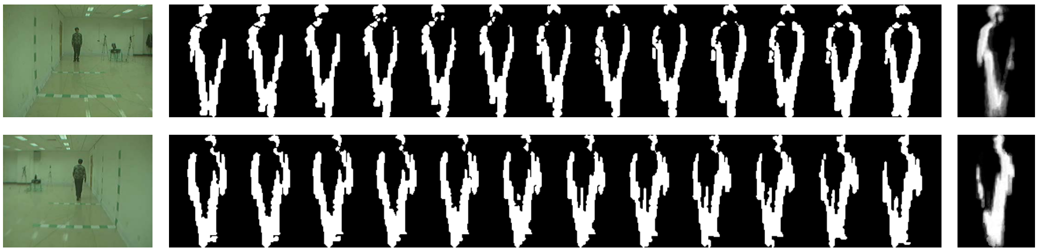


Figure 5.1: The segmentation results using GMM when the subject exists throughout the video sequences. Sequences on the top row are captured from 0° camera, while sequences on the bottom row are from 180°.

might be argued that the cross-quality gait recognition is worth studying. In order to perform robust gait recognition in open-set real-world cases, a quality-invariant gait recognition system has to be found. In this chapter, we studied the case when segmentation errors exist in gait gallery and probe data, which is caused by applying a set of segmentation algorithms on gallery and probe video sequences. We would like to focus on following situations:

- The gait data related to an individual to be recognized (i.e., the probe data) is not captured under ideal conditions, and therefore the associated gait silhouettes may be noisy and inaccurately segmented; whereas the stored gait data (gallery data) is captured in noise-free environments, or vice versa.

- The silhouettes extracted from gallery and probe data are obtained using different segmentation algorithms, or from different scenes, which may result in very different segmentation results[1].

The effect of the gait silhouette quality on the performance and practicability of appearance-based gait recognition algorithms are important issues but have not been intensively studied and only a limited number of solutions are reported in the

[1] A simple example is that a segmentation algorithm generates shadow-free gallery silhouettes, while another algorithm cannot generate shadow-free probe silhouettes. The shadow can then be considered as features (or noise) of the gallery silhouettes, thus affecting recognition accuracy.

literature. As is mentioned in the previous Section 2.3.2, Sarkar et al. discussed several cases when gait silhouette segmentation errors occur in the HumanID Gait Challenge Problem dataset due to the shadow of the individuals, varying lighting conditions and moving objects in the background [18]. Liu and Sarkar observed that the drop in gait silhouette segmentation quality may lead to a decrease in recognition accuracy [102]. They also observed that if gallery and probe gait sequences are captured under the same conditions, and are segmented by the same method, the recognition accuracy may be high even if the data quality is poor. Zhang et al. address the issue of poor recognition accuracy when low-resolution gait silhouettes are used [103]. They proposed to combine super-resolution with multi linear tensor-based learning without parameters (SRMTP) to overcome this problem. However, they only focused on the low-resolution problem, and tested their algorithm on an artificial dataset, thus the issue of silhouette quality remains unexplored in practical scenarios.

Hence we consider the situation in which the gait data related to an individual to be recognised (probe data) are not captured under ideal conditions, and therefore the associated gait silhouettes may be noisy and inaccurately segmented; whereas the stored gait data (gallery data) are well segmented, or vice versa. This is a common situation encountered in practice; for example, when the probe data is captured using CCTV cameras at low resolution and poor quality, but the gallery data is previously captured under ideal conditions and it is not feasible to re-capture the probe data under the same ideal conditions. Based on these scenarios, we employ various segmentation algorithms to generate different silhouette quality data using sequences from the CASIA Dataset B, and propose two methods to tackle this problem. In Guan et al.'s work [21], it is noticed that the effectiveness of majority voting lies on the number of weak classifiers that make the right decision. In this case, if the weak classifiers are generated using ViFS, they might have higher chance to make right decision, since ViFS reconstruct gallery features by minimising the distance between gallery feature set and probe feature. If the number of right decisions are above 50%, the majority voting could return the right decision of classification. Thus we propose a classier fusion strategy based on least square QR- decomposition (LSQR). Our approach uses GEI as gait template, as it is one of the most popular and efficient methods to represent gait features. We first create a dataset by employing different segmentation algorithms on gait video sequences to generate silhouettes with segmentation disparities. The dataset is divided into training, gallery and probe sets, where the training set is for discriminant learning. We project gallery and probe data into a discriminant subspace to generate gallery and probe feature sets. The gallery features are fused using LSQR, thus generating more gallery representations, which are considered as weak classifiers to match with probe features. The output of all classifiers goes through a majority voting process, where the voting result represents the final classification decision. The application of voting in obtaining valuable features from massive data (with redundancy) was addressed in the mid 1950s, and has been studied till recently [130]. It is suggested that majority voting scheme has been applied in the fusion of data obtained from multiple sources [131]. Classifier fusion has been studied by many literatures and is proved to be efficient in biometric field [21]. Local Fisher Discriminant Analysis (LFDA) is employed as the discriminant learning approach. As presented in Section 5.3.2, evaluation results show that our fusion strategy improves recognition accuracy compared to using only LFDA, or using a fusion strategy that assigns equal importance to all features.

We then extended the classifier fusion strategy to a global regression method. The classifier fusion strategy makes use of information from weak classifiers, and use voting to increase the probability of the right guess. However, we found out that the least square regression method takes advantage of global information, and can generates the optimal weight for each gallery vector at the same time. We employ two popular subspace learning methods, namely, Linear Discriminant Analysis (LDA) and Locality Preserving Projection (LPP), to confirm the improvement brought by regression approach. As a traditional discriminant learning approach, LDA shows great consistency in recent studies [21], while LPP as a classical manifold learning approach can preserve the local structure information. Despite that LFDA and its extension ViDP outperform LDA and LPP on cross-view gait recognition challenge, according to the experiment results on quality disparity problem, its performance is inferior to the results using LDA and LPP. Since it is recently proposed, the consistency of LFDA is yet to be testified. Experimental results show that the extended method attains higher recognition accuracy, making it a promising solution to reduce the negative effects of poor gait silhouette quality on individual recognition.

5.2 Dataset

In order to build a gait dataset containing silhouettes with different segmentation discrepancies, we combine background subtraction (BS), de-noising, and frame differentiation (FD), to generate different segmentation approaches. We also employ the Gaussian Mixture Model and Expectation Maximization (GMM & EM) segmentation method [18], and the Least Median of Squares (LMedS) segmentation method [100]. The segmented silhouettes obtained by each of these approaches is used to generate binary images (and GEIs) at a specific quality. The quality levels

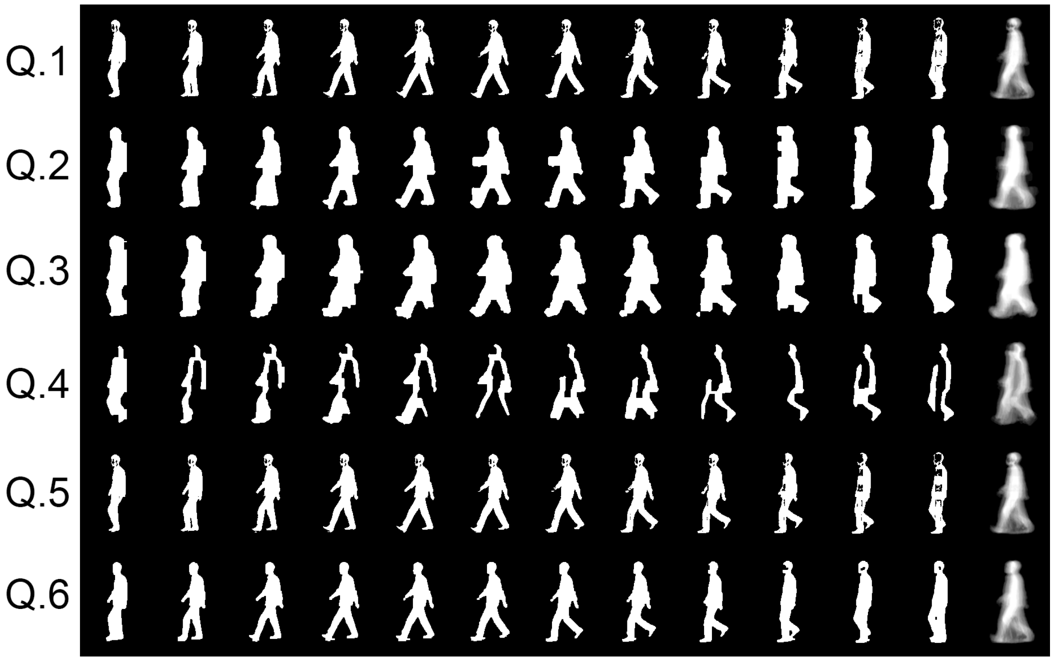


Figure 5.2: Samples of gait silhouette and corresponding GEIs with different qualities (Q.1 to Q.6) for the same subject. For each row, the first 11 images are the binary silhouettes obtained after segmentation, while the rightmost image is the corresponding GEI.

and the corresponding segmentation approaches used are listed in Table 5.1. The segmentation approaches are explained in the following paragraphs.

- *Approach 1*: A pixel is marked as foreground if $|I_t - B_t| > threshold$, where I_t refers to an image with both foreground and background objects and B_t contains only background objects. The threshold is set using Otsu's method, as it is the most commonly used tool in image segmentation [132].
- *Approach 2*: The background image is normalized to eliminate the negative effects of noise. Thus $|I_t - \hat{B}_t| > threshold$ where $\hat{B}_t = B_t / \sum p_{i,j}$; $p_{i,j}$ refers to the value of pixel $\{i, j\}$ in B_t. The threshold is set using Otsu's method. As the obtained foreground may comprise several disconnected regions, dilation

Table 5.1: Segmentation approaches for generating various data qualities, and the corresponding notation.

Quality	Segmentation Approach
Q.1	Approach 1: BS with Otsu's threshold
Q.2	Approach 2: Normalised BS plus dilation & erosion
Q.3	Approach 3: BS with small threshold (1/3 of Otsu's)
Q.4	Approach 4: FD plus dilation & erosion
Q.5	Approach 5: GMM & EM method
Q.6	Approach 6: LMedS method

and erosion operations are performed to generate the final foreground.

- *Approach 3*: A small threshold is used in order to introduce a distinct contrast in the segmented silhouettes and to include more background objects in the foreground, namely $|I_t - B_t| > threshold/3$. By setting the threshold to 1/3, more information on the background will be classified as foreground, thus add artificial segmentation error to the segmentation results.

- *Approach 4*: Frame difference (FD) approach is used to mark the moving foreground pixels, $I_t - I_{t-1} > threshold$, where the threshold is set using Otsu's method. In addition, dilation and erosion operations are used in order to connect the disconnected regions comprising the foreground.

- *Approach 5*: The GMM and EM method, as introduced by Sarkar et al. [18]. It is regarded as the baseline foreground detection algorithm for human gait recognition.

- *Approach 6*: The LMedS method, as is introduced in [100]. The gait silhouette images in CASIA Dataset B are generated using LMedS method. As the official segmentation results, they should be very close to the ground-truth results.

Using the distinct segmentation approaches tabulated in Table 5.1, each gait sequence can generate six sequences with different segmentation discrepancies. For each of these six sequences, we compute the corresponding GEI. Figure 5.2 shows sample GEIs with the six different qualities.

5.3 Solution 1: the Weak Classifier Ensemble Approach

It is mentioned in [133] that an efficient way of combining classifiers is to put them into groups and apply a different fusion strategy to each group. At the beginning phase of our research, we came up with a weak classifier assembling method, inspired by Guan's method in [21]. Using the dataset introduced in Section 5.2, we projected gallery and probe data into the discriminant subspace to generate gallery and probe feature sets. The gallery features were fused using Least Square QR-decomposition (LSQR) method, thus generate more gallery representations, which were considered as classifiers to match with probe features. The output of all classifiers went through majority voting process and generated the final classification decision. Local Fisher Discriminant Analysis (LFDA) was implemented as the feature enhancement approach. Experimental results show that the fusion based discriminant learning approach improves recognition accuracy compared to using only LFDA without fusion, or using equal-split fusion strategy, making it a suitable solution to reduce the negative effects of gait silhouette quality problem.

5.3.1 The Proposed Framework

Figure 5.3 shows the block diagram of the proposed weak classifier fusion approach.

We select k of the total n eigenvectors generated by PCA, when the sum of the k corresponding eigenvalues are above 99% of the sum of all eigenvalues. LFDA

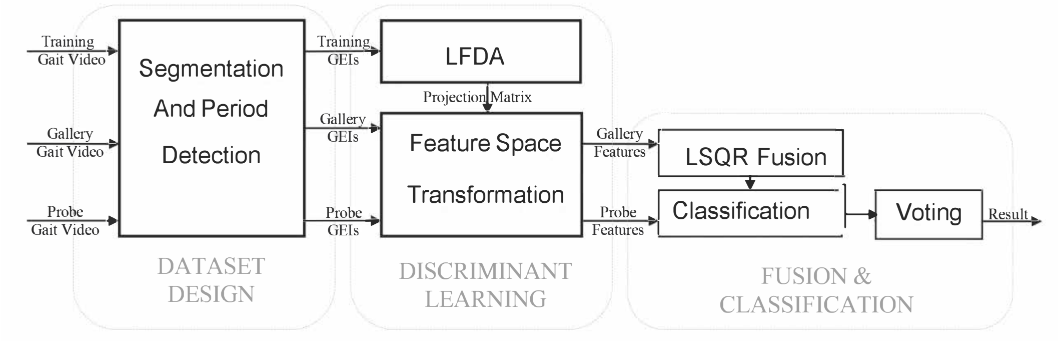


Figure 5.3: The block diagram of weak classifier fusion and voting approach.

is employed after reducing the dimensions of the training GEIs by using PCA. The generated transformation matrix W_{trans} is $W_{trans} = W_{LFDA}^{\top} W_{PCA}^{\top}$. We keep $(c-1)$ eigenvectors with k largest eigenvalues respectively for LFDA.

An important factor to consider during grouping is the level of diversity of classifier types. However, it is hard to acquire prior knowledge of the optimal strategy for grouping classifiers and applying fusion strategies. In this work, the gallery features with different segmentation discrepancies are fused in an exhaustive manner using a set of weights generated by LSQR, and each set of fused features is considered as one classifier. In other words, each distinct classifier is created by fusing gallery features with three different qualities. With 6 qualities in total, the number of generated fusion classifiers is $N_c = 6!/((6-3)! * 3!) = 20$. A set of weights are assigned to the three gallery features to be fused. For example, if a classifier comprises gallery features at qualities Q.1, Q.2 and Q.3, the features are fused as $g_f = w_1 * gallery_{Q.1} + w_2 * gallery_{Q.2} + w_3 * gallery_{Q.3}$, where $w_i, i \in \{1, 2, 3\}$ are the corresponding weights for the gallery features. The values of weights w_i are calculated as a vector by:

$$w = \arg\min_{w} \|gallery * w^{\top} - probe\|, \tag{5.1}$$

where

$$Gallerys = \begin{bmatrix} gallery_{Q.1} \\ gallery_{Q.2} \\ gallery_{Q.3} \end{bmatrix},$$

and $gallery_i, i \in 1, 2, 3$ are reshaped to column vectors. $Probe$ is the probe feature in column vector's form, too. Here $gallery$ and $probe$ are column vectors containing the gallery data at different qualities to be fused, and the corresponding probe data, respectively.

The set of gallery GEI vectors is denoted as $\mathbf{g}$, and the set of probe GEI vectors as $\mathbf{p}$. Following the subspace transformation processes in Section 2.4.1, the gallery feature sets $\{\hat{\mathbf{g}}\}$ and probe feature sets $\{\hat{\mathbf{p}}\}$ are obtained as follows:

$$\begin{aligned} &\{\hat{\mathbf{g}}\} : \hat{g_i} = W_{trans}G_i,\ i \in \{1, 2, ..., n_1\} \quad and \\ &\{\hat{\mathbf{p}}\} : \hat{p_j} = W_{trans}P_j,\ j \in \{1, 2, ..., n_2\} \end{aligned} \tag{5.2}$$

where n_1, n_2 are the total number of GEIs in gallery and probe data sets, respectively; G_i and P_i denote the GEI representations of gallery and probe data, respectively. The centroid of class l in $\{\hat{\mathbf{g}}\}$ is calculated as $mg_l = \frac{1}{n_l}\sum_{\hat{\mathbf{g}} \in \hat{\mathfrak{g}_l}} \hat{\mathbf{g}}$, where $\hat{\mathbf{g}_l}$ is the set of gallery feature vectors in class l. The centroid of class l in $\{\hat{\mathbf{p}}\}$ is calculated in the same way and is denoted as mp_l. The classifier is then defined as:

$$D(mg_l, mp_i) = \|mp_i - mg_l\|, i = 1, 2, ...c. \tag{5.3}$$

If $D(mg_l, mp_l) = \min_{i=1}^{c} D(mg_l, mp_i)$, the probe feature vector is assigned to the right class label l.

For each case, i.e. whenever a probe data comes, 20 classification results are

Table 5.2: The recognition rates in percentage (%) without discriminant learning. G: gallery data; p: probe data. The bold diagonal value indicates the recognition of using gallery and probe with same quality

G \ P	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6
Q.1	**85**	12	7	10	80	70
Q.2	12	**67**	17	8	10	35
Q.3	17	15	**78**	5	17	8
Q.4	15	8	5	**38**	18	15
Q.5	83	12	7	13	**83**	63
Q.6	58	25	5	10	43	**97**

obtained by the similarity measurement between the probe feature and 20 classifiers generated by the fusion approach. The results of the multiple classifiers will go through the majority voting process to obtain the final recognition rates.

5.3.2 Experimental evaluation

We use the gait sequences of CASIA Dataset B to generate the GEIs with 6 different silhouette qualities, as it is the only available dataset containing video sequences for us. CASIA Dataset B comprises video sequences for 124 individuals. We use the video sequences of first 62 individuals from the gallery dataset as the training data. The gallery and probe data are from the rest 62 individuals. The frame size of the gait video is 320×240, and the frame rate is 25 fps. All 6 segmentation algorithms produce binary silhouettes with the size of 128×88. As this work aims at studying the effect of gait silhouette quality on recognition, other factors that may influence the recognition performance are excluded. Therefore, only normal gait sequences are chosen from CASIA Dataset B, without the factors of carrying bags, different clothes, or different view angles.

We first evaluate the recognition rates of gallery and probe GEIs without discriminant learning or fusion. The recognition rates, in percentage, are shown in

Table 5.2. Two observations can be drawn from this table:

1. The entries in the main diagonal represent the matching results between gallery and probe data with the same segmentation errors. These values are generally the highest values among each column, suggesting when both gallery and probe data have same segmentation error, the best matching results are attained. However, matching cross gallery and probe with different segmentation errors obtains much lower result in most cases, with some exceptions, for example the matching rates cross Q.1 and Q.5 are almost the same. It indicates that different segmentation approaches might generate silhouettes with similar quality.

2. The entries outside the main diagonal show that the segmentation error disparity between gallery and probe data indeed decreases the recognition accuracy. In some cases, the matching rate between data segmented using the same approach can still be very low, which indicates that the segmentation approach may be inappropriate for the sequences (see for example Q.4 gallery matched with Q.4 probe).

Evaluation With Discriminant Learning

We measured the similarity between gallery and probe features generated by subspace transformation. The recognition rates are shown in Table 5.3. Note that by using dimensionality reduction plus a subspace learning method, the figures in Table 5.3 significantly improved comparing with the figures in Table 5.2.

Table 5.3: The recognition rates in percentage (%) using LFDA. Bold marks the diagonals, i.e. gallery and probe data have same quality.

G \ P	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6
Q.1	**95**	75	63.3	20	93.3	95
Q.2	85	**85**	83.3	30	78.3	91.7
Q.3	68.3	75	**95**	33.3	66.7	81.7
Q.4	48.3	46.7	70	**61.7**	56.7	68.3
Q.5	95	75	56.7	21.7	**95**	96.7
Q.6	88.3	66.7	65	23.3	85	**100**

Fusion Between Qualities

Based on the descriptions in Section 5.2, we fuse gallery features with different segmentation errors using weights computed by LSQR method. Measuring the distances between the fused gallery features (classifiers) and probe features with all kinds of segmentation errors generates a set of results, which are put into majority voting process to generate the final identification result. In addition, we would like to introduce the case where the particular type of segmentation error contained in probe data is not present in the gallery data. This attempts to represent the situation where the quality of the probe data is different from that of the gallery data. A comparison for different algorithms is tabulated in Table 5.4. The notations for this table are displayed as follows: DL(A): discriminant learning (lfda), and average through columns (column means of table 5.3); DL(H): the highest rates among each column of table 5.3; FDL(S): discriminant learning plus fusion, the weight is set split-equal ($w = 1/3$ for each of three gallery features to be fused); FDL: discriminant learning plus lsqr fusion, i.e. the proposed approach; FDL(I): incomplete fusion when the matching of gallery and probe data with same segmentation error is missing. The improvement of the fusion based discriminant learning approach is evident compared to the recognition performance using single classifier. The fu-

Table 5.4: The recognition rates in percentage (%) for probe data with 6 different quality. DL(A): average rates of lfda; DL(H): highest rates of lfda; FDL(S): fusion+lfda using split-equal weight; FDL: proposed approach; FDL(I): dealing with incomplete gallery data using proposed approach

Alg.	Probe						
	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Avg.
DL(A)	80	70.6	72.2	31.7	76.3	87	**68.3**
DL(H)	95	85	95	61.7	95	100	**88.6**
FDL(S)	90	78.3	83.3	33.3	88.3	96.7	**78.3**
FDL	95	85	90	58.3	95	98.3	**86.9**
FDL(I)	95	76.7	73.3	23.3	93.3	95	**76.1**

sion strategy using LSQR has better result than using equal-split fusion. Besides, when no gallery data has the same segmentation error as probe data, the proposed approach still shows promising performance (The result of this case is denoted as FDL(I) in Table 5.4).

Discussions

Experiment results suggests that the quality of the extracted binary silhouette images is particularly important for model-free gait recognition algorithms to perform accurately, since the segmentation errors might de-gradate the recognition rates. In [102], it is mentioned that the low quality silhouette may provide powerful features for gait recognition when both gallery and probe data have the same quality (acquired under same circumstances). However, through our analysis, an inaccurate segmentation could lead to very low recognition rate, even when employing discriminant learning methods, see for example results for probe data with Q.4 in Table 5.3.

One main shortcoming of the proposed method, as in any other discriminant learning method, is that subjects represented by very low quality silhouettes cannot be recognized accurately. Nevertheless, if the quality of gallery and probe data

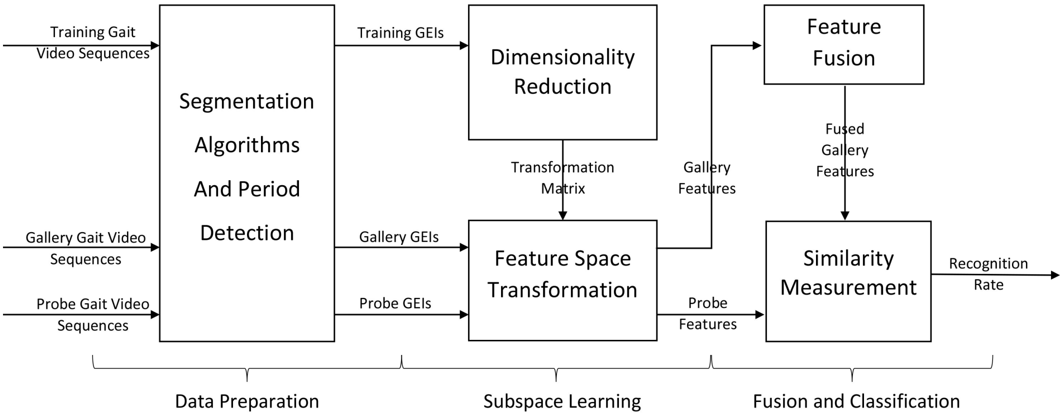


Figure 5.4: Block diagram of the proposed framework.

are acceptable, even if they are very different, our fusion based subspace learning approach can improve the matching performance to a promising level.

5.4 Solution 2: ViFS Based Quality-invariant Approach

5.4.1 The Proposed Framework

As argued in Section 2.2.1, *silhouette quality disparity* are summarised as global deformation challenge, identical to *gait recognition under arbitrary view.* Hence we assume that ViFS can still be effective in tackling quality disparity problem. In this section we introduce a quality-invariant gait recognition framework using ViFS. The block diagram of the proposed framework is shown in Fig. 5.4. We first generate the GEIs with different qualities for the training, gallery and probe data. Using the training data, we compute the transformation matrix corresponding to the subspace learning method. We fuse gallery data with a set of weights computed by least square fitting. Fused gallery data and probe data are transformed into fused gallery features and probe features in a lower dimension space. Finally, we measure the similarities between fused gallery features and probe features.

We use the same dataset as mentioned in Section 5.2. Different subspace learning methods may be used to project the data onto a feature subspace. In this work, to avoid singularity problems in computation, we employ PCA before implementing LDA, or LPP. The generated transformation matrix W_{trans} is then $W_{trans} = W_s^\top W_{PCA}^\top$, where $s \in \{LDA, LPP\}$.

The feature sets after projection are:

$$\begin{aligned} \{\hat{\mathbf{g}}\} &: \hat{g}_i = W_{trans} G_i \\ \{\hat{\mathbf{p}}\} &: \hat{p}_j = W_{trans} P_j \end{aligned} \tag{5.4}$$

where $i = \{1, 2, ..., n_1\}$, $j = \{1, 2, ..., n_2\}$, and n_1, n_2 are the total number of GEIs in gallery and probe data sets, respectively. The centroid of class l in $\{\hat{\mathbf{g}}\}$ is calculated as $\mathbf{mg_l} = \frac{1}{n_l} \sum_{\hat{\mathbf{g}} \in \hat{\mathbf{g}}_l} \hat{g}_i$, where $\hat{\mathbf{g}}_l$ is the set of gallery feature vectors in class l. The centroid of class l in $\{\hat{\mathbf{p}}\}$ is calculated in the same way and is denoted as mp_l. The classifier is then defined as:

$$D(\mathbf{mg_l}, \mathbf{mp_i}) = \|\mathbf{mp_i} - \mathbf{mg_l}\|, i = 1, 2, ...c. \tag{5.5}$$

If $D(\mathbf{mg_l}, \mathbf{mp_l}) = \min_{i=1}^{c} D(\mathbf{mg_l}, \mathbf{mp_i})$, the probe feature vector is assigned to the right class label.

5.4.2 Experimental Evaluation

In order to evaluate the framework, we use the gait sequences of CASIA B dataset to generate the GEIs at different qualities. CASIA B dataset comprises video sequences for 124 individuals. The frame size is 320×240, and the frame rate is 25 fps. As this work aims at studying the effect of gait silhouette quality on recognition, other factors that may influence the recognition performance are excluded. Therefore, only normal gait sequences are chosen form CASIA B, without the factors of carrying bags, different clothes, different view angles, etc.

Evaluation with subspace learning

Tables 5.5 and 5.6 tabulate the average matching rates in percentage, after using PCA+LDA and PCA+LPP, respectively. It is important to recall that in practical cases, the quality of the gallery and probe data may differ. It is then important for recognition algorithms to maintain a high accuracy even in this situation. Therefore, we measure the similarity between each individual in the probe data set against all individuals in the gallery data set for all qualities except for the quality of the probe data. This scenario corresponds to the empty entries in Table 5.5 and Table 5.6. These two tables show that by using dimensionality reduction plus a subspace learning method, matching rates can be considerably improved. Note that LPP can effectively deal with poor quality matching, i.e., qualities different from Q.6, while LDA appears to perform better than LPP with high quality matching, i.e., Q.6 data.

Evaluation with fusion strategy

We propose to further improve the performance of subspace learning methods by fusing gallery data before matching with probe data features. Specially, we fuse the multi GEI representation of the gallery subjects, which consists of various quality levels, into one GEI representation.

Before feature space transformation, we compute a set of weights to be used in fusion strategy. In this experiment, for each probe data of a specific quality, there are gallery data of 5 different qualities available for fusion. For example, for probe data of quality Q.1, we fuse gallery data of all qualities except quality Q.1. The set of weights for fusion are computed using least squares fitting. Specifically, we aim to find the combination of weights for the gallery data centroids, i.e., gallery data of different qualities, that best match the centroid of the probe data of a specific quality. Let us denote the set of GEI vectors in the gallery set as $\mathbf{G} = \{\mathbf{g_1}, \mathbf{g_2}, ..., \mathbf{g_n}\}$, for n different qualities. Let us also denote the probe GEI vector as $\mathbf{p}$. The set of weights $\mathbf{w} = \{w_1, w_2, ..., w_n\}$ for the gallery data of n different qualities is then computed as follows:

$$\mathbf{w} = \arg\min_{\mathbf{w}} \|\mathbf{G} * \mathbf{w}^\top - \mathbf{p}\|. \tag{5.6}$$

These weights are then used to fuse gallery data into $\mathbf{G_f}$:

$$\mathbf{G_f} = \sum_{i}^{n} \mathbf{g_i} * w_i, i \in \{1, 2, ..., n\}. \tag{5.7}$$

After fusion, the gallery data is projected into the feature space, where similarities are measured between probe and the fused gallery features. The results of this experiment are shown in Table 5.7. The first row 'LDA' shows the average matching rate with LDA only, while the second row 'LDAF' indicates the rates after applying LDA and feature fusion. It is shown that apart from the first column with quality 'Q.1', 'LDAF' outperforms 'LDA'. Recall in Table 5.5 that Q.3 has high matching rate against Q.2, while Q.5 has high matching rate against Q.1. Thus when fusing gallery features of 5 different quality, ViFS is able to assign higher weight to those with good matching rates (Q.2 and Q.2 respectively, in this case), and eventually generate a reconstruction that preserve most useful features. Same mechanism happens to the case of 'LPP' and 'LPPF'. Generally, by applying ViFS

Table 5.5: Matching rates between gallery data (G) and probe data (P) using PCA+LDA (%)

G \ P	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6
Q.1		73.3	63.3	28.3	93.3	95
Q.2	85		86.7	31.7	75	90
Q.3	71.7	73.3		28.3	70	83.3
Q.4	61.7	51.7	71.7		58.3	70
Q.5	95	73.3	63.3	23.3		96.7
Q.6	85	70	66.7	28.3	85	

Table 5.6: Matching rates between gallery data (G) and probe data (P) using PCA+LPP (%)

G \ P	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6
Q.1		71.7	75	28.3	96.7	93.3
Q.2	83.3		85	35	83.3	93.3
Q.3	78.3	73.3		26.7	75	83.3
Q.4	66.7	53.3	63.3		66.7	65
Q.5	96.7	68.3	68.3	26.7		93.3
Q.6	88.3	65	66.7	20	91.7	

to fuse data of different quality, the average matching rate has 10% promotion.

Discussions

In cases when the data quality in the gallery set is different from that of the data in the probe set, the performance of recognition algorithms may be poor, making it hard to chose a dependable classifier. The fusion strategy proposed in this work finds the combination of gallery data that has a minimum distance to the probe data. This is done by finding a set of weights using least square fitting, which is efficient and parameter-free. In our experiment design, we assume that the silhouette quality of probe data does not match any of the quality in gallery data, which would be more frequently occurred for practical cases. As show in Table 5.7, this strategy can considerably improve recognition performance under such circumstances. It

Table 5.7: Average matching rate of six different quality gallery data (G) and probe data (P) using PCA+LDA (LDA) and PCA+LPP (LPP), and the matching rate of fused gallery data using PCA+LDA with feature fusion (LDAF) and PCA+LPP with feature fusion (LPPF)(%)

G \ P	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Average
LDA	81	72	74	32	79	88	**68.2**
LDAF	75	78.3	93.3	35	93.3	95	**78.3**
LPP	84	72	75	31	85	90	**69.4**
LPPF	78.3	75	90	31.7	95	95	**77.5**

is important to mention that by introducing least square fitting, the multi-quality gallery data is fused to best fit the probe data, which is similar to the case where gallery and probe data are equally segmented, i.e. the diagonal data in Table 5.2.

5.5 Summary

In this section, we studied the performance of GEI-based gait recognition algorithms when a disparity in quality between gallery and probe data exists. The motivation is to tackle the problem where gallery and probe data are segmented using different algorithms. We first evaluate the matching result of gallery and probe GEIs with no dimensionality reduction or subspace learning method.

It is suggested that different segmentation algorithms might generate similar features, while in most cases different segmentation algorithm or de-noising process could lead to large quality disparity and thus influence the accuracy of recognition. To this end, we generated gait silhouettes with different segmentation discrepancies in order to represent different levels of data qualities. To perform recognition, we study the use of subspace learning methods after dimensionality reduction by PCA. Simulation experiments on the CASIA B dataset using LDA and LPP indicate that gait recognition is indeed affected if the quality of the probe data set differs from that of the gallery data set. Results also suggest that important improvements in matching rate may be attained when subspace learning methods are used, since the feature subspace finds the best projection to match probe with gallery features of the same quality level. The paper also presented a fusion strategy that fuses gallery data of different qualities before feature space transformation. Experiments showed that this fusion strategy, which employs a number of weights, can further improve matching rates.

A classifier fusion strategy in conjunction with discriminant learning was proposed to tackle the negative impact of quality disparity on matching rate. Specially, we proposed to generate weights by using LSQR to fuse gallery features and generate several classifiers. We then proposed to use majority voting to compute the final classification result. Experimental results on the CASIA Dataset B suggested that this approach provides better performance than the case of using a single classifier and the case of employing fusion with equal weights. Furthermore, we discovered that the least square regression method on global information achieves better performance than ensemble of weak classifiers.

Chapter 6

Conclusion

In this thesis, we studied the silhouette global deformation challenge of gait recognition caused by silhouette quality disparity or camera view point variation. We analysed the limitations of existing machine learning techniques on hard deformation problems, and proposed a robust deformation-invariant framework to tackle this problem. We prepared specific solutions for different scenarios. For fast implementation with sufficient gallery data available, we proposed to use a linear and parameter-free framework as described in Chapter 3. On the other hand, when only a few gallery images are available, we proposed a CNN based framework to achieve deformation-invariant identification as described in Chapter 4. The proposed framework achieved the state-of-the-art results on mainstream datasets.

6.1 Thesis Contributions

The motivation of this thesis is to make gait recognition a reliable technique in remote human identification. At the present stage, camera view-point changes remains one of the unsolved challenges for gait recognition, while silhouette quality is another factor that hinder the performance of gait approaches, and yet less studied due to the lack of pertinence datasets. We summarise these two problems as one challenge, i.e. gait global deformation challenge, since both of these two factors cause global deformation on gait silhouette, and it is very difficult to establish a feature space mapping between a normal silhouette and a seriously deformed silhouette. Our objective is to develop a fast and accurate framework that is robust to silhouette deformation.

1. In Chapter 3, we proposed a multi-view feature selector, ViFS, and developed a robust framework based on ViFS and subspace feature enhancers to tackle the multi-view gait recognition problem. We tested the proposed framework on CASIA Dataset B, and evaluated the effect of LDA and SLPP as feature enhancers. The results indicated that if the gallery set covers 4 or more view angles, the matching accuracy of the proposed framework with a probe set containing different view angles can exceed 98%, on average. Furthermore, the proposed framework is parameter-free and has low computational cost, which indicates great potential for practical applications.

2. In Chapter 4, we explored the possibility of combining CNN feature maps with the proposed ViFS method. The experiment results suggest that CNN+ViFS has satisfying results when gallery view angle is equal or less than 2 views. The ViFS can successfully select view-invariant features from frontal or back view gallery data, while its effect is attenuated when gallery data from various lateral views exists. Our conclusion is that the hierarchical non-linearity feature maps extracted from CNN are not always compatible with the linear regression based ViFS method. Besides, the feature-level fusion is unstable comparing with image-level fusion.

3. Chapter 5 studies the performance of GEI-based gait recognition algorithms when a disparity in quality between gallery and probe data exists. To this end, we generate gait silhouettes with different SEs in order to represent different levels of data qualities and commonly encountered segmentation inaccuracies. A classifier fusion strategy in conjunction with discriminant learning methods is proposed to tackle the negative impact of quality disparity on matching rate. Specially, we automatically generate weights by using LSQR method. Experimental results on the CASIA Dataset B suggest that this strategy provides better performance than normal fusion strategy. However, the limitation of this paper is also obvious. The silhouette quality contains other aspects, for example illumination changes or shadow. Although our experiment design can simulate these cases to some extent, it is still necessary to perform this experiment on video sequences containing these variances. Unfortunately, there is no available database satisfying this requirements.

6.2 Limitations

1. For recognition under arbitrary view, the performance of the proposed framework relies on the effectiveness of feature enhancers, as well as the sufficient gallery data. As mentioned in Chapter 3, our framework has limited effect when the gallery set contains only single view data. Under such circumstances, the performance mainly depend on the enhancers, where CNN is proved to be most effective. In order to maximise its effect, we have to develop new approaches to establish a robust mapping between gallery and probe templates with large view variation (e.g. gallery from 0° and probe from 90°).

2. For silhouette quality problem, we only testify our method on a dataset with limited covariate factors, while its net effect on the real world scenario remains untested. Furthermore, despite that the modelling of segmentation and post-processing methods is easy, it might be difficult to model the shape deformation caused by scenario or camera variation. In addition, there are other factors that are seldom studied, for example the camera could be on top of the subject or at a lower position, resulting severe shape deformation on the gait silhouette. For these special problems, we require a dataset containing all the variations to evaluate the existing algorithms, as well as developing new approaches to overcome the covariate factors challenges.

6.3 Future Work

1. The emerging deep learning tools provide possibilities to overcome the weakness of the proposed ViFS method. The Generative Adversarial Network (GAN) is able to create new images from large-scale training set, thus one possible implementation is to train the network with gait templates from different views, and feed the network with gallery image from single view to generate images from other views. In such case the proposed ViFS framework can be well established and ready for probe image from arbitrary views.

2. For silhouette quality problem, we plan to evaluate the proposed framework on outdoor dataset, and analyse the ROC curve as well as equal error rate of the existing algorithms to establish benchmark for silhouette quality disparity problem under different cases. With sufficient experiment and analysis, more interesting points might be explored and help gait recognition to be applicable.

Bibliography

[1] J. Tang, J. Luo, T. Tjahjadi, and F. Guo, “Robust arbitrary-view gait recognition based on 3d partial similarity matching,” *IEEE Transactions on Image Processing*, vol. 26, no. 1, pp. 7–22, Jan 2017.

[2] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, “A comprehensive study on cross-view gait based human identification with deep cnns,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2016.

[3] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *Proceedings of European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 818–833.

[4] A. K. Jain, P. Flynn, and A. A. Ross, *Handbook of Biometrics.* Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[5] M. S. Nixon, T. Tan, and R. Chellappa, *Human identification based on gait.* Springer Science & Business Media, 2006, vol. 4.

[6] A. K. Jain, S. Prabhakar, and A. Ross, “Fingerprint matching: Data acquisition and performance evaluation,” *Dept. of Computer Science, Michigan State Univ., East Lansing, Tech. Rep. MSU-CPS-99–14*, 1999.

[7] J. Ashbourn, *Biometrics in the New World: The Cloud, Mobile Technology and Pervasive Identity.* Springer Publishing Company, Incorporated, 2014.

[8] C. Halder, S. M. Obaidullah, and K. Roy, "Effect of writer information on bangla handwritten character recognition," in *Proceedings of National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, Dec 2015.

[9] J. L. Wayman, *Biometric Verification/Identification/Authentication/Recognition: The Terminology.* Boston, MA: Springer US, 2009, pp. 153–157.

[10] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, and G. Wang, "Recent advances in convolutional neural networks," *Computing Research Repository*, vol. abs/1512.07108, 2015.

[11] S. Srinivas, R. K. Sarvadevabhatla, K. R. Mopuri, N. Prabhu, S. S. Kruthiventi, and R. V. Babu, "A taxonomy of deep convolutional neural nets for computer vision," *Frontiers in Robotics and AI*, vol. 2, p. 36, 2016.

[12] P. N. Druzhkov and V. D. Kustikova, "A survey of deep learning methods and software tools for image classification and object detection," *Pattern Recognition and Image Analysis*, vol. 26, no. 1, pp. 9–15, 2016.

[13] A. K. Jain, S. C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," in *Biometric Authentication*, 2004, pp. 731–738.

[14] D. Reid, S. Samangooei, C. Chen, M. Nixon, and A. Ross, "Soft biometrics for surveillance: an overview," *Machine learning: Theory and Applications. Elsevier*, pp. 327–352, 2013.

[15] A. A. Ross, K. Nandakumar, and A. Jain, *Handbook of multibiometrics.* Springer Science & Business Media, 2006, vol. 6.

[16] T. K. Lee, M. Belkhatir, and S. Sanei, "A comprehensive review of past and present vision-based techniques for gait recognition," *Multimedia Tools and Applications*, vol. 72, no. 3, pp. 2833–2869, 2014.

[17] D. Barrett, "One surveillance camera for every 11 people in britain, says cctv survey," *The Telegraph*, vol. 10, 2013.

[18] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162–177, 2005.

[19] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[20] B. Kulis, "Metric learning: A survey," *Foundations and Trends in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2012.

[21] Y. Guan, C.-T. Li, and F. Roli, "On reducing the effect of covariate factors in gait recognition: a classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 7, pp. 1521–1528, 2015.

[22] X. Zhou and B. Bhanu, "Feature fusion of side face and gait for video-based human identification," *Pattern Recognition*, vol. 41, no. 3, pp. 778 – 795, 2008, part Special issue: Feature Generation and Machine Learning for Robust Multimodal Biometrics.

[23] D. Muramatsu, H. Iwama, Y. Makihara, and Y. Yagi, “Multi-view multi-modal person authentication from a single walking image sequence,” in *Proceedings of the International Conference on Biometrics (ICB)*, 2013, pp. 1–8.

[24] P. K. Larsen, E. B. Simonsen, and N. Lynnerup, “Gait analysis in forensic medicine,” *Journal of Forensic Sciences*, vol. 53, no. 5, pp. 1149–1153, 2008.

[25] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon, “On using gait in forensic biometrics,” *Journal of Forensic Sciences*, vol. 56, no. 4, pp. 882–889, 2011.

[26] Aristotle, *On the motion of animals.* classics.mit.edu, 350 B.C.E. [Online]. Available: http://classics.mit.edu/Aristotle/motion_animals.html

[27] M. P. Murray, A. B. Drought, and R. C. Kory, “Walking patterns of normal men,” *J Bone Joint Surg Am*, vol. 46, no. 2, pp. 335–360, 1964.

[28] M. P. Murray, “Gait as a total pattern of movement: Including a bibliography on gait.” *American Journal of Physical Medicine & Rehabilitation*, vol. 46, no. 1, pp. 290–333, 1967.

[29] G. Johansson, “Visual perception of biological motion and a model for its analysis,” *Perception and Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[30] J. E. Cutting and L. T. Kozlowski, “Recognizing friends by their walk: Gait perception without familiarity cues,” *Bulletin of the Psychonomic Society*, vol. 9, no. 5, pp. 353–356, 1977.

[31] H. Murase and R. Sakai, “Moving object recognition in eigenspace representation: gait analysis and lip reading,” *Pattern Recognition Letters*, vol. 17, no. 2, pp. 155–162, 1996.

[32] J. Little and J. Boyd, “Recognizing people by their gait: the shape of motion,” *Videre: Journal of Computer Vision Research*, vol. 1, no. 2, pp. 1–32, 1998.

[33] C. BenAbdelkader, R. Cutler, and L. Davis, “Motion-based recognition of people in eigengait space,” in *Proceedings of the 5th IEEE International Conference onAutomatic Face and Gesture Recognition*, 2002, pp. 267–272.

[34] I. R. Vega and S. Sarkar, “Statistical motion model based on the change of feature relationships: human gait-based recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1323–1328, 2003.

[35] S. D. Mowbray and M. S. Nixon, “Automatic gait recognition via fourier descriptors of deformable objects,” in *Proceedings of the International Conference on Audio-and Video-Based Biometric Person Authentication*, 2003, pp. 566–573.

[36] J. B. Hayfron-Acquah, M. S. Nixon, and J. N. Carter, “Automatic gait recognition by symmetry analysis,” *Pattern Recognition Letters*, vol. 24, no. 13, pp. 2175–2183, 2003.

[37] Z. Liu and S. Sarkar, “Simplest representation yet for gait recognition: Averaged silhouette,” in *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 4, 2004, pp. 211–214.

[38] C.-S. Lee and A. Elgammal, “Gait style and gait content: bilinear models for gait recognition using gait re-sampling,” in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 147–152.

[39] A. Kale, A. Sundaresan, A. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of humans using gait," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1163–1173, 2004.

[40] T. Kobayashi and N. Otsu, "Action and simultaneous multiple-person identification using cubic higher-order local auto-correlation," in *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 4, 2004, pp. 741–744.

[41] M. Hofmann and G. Rigoll, "Improved gait recognition using gradient histogram energy image," in *Proceedings of 19th IEEE International Conference on Image Processing*, 2012, pp. 1389–1392.

[42] H. El-Alfy, I. Mitsugami, and Y. Yagi, "A new gait-based identification method using local gauss maps," in *Proceedins of the Asian Conference on Computer Vision*, 2014, pp. 3–18.

[43] Y. Chew-Yean and M. Nixon, "Model-based gait recognition," *Enclycopedia of Biometrics*, pp. 633–639, 2009.

[44] S. A. Niyogi and E. H. Adelson, "Analyzing gait with spatiotemporal surfaces," in *Proceedings of the 1994 IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, 1994, pp. 64–69.

[45] D. Cunado, J. M. Nash, M. S. Nixon, and J. N. Carter, "Gait extraction and description by evidence-gathering," in *Proceedings of International Conference on Audio and Video Based Biometric Person Authentication*, 1999, pp. 43–48.

[46] C. Yam, M. S. Nixon, and J. N. Carter, “Automated person recognition by walking and running via model-based approaches,” *Pattern Recognition*, vol. 37, no. 5, pp. 1057–1072, 2004.

[47] I. Bouchrika and M. S. Nixon, “Exploratory factor analysis of gait recognition,” in *Proceedings of 8th IEEE International Conference on Automatic Face and Gesture Recognition*, 2008, pp. 1–6.

[48] L. Wang, H. Ning, T. Tan, and W. Hu, “Fusion of static and dynamic body biometrics for gait recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 2, pp. 149–158, 2004.

[49] R. Urtasun and P. Fua, “3d tracking for gait characterization and recognition,” in *Proceedings of 6th IEEE International Conference on Automatic Face and Gesture Recognition*, 2004, pp. 17–22.

[50] G. Zhao, G. Liu, H. Li, and M. Pietikainen, “3d gait recognition using multiple cameras,” in *Proceedings of 7th International Conference on Automatic Face and Gesture Recognition*, 2006, pp. 529–534.

[51] D. Ioannidis, D. Tzovaras, I. G. Damousis, S. Argyropoulos, and K. Moustakas, “Gait recognition using compact feature extraction transforms and depth information,” *IEEE Transactions on Information Forensics and security*, vol. 2, no. 3, pp. 623–630, 2007.

[52] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes, “Gait energy volumes and frontal gait recognition using depth images,” in *Proceedings of International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–6.

[53] H. Nakajima, I. Mitsugami, and Y. Yagi, "Depth-based gait feature representation," *Information and Media Technologies*, vol. 8, no. 4, pp. 1085–1089, 2013.

[54] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll, "The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits," *Journal of Visual Communication and Image Representation*, vol. 25, no. 1, pp. 195–206, 2014.

[55] T. H. Lam and R. S. Lee, "A new representation for human gait recognition: motion silhouettes image (msi)," in *Proceedings of International Conference on Biometrics (ICB)*, 2006, pp. 612–618.

[56] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan, "Human identification using temporal information preserving gait template," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2164–2176, Nov 2012.

[57] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *Proceedings of 3rd International Conference on Crime Detection and Prevention (ICDP)*, Dec 2009, pp. 1–6.

[58] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1511–1521, 2012.

[59] Y. Makihara, D. S. Matovski, M. S. Nixon, J. N. Carter, and Y. Yagi, "Gait recognition: Databases, representations, and applications," *Wiley Encyclopedia of Electrical and Electronics Engineering*, 2015.

[60] E. Zhang, Y. Zhao, and W. Xiong, "Active energy image plus 2dlpp for gait recognition," *Signal Processing*, vol. 90, no. 7, pp. 2295–2302, 2010.

[61] Y. Liu, R. T. Collins, and Y. Tsin, "A computational model for periodic pattern perception based on frieze and wallpaper groups," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 3, pp. 354–371, March 2004.

[62] S. Lee, Y. Liu, and R. Collins, "Shape variation based frieze pattern for robust gait recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[63] M. A. Hossain, Y. Makihara, J. Wang, and Y. Yagi, "Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control," *Pattern Recognition*, vol. 43, no. 6, pp. 2281–2291, 2010.

[64] C. Luo, W. Xu, and C. Zhu, "Robust gait recognition based on partitioning and canonical correlation analysis," in *Proceedings of IEEE International Conference on Imaging Systems and Techniques (IST)*, Sept 2015, pp. 1–5.

[65] G. Ma, Y. Wang, and L. Wu, "Subspace ensemble learning via totally-corrective boosting for gait recognition," *Neurocomputing*, Mar 2016.

[66] M. Goffredo, J. N. Carter, and M. S. Nixon, "Front-view gait recognition," in *Proceedings of IEEE International Conference on Biometrics: Theory, Applications and Systems BTAS*, 2008, pp. 1–6.

[67] S. Yu, D. Tan, and T. Tan, “A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition,” in *Proceedings of 18th International Conference on Pattern Recognition (ICPR)*, vol. 4, 2006, pp. 441–444.

[68] Y. Makihara, H. Mannami, A. Tsuji, M. A. Hossain, K. Sugiura, A. Mori, and Y. Yagi, “The ou-isir gait database comprising the treadmill dataset,” *IPSJ Transactions on Computer Vision and Applications*, vol. 4, no. 0, pp. 53–62, 2012.

[69] M. Sandau, R. V. Heimbrger, K. E. Jensen, T. B. Moeslund, H. Aans, T. Alkjr, and E. B. Simonsen, “Reliable gait recognition using 3d reconstructions and random forests an anthropometric approach,” *Journal of Forensic Sciences*, vol. 61, no. 3, pp. 637–648, 2016.

[70] A. Laurentini, “The visual hull concept for silhouette-based image understanding,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 150–162, 1994.

[71] G. Shakhnarovich, L. Lee, and T. Darrell, “Integrated face and gait recognition from multiple views,” in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, 2001, pp. I–439–I–446 vol.1.

[72] R. D. Seely, S. Samangooei, M. Lee, J. N. Carter, and M. S. Nixon, “The university of southampton multi-biometric tunnel and introducing a novel 3d gait dataset,” in *Proceedings of IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2008, pp. 1–6.

[73] G. Ariyanto and M. S. Nixon, "Model-based 3d gait biometrics," in *Proceedings of International Joint Conference on Biometrics (IJCB)*, 2011, pp. 1–7.

[74] B. Kwolek, T. Krzeszowski, A. Michalczuk, and H. Josinski, "3d gait recognition using spatio-temporal motion descriptors," in *Proceedings of Asian Conference on Intelligent Information and Database Systems*, 2014, pp. 595–604.

[75] F. Ahmed, P. P. Paul, and M. L. Gavrilova, "Dtw-based kernel and rank-level fusion for 3d gait recognition using kinect," *The Visual Computer*, vol. 31, no. 6-8, pp. 915–924, 2015.

[76] W. Kusakunniran, Q. Wu, J. Zhang, Y. Ma, and H. Li, "A new view-invariant feature for cross-view gait recognition," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 10, pp. 1642–1653, 2013.

[77] M. Goffredo, I. Bouchrika, J. Carter, and M. Nixon, "Self-calibrating view-invariant gait biometrics," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 4, pp. 997–1008, Aug 2010.

[78] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1700–1715, 2007.

[79] M. Hu, Y. Wang, Z. Zhang, J. Little, and D. Huang, "View-invariant discriminative projection for multi-view gait-based human identification," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 12, pp. 2034–2045, Dec 2013.

[80] A. Mansur, Y. Makihara, D. Muramatsu, and Y. Yagi, "Cross-view gait recognition using view-dependent discriminative analysis," in *Proceedings of IEEE International Joint Conference on Biometrics (IJCB)*, Sept 2014, pp. 1–8.

[81] H. Hu, "Multiview gait recognition based on patch distribution features and uncorrelated multilinear sparse local discriminant canonical correlation analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 4, pp. 617–630, April 2014.

[82] S. Zheng, J. Zhang, K. Huang, R. He, and T. Tan, "Robust view transformation model for gait recognition," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Sept 2011, pp. 2073–2076.

[83] D. Muramatsu, A. Shiraishi, Y. Makihara, M. Uddin, and Y. Yagi, "Gait-based person recognition using arbitrary view transformation model," *IEEE Transactions on Image Processing*, vol. 24, no. 1, pp. 140–154, Jan 2015.

[84] D. Muramatsu, Y. Makihara, and Y. Yagi, "View transformation model incorporating quality measures for cross-view gait recognition," *IEEE Transactions on Cybernetics*, vol. PP, no. 99, pp. 1–1, 2015.

[85] S. Huang, A. Elgammal, J. Lu, and D. Yang, "Cross-speed gait recognition using speed-invariant gait templates and globality–locality preserving projections," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 10, pp. 2071–2083, 2015.

[86] H. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Gait recognition through mpca plus lda," in *2006 Biometrics Symposium: Special Session on Research at the Biometric Consortium Conference*, Sept 2006, pp. 1–6.

[87] ——, “Boosting lda with regularization on mpca features for gait recognition,” in *Proceedings of Biometrics Symposium*, Sept 2007, pp. 1–6.

[88] Z. Fan, J. Jiang, S. Weng, Z. He, and Z. Liu, “Human gait recognition based on discrete cosine transform and linear discriminant analysis,” in *Proceedings of IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Aug 2016, pp. 1–6.

[89] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, “Which reference view is effective for gait identification using a view transformation model?” in *Proceedings of Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, June 2006, pp. 45–45.

[90] W. Kusakunniran, Q. Wu, J. Zhang, H. Li, and L. Wang, “Recognizing gaits across views through correlated motion co-clustering,” *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 696–709, Feb 2014.

[91] D. Zhang, Y. Wang, Z. Zhang, and M. Hu, “Ethnicity classification based on fusion of face and gait,” in *Proceedings of IAPR International Conference on Biometrics (ICB)*, March 2012, pp. 384–389.

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

[93] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” in *Proceeding of International Conference on Learning Representations (ICLR)*, 2013, p. 16.

[94] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[95] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[96] M. Alotaibi and A. Mahmood, "Improved gait recognition based on specialized deep convolutional neural networks," in *Proceedings of IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, 2015, pp. 1–7.

[97] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Geinet: View-invariant gait recognition using a convolutional neural network," in *Proceedings of International Conference on Biometrics (ICB)*, June 2016, pp. 1–8.

[98] A. Hicklin and R. Khanna, "The role of data quality in biometric systems," *White Paper. Mitretek Systems*, pp. 1–77, Feb 2006.

[99] D. Matovski, M. Nixon, S. Mahmoodi, and T. Mansfield, "On including quality in applied automatic gait recognition," in *Proceedings of 21st International Conference on Pattern Recognition (ICPR)*, Nov 2012, pp. 3272–3275.

[100] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1505–1518, 2003.

[101] R. Chellappa, A. Veeraraghavan, and N. Ramanathan, "Gait biometrics, overview," in *Encyclopedia of Biometrics*, S. Li and A. Jain, Eds. Springer US, 2009, pp. 628–633.

[102] Z. Liu and S. Sarkar, "Effect of silhouette quality on hard problems in gait recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 35, no. 2, pp. 170–183, 2005.

[103] J. Zhang, J. Pu, C. Chen, and R. Fleischer, "Low-resolution gait recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 4, pp. 986–996, Aug 2010.

[104] Y. Guan, Y. Sun, C.-T. Li, and M. Tistarelli, "Human gait identification from extremely low-quality videos: an enhanced classifier ensemble method," *IET Biometrics*, vol. 3, no. 2, pp. 84–93, 2014.

[105] N. Akae, A. Mansur, Y. Makihara, and Y. Yagi, "Video from nearly still: An application to low frame-rate gait recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012, pp. 1537–1543.

[106] Y. Guan, C.-T. Li, and S. D. Choudhury, "Robust gait recognition from extremely low frame-rate videos," in *Proceedings of International Workshop on Biometrics and Forensics (IWBF)*, 2013, pp. 1–4.

[107] A. Webb and K. Copsey, *Statistical Pattern Recognition*, 3rd ed. Chichester: Wiley, 2011.

[108] X. He and P. Niyogi, "Locality preserving projections," *Advances in Neural Information Processing Systems(NIPS)*, vol. 16, pp. 234–241, 2003.

[109] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, pp. 328–340, March 2005.

[110] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.

[111] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *The Journal of Machine Learning Research*, vol. 8, pp. 1027–1061, 2007.

[112] J. Yang, D. Zhang, A. F. Frangi, and J.-y. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131–137, 2004.

[113] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[114] Z. Wu, Y. Huang, Y. Yu, L. Wang, and T. Tan, "Early hierarchical contexts learned by convolutional networks for image segmentation," in *Proceedings of International Conference onPattern Recognition (ICPR)*, Aug 2014, pp. 1538–1543.

[115] C. Song, Y. Huang, Z. Wang, and L. Wang, "1000fps human segmentation with deep convolutional neural networks," in *Proceedings of Asian Conference on Pattern Recognition (ACPR)*, Nov 2015.

[116] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440.

[117] J. Luo, J. Tang, T. Tjahjadi, and X. Xiao, "Robust arbitrary view gait recognition based on parametric 3d human body reconstruction and virtual posture synthesis," *Pattern Recognition*, 2016.

[118] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning.* Springer series in statistics Springer, Berlin, 2001, vol. 1.

[119] D. Muramatsu, Y. Makihara, and Y. Yagi, "View transformation model incorporating quality measures for cross-view gait recognition," *IEEE Transactions on Cybernetics*, vol. 46, no. 7, pp. 1602–1615, July 2016.

[120] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of ACM international conference on Multimedia*, 2014, pp. 675–678.

[121] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[122] S. Ren, K. He, R. Girshick, X. Zhang, and J. Sun, "Object detection networks on convolutional feature maps," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2016.

[123] G. Ma, M. Liu, L. Wu, and H. R. Karimi, “Local patch-based subspace ensemble learning via totally-corrective boosting for gait recognition,” in *Proceedings of International Conference on Machine Learning, ICML Workshop*, 2013.

[124] X. Yang, J. Dai, Y. Zhou, and J. Yang, “Gabor-based discriminative common vectors for gait recognition,” in *Proceedings of Congress on Image and Signal Processing (CISP)*, vol. 4, May 2008, pp. 191–195.

[125] H. Hu, “Enhanced gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1274–1286, July 2013.

[126] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[127] “Baidu people segmentation dataset,” 2013, http://www.cbsr.ia.ac.cn/users/ynyu/dataset/.

[128] H. Chen, D. Ni, J. Qin, S. Li, X. Yang, T. Wang, and P. A. Heng, “Standard plane localization in fetal ultrasound via domain transferred deep neural networks,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 5, pp. 1627–1636, Sept 2015.

[129] T. Zeng and S. Ji, “Deep convolutional neural networks for multi-instance multi-task learning,” in *Proceedings of IEEE International Conference on Data Mining (ICDM)*, Nov 2015, pp. 579–588.

[130] B. Parhami, "A taxonomy of voting schemes for data fusion and dependable computation," *Reliability Engineering & System Safety*, vol. 52, no. 2, pp. 139–151, 1996.

[131] J. Kittler and F. Alkoot, "Sum versus vote fusion in multiple classifier systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 1, pp. 110–115, 2003.

[132] M. Sezgin *et al.*, "Survey over image thresholding techniques and quantitative performance evaluation," *Journal of Electronic imaging*, vol. 13, no. 1, pp. 146–168, 2004.

[133] D. Ruta and B. Gabrys, "An overview of classifier fusion methods," *Computing and Information systems*, vol. 7, no. 1, pp. 1–10, 2000.