UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering and Applied Science

Department of Electronics and Computer Science

AUTOMATIC GAIT RECOGNITION VIA STATISTICAL APPROACHES

by

Pingsheng Huang

A Doctoral Thesis submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy at the University of Southampton

March 1999

SUPERVISORS: Prof. Chris Harris and Dr. Mark Nixon

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND APPLIED SCIENCE
DEPARTMENT OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

AUTOMATIC GAIT RECOGNITION VIA STATISTICAL APPROACHES

by Pingsheng Huang

Biometrics are methods to automatically recognise a person by physiological or behavioural characteristics. Examples of human traits used currently for biometric recognition include fingerprints, speech and face. Gait is a new biometric aimed to recognise subjects by the way they walk. The research objective presented in this thesis is to develop a detection and recognition system that is capable of automatically recognising humans by their gait, using computer vision techniques.

Human gait is an articulated motion comprising different movements of individual body parts. The problem in recognising human gait arises from the non-rigidity of the human body. Therefore, in order to interpret these complex movements it is necessary to represent knowledge about the shape and its movement. Intuitively, recognising humans by their gait depends on how the silhouette of individual subjects changes, either spatially or temporally. As such, template matching along the gait sequence is chosen to recognise different gaits. The features used are: spatial templates, which are the human silhouettes extracted from each scene image; temporal templates, which incorporate temporal information from the computation of optical flow between two consecutive silhouettes; and extended features, which combine spatial with temporal information into a single feature.

Statistical approaches are used for feature extraction. Eigenspace transformation - also called Principal Component Analysis (PCA), has been proven to be one of the most potent metrics in automatic face recognition, but without using data analysis to increase classification capability. The new method combines Canonical Analysis (CA) with the eigenspace approach not only to reduce the data dimensionality but also to optimise the class separability of different gait sequences. This combined approach is used to project each of these types of template into a canonical space and gait recognition is achieved in this subspace using a measure of accumulated distance as the metric. Based on this method for feature extraction, this thesis presents the experimental results and their analysis of different template features in gait recognition. In comparison with other approaches to gait recognition, the new approach appears to provide promising results. As such, this thesis describes a potent approach to data analysis for gait recognition. Future work will concentrate on establishing more precisely the attributes and ramifications of this work, whilst aiming to develop the technique still further.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisors, Prof. Chris Harris and Dr. Mark Nixon, for their invaluable advice and encouragement. Furthermore, I would like to express my deepest gratitude towards my examiners, Dr. Adam Prügel-Bennett and Dr. Roger Boyle, for their constructive criticisms.

Also, I have benefited from Dr. Hiroshi Murase at NTT Basic Research Laboratories, Japan, for helpful discussions on his eigenspace approach in gait recognition; Dr. Daniel Swets at Augustana College, USA, for his detailed explanations of the DKL approach used in image retrieval; and Dr. Jeffrey Boyd at University of Calgary, Canada, for providing useful UCSD gait data and suggestions.

I would also like to thank my parents and in-laws for their constant support and care to Pei-Ru and Hsin-Ya in Taiwan during my absence. Above all, I would like to express my endless memory to my mother-in-law who passed away during the first year of my study.

Finally, I would like to dedicate this thesis to my beloved wife, Pei-Ru, and lovely daughter, Hsin-Ya. Without their love and encouragement, I would not have finished my study.

Chapter 1

Introduction

1.1 Recognising Individuals via Biometrics

Automatic or machine-based human identification is essential in many tasks for reasons of security, including cash withdrawal from ATM machines, access control in buildings and credit card transactions. However, many of them use smart cards with correct passwords as entry keys and their major limitation is that the card presenter may not be the original owner, but knows the password, or the correct user forgets the password. Furthermore, smart cards are not applicable to applications such as criminal identification. With the growing importance of applications requiring human identification, the demand for adequate security measures has increased dramatically. In response to this demand, new technologies are being introduced aimed to ensure that the requisite level of security can be achieved. One of these technologies is often referred to as biometrics. Biometrics use certain traits, such as fingerprints, face and handwritten signatures, to verify or recognise the identity of an individual. Applications using biometrics include data security, physical access, and customer identification. Using human traits as features and pattern recognition techniques to identify people, biometrics can be denoted as a recognition technique that has several applications: verification of a person identity, authentication, authorisation and security.

Individuals can be identified by humans using physical characteristics or behavioural traits. Examples of the first source are faces, fingerprints, hand geometry, retina or iris. Handwritten signatures, voice and gait patterns are examples of the latter. The application of biometrics is based on the empirically verifiable notion that nature does not repeat itself across the species and therefore certain traits are unique. With the introduction of biometrics there is a shift in the focus from knowledge-based recognition (e.g., with a personal identification number) or token-based recognition (e.g., with a key) towards the recognition of a physical or behavioural trait. In the case of smart card technology, the presence of such a trait is combined with possession of a smart card. An important advantage of biometrics lies in the fact that physical or behavioural traits

cannot be transferred to other individuals. Moreover, automated verification has some advantages over verification processes conducted by humans; automated verification could be more reliable and consistent, since machines do not get tired and are not affected by psychological defects.

Two phases can be distinguished in biometrics. In the first phase, the enrollment phase, a certain characteristic of an individual (e.g., a mugshot) is taken by a camera. This is continued by the signal, the so-called 'training data', being processed and formed into a template. This process is repeated to allow for determining an estimate of the variance of the parameters for a particular subject. In the second phase, the verification phase, again a characteristic is measured. The data obtained, the 'test data', is compared with the stored template. If the match falls within a specified range of values, the match is valid.

Gait is an emergent biometric aimed essentially to recognise people by the way they walk. Gait's advantages are that it requires no subject contact, like automatic face recognition, and that it is less likely to be obscured than other biometrics. Gait has allied subjects including medical studies, psychology, human body modelling and motion tracking. These lend support to the view that gait has clear potential as a biometric, especially in conjunction with other independent invariant biometrics. Clearly, vision-based methods offer an obvious means of abstracting individual gait characteristics or signatures from a sequence of appropriate images.

1.2 Scope of This Thesis

This thesis presents a contribution to automatic gait recognition [1] which is in the general field of computer understanding of image sequences. The information revealed in single and multiple images is sufficient for humans to comprehend the events being photographed in nearly all situations. However, developing an algorithm which can obtain a similar understanding to that achieved by a human (given the same data) is a very difficult task, especially for recognising individuals by their gait. Motion-based recognition is an approach that extracts motion information from a sequence of images for the purpose of recognition. Recognising humans by their gait is easily performed by a human visual system, however it can be also realised by extracting features from individual silhouettes of different walking subjects. Those features must be common to all individual gait sequences and yet be distinguishable between different subjects. Template matching is a simple technique in image processing to measure the similarity of two identical-size images by directly computing their correlation between corresponding pixels. Based on template matching, calculating the correlation between gait sequences appears to be the simplest and most straightforward approach for gait recognition.

The features proposed in this thesis are templates; statistical approaches are used to extract prominent information from the feature templates for recognition. Since the

main concern in the gait sequence is the spatial and temporal information change in the body's pose, we propose two feature templates which include the spatial templates and the temporal templates. Spatial templates are actually the human spatial shapes extracted from each gait sequence, and temporal templates are extracted from temporal changes of two consecutive human shapes, by calculating the optical flow. Based on the assumption of one single subject in the scene, a simple segmentation technique is used to extract each template which covers only the area of human body. Templates are extracted from image frames and each gait sequence is then converted into a template sequence. Templates are rescaled to the same size and then considered as individual samples for further statistical analysis.

Based on the conventional stages used in automatic understanding of image sequences and most notably at the lower levels of understanding, we make the following assumptions for extracting the feature templates:

- A static camera and only one walking subject appears in any scene.

  For vision problems in the real world, finding the three dimensional motion of the camera relative to the world is a necessary stage. Since the objective of this thesis is to recognise human gait automatically, we are mainly concerned with the motions related to the human body, but not the camera motion. To simplify the preprocessing problem introduced by camera motion, this thesis assumes the camera is static. Further, there is only one walking subject in any scene and the background scene is static to ease extraction. Extracting gait from a scene containing multiple subjects and viewed by a moving camera is the subject of future work which is outside the scope of this thesis.

- Each subject is walking laterally to the camera.

  It is also essential to segment images into areas which are the projections of objects in the world where objects have different motion from the world and each other. In this thesis, human outline segmentation (extraction of spatial templates) can be achieved, based on the assumption of one subject walking laterally, by simply subtracting the image with the subject from the background image. The segmentation of optical flow (extraction of temporal templates) is simplified by bounding only the outline area.

- Only the 2D projection of the subject is considered.

  The last stage is to find the three dimensional structure of the world and any moving objects in it. Since the objective of this thesis is recognition rather than tracking, 3D structure recovery which is computational intensive is not considered here. Because of the assumption of walking front to parallel, only the 2D projection of the subject needs to be considered.

After extracting feature templates from each gait sequence which are then converted

into a template sequence, based on template matching, gait recognition can be simply achieved by calculating the correlation between template sequences. However, the problem of the "curse of dimensionality" occurs in image matching (which is computationally intensive, especially for image sequence matching), dimensionality reduction is a non-trivial issue. Therefore, Principal Component Analysis (PCA) is used to reduce the image dimensions whilst simultaneously maintaining minimum information loss. After PCA, each template is projected from a high-dimensional image space into a point in a low-dimensional eigenspace. A gait sequence is then represented by a point trajectory in the eigenspace.

Although eigenspace representation maintains the maximum information, there is no class information which is necessary for recognising different subjects in each projected vector. Hence, Canonical Analysis (CA), as used to increase the class separability, is applied to these projected vectors in the eigenspace. After CA, each gait sequence in the eigenspace is further projected into a canonical space. A gait sequence becomes a trajectory cluster in this new space. Gait recognition is actually achieved in the canonical space. Therefore, the combination of PCA and CA is used to reduce image dimensionality and to optimise the class separability of different gait sequences, simultaneously.

In the enrollment phase, all the training gait sequences from different subjects are analysed by PCA and CA and then projected into the canonical space. The cluster centroids of different subjects are calculated and stored in the database for further matching with unknown gait sequences. In the verification phase, each test gait sequence is projected into the canonical space and gait recognition is accomplished in the canonical space by selecting the minimum *accumulated distance* of test gait sequences to the training centroids of different subjects in the database.

Based on the proposed combined approach for template transformation, two gait recognition systems using spatial templates and temporal templates are presented in Chapter 4 and Chapter 5 of this thesis. Furthermore, by incorporating the spatial and temporal information from spatial and temporal templates, extended features which concatenate the feature vectors of spatial and temporal templates in the canonical space are proposed in Chapter 6 for gait recognition. In summary, we mainly focus on the recognition of individuals using gait sequences and the contributions of this thesis are:

1. an approach to recognise people by their gait using statistical means;
2. the combination of Principal Component Analysis with Canonical Analysis for feature extraction in automated gait recognition;
3. the use of spatial silhouette features and temporal features from optical flow for gait recognition; and
4. the use of extended features which combine spatial with temporal information, for gait recognition.

Publications related to this thesis are listed before the References.

1.3 Thesis Overview

The remainder of this thesis is arranged as follows.

- **Chapter 2: Gait: Description and Recognition**

  Chapter 2 reviews previous work relevant to automatic gait recognition. The review starts with earlier research in medical and psychological studies and then describes relevant and current approaches developing in computer vision. All these allied researches either lend support to the potential for gait as a biometric, or suggest that its analysis can be achieved by computer vision. This leads to the feasibility of automatic gait recognition. Finally, this chapter summarises the lessons learned from previous work and the aims of this research are outlined.

- **Chapter 3: Feature Extraction - Eigenspace and Canonical Space Transformation**

  Chapter 3 describes the techniques of feature extraction. Since template matching is used for gait recognition and statistical techniques are adopted for feature extraction from templates, gait recognition is achieved in the canonical space. Therefore, the core process of this thesis is feature extraction accomplished by the combination of PCA and CA. Eigenspace Transformation (EST) is generated by PCA and Canonical Space Transformation (CST) is produced by CA. Each template can be projected into the canonical space by the combination of EST and CST. This method can be used to reduce the data dimensionality and optimise the class separability of different gait sequences simultaneously. The theory of PCA and CA is introduced in this chapter, their computational considerations and modifications suitable to gait recognition are also discussed.

- **Chapter 4: Statistical Gait Recognition Using Spatial Templates**

  The objective of gait recognition is to recognise people by the way they walk. Thus, the main concern is the changes of human shape without regard to the clothes worn or to differing background. Therefore, spatial templates which are binary images of human silhouettes extracted from each scene image are used as features for gait recognition in this chapter. Based on the techniques of feature extraction in Chapter 3, this chapter presents a gait recognition system using spatial templates as features. Gait recognition is achieved in the canonical space using the accumulated distance measure as a metric.

  Performance results in two experiments achieved by the proposed system both show 100% recognition rates. In comparison with the results of two other approaches in gait recognition, the proposed system also provides better results.

- **Chapter 5: Incorporating Temporal Information**

  Motion information can be extracted from the change of two consecutive human silhouettes. Extracting motion information from gait sequences becomes a useful clue to distinguish different subjects. This chapter describes a gait recognition system using temporal templates as features. Temporal templates are extracted from the optical flow field between two consecutive human silhouettes. Temporal information is incorporated from optical-flow changes between two consecutive spatial templates into temporal templates which represent the distribution of velocity magnitudes in each pixel. The techniques used for feature extraction and recognition are the same as presented in Chapter 4.

  By comparing the performance with two other approaches, the proposed system still achieves the best result in gait recognition. The analysis and comparison of recognition performance for three types of temporal templates and spatial templates is also discussed.

- **Chapter 6: Combining Spatial with Temporal Information**

  Chapter 6 concerns the extended feature vectors which combine the feature vectors of the spatial and the temporal templates in the canonical space. This fuses the spatial and temporal information into a single feature. Their performance in gait recognition is evaluated and discussed. Spatial and temporal templates have either only spatial or temporal information of human gait though they have achieved promising results for gait recognition. The hypothesis used in this chapter is that gait recognition could be more robust and accurate if the spatial and the temporal information can be integrated together.

  Experimental results show that the proposed extended features achieve better performance than using spatial or temporal templates alone. This shows that including spatial with temporal information into a single feature can increase the robustness of gait recognition using statistical techniques.

  For the extraction of temporal templates, the computation of the optical flow field is achieved by region-based techniques. The quality of temporal templates is affected by selected patch sizes. The analysis of performance affected by using different patch sizes to extract temporal templates is discussed.

- **Chapter 7: Conclusions and Future Work**

  Finally, this chapter briefly summarises the work presented in this thesis. Deficiencies of the proposed approaches are raised and discussed. Future extensions to improve the performance of gait recognition are also described.

Chapter 2

Gait: Description and Recognition

2.1 Introduction

Biometrics are a range of methods to automatically recognise a person based on physiological or behavioural characteristics. Examples of human traits used currently for biometric recognition include fingerprint, speech, face. Fingerprint and face recognition have already been used in commercial and law enforcement applications. Handwritten signatures are used for verification. There has been also growing interest in biomechanical analysis of the human gait in recent years. Gait is a new biometric generated to recognise human subjects by the way they walk. In many applications of person identification, many established biometrics, such as face and fingerprint, can be obscured. The face may be hidden or at low resolution and the palm can often not be seen. Gait is attractive since it requires no subject contact, in common with automatic face recognition and other biometrics. Apart from perceptibility, another attraction of using gait is that motion can be hard to disguise. Comparing the information provided by gait with information generated by other biometrics, gait offers not only the information on the disposition of the human body during walking but also motion information from the change of silhouettes. As such, gait patterns require further study; there are limits to the use of gait as a biometric and their detailed study awaits the development of new techniques to find and describe the moving body's shape.

Motion plays an important role in the human visual system. It provides an aid to recognising an object with different motions or different objects with the same motion. For gait recognition, the recognition process does not rely on what the person wears or how the background changes. One clue we can use is the spatial and temporal changes of the human shape during walking. There are two main stages in automatic gait recognition. The first stage consists of determining an appropriate representation of the subjects or their motions from the image sequence. The second stage consists of

matching the data derived from some unknown gait sequences with those derived from training sequences of different subjects which have been stored in a database. Manual design of a feature set is appealing because such a feature set can be very efficient and it exploits prior knowledge (e.g.phenomenological). When designed properly, a very small number of parameters for each object can be sufficient to capture the distinguishing characteristics from among the objects to be recognised. However, unlike the human face, there are few individual features which can be extracted from human shapes during walking for discriminating different subjects. Consider for example the contrast between the rich lexicon of words used to describe a face and its expressions with the limited vocabulary used to describe the lower torso and its motion. The alternative approach is to consider the property of the spatio-temporal gait pattern as a whole, rather than extracting features from individual body parts.

As described in the previous chapter, this thesis proposes a statistical approach to extract the feature vectors from templates containing a body's silhouette. Our own experience shows that it is easy to recognise friends from the way they walk, even at the distance where their facial features are unrecognisable. Therefore, the features provided by gait have the potential to distinguish different subjects. However, gait is a sophisticated motion, being highly articulated and the rotations and twists of body parts occur in nearly every movement. Furthermore, various parts of the body continually move into and out of occlusion. Although humans find it possible to recognise a friend by gait through experience, automatic gait recognition by computer vision is still a challenging issue.

This chapter reviews previous work relevant to automatic gait recognition. Section 2.2 concerns various researches in medical studies for gait analysis. In order to collect gait patterns, most of them use external markers attached to the subjects. The objective of gait analysis is to provide information to clinicians in recognising pathological abnormalities. Section 2.3 describes different researches in psychological studies showing that people do have the ability to recognise different human motions, even different subjects, using their visual perception. Psychological studies in gait support the view that gait can indeed be used for recognition. Different representations of the human body and its motion for analysis are explained in Section 2.4. In order to analyse human motion, various human models have been discussed. The selection of appropriate human models is essential to represent the human body and to assist the analysis of human motion. Also, adequate models of human motion are vital to recognise different types of motions. Section 2.5 discusses various techniques for tracking human motion in computer vision. Model-based and feature-based methods are described. Model-based techniques provide the prior model and the position of the subject, a tracking procedure is then used to predict the next movement of the subject. Feature-based approaches achieve the tracking by using low-level image features, such as points, edges, contours, and regions in an image sequence. Researches for the detection and recognition of human motion are examined in Section 2.6. The detection of a human body and its motion is an es-

sential step before the analysis of articulated body motion. Human motion recognition needs to recognise the behaviour performed by articulated objects in the temporal space. Current approaches to automatic gait recognition are given in Section 2.7, prior to a summary concerning the potential for gait as a biometric and the aims of this research in Section 2.8.

2.2 Medical Studies

Typical clinical use of gait data takes the form of case-by-case analyses with concerted efforts in understanding the biomechanical significance of particular deviations from normal gait patterns or values [2]. The aim of medical research has been to classify the components of gait for the treatment of pathologically abnormal patients. By far the most prevalent method of collecting information associated with the position of body segments and joints is through the use of external markers placed on the subject [2].

Selecting sixty subjects in five age groups ranging from 20 to 65, Murray *et al.* [3] produced standard gait patterns of normal men which were compared with the patterns of pathologically abnormal patients [4]. The data were collected from the reflective markers attached to the subjects. This is simple and typical for most of the data collection systems used in the medical field, but they are not suitable for automated identification purposes. Gait was considered by Murray as "a total walking cycle" - the action of walking can be thought of as a periodic signal. The following terms are used to describe the gait cycle, as given earlier [3]:

- Duration of a walking cycle: the time interval between successive heel-strikes (on the floor) of the same foot
- Duration of stance: the duration when each foot is in contact with the floor
- Duration of swing: the period when one foot is off the floor moving forward while the opposite foot is in contact with the floor (single-limb support)
- Duration of double-limb support: the time interval when both feet are in contact with the floor at the same time
- Stride length: the linear distance in the plane of progression between successive points of contact of the same foot
- Step length: the distance between successive contact points of alternate feet

Figure 2.1 illustrates the terms described for the left foot. Each leg has two distinct periods: a stance phase, when the foot is in contact with the floor, and a swing phase, when the foot is off the floor moving forward to the next step. The results reveal the consistency of performance of each of the subjects with respect to successive elements

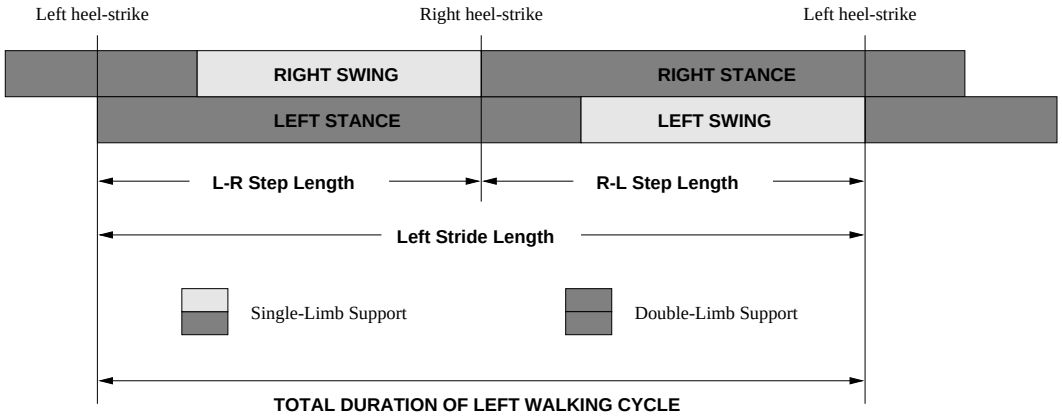


Figure 2.1: Schematic diagram of a left walking cycle showing the temporal relationships of used terms

of gait in one walking trial and in repeated trials. Murray *et al*'s work [3] suggests that if all gait movements were considered, gait is unique. Kairento [5] applied the same data collection system as in [3] to nine subjects (five healthy ones and four invalids), their walking patterns were compared using biomechanical analysis. According to the angular accelerations and moments measured from different body segments of subjects, the results show that the motion is symmetrical for healthy people but not for the ill subjects and the results are reproducible even though the tests were taken at different times.

In order to process gait data quickly and identify the functional deficiencies of a patient, classification methods are needed to characterise a patient's gait and direct the clinician's analysis to the movement abnormalities. The problem of classifying gait disorders is a problem of mapping a multivariate temporal pattern to the most likely known disorder. As such, two approaches have been used for gait data analysis [6]: statistical techniques and artificial intelligence techniques (e.g. knowledge-based systems and neural networks).

There is an extensive literature on studies of gait for medical use [7], none of which is concerned primarily with biometrics. Intuitively, measurements by gait researchers could prove to be of benefit in biometrics, though there is natural concern that the markers used do not realistically capture individual characteristics. Using gait as a biometric concerns its derivation by computer vision, for this is one way it can satisfy its purpose in automatic gait recognition. Some insight into gait as a biometric can however be drawn from psychology.

2.3 Psychological Studies of Gait

Humans have a remarkable ability to recognise different kinds of motions. In the earliest studies of gait perception, Johansson [8] published results from a series of psychophysi-

cal experiments which proved that people could recognise different human movements, such as walking, running and dancing, from Moving Light Displays (MLDs) attached to each walker's joints. A MLD isolates and presents geometric evidence of motion independent of such factors as texture, color and lighting. MLDs have also been used in motion analysis [9] (e.g. motion tracking and recognition) and investigated for its interpretation [10]. Although test subjects could not identify humans when the lights were stationary, they invariably identified the moving clusters as representing human movements. Later, Dittrich's work [11] demonstrated the ability of humans to recognise human movements which include locomotory actions (e.g. walking, going upstairs, jumping and leaping), instrumental actions (e.g. hammering, ball bouncing, stirring and box lifting) and social actions (e.g. dancing, boxing, greeting and threatening). Locomotory actions were recognised better and faster than social and instrumental actions. More recently Binham [12] has shown that point light displays are sufficient for the discrimination of different types of object motion which include 6 rigid-body events and 3 nonrigid events. As such, human vision appears adept at perceiving human motion, even when viewing a display of light points. Indeed, the redundancy involved in the light point display might provide an advantage for motion perception [13] and could perhaps offer improved performance over video images.

Studies in perception have not concentrated on motion alone, but have also addressed the human's ability to discriminate gender, again using point light displays. One early study [13] showed how gender could be perceived, and how accuracy was improved by inclusion of height information [14]. The ability to perceive gender has been attributed to anatomical differences which result in greater shoulder swing for men, and more hip swing for women [15]. Indeed, a torso index (the hip shoulder ratio) has been shown to discriminate gender [16] where the identification of gender by motion of the centre of moment was also suggested. However, gender identification would appear to be less demanding than person identification. It has been shown how subjects could recognise themselves and their friends [17] without familiarity cues. Viewers recognised their friends from the video-taped images. This psychological experiment shows that it is possible for humans to identify a subject from the paths of MLDs. Later, this is explained by considering gait as a synchronous, symmetric pattern of movement from which identity can be perceived [18]. As such, these studies encourage the view that gait can indeed be used as a biometric.

Surprisingly, research into the psychology of gait has not received much attention, especially using video, in contrast with the enormous attention paid to face recognition. One recent study [19], using video rather than point light displays, has shown that humans can indeed recognise people by their gait, and learn their gait for purposes of recognition. The study concentrated on determining whether illumination or length of exposure could impair the ability of gait perception. The study confirmed that, even under adverse conditions, gait could still be perceived. Clearly, psychological studies support the view that gait can indeed be used for recognition. Prior to a study of au-

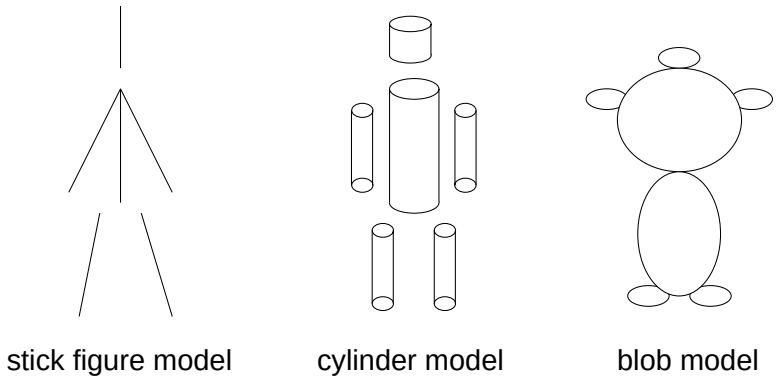


Figure 2.2: Human body models

tomatic recognition, we shall consider some of the (many) approaches to human body and motion modelling, for these are of potential benefit in recognition. Indeed, some of the approaches have found deployment in automatic gait recognition.

2.4 Modelling of Human Body and Its Motion

2.4.1 Modelling of Human Body

The human body has a well defined structure that can be connected by a natural hierarchy of parts. Many studies have considered human motion extraction and tracking, though not for recognition purposes. The selection of appropriate body models is important to efficiently recognise human shapes from images and properly analyze human motion. Generalised cylinder models and stick figure models are commonly used for three-dimensional tracking, and the ribbon model and the blob model are also used but are not so popular. Figure 2.2 shows three different examples of models: a stick figure; a cylinder; the blob model.

The original 3D human body model was proposed by Marr and Nishihara [20]. They proposed the model of a generalised cone which is the surface swept out by moving a cross-section of constant shape but smoothly varying size along an axis. Generalised cylinders are the simplified case of generalised cones that have a cross-section of constant shape and size. Based on Marr and Nishihara's model, Hogg [21, 22] and Rohr [23, 24] have proposed individual models with 14 elliptical cylinders representing the head, torso and 3 primitives for each arm and leg. Hogg [21, 22] proposed an approach by first generating various poses of the given 3D model, then projecting them into the image, and matching with image edge data. The search procedure will be computationally intensive when the subject's overall position is unknown. An alternative approach proposed by Rohr [24] uses data from medical motion studies for his motion model and improves Hogg's approach by introducing a Kalman filter to estimate the model parameters. This method only worked correctly when the human motion was well represented by the

pattern of medical data. Akita [25] used a simplified human body model consisting of six cylinders: two arms, two legs, the torso and the head where only one cylinder was used for each arm and leg. The set of axes of these cylinders is called a stick figure and the key frame sequence of stick figures was proposed to model the rough movement of the body. However, no joints are explicitly defined.

Stick figure models connect sticks at joints to represent the human body and have the advantage of simplicity. Lee and Chen's model [26, 27] uses 14 joints and 17 segments. The length of each segment and the coordinates of the joints are included in the model parameters which are used for tracking. This model involved locating the coordinates of joints and finding the joint angles in 3D space. Guo *et al* [28, 29] represent the human body structure in the silhouette by a 2D stick figure model which has ten sticks articulated with six joints. After extracting the human silhouette, the skeleton of the silhouette is matched to the stick figure model by generating a potential field around the skeleton. The correspondence is achieved by searching for the optimal pose of the figure model with the minimal energy in the potential field. Concentrating on the lower limbs of a walking person, Bharatkumar *et al* [30] extracted the stick figures by applying the medial axis transformation to the human silhouette. The reference sequence of a stick model was derived by averaging the 3D kinematic data from subjects with markers over the joints of their limbs. However, the stick models of above two systems are transformed from a thresholded silhouette image which in general is not robust in a complex scene.

The blob model was used in human motion tracking [31, 32]. The person is modelled as a connected set of blobs, each of which serves as one class. Each blob has a spatial and colour Gaussian distribution and a support map that indicates which pixels are members of the blobs. O'Rourke and Badler [33] proposed the overlapping sphere model which consists of 24 segments and 25 joints where those segments and joints are linked together into a tree-structured skeleton. The "flesh" of each segment is defined by a collection of spheres located at fixed positions within the segment's co-ordinate system. At the same time, angle limits and collision detection are incorporated in the motion restrictions of the human model. This method was only applied to synthesised image sequences.

Kurakake and Nevatia [34] treat the human body as an articulated object having parts that can be considered as almost rigid and connected through articulations. They use the ribbon which is the two-dimensional version of the generalised cylinder to represent the parts. Since the ribbon extraction is achieved after the operation of edge detection, the performance relies on the quality of edge detector and the scene complexity. Furthermore, the problem of one-to-many correspondence needs to be solved in ribbon matching between consecutive frames.

Although these structural models can represent the human body well, they need to be modified according to different applications and are mainly used in human motion tracking. Based on the selected models of human body, the modelling of human motion can be realised and is presented in next section.

2.4.2 Modelling of Human Motion

Amongst the current research, human motion can be defined by the different gestures of body motion, different athletic sports (tennis, ballet) or human walking or running. The analysis varies according to different motions. There are two main methods to model human motion. The first is model-based: after the human body model is selected, the 3-D structure of the model is recovered from image sequences with [10, 26] or without [21, 24, 25, 29] moving light displays. The second emphasises determining features of motion fields without structural reconstruction [35, 36, 37].

Ideas from human motion studies [3] can be used for modelling the movement of human walking. Hogg [22] and Rohr [23] use flexion/extension curves for the hip, knee, shoulder and elbow joints in their walking models. Therefore, human motion can be modelled using joint angles. Guo *et al.* [29] use joint angles between different sticks as features of different walking persons. Joint angles are also used by Bharatkumar *et al* [30] to represent the walking cycle of lower limbs in human walking and compare it to the kinematic model.

A different approach for the modelling of motion was taken by Akita [25], who used a sequence of stick figures, called a key frame sequence, to model rough movements of the body. In his key frame sequence of stick figures, each figure represents a different phase of body posture. The key frame sequence is determined in advance and referred to in the prediction process. In order to find out the interpretation tree of human body and reduce its computational complexity, Chen and Lee [26] applied general walking-model constraints from walking motion knowledge to eliminate the number of infeasible solutions.

Campbell and Bobick [38] proposed techniques for representing movements based on space curves in subspaces of a "phase space", a symbolic description that translates the continuous domain of human motion into a discrete sequence of symbols. The phase space has axes of joint angles and torso location and attitude, and the axes of the subspaces are subsets of the axes of the phase space.

Other approaches that are different from those above consider the properties of the spatio-temporal pattern as a whole. Polana and Nelson [35, 39] define *temporal textures* to be the motion patterns of indeterminate spatial and temporal extent, *activities* to be motion patterns which are temporally periodic but are limited in spatial extent, and *motion events* to be isolated simple motions that do not exhibit any temporal or spatial repetition. Little and Boyd's approach [36, 40] is similar to Polana and Nelson's, but they derive dense 2-D optical flow of the person and derive a series of measures of the position of the person and the distribution of the flow. The frequency and phase of these periodic signals are determined and used as features of the motions. Davis and Bobick [41] proposed *temporal templates* including *Motion-Energy Images* (MEIs) and *Motion-History Images* (MHIs) which are actually binary cumulative motion images and

temporal-history motion images (at each pixel) through an entire sequence. Niyogi and Adelson [42, 43] used the spatio-temporal edge of the body boundary in a spatio-temporal volume (XYT), a characteristic "braided" pattern is used to describe the human walking pattern. However, apart from Little and Boyd's approach [36, 40], the above approaches are only used either for recognising some specific motions or for tracking purposes.

Using statistical means, Murphy *et al.* [44] analyzed image frames from the walking sequence with a walking subject and proposed that human motion analysis becomes tractable by using eigenspace representation. The same idea is used by Murase and Sakai [37] who extract the silhouette movements of the human body from sequences and project them into a parametric eigenspace. The motion of human walking is represented by a trajectory in the eigenspace. Also, Black *et al.* [45] apply eigenspace analysis to the optical flow field of walking patterns and use the time history of the first two components to represent the gait motion pattern of two legs.

Basically, model-based approaches can represent human motion well. However, model recovery from the images requires the matching from images into models and the problem of model reconstruction is computationally intensive. Furthermore, motion characteristics extracted from the recovered inaccurate features, such as joint angles, are more suitable to distinguish different motions (e.g. running and walking) performed by subjects, rather than recognising different subjects performed the same motion (e.g. gait). Avoiding the inaccuracy incurred by the model recovery, feature-based methods use the change of low-level image features, such as region templates, to represent human motion. This appears better to maintain the motion information for recognition. However, there are still redundancies in those features. Statistical means can be used to extract prominent information from features and the objective of gait recognition can be achieved.

2.5 Human Motion Tracking

There has been a number of approaches to tracking humans in scenes, more for security applications rather than for recognition. Typically, they are divided into two groups: the model-based approaches and the feature-based approaches which use low-level image features, such as points, edges, contours, and regions in an image sequence. There has been much progress since, and we shall review only some of the more recent work.

2.5.1 Model-based Tracking

Given the prior model and position of the subject, the objective of a tracking procedure is to predict the next movement of the subject. In general, a human model has two sets of parameters: one is the shape parameters which vary between different subjects; the other is the control parameters which define the pose of a subject. Basically, there

are two stages in model-based tracking methods. First is the initialisation which is to recover the shape and the initial pose of an object. Second, the tracking, whose goal is to predict the pose of the subject at a specific time. Model-based tracking methods typically involve a search process in the control parameter space in which the projection of best fit to the image data defines the optimal pose.

A model-based approach was used in one of the earliest tracking studies [21]. The WALKER model mapped images into a description in which a person was represented using a series of hierarchical levels. The tracking is done through a function, TRACK, which searches the optimal parameter assignment for the current projected image. The performance of the system was illustrated by superimposing the machine-generated picture over the original photographic images. Rohr's [24] approach used a linear regression method to calculate the initial estimate of the subject posture and then used a Kalman filter approach to estimate the model parameters incrementally. The generalised cylinder model and Kalman filter are also use by Assereto *et al.* [46] to track a walking person. Watchter and Nagel [47] use an iterated extend Kalman filter (IEKF) to determine the parameters characterising the degrees of freedom according to the human model of generalised cylinders. Similar to Hogg's and Rohr's model, Cheng and Moura [48] proposed a human model to track the walking subject. The tracking is achieved by a gradient-based method for estimating the motion of the body parts between consecutive frames.

In Akita's [25] method, the estimation of current body position is done by applying a correlation technique to find the correspondence in the previous frame. A sequence of key frames which represented the different postures was used to provide the rough position of the person when occlusion occurs and the correlation is not applicable.

Another recent approach, the "person finder" Pfinder system [32], has been aimed to solve the problem of tracking a single person given a fixed-camera. The system uses the blob description of human motion which uses coherent connected regions. The statistics of the blobs are then recursively updated to combine present information with prior knowledge. The system then learns the scene of the fixed camera and then detects a person as a large deviation from that scene. Then, the person can be tracked through an image sequence. The system is not aimed at recognition and applications include real-time interface devices and video games.

A new method for the 3-D model-based tracking of human body uses multiple views to avoid occlusion of body parts [49]. Available parts are then tracked between frames of a video sequence in a model aimed to minimise the difference between the human model and the imaged views. Initial results were presented showing how humans could be tracked in the presence of severe occlusion.

2.5.2 Feature-based Tracking

Recovering the model positions from images requires the mapping from images into models. This is a model reconstruction problem which is rather difficult to compute. An alternative approach to model-based tracking is to consider the low-level features in the images. One approach to restrict the solution space is to learn the subject representation from a training set.

Cootes *et al.* [50] proposed a point distribution model in which a shape is represented by a set of labelled points. This set of labeled points was manually generated from a set of training shape images in various positions. The shapes were aligned and the deviations from the mean were analyzed using principal component analysis (PCA). The M most significant modes of variations, represented by the M largest eigenvectors, gave a compact representation of the object shape. This model has been successfully modified and used by Baumberg and Hogg [51, 52] for tracking the contour of walking people. Instead of manually generating the labelled points of the human shape, those points were automatically produced by the B-spline method. The learning process is the same as in the approach of Cootes *et al.* [50]. The tracking process was achieved by searching the best matches in the model subspace using the Kalman filter. The results show that a good representation and learning scheme can significantly remove the difficulties in tracking complex objects. The tracking process has been also implemented by nonlinear neural networks [53]. Based on the same shape model, an improved model [54] has been derived to automatically describe the shape of a moving pedestrian from the vibration modes of the finite element method (FEM) [55]. However, the tracking only works well when the postures and views are well represented in the training set. An active-contour tracking approach - *active rays*, was used by Denzler and Niemann [56] for tracking moving pedestrians. Since the human contour cannot be well represented by the extracted contour points, it can be used for tracking rather than for recognition.

Parameterised optical flow has actually been used to track articulated motion in an image sequence [57]. Limbs were represented as a set of connected cardboard patches where analyzed motion was constrained to enforce articulated motion. The approach was demonstrated to track humans walking, over long image sequences. The possibility of recognition from these data was noted, but not explored greatly. One system used 3-D planar projections to achieve better tracking than contemporaneous 2-D trajectory-based systems [58]. The system was based on detecting and segmenting optical flow from within a central region. Then 3-D planar geometry was used with an active camera system to ensure focus on the central region. Extended Kalman filters were used to analyze the trajectories and the system was shown to successfully track moving objects, including people and pursuit performance was shown to improve on 2-D performance.

Kurakake and Nevatia [34] used a image region called a *ribbon* which was described earlier in Section 2.4.1 to model the parts of the human shape. The tracking is achieved by searching corresponding ribbon in the next frame which has the most similar width

and axis angle to the current ribbon. Mainly based on the analysis of static images after edge detection, the scene can only contain a simple background and one person without wearing complex clothes.

Most tracking approaches naturally lack the accuracy required for recognition since that was not their original purpose. However, it would seem reasonable to assume that tracking procedures could be deployed to develop a gait signature.

2.6 Human Detection and Motion Recognition

Under the assumption that object structure information can be used for tracking or recognition, traditional approaches [59, 60] for dynamic scene analysis attempt to recover the structure of objects using a sequence of images. If the reconstruction of structure is successful, the measured parameters (e.g. joint angles) from the human body shape can provide rich and useful representations. However, in modelling humans, the recovery process should not be sensitive to clothing and the change of viewpoint, especially for recognition purposes. Furthermore, the human body is composed of a large number of parts which can move non-rigidly with respect to one another and the recovery of shape and motion parameters normally involves an expensive search procedure in a high dimensional parameter space. Therefore, any approach that attempts to recover a full three-dimensional model will have considerable difficulty. The alternative approaches are to use low-level features, such as points, edges, contours and regions in the image sequence for recognition. Moreover, the use of motion information for recognition is often more prominent than its use in reconstruction. The problem introduced by the incompleteness of the knowledge can be reduced with a learning procedure from the training set. Learning consists of summarising a set of motion models or motion trajectories by representing the variance of the motion in the space of measurements. Two widely used methods, Hidden Markov Models and Eigenspace decomposition, are reviewed in Section 2.6.2.

2.6.1 Human Segmentation and Detection

A practical application in motion segmentation is to extract pedestrians walking across a street or to segment a human body from the background. This task is an essential step before the analysis of articulated body motion.

Leung and Yang [61] segmented human bodies from the stationary background by classifying coincidence edges which were edges in both the difference picture and the current frame. However, their method can only handle a scene containing one person exercising at a stationary position, not suitable for a walking subject. Shio and Sklansky [62] described a method based on intensity, motion and an object model - i.e. a model of the image of a person in motion. The people segmentation is actually achieved by apply-

ing their region splitting and merging procedures to the optical flow field. Subregions are generated by splitting each region according to the quantised direction of motion. These subregions are merged to groups according to a coherent average motion. Meyer *et al.* [63] segmented the different parts of a walking subject by computing the centroid of each part in the optical flow field and applying an active contour method for segmentation. However, this segmentation result is only used for human model fitting due to a rough segmentation. Oren *et al.* [64] proposed an approach for detecting pedestrians from a static scene. They convert each image template with a pedestrian into a wavelet template and use the "bootstrapping" method [65] to train the database. The detection is accomplished by matching the test template to the database. Since the pedestrian is bounded in a rectangular window, the exact human outline can not be extracted.

All of these methods are limited in their accuracy and generality. If the background image and the background motion are available, the problem of segmenting moving persons from the background can be partially solved by image subtraction. However, this approach can only yield coarse segmentation, and the segmented regions do not guarantee to be coherent. Thus, further processing is needed. Therefore, segmenting a moving person from the background and/or other persons is not a simple problem. So far, current approaches can only provide rough segmentations.

2.6.2 Human Motion Recognition

Human motion recognition is to recognise the behaviour performed by articulated objects in the temporal space. Basically, this problem can be divided into two parts. The first one is to recognise the action performed by a person among a database of human actions. The other one is to define motion as a sequence of the parameters in the configuration space. In order to recognise activities, we have to develop a sufficient movement description in the environment so that these recovered descriptions can be matched to stored descriptions.

Polana and Nelson [35] use a spatio-temporal derivative method to compute the normal flow magnitude at each pixel position between consecutive frames. Those pixels where significant motion is present are marked, and the centroid of the marked pixels is computed in each frame. The motion is represented by fitting a linear trajectory to the sequence of centroids. A periodicity measure is computed from this trajectory and its low-frequency components are used to classify periodic and non-periodic activities. Although such an approach probably has the ability to reduce noise, the low-frequency components cannot efficiently represent the original patterns.

The phase space is a Euclidean space with axes for all the independent variables of a system and their time derivatives. Each point in phase space represents a state of the system, and as the system evolves over time it moves along a phase trajectory. To identify the different ballet actions, Campbell and Bobick [38] proposed a phase plane

method with axes including the torso position, attitude parameters and joint angles. At each time instance, the input tracking data is converted to a point in the phase space. This point will be accepted as part of a recognised movement if it is within a threshold distance of each of the 2D-projection space curves used to define the move. Motions are represented by a collection of 2D-projection curves in the phase plane. However, motion trajectories in the phase space are typically complex curves which are hard to recognise due to the variation in estimated motions.

Recently, Hidden Markov Models (HMMs) [66] have been applied successfully in speech recognition [67]. HMMs are a stochastic model with state transition and characterised by the learning ability which is achieved by optimising the input time sequence data automatically. Each state in the HMMs is assigned a symbol and a probability of state transition from one state to another state. An unknown event is recognised by matching the output symbol sequence to the training sequences. Based on HMMs, Yamato *et al.* [68] proposed an approach to recognise human motion. A sequence of moving images is converted into a sequence of feature vectors which are further converted into a symbol sequence by vector quantisation. Different motion categories (tennis actions) are learned by the proposed HMM which is used to recognise the symbol sequence of unknown actions. Since the recognition is based on the unknown symbol sequence converted from brightness patterns, the method was sensitive to the human shape.

Motivated by Kirby and Sirovich's [69] approach in face recognition, eigenspace representation has been used to recognising human motions. Black *et al.* [70] use the eigenspace decomposition to find the dominant curve components of motion curves measured from different parts of each subject performing the same activity. This can be further used for the recognition of different human motions. Also, the principal components of eigenspace representation are used to represent the lip images and to recognise the lipreading [37, 71]. Furthermore, eigen decomposition has been also successfully applied in tracking, motion segmentation and pattern recognition. However, these methods require that the subjects or their motion should be well approximated by the training set.

In order to recognise 3 different motions (i.e. walking, running and unknown), Guo *et al.* [28] use neural networks as the classifier and take frequency components of Discrete Fourier Transform (DFT) as inputs which are computed from the time sequences of 5 human model parameters (e.g. joint angles). Also, neural networks are used to recognise the walking speed of subjects [72] by taking inputs from the measured values of the right single-support phase, left single-support phase and double-support phase. Ushida *et al.* [73] use a fuzzy system [74] to classify three different tennis motions (i.e. forehand, backhand and smash). The recognition of motion patterns are defined by the fuzzy inference rules.

However, current approaches mainly focus on the analysis and the recognition of different human actions rather than recognising subjects. Nevertheless, each of the allied

subjects continues to support the notion that gait can be used as a biometric. The physical characteristics of gait are established and viewed to be characteristic, humans can perceive gait and gait can be modelled and extracted by computer vision techniques. Much of this work has been of benefit to the approaches of automatic gait recognition.

2.7 Automatic Gait Recognition

In what was perhaps the earliest approach to automatic recognition by gait, Niyogi and Adelson [42, 43] distinguish different subjects by extracting their gait signatures from the walking patterns in a spatio-temporal volume (XYT). Here, in the XT dimensions (translation and time), the motions of the head and of the legs have different patterns. These patterns were processed to determine the body motion's bounding contours and then a five stick model was fitted. The gait signature was derived by normalising the fitted model for velocity and then by using linear interpolation to derive normalised gait vectors. This was then applied to a database of 26 sequences of five different subjects, taken at different times during the day. Depending on the values used for the weighting factors in a Euclidean distance metric, the correct classification rate varied from nearly 60% to just over 80%, a promising start indeed.

Later, optical flow was used to derive a gait signature [36, 40]. This did not aim to use a model of a human walking, but more to describe features of an optical flow distribution. The optical flow was filtered to produce a set of moving points together with their flow values. The geometry of the set of points was then measured using a set of basic measures and further information was derived from the flow information. Then, the periodic structure of the sequence was analyzed to show several irregularities in the phase differences; measures including the difference in phase between the centroid's vertical component and the phase of the weighted points were used to derive a gait signature. Experimentation [40] on a limited database with 6 subjects and 7 sequences each showed how people could be discriminated with these measures, appearing to classify all subjects correctly. The best recognition rate achieved is 95.2% for all the 42 sequences.

Another approach was aimed more at generic object-motion characterisation [37], using gait as an exemplar of their approach. The approach was similar in function to spatio-temporal image correlation, but used the parametric eigenspace approach to reduce computational requirement and to increase robustness. The approach first derived body silhouettes by subtracting adjacent images, with further processing to reduce noise. Then, the images were projected into eigenspace, a well established approach in automatic face recognition. Eigenvalue decomposition was then performed on the sequence of silhouettes where the order of the eigenvectors corresponds to frequency content. Recognition from a database of 10 sequences of seven subjects showed classification rates of 100% for 16 eigenvectors and 88% for eight, compared with 100% for the (more computationally demanding) spatio-temporal correlation approach. Further, the approach

appears robust to noise in the input images.

Using a model-based approach, the gait signature is derived from the spectra of measurements of the thigh's orientation [75]. Two legs are considered as an interlinked penduli and the phase-weighted Fourier magnitude spectra is used as the feature to recognise different subjects. The frequency feature came from the changes of thigh angles by the edge detection of legs. This was demonstrated to achieve a recognition rate of 90% on a database of 10 subjects with 4 sequences each, using the k-nearest neighbor rule.

By using MLDs attached to the joints of subjects and measuring the trajectories of 13 points, Lakany and Hayes [76] used neural networks to recognise walking people. The feature vectors are extracted from the frequency components of 13 trajectories after the analysis of 2D fast Fourier transform (FFT). The performance showed the recognition rate ranging from 95% to 100% for 4 subjects with 15 gait cycles each. However, they need to use auxiliary materials attached on each subject's joints for data acquisition.

Contemporaneously, the nature of gait has been recognised by "probabilistic decomposition of human dynamics at multiple abstractions" [77] where the dynamics of gait in video sequences were recognised by learned HMMs using the features from the translation and angular velocities of the limb segments. Also, HMMs are used for recognising different types of gait (i.e. walking, limping, hopping and running) by the trajectories of tracked body parts [78]. The feature vector was extracted from optic flow and from trajectory information and then classified by use of HMMs, showing good gait discrimination. By modelling body parts and the background as mixture densities, an improved version [79] with the position locating to different parts (i.e. head, trunk and leg) using more gait data is presented.

The majority of current approaches are motion-based, combining the image sequence by its motion or by statistical analysis. This appears to be a feasible solution for gait recognition. As such, based on this direction, the aims of this research are presented in next section.

2.8 Summary and Aims

Gait is a potential biometric since humans can perceive it. It is attractive because it requires no contact and is less likely to be concealed. Allied studies in physiology suggest that gait can be modelled and is unique, as supported by psychological studies. A number of approaches have modelled the body and tracked it through image sequences, though not for recognition. All these allied researches either lend support to the potential for gait as a biometric, or suggest that its analysis can be achieved by computer vision. Indeed, a number of approaches have already shown that it is possible to recognise people by their gait. Naturally, this is more exploratory work than established system development, but results suggest that further development is warranted.

The aims of this thesis are to develop automatic systems for gait recognition by computer vision without attaching any auxiliary material to the subjects. As described in Section 2.6 and in Section 2.7, Hidden Markov Models and Eigenspace decomposition have been successfully used in recognition for human actions or gait. Based on the idea of learning from examples, the problem introduced by the incompleteness of the knowledge in the sample space can be reduced by the training procedure. Thus, statistical techniques appear to have potential for automatic gait recognition. Motivated by Murase and Sakai's [37] approach, the systems used to recognise people by their gait are discussed in this thesis. In this thesis, we propose different feature templates as features and a new statistical approach for feature extraction. Gait recognition is achieved in a low-dimensional space, the canonical space, rather than in the image space, by distance accumulation.

Murase and Sakai [37] proposed a statistical method using the parametric eigenspace representation to represent gait patterns and to recognise different human gait. This can be applied to various object sets, and is robust to relatively clean environments. The details have been described in Section 2.7. Eigenspace representation, also called "eigenfeatures", is calculated from *Principal Component Analysis* (PCA). PCA, also known as the *Karhunen-Loève Transform* (KLT), is a mature statistical approach which has been applied for face recognition [69, 80, 81, 82], lip reading [37, 71], gait recognition [37], image motion modeling [44, 45], 3-D object recognition [83, 84, 85, 86], 3-D object detection [87], visual learning [88, 89, 90] and recognising partially occluded objects [91], respectively. In order to maximise the information content of fused data in the eigenspace, an approach [92] to calculate the scaling factors of individual features has been proposed. An eigenfeature, however, may represent aspects of the imaging process which are unrelated to recognition (for example, the illumination direction [93]). Moreover, an increase or decrease in the number of eigenfeatures that are used does not necessarily lead to an improved recognition rate. PCA utilises the eigenvectors of the sample scatter matrix associated with the largest eigenvalues. Those vectors are in the direction of the major variations of the samples and can be used as a transformation basis to describe the image samples. We called this basis the *Eigenspace Transformation* (EST) matrix and the space which this basis spans the *Eigenspace*. In this thesis, EST is used in the first stage of feature extraction to reduce the dimensionality of feature templates. Class separation is considered in the second stage.

Since the early 1960s [94], automatic face recognition has been investigated and conducted on various aspects in psychophysics, neurosciences and engineering over the past 20 years. Two survey papers by Samal and Iyengar [95], and Chellapa *et. al.* [96] discussed a variety of approaches ranging from Karhunen-Loève expansion [69, 80], feature matching [97, 98], and neural networks [99] in face recognition, using different sources such as video, profile and range imagery. Face image representations based on PCA have been used successfully for recognition [69, 80, 81, 82]. However, since PCA is computed from the global covariance matrix of the full set of image data, it is not sensitive to class

structure in the data. In order to increase the discriminatory power of various facial features, Etemad and Chellappa [100, 101] applied Linear Discriminant Analysis (LDA) - also called Canonical Analysis (CA), directly to face images. CA can be used to optimise the class separability of different face classes and improve the classification performance. The features are obtained by maximising between-class and minimising within-class variations. Unfortunately, this approach has high computational cost. Hence, it can only be tested with small images. Moreover, the within-class covariance matrix obtained via CA alone may be singular. CA has also been used in the classification of remote-sensing data [102]. In this thesis, we call this approach *canonical space transformation* (CST) and CST is used in the second stage of feature extraction for separating different gait classes belong to different subjects.

In this thesis, the combination of EST and CST is used for gait feature extraction and this combination reduces data dimensionality and optimises the class separability of different gait sequences, simultaneously. This statistical approach has been also successfully applied in image retrieval [103] and contemporaneously, in face recognition [104]. Based on this technique of feature extraction, automatic gait recognition is achieved by proposing different feature templates and a measure of distance accumulation in the canonical space. They are presented in the remaining part of this thesis.

Chapter 3

Feature Extraction - Eigenspace and Canonical Space Transformation

3.1 Introduction

Image analysis is often very difficult due to the large amounts of information embedded in an image, especially when noise or redundant information are involved. For image processing applications, feature extraction plays the key role in extracting important features from the image data, from which the description, interpretation and understanding of the scene can be achieved. Different features ranging from low-level to high-level need to be extracted from the image for different applications.

Template matching which calculates the correlation between images is a common technique used in image matching problems. However, the "curse of dimensionality" is the first problem caused by the computation involved in image correlation and this problem becomes more serious when matching image sequences, e.g. gait sequences, especially when matching one sequence to a large database. For example, the computation involved with matching two images each of size 512×512 pixels is of order $O(10^5)$ and is $O(10^7)$ for two sequences with 100 images each. Therefore, to reduce the image dimensions whilst simultaneously maintaining minimum information loss is not trivial for template matching. The two features used in this thesis are, spatial templates: the human spatial shapes extracted from each gait sequence; and temporal templates, which are extracted from temporal changes of two consecutive human shapes during walking by calculating the optical flow. Thus, human gait becomes a sequence of moving template images and gait recognition involves matching the test sequence with the training sequences in the database.

Principal Component Analysis (PCA), also known as the *Karhunen-Loève Transform*

(KLT), is a mature statistical approach used for image dimensionality reduction. PCA uses the eigenvectors of the sample scatter matrix associated with the largest eigenvalues. Those vectors are in the direction of the major variations of the samples and can be used as a transformation basis to describe the image samples. We call this basis the *Eigenspace Transformation* (EST) matrix and the space which this basis spans the *Eigenspace*. Since PCA captures the major variations in the training data, projected vectors can express and approximate the samples well, where the reconstruction is very close to the original. However, an increase or decrease in the number of eigenfeatures used does not necessarily lead to an improved recognition rate. Therefore, PCA was selected as the first stage of feature extraction and applied to the training template images for the purpose of dimensionality reduction in gait recognition.

Although EST is well-suited for object representation, projected features are not necessarily good for discriminating different classes among samples [104]. In order to separate different classes further in the feature space and to increase the recognition rate of template matching, another statistical approach - *Canonical Analysis* (CA), is selected as the second stage of feature extraction and applied to the projected vectors in the eigenspace for the purpose of class separation in gait recognition. CA is also known as *Linear Discriminant Analysis* (LDA) and has been applied to remote sensing [105, 106] and in face recognition [101]. In canonical analysis, the between-class scatter is maximised while minimising the within-class scatter. The generated eigenvectors are used to project samples of each class to a space where each class is clustered more tightly and the separation between class means is increased. We call this space the *Canonical Space* and the set of generated eigenvectors the *Canonical Space Transformation* (CST) matrix. Features obtained using this projection optimally discriminate among the classes represented in the training set, in the sense of a linear transformation [107].

This chapter describes the theory for the two stages of feature extraction. Basically we combine the two transformations - EST and CST, to project template images from the high-dimensional image space to a low-dimensional eigenspace and then into a canonical space. Recognition is performed via the canonical space. Figure 3.1 illustrates the processing stages that generate the feature vectors for recognition by eigenspace transformation and canonical space transformation. Each template image can be treated as a high dimensional feature vector by concatenating the rows of the image together using each pixel as a single feature. Each individual template image is ultimately transformed to a one-dimensional canonical vector which is a point in the canonical space.

In the following sections, Section 3.2 describes the basic theory of PCA before the implementation of EST in Section 3.3 and Section 3.4. The basic theory of CA is introduced in Section 3.5 and Section 3.6 before the presentation of CST in Section 3.7. Section 3.8 describes the integration of EST with CST prior to conclusions which can be drawn on this approach, Section 3.9.

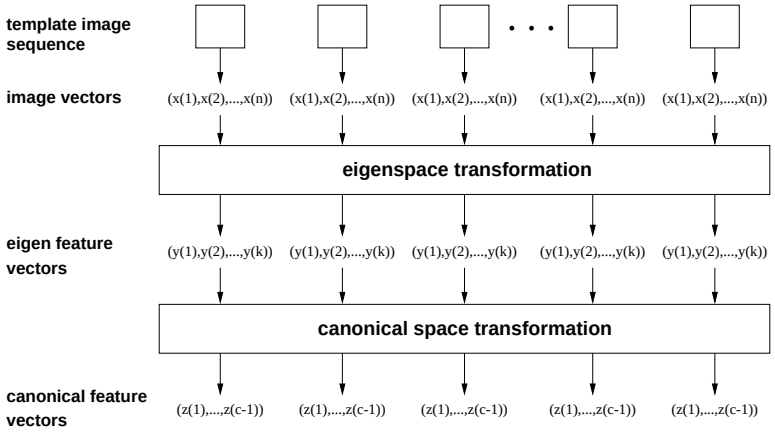


Figure 3.1: Projection of template images by EST and CST

3.2 Principal Component Analysis

PCA is widely used in image compression and pattern recognition [108]. Basically, it is used to reduce the dimensionality of an input space by mapping the data from a highly correlated high-dimensional space to an uncorrelated low-dimensional space whilst simultaneously maintaining the minimum mean-square error of information loss. PCA uses the eigenvalues and eigenvectors generated by the data covariance matrix to rotate the original data coordinates along the direction of maximum variance. Figure 3.2 shows a two-dimensional example in which dimensionality reduction can be achieved by projecting all data points from a two-dimensional space into a one-dimensional principal axis. Clearly, the data are distributed along this axis, hence rotation resulting in using the axis as the principal basis maximises potential classification rates.

Following [109] and [108], the PCA technique is described next. For a real arbitrary vector $\mathbf{x} = [x_1, \cdots, x_n]^T$ with n components, the covariance matrix $\Sigma_{\mathbf{x}}$ of $\mathbf{x}$ can be represented by

$$\Sigma_{\mathbf{x}} = E\{[\mathbf{x} - E(\mathbf{x})][\mathbf{x} - E(\mathbf{x})]^T\} \tag{3.1}$$

This matrix $\Sigma_{\mathbf{x}}$ is symmetric and positive definite [110]. Thus, the eigenvalues of $\Sigma_{\mathbf{x}}$ are all positive real numbers. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$ be the eigenvalues of $\Sigma_{\mathbf{x}}$ in non-increasing order. There exists an orthogonal matrix $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n]$ which can diagonalise the matrix $\Sigma_{\mathbf{x}}$ by

$$\mathbf{U}^T \Sigma_{\mathbf{x}} \mathbf{U} = \mathbf{D}_\lambda = Diag(\lambda_1, \lambda_2, \cdots, \lambda_n) \tag{3.2}$$

where λ_i $(i = 1, \cdots, n)$ and $\mathbf{u}_i$ $(i = 1, \cdots, n)$ are eigenvalues and associated eigenvectors

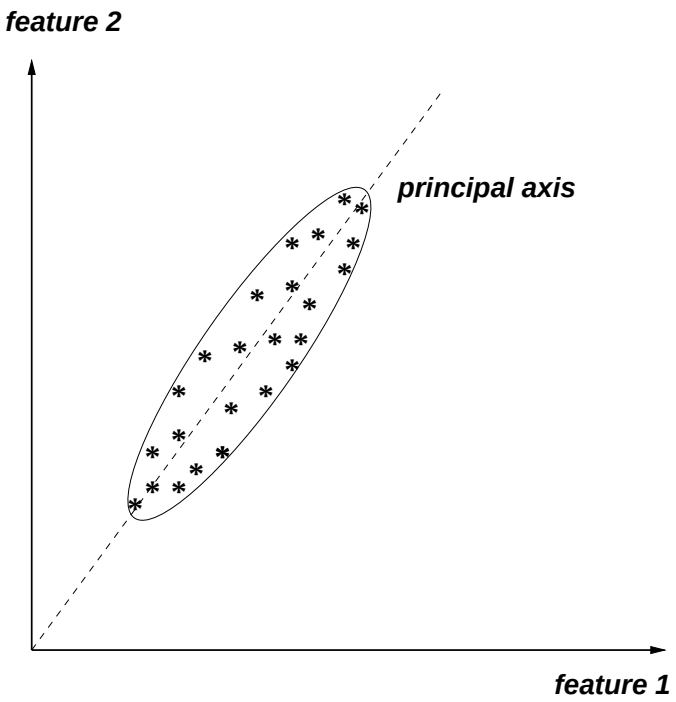


Figure 3.2: Dimensionality reduction by PCA

of matrix $\Sigma_{\mathbf{x}}$. Suppose another random vector $\mathbf{y} = [y_1, \cdots, y_n]^T$ can be generated by

$$\begin{aligned} \mathbf{y} &= \mathbf{U}^T\mathbf{x} \\ &= [\mathbf{u}_1, \cdots, \mathbf{u}_n]^T\mathbf{x}. \end{aligned} \tag{3.3}$$

Then, each component y_i of $\mathbf{y}$ can be computed by

$$y_i = \mathbf{u}_i^T\mathbf{x} \qquad \mathrm{i} = 1, \cdots, n. \tag{3.4}$$

Since $\mathbf{U}\mathbf{U}^T = \mathbf{I}$, the inverse transform of $\mathbf{y}$ is given by

$$\begin{aligned} \mathbf{x} &= (\mathbf{U}^T)^{-1}\mathbf{y} \\ &= (\mathbf{U}^{-1})^{-1}\mathbf{y} \\ &= \mathbf{U}\mathbf{y}, \end{aligned} \tag{3.5}$$

it follows that

$$\|\mathbf{x}\|^2 = \mathbf{x}^T\mathbf{x} = \mathbf{y}^T\mathbf{U}^T\mathbf{U}\mathbf{y} = \mathbf{y}^T\mathbf{y} = \|\mathbf{y}\|^2 \tag{3.6}$$

Thus an orthogonal transformation preserves the signal energy or, equivalently, the length of the vector $\mathbf{x}$ in the n-dimensional vector space [111]. This means each orthogonal transformation is simply a rotation of the vector $\mathbf{x}$ in the n-dimensional vector

space. According to Equation (3.5), the covariance matrix of $\mathbf{y}$ is represented by

$$\begin{aligned}\Sigma_{\mathbf{y}} &= E\{[\mathbf{y} - E(\mathbf{y})][\mathbf{y} - E(\mathbf{y})]^T\} \\ &= E\{[\mathbf{U}^T\mathbf{x} - E(\mathbf{U}^T\mathbf{x})][\mathbf{U}^T\mathbf{x} - E(\mathbf{U}^T\mathbf{x})]^T\} \\ &= \mathbf{U}^T\Sigma_{\mathbf{x}}\mathbf{U} \\ &= \mathbf{D}_\lambda. \end{aligned} \tag{3.7}$$

Since $\mathbf{D}_\lambda$ is a diagonal matrix, the components $y_i = \mathbf{u}_i{}^T\mathbf{x}$ $(i = 1, \cdots, n)$ of $\mathbf{y}$ are uncorrelated with each other. The variance of y_i equals λ_i. Due to the order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, y_1 is called the *first principal component* with the maximum variance λ_1, y_2 is called the *second principal component* with the second maximum variance λ_2, etc. Since the orthogonal transformation is energy preserving [111], this is given by

$$\sum_{i=1}^{n} \sigma_y(i)^2 = tr(\mathbf{U}^T\Sigma_{\mathbf{x}}\mathbf{U}) = tr(\Sigma_{\mathbf{x}}) = \sum_{i=1}^{n} \sigma_x(i)^2 \tag{3.8}$$

in which tr is the trace of matrix, $\sigma_x(i)^2$ and $\sigma_y(i)^2$ are the individual variances of random variable x_i and y_i $(i = 1, \cdots, n)$. The variance summation of y_i $(i = 1, \cdots, n)$ equals the variance summation of the original random variable x_i $(i = 1, \cdots, n)$ and can be written as

$$\sum_{i=1}^{n} \sigma_x(i)^2 = \sum_{i=1}^{n} \sigma_y(i)^2 = \sum_{i=1}^{n} \lambda_i. \tag{3.9}$$

The principal components with smaller variance can be discarded without affecting the total variance much.

For PCA, the goal is to find a linear transformation which is used to project an arbitrary vector $\mathbf{x}$ in a n-dimensional space to another vector $\mathbf{y} = [y_1, \cdots, y_K]^T$ in a K-dimensional space, where $K < n$. From Equation (3.5), $\mathbf{x}$ can be represented without error by the summation of n orthogonal vectors as

$$\mathbf{x} = \mathbf{U}\mathbf{y} = \sum_{i=1}^{n} y_i\mathbf{u}_i \tag{3.10}$$

where $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_n]$ and $\mathbf{y} = [y_1, \cdots, y_n]^T$. Each component y_i of $\mathbf{y}$ can be computed by Equation (3.4). The columns of $\mathbf{U}$ form an orthogonal set and span the n-dimensional space containing $\mathbf{x}$, that is,

$$\mathbf{u}_i^T\mathbf{u}_j = \begin{cases} 1 & \text{for } i = j, \\ 0 & \text{for } i \neq j. \end{cases} \tag{3.11}$$

If we take only K of the n vectors from this basis $\mathbf{U}$ to approximate $\mathbf{x}$, and replace the remaining coefficients by constants c_i in which $i = K+1, \cdots, n$. Then, the approximation

to the arbitrary vector $\mathbf{x}$, is $\tilde{\mathbf{x}}$ as given by

$$\tilde{\mathbf{x}} = \sum_{i=1}^{K} y_i \mathbf{u}_i + \sum_{i=K+1}^{n} c_i \mathbf{u}_i, \tag{3.12}$$

and the approximation error is given by

$$\begin{aligned} \mathbf{x} - \tilde{\mathbf{x}} &= \sum_{i=1}^{n} y_i \mathbf{u}_i - (\sum_{i=1}^{K} y_i \mathbf{u}_i + \sum_{i=K+1}^{n} c_i \mathbf{u}_i) \\ &= \sum_{i=K+1}^{n} (y_i - c_i)\mathbf{u}_i. \end{aligned} \tag{3.13}$$

In order to achieve the best approximation by the K orthogonal basis vectors, we need to choose c_i so as to minimise the mean-square error ξ_K^2 which is given by

$$\begin{aligned} \xi_K^2 &= E\{\|\mathbf{x} - \tilde{\mathbf{x}}\|^2\} \\ &= E\{\sum_{i=K+1}^{n} \sum_{j=K+1}^{n} [(y_i - c_i)\mathbf{u}_i]^T [(y_j - c_j)\mathbf{u}_j]\} \\ &= \sum_{i=K+1}^{n} E[(y_i - c_i)^2] \end{aligned} \tag{3.14}$$

in which we have used the orthogonal equation (3.11) to simplify the equation. The optimal values for c_i can be obtained by computing the partial derivative of ξ_K^2 with respect to c_i. That is to compute the equation

$$\begin{aligned} \frac{\partial}{\partial c_i} E[(y_i - c_i)^2] &= -2[E(y_i) - E(c_i)] \\ &= -2[E(y_i) - c_i] \\ &= 0 \end{aligned} \tag{3.15}$$

in which we find $c_i = E(y_i) = \mathbf{u}_i^T E(\mathbf{x})$ by equation (3.4) since $E(\mathbf{u}_i{}^T \mathbf{x}) = \mathbf{u}_i{}^T E(\mathbf{x})$. Therefore, the criterion function can be rewritten as

$$\begin{aligned} \xi_K^2 &= \sum_{i=K+1}^{n} E[(y_i - E(y_i))^2] \\ &= \sum_{i=K+1}^{n} \mathbf{u}_i^T E[(\mathbf{x} - E(\mathbf{x}))(\mathbf{x} - E(\mathbf{x}))^T]\mathbf{u}_i \\ &= \sum_{i=K+1}^{n} \mathbf{u}_i^T \Sigma_{\mathbf{x}} \mathbf{u}_i \end{aligned} \tag{3.16}$$

where $\Sigma_{\mathbf{x}}$ represents the covariance matrix of $\mathbf{x}$. The optimal choices for the $\mathbf{u}_i$ elements

$T = \lambda_1 + \lambda_2 + \cdots + \lambda_k + \cdots + \lambda_n$						
eigenvalue	λ_1	λ_2	$\cdots$	λ_k	$\cdots$	λ_n
percentage in total variance	$\frac{\lambda_1}{T}$	$\frac{\lambda_2}{T}$	$\cdots$	$\frac{\lambda_k}{T}$	$\cdots$	$\frac{\lambda_n}{T}$
accumulated percentage	$\frac{\sum_{i=1}^{1}\lambda_i}{T}$	$\frac{\sum_{i=1}^{2}\lambda_i}{T}$	$\cdots$	$\frac{\sum_{i=1}^{k}\lambda_i}{T}$	$\cdots$	$\frac{\sum_{i=1}^{n}\lambda_i}{T}$

Table 3.1: Relationship between eigenvalue and total variance.

are those which satisfy

$$\Sigma_{\mathbf{x}}\mathbf{u}_i = \lambda_i\mathbf{u}_i, \tag{3.17}$$

where λ_i and $\mathbf{u}_i$ are eigenvalues and eigenvectors of $\Sigma_{\mathbf{x}}$ and λ_i are in non-increasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$.

By substituting equation (3.17) into (3.16), the minimum mean-square error becomes

$$\xi^2_{K(min)} = \sum_{i=K+1}^{n} \lambda_i. \tag{3.18}$$

Thus, the minimum mean-square error is obtained by choosing the first K eigenvectors of $\Sigma_{\mathbf{x}}$ to optimally approximate the original vector $\mathbf{x}$. This linear transformation used to reduce the dimensionality is called the Principal Component Analysis or Karhunen-Loève Transform. Each of the eigenvectors $\mathbf{u}_i$ is called a *principal component* with the variance λ_i. Generally, we will use equation (3.18) as the approximation of data loss, and Table 3.1 as the strategy to select the best K. In Table 3.1, the value of total variance $T = \sum_{i=1}^{n} \lambda_i$ and the accumulated percentage P of K eigenvalues in total variance is given by

$$P = \frac{\sum_{i=1}^{K} \lambda_i}{T}. \tag{3.19}$$

P can be considered as the percentage of information retained by using K eigenvectors to approximate the vector $\mathbf{x}$. The percentage of information loss can be represented as

$$\frac{\sum_{i=K+1}^{n} \lambda_i}{T} = 1 - P. \tag{3.20}$$

3.3 Eigenspace Transformation (EST)

EST is widely used in face recognition and is already used in gait recognition [37]. For gait recognition, the way to recognise a given human walking sequence is to measure the similarities between unknown gait sequences and training sequences by spatio-temporal

image correlation. Spatio-temporal correlation is an extension of 2-D image correlation to 3-D correlation, in both the space and the time domains. The computational cost is very high if each image is large. Therefore, it is necessary to reduce the dimensionality of input images prior to using template matching to recognise different people. Based on the theory of PCA in Section 3.2, the eigenvectors of the data covariance matrix form an orthogonal basis and only the most significant ones are required to span a smaller subspace. By using the orthogonal basis of this subspace, each image can be transformed to a low-dimensional subspace and represented by a vector in this subspace [44]. Each input image can be mapped to a point in the eigenspace which is a vector with k elements, if the number of selected principal eigenvectors is k. Therefore, the implementation of spatio-temporal image correlation is converted from pixel-level comparisons to the distance measures between different parameter sets.

Assume that there are c training classes used for learning. $\mathbf{x'}_{i,j}$ is the j-th template image in class i, and N_i is the number of template images in the i-th class. The number of total training template images is $N_T = N_1 + N_2 + \cdots + N_c$. This training set is represented by

$$[\mathbf{x'}_{1,1}, \cdots, \mathbf{x'}_{1,N_1}, \mathbf{x'}_{2,1}, \cdots, \mathbf{x'}_{c,N_c}] \tag{3.21}$$

where each sample $\mathbf{x'}_{i,j}$ is a template image with n pixels. This is actually stored as a one-dimensional column vector with n elements. Firstly, the brightness of each template image is normalised by

$$\mathbf{x}_{i,j} = \frac{\mathbf{x'}_{i,j}}{\|\mathbf{x'}_{i,j}\|}. \tag{3.22}$$

After normalisation, the mean of the total set of images is given by

$$\mathbf{m_x} = \frac{1}{N_T} \sum_{i=1}^{c} \sum_{j=1}^{N_i} \mathbf{x}_{i,j}. \tag{3.23}$$

By subtracting the mean from each image, the training set can be described by an $n \times N_T$ matrix $\mathbf{X}$, where each image $\mathbf{x}_{i,j}$ forms one column of $\mathbf{X}$, that is

$$\begin{aligned} \mathbf{X} &= [\mathbf{x}_{1,1} - \mathbf{m_x}, \cdots, \mathbf{x}_{1,N_1} - \mathbf{m_x}, \mathbf{x}_{2,1} - \mathbf{m_x}, \cdots, \mathbf{x}_{c,N_c} - \mathbf{m_x}] \\ &= [\Psi_1, \cdots, \Psi_{N_1}, \Psi_{N_1+1}, \cdots, \Psi_{N_T}]. \end{aligned} \tag{3.24}$$

If the rank of the matrix $\mathbf{X}\mathbf{X}^T$ is K, then the K nonzero eigenvalues of $\mathbf{X}\mathbf{X}^T$, $\lambda_1, \ldots, \lambda_K$, and their associated eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_K$ satisfy the fundamental eigenvalue relationship

$$\mathbf{R}\mathbf{e}_i = \lambda_i \mathbf{e}_i, \qquad \text{i=1, } \ldots \text{,K} \tag{3.25}$$

where $\mathbf{R}$ is a square, symmetric $n \times n$ covariance matrix computed as

$$\mathbf{R} = \mathbf{X}\mathbf{X}^T \tag{3.26}$$

or

$$\mathbf{R} = \sum_{i=1}^{c} \sum_{j=1}^{N_i} (\mathbf{x}_{i,j} - \mathbf{m}_\mathbf{x})(\mathbf{x}_{i,j} - \mathbf{m}_\mathbf{x})^T. \tag{3.27}$$

The K eigenvectors are used as an orthogonal basis to span a new vector space. As such, each image can be projected to a point in this K-dimensional space. According to the previous section, the image data can be approximated by taking only the $k \leq K$ largest eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ and their associated eigenvectors $\mathbf{e}_1, \ldots, \mathbf{e}_k$. This partial set of k eigenvectors spans an eigenspace in which $[\mathbf{y}_{1,1}, \ldots, \mathbf{y}_{c,N_c}]$ are the projections of the original vectors $[\mathbf{x}_{1,1}, \ldots, \mathbf{x}_{c,N_c}]$ according to

$$\mathbf{y}_{i,j} = [\mathbf{e}_1, \cdots, \mathbf{e}_k]^T \mathbf{x}_{i,j} \tag{3.28}$$

where the matrix $[\mathbf{e}_1, \cdots, \mathbf{e}_k]^T$ is called the *eigenspace transformation matrix*.

The parametric eigenspace approach used here is based upon the global covariance matrix of the full set of image data and thus is not sensitive to class structure in the data. There will be a separability problem when the number of walking sequences for different people is greatly increased. Accordingly, the eigenspace transformation is used for reducing image dimensionality in the first stage for gait recognition. By considering the within-class and between-class covariance matrix, canonical analysis is able to increase the differences between the samples belonging to different classes. This will be described in Section 3.5 which follows after the computational considerations have been discussed.

3.4 Computational Considerations of EST

In order to solve equation (3.25), we need to calculate the eigenvalues and eigenvectors of the $n \times n$ matrix $\mathbf{X}\mathbf{X}^T$ which is computationally intractable for typical image sizes. If the number of training images is less than the number of pixels in each image ($N_T < n$), the number of meaningful eigenvectors will be only $(N_T - 1)$ rather than n and the associated eigenvalues of remaining $(n - N_T + 1)$ eigenvectors will be zero. Therefore, it is only necessary to calculate the $(N_T - 1)$ eigenvalues instead of n, together with their associated eigenvectors.

There are mainly three approaches [112, 80, 113] which compute the eigenvectors of an $N_T \times N_T$ matrix first and then reconstruct their corresponding eigenvectors in the original matrix $\mathbf{X}\mathbf{X}^T$ by a specific reconstruction. This reduces the size of the covariance matrix from $O(n^2)$ to $O(N_T^2)$ which involves a smaller computational cost and is very

attractive, when the number of training images is much smaller than the number of pixels in each image ($N_T \ll n$). Even with considerable availability of memory, the computation of the eigenvalues of $n \times n$ covariance matrix can still impose a problem which exceeds commonly-available resource. As such, and for simplicity, stability and suitability, we choose the second approach to compute the principal eigenvectors for our training set. These approaches are described below.

3.4.1 Eigenface Approach

According to the eigenface approach [80] in face recognition, the eigenvectors of matrix $\mathbf{X}^T\mathbf{X}$ can be obtained by the relation

$$\mathbf{X}^T\mathbf{X}\tilde{\mathbf{e}}_i = \tilde{\lambda}_i\tilde{\mathbf{e}}_i \tag{3.29}$$

By multiplying both sides by $\mathbf{X}$, the equation above becomes

$$\mathbf{X}\mathbf{X}^T\mathbf{X}\tilde{\mathbf{e}}_i = \tilde{\lambda}_i\mathbf{X}\tilde{\mathbf{e}}_i \tag{3.30}$$

where we can find that $\mathbf{X}\tilde{\mathbf{e}}_i$ are the eigenvectors of matrix $\mathbf{X}\mathbf{X}^T$. Following this analysis, we compute the eigenvectors $\tilde{\mathbf{e}}_i$ $(i = 1, \cdots, N_T)$ of equation (3.29) first and then the first N_T eigenvalues and eigenvectors of matrix $\mathbf{X}\mathbf{X}^T$ can be reconstructed by

$$\begin{cases} \lambda_i = \tilde{\lambda}_i \\ \mathbf{e}_i = \sum_{k=1}^{N_T} \tilde{e}_{i,k}\Psi_k \end{cases} \tag{3.31}$$

where $i = 1, \ldots, N_T$ and $\tilde{e}_{i,k}$ is the k-th element in vector $\tilde{\mathbf{e}}_i$. Ψ_k is the k-th image in the training set where $\Psi_k, k = 1, \ldots, N_T$.

3.4.2 Singular Value Decomposition (SVD) Algorithm

Based on *singular value decomposition* theory [112, 113], we can compute another matrix $\tilde{\mathbf{R}}$ instead of $\mathbf{R}$, that is

$$\tilde{\mathbf{R}} = \mathbf{X}^T\mathbf{X} \tag{3.32}$$

in which the matrix size is $N_T \times N_T$ and is much smaller than $n \times n$ in practical problems. Suppose the matrix $\tilde{\mathbf{R}}$ has eigenvalues $\tilde{\lambda}_1, \ldots, \tilde{\lambda}_{N_T}$ and associated eigenvectors $\tilde{\mathbf{e}}_i, \ldots, \tilde{\mathbf{e}}_{N_T}$ which are related to those in $\mathbf{R}$ by

$$\begin{cases} \lambda_i = \tilde{\lambda}_i \\ \mathbf{e}_i = \tilde{\lambda}_i^{-\frac{1}{2}}\mathbf{X}\tilde{\mathbf{e}}_i \end{cases} \tag{3.33}$$

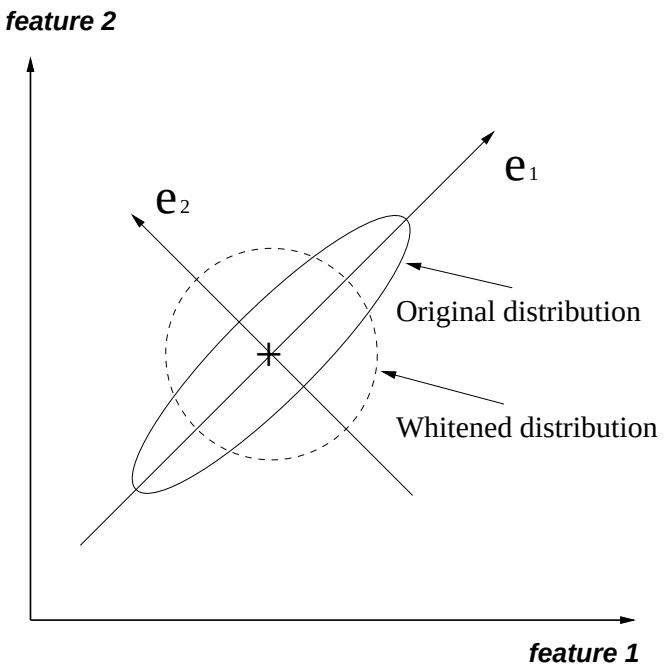


Figure 3.3: Comparison between with and without whitening transformation

where $i = 1, \ldots, N_T$. This is similar to the previous approach and the difference is that this approach adds a normalisation term, $\tilde{\lambda}_i^{-\frac{1}{2}}$, to normalise each original eigenvector during reconstruction. This is called the *whitening transformation* [108] which converts the covariance matrix $\Sigma_{\mathbf{y}}$ in Equation (3.7) equal to the unit matrix $\mathbf{I}$ by

$$\begin{aligned} \Sigma_{\mathbf{y}} &= \mathbf{D}_\lambda{}^{-\frac{1}{2}} \mathbf{U}^T \Sigma_{\mathbf{x}} \mathbf{U} \mathbf{D}_\lambda{}^{-\frac{1}{2}} \\ &= \mathbf{D}_\lambda{}^{-\frac{1}{2}} \mathbf{D}_\lambda \mathbf{D}_\lambda{}^{-\frac{1}{2}} \\ &= \mathbf{I} \end{aligned} \tag{3.34}$$

where $\mathbf{D}_\lambda{}^{-\frac{1}{2}} = Diag(\lambda_1^{-\frac{1}{2}}, \cdots, \lambda_{N_T}^{-\frac{1}{2}})$. After the whitening transformation, the scales of the principal components are changed proportionally to $\lambda_i^{-\frac{1}{2}}$. Thus, this approach is more stable than the previous approach. The effect of the whitening transformation to projected data in the eigenspace is shown in Figure 3.3.

3.4.3 Spatial Temporal Adaptive (STA) Algorithm

This algorithm proposed by Murase and Lindenbaum [113] uses the coefficients from Discrete Cosine Transform (DCT) [114] instead of image pixel data to reconstruct the original eigenvectors. Let $\mathbf{U}$ be a DCT transformation matrix in which its column vectors are orthogonal to each other, the training set $\mathbf{X}$ can be converted to

$$\mathbf{Y} = \mathbf{U}\mathbf{X} \tag{3.35}$$

in which $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. By using equation (3.35), equation (3.33) can be rewritten by

$$\begin{aligned} \mathbf{e}_i &= \tilde{\lambda}_i^{-\frac{1}{2}}\mathbf{X}\tilde{\mathbf{e}}_i \\ &= \tilde{\lambda}_i^{-\frac{1}{2}}\mathbf{U}^{-1}\mathbf{Y}\tilde{\mathbf{e}}_i \\ &= \mathbf{U}^{-1}(\tilde{\lambda}_i^{-\frac{1}{2}}\mathbf{Y}\tilde{\mathbf{e}}_i) \end{aligned} \tag{3.36}$$

where $\mathbf{Y}$ is the DCT of $\mathbf{X}$, and $\mathbf{U}^{-1}$ denotes an inverse DCT operation. Therefore, the original eigenvectors can be obtained by

$$\begin{cases} \lambda_i = \tilde{\lambda}_i \\ \mathbf{e}_i = \mathbf{U}^{-1}(\tilde{\lambda}_i^{-\frac{1}{2}}\mathbf{Y}\tilde{\mathbf{e}}_i) \end{cases} \tag{3.37}$$

where $i = 1, \ldots, N_T$.

Although this approach is efficient for natural and large grey-level images by using DCT coefficients for the reconstruction of eigenvectors instead of using original images [113], it involves more computation for the DCT and inverse DCT when compared to the previous approach in Section 3.4.2. Moreover, its efficiency is achieved by using a small number of DCT coefficients to approximate each image. This will result in the information loss of each eigenvector which is reconstructed from this partial set of DCT coefficients. Therefore, it is not adopted for our template images.

3.5 Canonical Analysis

Canonical Analysis (CA) - also called Linear Discriminant Analysis (LDA), is used after PCA for class separation by maximising between-class and minimising within-class variations. A comparison of PCA with CA is shown in Figure 3.4. According to Figure 3.4, data points in two classes are not separable after projection into the principal axis using PCA. After using CA, the two classes have the largest possible separation between their means after projection onto that axis and appear as small as possible in their individual spreads (i.e. the clusters have minimum variance).

A brief introduction to canonical analysis is described as follows [108]. Suppose that the vector set $\{\Phi_1, \Phi_2, \cdots, \Phi_c\}$ represents the c classes of training vectors and $\mathbf{y}_{i,j}$ is the j-th vector in class Φ_i $i = 1, \cdots, c$. In this thesis, $\mathbf{y}_{i,j}$ represents the projected vector of each template image in the eigenspace. The mean of the total vectors is given by

$$\mathbf{m}_\mathbf{y} = \frac{1}{N_T}\sum_{i=1}^{c}\sum_{j=1}^{N_i}\mathbf{y}_{i,j}, \tag{3.38}$$

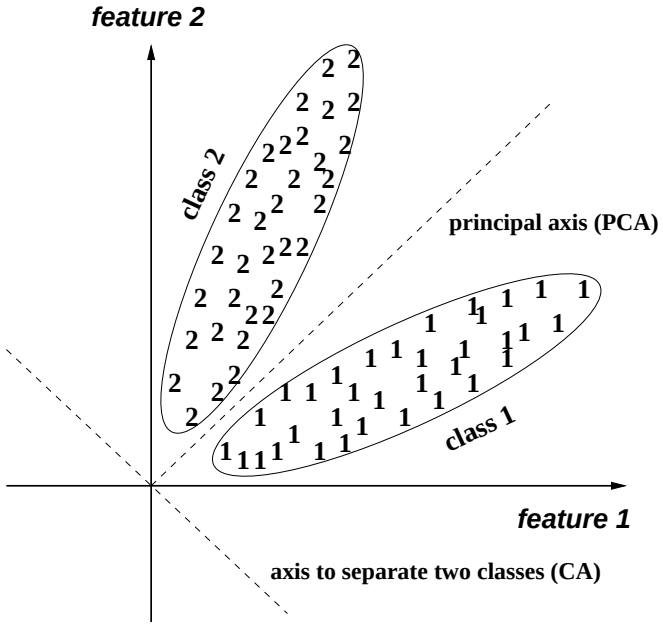


Figure 3.4: Comparison of PCA and CA in class separation

and the mean of vectors in the i-th class is represented by

$$\mathbf{m}_i = \frac{1}{N_i} \sum_{\mathbf{y}_{i,j} \in \Phi_i} \mathbf{y}_{i,j}. \tag{3.39}$$

To handle data with any number of dimensions it is necessary to define average data scatter within the classes, and the scatter of the classes themselves around the multispectral space, by covariance matrices [115]. Let the *total scatter matrix* $\mathbf{S_t}$, the *within-class scatter matrix* $\mathbf{S_w}$ and the *between-class scatter matrix* $\mathbf{S_b}$ be defined by

$$\mathbf{S_t} = \frac{1}{N_T} \sum_{i=1}^{c} \sum_{j=1}^{N_i} (\mathbf{y}_{i,j} - \mathbf{m_y})(\mathbf{y}_{i,j} - \mathbf{m_y})^T \tag{3.40}$$

$$\mathbf{S_w} = \frac{1}{N_T} \sum_{i=1}^{c} \sum_{\mathbf{y}_{i,j} \in \Phi_i} (\mathbf{y}_{i,j} - \mathbf{m}_i)(\mathbf{y}_{i,j} - \mathbf{m}_i)^T \tag{3.41}$$

$$\mathbf{S_b} = \frac{1}{N_T} \sum_{i=1}^{c} N_i (\mathbf{m}_i - \mathbf{m_y})(\mathbf{m}_i - \mathbf{m_y})^T \tag{3.42}$$

where $\mathbf{S_w}$ in (3.41) represents the scatter of samples around their respective class expected vectors and $\mathbf{S_b}$ in (3.42) stands for the scatter of the expected vectors around the global mean $\mathbf{m_y}$.

Suppose that there is a projection basis $\mathbf{W}$ which can project each vector, $\mathbf{y}_{i,j}$, to another space by

$$\mathbf{z}_{i,j} = \mathbf{W}^T \mathbf{y}_{i,j}$$

where $\mathbf{z}_{i,j}$ is a point in the new space. The objective is to maximise the class separability between classes. Therefore, we have to find the matrix $\mathbf{W}$ which can simultaneously minimise within-class distances and maximise between-class distances, that is to maximise the criterion function which is known as the *generalised Fisher linear discriminant function* [107]

$$\mathbf{J}(\mathbf{W}) = \frac{\mathbf{W}^T\mathbf{S_b}\mathbf{W}}{\mathbf{W}^T\mathbf{S_w}\mathbf{W}}. \tag{3.43}$$

Let $\mathbf{w}$ be one particular column vector in $\mathbf{W}$ which is called a *canonical axis* and the classes will be optimally separated along this axis by maximising

$$\lambda = \frac{\mathbf{w}^T\mathbf{S_b}\mathbf{w}}{\mathbf{w}^T\mathbf{S_w}\mathbf{w}} \tag{3.44}$$

where λ is the ratio of variances. The ratio of variances in the new space is maximised by the selection of a feature $\mathbf{w}$ if

$$\frac{\partial\lambda}{\partial\mathbf{w}} = 0. \tag{3.45}$$

According to [115], this equation can be solved by the following procedures. Noting the identity that $\frac{\partial}{\partial\mathbf{x}}\{\mathbf{x}^TA\mathbf{x}\} = 2A\mathbf{x}$ then

$$\begin{aligned}\frac{\partial\lambda}{\partial\mathbf{w}} &= \frac{\partial}{\partial\mathbf{w}}\{(\mathbf{w}^T\mathbf{S_b}\mathbf{w})(\mathbf{w}^T\mathbf{S_w}\mathbf{w})^{-1}\} \\ &= 2\mathbf{S_b}\mathbf{w}(\mathbf{w}^T\mathbf{S_w}\mathbf{w})^{-1} - 2\mathbf{S_w}\mathbf{w}(\mathbf{w}^T\mathbf{S_b}\mathbf{w})(\mathbf{w}^T\mathbf{S_w}\mathbf{w})^{-2} \\ &= 0. \end{aligned} \tag{3.46}$$

Thus,

$$\mathbf{S_b}\mathbf{w} - \mathbf{S_w}\mathbf{w}(\mathbf{w}^T\mathbf{S_b}\mathbf{w})(\mathbf{w}^T\mathbf{S_w}\mathbf{w})^{-1} = 0 \tag{3.47}$$

which can be represented as

$$(\mathbf{S_b} - \lambda\mathbf{S_w})\mathbf{w} = 0 \tag{3.48}$$

where $\lambda = (\mathbf{w}^T\mathbf{S_b}\mathbf{w})(\mathbf{w}^T\mathbf{S_w}\mathbf{w})^{-1}$. Equation (3.48) needs to be solved for the unknowns λ and $\mathbf{w}$. This canonical axis will be in the direction of $\mathbf{w}$ and λ will give the associated ratio of between-class to within-class variance along that axis. Let $\mathbf{W}^*$ be the optimal solution for the projection basis and λ_i be its i-th largest eigenvalue with the eigenvector $\mathbf{w}_i^*$. Equation (3.48) can be written as

$$\mathbf{S_b}\mathbf{w}_i^* = \lambda_i\mathbf{S_w}\mathbf{w}_i^* \tag{3.49}$$

or equivalently,

$$\mathbf{S}_\mathbf{w}^{-1}\mathbf{S}_\mathbf{b}\mathbf{w}_i^* = \lambda_i \mathbf{w}_i^* \tag{3.50}$$

provided that $\mathbf{S}_\mathbf{w}$ is nonsingular. Equation (3.49) is called a *generalised eigenvalue equation*.

Note that in a c-class classification problem, the rank of $\mathbf{S}_\mathbf{b}$ is not greater than $(c-1)$. Therefore, there are at most $(c-1)$ nonzero eigenvalues. The larger the eigenvalue is, the better the discrimination that can be achieved. That is, the eigenvector associated with each nonzero eigenvalue can be treated as a discriminant vector with the discrimination ability determined by the magnitude of the corresponding eigenvalue. By using these $(c-1)$ eigenvectors the c classes can be separated effectively.

3.6 Computational Considerations of CA

According to the computational considerations [103], if $\mathbf{S}_\mathbf{w}^{-1}\mathbf{S}_\mathbf{b}$ is not symmetric, the solution of this generalised eigenvalue problem will be unstable. Based on the technique of simultaneous diagonalisation [108], Swets and Weng [103] proposed an approach which can simultaneously diagonalise those two matrices, $\mathbf{S}_\mathbf{w}$ and $\mathbf{S}_\mathbf{b}$, and produce a stable eigensystem computation procedure. The approach is described here.

Suppose that $\mathbf{S}_\mathbf{w}$ can be represented by $\mathbf{S}_\mathbf{w} = \mathbf{V}\Lambda\mathbf{V}^T$, in which $\mathbf{V}$ is orthogonal and Λ is diagonal. Therefore, $(\mathbf{V}\Lambda^{-\frac{1}{2}})^T\mathbf{S}_\mathbf{w}\mathbf{V}\Lambda^{-\frac{1}{2}} = \mathbf{I}$. If we compute $\mathbf{U}$ and Σ and let $(\mathbf{V}\Lambda^{-\frac{1}{2}})^T\mathbf{S}_\mathbf{b}\mathbf{V}\Lambda^{-\frac{1}{2}} = \mathbf{U}\Sigma\mathbf{U}^T$, where $\mathbf{U}$ is orthogonal and Σ is diagonal. Then $\mathbf{S}_\mathbf{w}$ and $\mathbf{S}_\mathbf{b}$ are represented as

$$\mathbf{S}_\mathbf{b} = \mathbf{V}\Lambda^{\frac{1}{2}}\mathbf{U}\Sigma\mathbf{U}^T\Lambda^{\frac{1}{2}}\mathbf{V}^T \tag{3.51}$$

$$\mathbf{S}_\mathbf{w} = \mathbf{V}\Lambda^{\frac{1}{2}}\mathbf{U}\mathbf{I}\mathbf{U}^T\Lambda^{\frac{1}{2}}\mathbf{V}^T. \tag{3.52}$$

Since

$$\mathbf{S}_\mathbf{w}^{-1} = \mathbf{V}\Lambda^{-1}\mathbf{V}^T, \tag{3.53}$$

we define $\nabla = \mathbf{V}\Lambda^{-\frac{1}{2}}\mathbf{U}$ where ∇ can diagonalise $\mathbf{S}_\mathbf{w}$ and $\mathbf{S}_\mathbf{b}$ simultaneously. Using equation (3.51) and equation (3.53), the generalised eigenvalue equation can be solved by

$$\begin{aligned}\mathbf{S}_\mathbf{w}^{-1}\mathbf{S}_\mathbf{b} &= \mathbf{V}\Lambda^{-1}\mathbf{V}^T\mathbf{V}\Lambda^{\frac{1}{2}}\mathbf{U}\Sigma\mathbf{U}^T\Lambda^{\frac{1}{2}}\mathbf{V}^T \\ &= \mathbf{V}\Lambda^{-\frac{1}{2}}\mathbf{U}\Sigma\mathbf{U}^T\Lambda^{\frac{1}{2}}\mathbf{V}^T \\ &= \nabla\Sigma\nabla^{-1}. \end{aligned} \tag{3.54}$$

In equation (3.54), the eigenvalues and eigenvectors of $\mathbf{S}_\mathbf{w}^{-1}\mathbf{S}_\mathbf{b}$ are included in Σ and ∇.

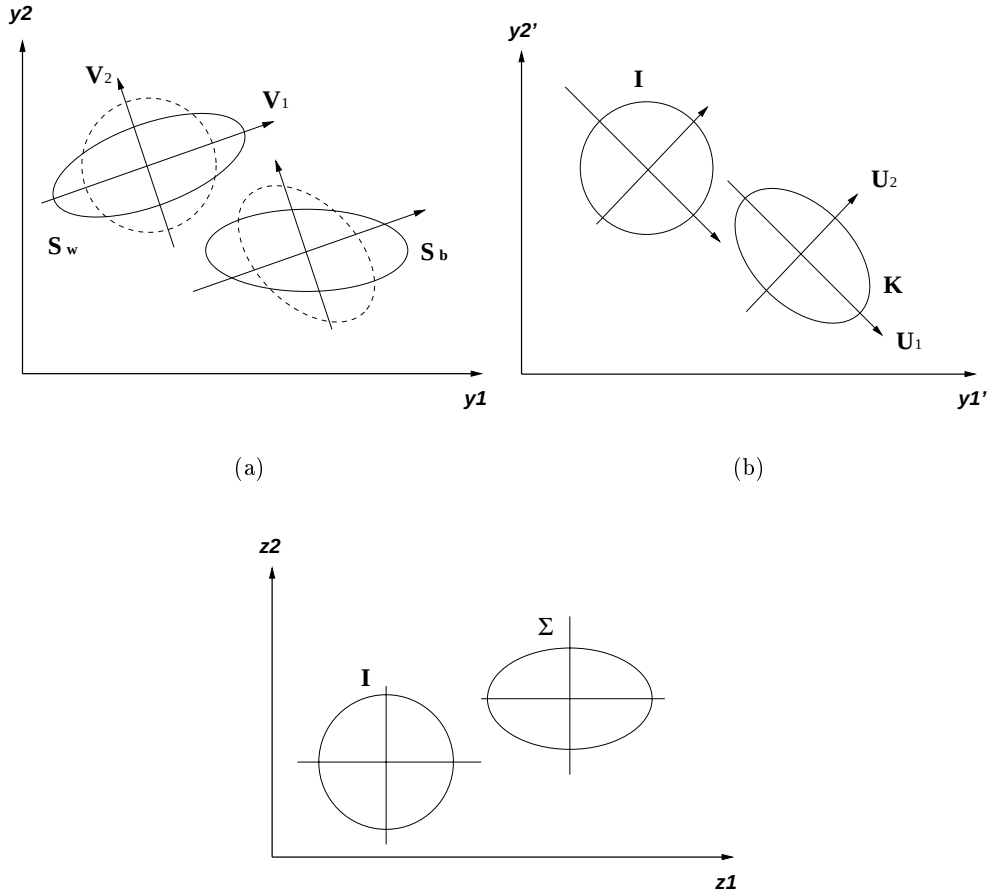


Figure 3.5: Simultaneous diagonalisation

According to [108], a two-dimensional example of this process can be drawn in Figure 3.5 in which $\mathbf{K} = \mathbf{U}\Sigma\mathbf{U}^T$. In the first step shown in Figure 3.5(a), $\mathbf{S_w}$ has been whitened and diagonalised by $(\mathbf{V}\Lambda^{-\frac{1}{2}})^T$ which has also transformed $\mathbf{S_b}$ to $\mathbf{K}$. Figure 3.5(b) shows the second step that $\mathbf{K}$ can be diagonalised by $\mathbf{U}$ which also rotates the whitened $\mathbf{S_w}$. Combining the first and the second step, $\mathbf{S_w}$ and $\mathbf{S_b}$ have been diagonalised to $\mathbf{I}$ and Σ simultaneously by ∇ which is shown in Figure 3.5(c). By using this technique to diagonalise $\mathbf{S_w}$ and $\mathbf{S_b}$ simultaneously, the generalised eigenvalue equation in Equation (3.49) can be solved at the same time. This can also avoid the instability resulted in by calculating the inverse of $\mathbf{S_w}$, therefore, this method is used in this thesis.

3.7 Canonical Space Transformation (CST)

The parametric eigenspace approach in Section 3.3 is based upon the global covariance matrix of the full set of image data and generates the eigenvalues and eigenvectors to rotate the original data coordinates along the direction of maximum variance. The

reason it often works well as a feature reduction tool is that classes are frequently distributed in the direction of maximum data scatter [115]. When good class separation cannot be achieved by the eigenspace transformation, canonical analysis is the technique which can be used to generate a transformed set of feature axes and optimise the class separability. By using the features derived from canonical analysis to further process the transformed data in eigenspace, the classification performance should, therefore, be improved. Therefore, canonical analysis is used in the second stage after PCA for gait recognition.

After the generalised eigenvalue equation in Equation (3.49) is solved, we will obtain $(c-1)$ nonzero eigenvalues and their corresponding eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_{c-1}$ that create another orthogonal basis and span a $(c-1)$-dimensional canonical space. The length of each eigenvector $\mathbf{v}_i$ $(i = 1, \cdots, c-1)$ equals to the length of $\mathbf{e}_j$ $(j = 1, \cdots, k)$ in Equation (3.28). By using this basis, each point in eigenspace can be projected to another point in this canonical space by

$$\mathbf{z}_{i,j} = [\mathbf{v}_1, \cdots, \mathbf{v}_{c-1}]^T \mathbf{y}_{i,j}, \tag{3.55}$$

where $\mathbf{z}_{i,j}$ represents the new point and $[\mathbf{z}_{i,1}, \ldots, \mathbf{z}_{i,N_i}]$ is the new trajectory in canonical space. We call the matrix $[\mathbf{v}_1, \cdots, \mathbf{v}_{c-1}]$ the *canonical space transformation matrix*. According to this analysis, different classes will be greatly separated. That means that canonical analysis is useful to separate different samples of human gait even when the number of samples of different gait sequence is increased.

3.8 Combination of EST and CST

By merging equation (3.28) and equation (3.55), each template image can be projected into one point in the new $(c-1)$-dimensional space by

$$\begin{aligned} \mathbf{z}_{i,j} &= [\mathbf{v}_1, \cdots, \mathbf{v}_{c-1}]^T \mathbf{y}_{i,j} \\ &= [\mathbf{v}_1, \cdots, \mathbf{v}_{c-1}]^T [\mathbf{e}_1, \cdots, \mathbf{e}_k]^T \mathbf{x}_{i,j} \\ &= \mathbf{H}\mathbf{x}_{i,j}. \end{aligned} \tag{3.56}$$

in which $\mathbf{H} = [\mathbf{v}_1, \ldots, \mathbf{v}_{c-1}]^T [\mathbf{e}_1, \ldots, \mathbf{e}_k]^T$. The eigenspace transformation matrix $[\mathbf{e}_1, \ldots, \mathbf{e}_k]$ and canonical space transformation matrix $[\mathbf{v}_1, \ldots, \mathbf{v}_{c-1}]$ are then combined as one single transformation matrix $\mathbf{H}$ which is a $(c-1) \times k$ matrix. This new transformation matrix $\mathbf{H}$ can be used to directly project each template image into one point in the new $(c-1)$-dimensional canonical space which is different from the eigenspace. Meanwhile, it inherits the merits of PCA and CA which can not only reduce the data dimensionality by optimal approximation, but also achieve the best class separability simultaneously. Figure 3.6 shows 3 figures for the comparison of EST and the combined approach, EST and CST, when applied to a data set with 3 classes and 10 points each.

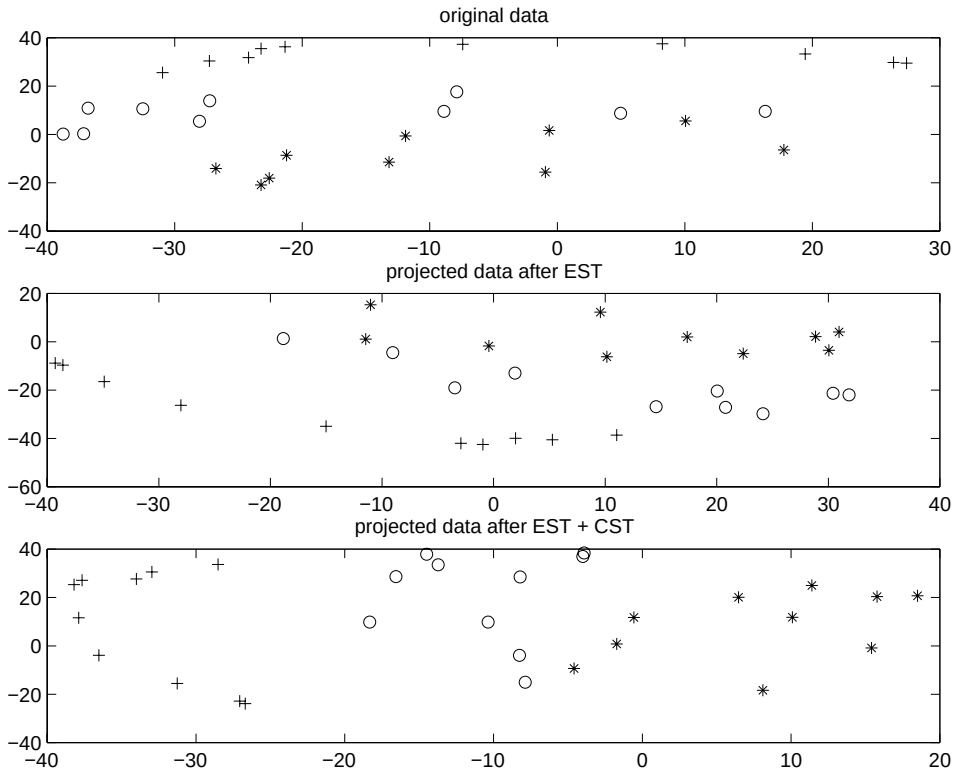


Figure 3.6: The comparison of EST and the combined approach - EST and CST

In those 3 figures, 3 different symbols, '*', 'o' and '+', represent the data points of 3 different classes. The distribution of original data is shown in the upper figure. The figure in the middle shows the data distribution after EST only. The distribution after EST and CST is shown in the bottom figure. Clearly, the class separation in the bottom figure by EST and CST is better than the middle figure where only the EST is applied since the classes become much more compact and the boundaries between them are clearer.

3.9 Conclusions

In the previous sections, the theory of PCA and CA, and their combination in the two stages of feature extraction has been described. The combined approach is used later in gait recognition. Basically, two transformations, eigenspace transformation and canonical space transformation, are combined to project each template image from the high-dimensional image space to a low-dimensional eigenspace and then a canonical space. The first stage performs dimensionality reduction by eigenspace transformation. The second stage achieves class separability by canonical space transformation. The recognition is done in the canonical space. In this chapter, we have concentrated on the feature vector extraction from each template image using statistical transformation which combines eigenspace transformation and canonical space transformation. These

give mechanisms which are appropriate for reducing the data dimension associated with processing large sequences of image data. For example, the size of each template image in the experiments of the next chapter is $64 \times 64 = 4096$. This requires pixel operations of $4096 \times 100 = 409,600$ for the template matching of two sequences with 100 templates each. By using the combined approach of EST and CST, the size of each template is reduced from 4096 to a 5-dimensional vector for 6 classes in canonical space. The pixel operations are greatly reduced to $5 \times 100 = 500$ which is 800 times less (and hence faster than) than the original templates. As such, this affords a mechanism not only to facilitate gait recognition by statistical means but also to decrease the recognition time drastically. Depending on different applications, various distance measures are used for recognition. The main objective of this thesis is to investigate how to recognise humans by their gait. Automatic gait recognition via statistical approaches is described in the following chapters using various template features. In order to compare the performance between PCA and the combined approach, another biometric application, face recognition, is also presented in the next chapter.

Chapter 4

Statistical Gait Recognition Using Spatial Templates

4.1 Introduction

This chapter describes a gait recognition system [116, 117, 118, 1, 119] using the statistical approach - EST and CST for feature extraction, as proposed in the previous chapter. This method uses inherent facilities of these techniques to reduce the data dimensionality and optimise the class separability of different gait sequences simultaneously. Based on the hypothesis that gait recognition by the human visual system depends on the spatial changes of human body and motivated by the EST approach for gait recognition [37], we choose template matching for gait recognition. The objective of gait recognition is to recognise people by the way they walk. Thus, the only concern is the changes of human shape without regard to the clothes worn or to differing background. Therefore, we propose to use spatial templates which are binary images of human silhouettes extracted from each scene image. These templates are rescaled to be image templates all with the same size. This will eliminate the redundancy introduced by irrelevant background data, by scale changes during walking and by different clothing worn by the subject. Besides, using templates with same size is consistent with application of the statistical approaches of PCA and CA. Template matching has been successfully applied for gait recognition using *spatio-temporal correlation* in the eigenspace [37]. However, the eigenspace approach can reduce image dimensionality without considering class separability, so is not optimal for discriminating different gaits. Moreover, the whole gait sequence of all training subjects has to be retained in the database for further matching. This requires a massive storage space. In this chapter, we propose the statistical approach - EST and CST for feature extraction and *accumulated distance* to training centroids in the canonical space for recognition. This will eliminate matching problems caused by velocity changes and phase shifts. Only training centroids need to be retained in the database.

Before the feature extraction by EST and CST, preprocessing is needed and applied to each gait sequence which will be converted to a template sequence. Template extraction is done by simple subtraction and thresholding in [37]. In order to increase the robustness of segmentation, we choose region growing [120] after subtraction instead. After preprocessing, all the template sequences from different subjects for training are analysed by PCA and CA. This will generate a transformation matrix which can project each template sequence into the canonical space with separate clusters for different subjects. Matching a test gait sequence to a training subject is done by choosing the minimum *accumulated distance* to training centroids in the canonical space. Therefore, recognition of human gait becomes much more accurate and robust in this new space by projecting each image template from the high-dimensional image space to a low-dimensional canonical space. Similar approaches have been also successfully used in image retrieval from a database [103], face recognition [104] and data classification in remote sensing [102]. This statistical approach is also applied for feature extraction in face recognition which is also investigated in this chapter. Since the original grey-level images are used for face recognition, no further preprocessing is needed after brightness normalisation for face images.

Basically, this chapter is organised as follows. Section 4.2 briefly outlines the gait recognition system using spatial templates. In Section 4.3, how spatial templates are extracted from a gait sequence in the preprocessing stage is explained. In Section 4.4 we describe the basic theories of our approach inherited from previous chapter for feature extraction and how to project each template into the canonical space by EST and CST. Then, Section 4.6 shows experimental results for gait recognition using EST and CST. The comparison of the EST approach, the frequency approach of Little and Boyd and our combined approach are also shown here. Our approach is also applied to face recognition for comparison shown in Section 4.7. Results are discussed in Section 4.8, prior to conclusions in Section 4.9.

4.2 System Overview

Basically, the recognition process is comprised of five steps:

1. generating a sequence of silhouette images, spatial templates, by extracting the shape of a walking person from the background;
2. projecting the template images into the eigenspace;
3. projecting each point in the eigenspace into the canonical space;
4. in the canonical space, calculating the accumulated distance of each test template sequence to each centroid of reference patterns in a database; and then

5. the class label is derived from the minimum distance of the centroid to the training data centroids.

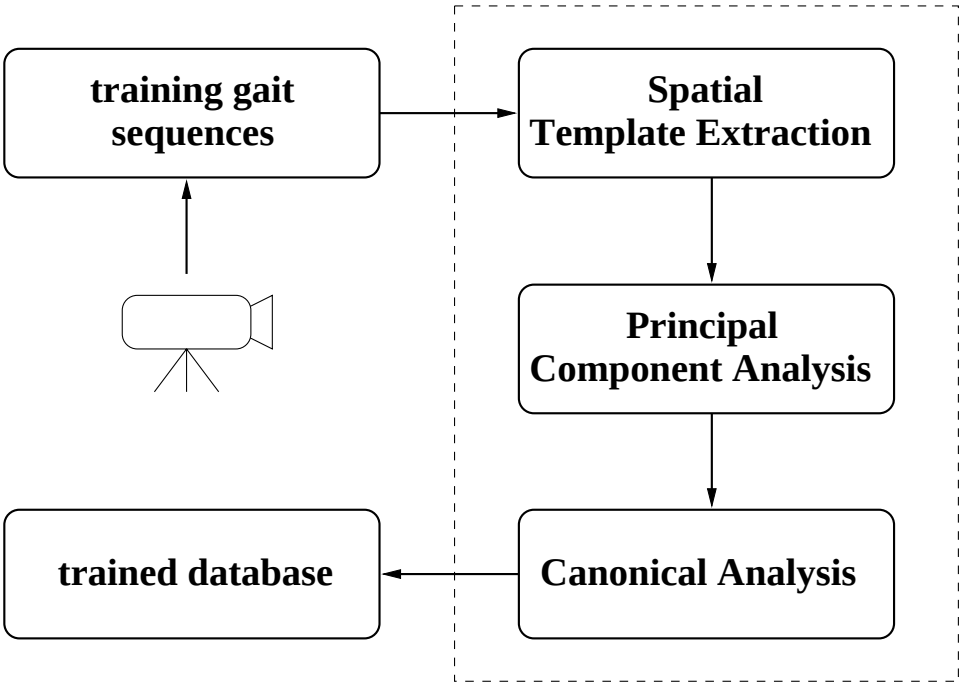
Figure 4.1 gives two block diagrams of the gait recognition system for the steps in the training and the test procedures. Preprocessing of spatial template extraction is needed in both the training and the test procedure. Figure 4.1(a) describes the steps in the training procedure. After applying principal component analysis to spatial templates, an eigenspace transformation matrix is generated and each spatial template from training sequences is projected into the eigenspace. Applying Canonical analysis to projected vectors in the eigenspace generates a canonical space transformation matrix which converts each point in the eigenspace to the canonical space. The projected vectors of all training templates in the canonical space and the matrix combining the eigenspace transformation matrix with the canonical space transformation matrix are included in the trained database for further recognition.

Figure 4.1(b) shows how to recognise a gait sequence in the test procedure. By referring to the trained database generated in the training procedure, each test gait sequence can be recognised after spatial templates are extracted and projected into the canonical space. The distance measures for recognition is described in Section 4.5. The block diagram of template projection in Figure 4.1(b) is shown in Figure 4.2.

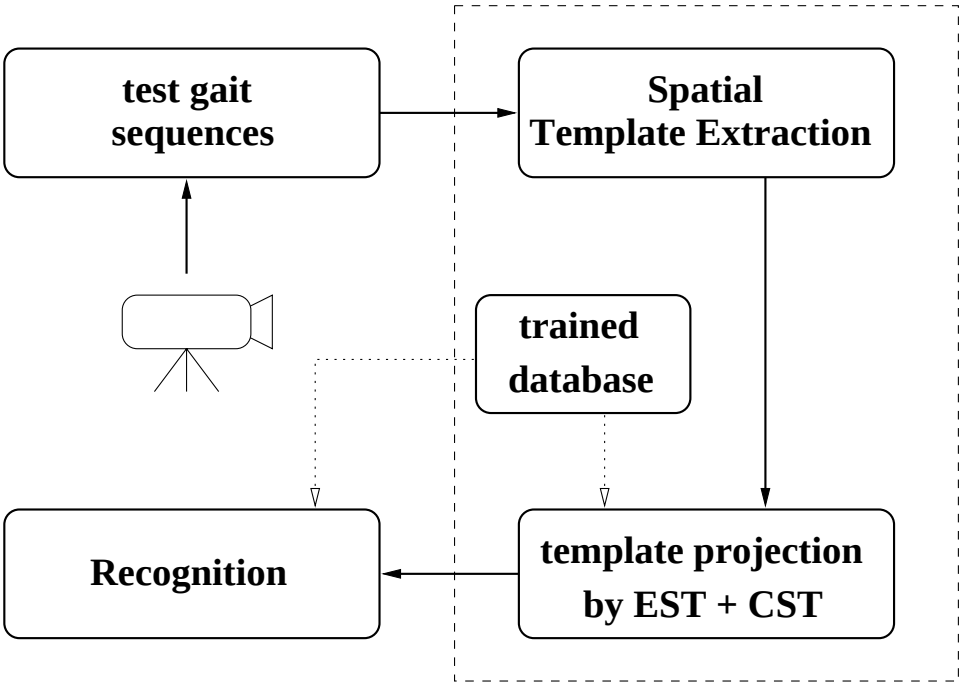
4.3 Preprocessing - Spatial Template Extraction

The aim of preprocessing is to improve the image data so as to suppress unwanted distortions or to enhance some image features important in later analysis. Geometric transformations of images (e.g. rotation, scaling, translation) are also classified among preprocessing methods in [121] since similar techniques are used. Motivated by [37], spatial templates are used for gait recognition in this chapter. Spatial templates are binary images of human silhouettes extracted from each image and rescaled to be image templates, here all with the same size of 64×64 pixels.

We make two assumptions concerning the human walking sequences. The first is that one single subject is walking laterally before a static camera, and secondly, that the body is not occluded by other objects. Naturally, to isolate the human silhouette, we can simply subtract the background from each image. To obtain an approximate background image for a walking sequence, a mean image is computed by averaging grey-level values for each pixel position over the entire image sequence. This mean image is then subtracted from each image to extract the desired silhouette. Obviously, the difference image thus obtained is not binarised. In [37] only difference patterns were used, instead a region growing technique [120] is applied here to obtain a binary image. From this binarised image, the bounding area and centroid of the human body are obtained. They are used in the normalisation of the segmented human body into each template frame. Finally, the position and size of the silhouette are normalised and fitted into a 64×64 template



(a) Training diagram



(b) Test diagram

Figure 4.1: Block diagrams of training and test for spatial templates

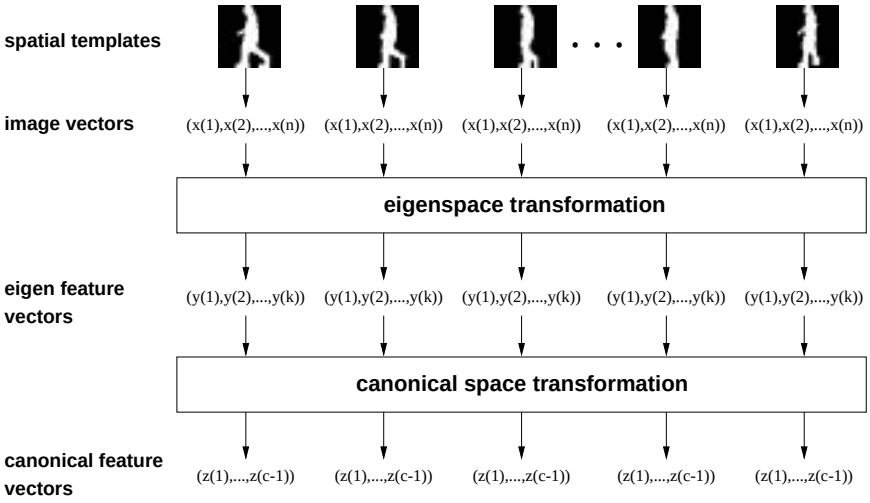


Figure 4.2: Projection of spatial templates by EST and CST

frame. The aspect ratio is kept constant when the size is normalised. Figure 4.3 shows an original human walking image in a sequence, the background image, the preprocessed image and the extracted spatial template. Sample spatial templates extracted from this sequence in order are illustrated in Figure 4.4.

4.4 Feature Extraction

After the extraction of spatial templates from all training sequences, the second step is feature extraction. Essentially, we combine two transformations - EST based on PCA and CST based on CA, for the feature extraction from spatial templates. Spatial templates in a high-dimensional image space are converted to a low-dimensional eigenspace by EST. The vectors obtained are further projected to a canonical space by CST. Recognition is actually accomplished in the canonical space. Projection stages can be referred to Figure 4.2. Eventually, each template image is converted to a one-dimensional canonical vector. Patently, the reduced dimensionality results in concomitant decrease in computational cost.

Assume that there are c training classes which represent different subjects for training. For gait recognition, each class provides a walking sequence with training spatial templates from an individual subject. For face recognition, each class gives a different number of training images for different experiments. Suppose $\mathbf{x}_{\mathbf{s}(i,j)}'$ is the j-th spatial template in class i, and N_i is the number of template images in i-th class. Note that $\mathbf{x}_{\mathbf{s}(i,j)}'$ is a column vector with each pixel scanned from the top row to the bottom row and placed in each corresponding element. The total number of training images is $N_T = N_1 + N_2 + \cdots + N_c$. This training set is represented by

$$[\mathbf{x}_{\mathbf{s}(1,1)}', \cdots, \mathbf{x}_{\mathbf{s}(1,N_1)}', \mathbf{x}_{\mathbf{s}(2,1)}', \cdots, \mathbf{x}_{\mathbf{s}(c,N_c)}'] \tag{4.1}$$

(a) Original image

(b) Background image

(c) Subtracted image

(d) Extracted spatial template

Figure 4.3: Human silhouette extraction

(a) $t+1$ (b) $t+2$ (c) $t+3$ (d) $t+4$ (e) $t+5$ (f) $t+6$

(g) $t+7$ (h) $t+8$ (i) $t+9$ (j) $t+10$ (k) $t+11$ (l) $t+12$

Figure 4.4: Sample spatial templates of a subject

where each sample $\mathbf{x}_{\mathbf{s}(i,j)}'$ is a spatial template with n pixels.

At first, the brightness of each sample template is normalised by

$$\mathbf{x}_{\mathbf{s}(i,j)} = \frac{\mathbf{x}_{\mathbf{s}(i,j)}'}{\|\mathbf{x}_{\mathbf{s}(i,j)}'\|}. \tag{4.2}$$

After normalisation, the mean template image for the full training set is given by

$$\mathbf{m}_{\mathbf{x_s}} = \frac{1}{N_T}\sum_{i=1}^{c}\sum_{j=1}^{N_i}\mathbf{x}_{\mathbf{s}(i,j)}. \tag{4.3}$$

By subtracting the mean from each template image, the image set can be described by a $n \times N_T$ matrix $\mathbf{X_s}$, with each image $\mathbf{x}_{\mathbf{s}(i,j)}$ forming one column of $\mathbf{X_s}$, that is

$$\mathbf{X_s} = [\mathbf{x}_{\mathbf{s}(1,1)} - \mathbf{m}_{\mathbf{x_s}}, \cdots, \mathbf{x}_{\mathbf{s}(1,N_1)} - \mathbf{m}_{\mathbf{x_s}}, \cdots, \mathbf{x}_{\mathbf{s}(c,N_c)} - \mathbf{m}_{\mathbf{x_s}}]. \tag{4.4}$$

Based on the basic theory described in the previous chapter, we only show the important equations for EST and CST in the following section. Theoretical details can be referred to the previous chapter.

4.4.1 Eigenspace Transformation (EST)

EST uses the eigenvalues and eigenvectors generated by the data covariance matrix to rotate the original data coordinates along the direction of maximum variance. If the rank of the matrix $\mathbf{X_s}\mathbf{X_s}^T$ is K, then the K nonzero eigenvalues of $\mathbf{X_s}\mathbf{X_s}^T$, $\lambda_{s(1)}, \ldots, \lambda_{s(K)}$, and their associated eigenvectors $\mathbf{e}_{\mathbf{s}(1)}, \ldots, \mathbf{e}_{\mathbf{s}(K)}$ satisfy the fundamental eigenvalue relationship

$$\lambda_{s(i)}\mathbf{e}_{\mathbf{s}(i)} = \mathbf{R_s}\mathbf{e}_{\mathbf{s}(i)}, \qquad \text{i=1, } \ldots \text{,K,} \tag{4.5}$$

where $\mathbf{R_s}$ is a square, symmetric $n \times n$ matrix derived from $\mathbf{X_s}$ and its transpose $\mathbf{X_s}^T$ by

$$\mathbf{R_s} = \mathbf{X_s}\mathbf{X_s}^T. \tag{4.6}$$

In order to solve equation (4.5), we need to calculate the eigenvalues and eigenvectors of the $n \times n$ matrix $\mathbf{X_s}\mathbf{X_s}^T$ which is computationally intractable for typical image sizes. Based on *singular value decomposition* theory [112] which has also been explained in the previous chapter, we can compute another matrix $\tilde{\mathbf{R}}_\mathbf{s}$ instead, that is

$$\tilde{\mathbf{R}}_\mathbf{s} = \mathbf{X_s}^T\mathbf{X_s}, \tag{4.7}$$

in which the matrix size is $N_T \times N_T$ and is much smaller than $n \times n$ in practical problems.

Suppose the matrix $\tilde{\mathbf{R}}_{\mathbf{s}}$ has nonzero eigenvalues $\tilde{\lambda}_{s(1)}, \ldots, \tilde{\lambda}_{s(K)}$ and associated eigenvectors $\tilde{\mathbf{e}}_{\mathbf{s}(i)}, \ldots, \tilde{\mathbf{e}}_{\mathbf{s}(K)}$ which are related to those in $\mathbf{R}_{\mathbf{s}}$ by the following two equations

$$\begin{cases} \lambda_{s(i)} = \tilde{\lambda}_{s(i)} \\ \mathbf{e}_{\mathbf{s}(i)} = \tilde{\lambda}_{s(i)}^{-\frac{1}{2}} \mathbf{X}_{\mathbf{s}} \tilde{\mathbf{e}}_{\mathbf{s}(i)} \end{cases} \tag{4.8}$$

where $i = 1, \ldots, K$. The K eigenvectors are used as an orthogonal basis to span a new vector space. Using this basis, each template image can be projected to a single point in this K-dimensional space.

According to the theory of PCA, each image can be approximated by taking only the $k \leq K$ largest eigenvalues $\lambda_{s(1)} \geq \lambda_{s(2)} \geq \cdots \geq \lambda_{s(k)}$ and their associated eigenvectors $\mathbf{e}_{\mathbf{s}(1)}, \ldots, \mathbf{e}_{\mathbf{s}(k)}$. This partial set of k eigenvectors spans an eigenspace in which $\mathbf{y}_{\mathbf{s}(i,j)}$ are the points that are the projections of the original images $\mathbf{x}_{\mathbf{s}(i,j)}$ by the equation

$$\mathbf{y}_{\mathbf{s}(i,j)} = [\mathbf{e}_{\mathbf{s}(1)}, \cdots, \mathbf{e}_{\mathbf{s}(k)}]^{\mathbf{T}} \mathbf{x}_{\mathbf{s}(i,j)}, \tag{4.9}$$

where $i = 1, \ldots, c$ and $j = 1, \ldots, N_c$. We called this matrix $[\mathbf{e}_{\mathbf{s}(1)}, \cdots, \mathbf{e}_{\mathbf{s}(k)}]^{\mathbf{T}}$ the *eigenspace transformation matrix*. After this transformation, each original template image $\mathbf{x}_{\mathbf{s}(i,j)}$ can be approximated by the linear combination of these k eigenvectors and $\mathbf{y}_{\mathbf{s}(i,j)}$ is a one-dimensional vector with k elements which are their associated coefficients. We call these eigenvectors *eigengaits* in gait analysis and *eigenfaces* in face recognition. The samples will be shown in the experimental results.

For a gait sequence, the sequence of spatial templates represents a sequential movement of human walking which can be represented as a trajectory in the eigenspace. An example of a template sequence in the eigenspace is shown in Figure 4.5. For display purposes, only a 3-dimensional eigenspace is shown here. This figure shows that human walking is actually a periodic movement reflected by the trajectory which circles around a central point in the eigenspace.

4.4.2 Canonical Space Transformation (CST)

Suppose $\{\Phi_{s(1)}, \Phi_{s(2)}, \cdots, \Phi_{s(c)}\}$ represents the classes of projected vectors by eigenspace transformation and $\mathbf{y}_{\mathbf{s}(i,j)}$ is the j-th vector in class i. The mean vector of the entire set is given by

$$\mathbf{m}_{\mathbf{sy}} = \frac{1}{N_T} \sum_{i=1}^{c} \sum_{j=1}^{N_i} \mathbf{y}_{\mathbf{s}(i,j)} \tag{4.10}$$

and the mean vector of the i-th class is represented by

$$\mathbf{m}_{\mathbf{s}(i)} = \frac{1}{N_i} \sum_{\mathbf{y}_{\mathbf{s}(i,j)} \in \Phi_{s(i)}} \mathbf{y}_{\mathbf{s}(i,j)}. \tag{4.11}$$

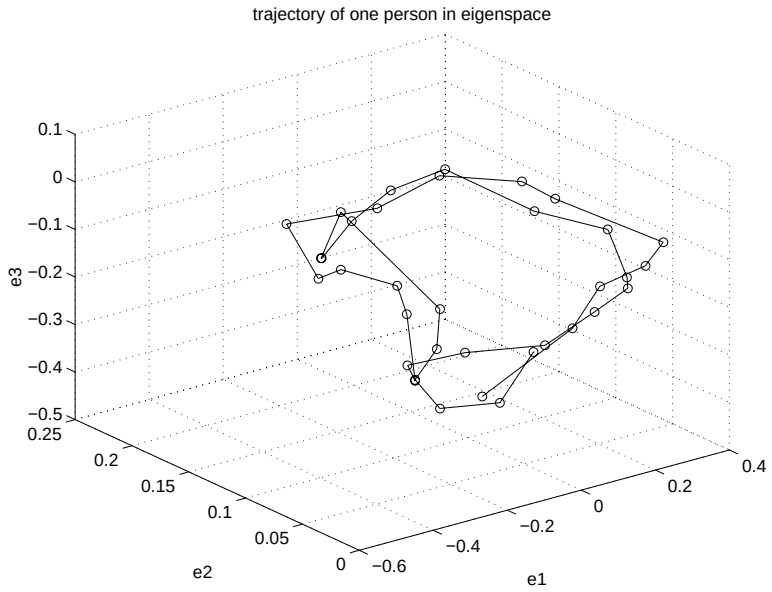


Figure 4.5: One trajectory in eigenspace using spatial templates

Let $\mathbf{S_{sw}}$ denote *within-class matrix* and $\mathbf{S_{sb}}$ denote *between-class matrix*, then

$$\begin{aligned}\mathbf{S_{sw}} &= \frac{1}{N_T}\sum_{i=1}^{c}\sum_{\mathbf{y_{s}}_{(i,j)}\in\Phi_{s(i)}}(\mathbf{y_{s}}_{(i,j)}-\mathbf{m_{s}}_{(i)})(\mathbf{y_{s}}_{(i,j)}-\mathbf{m_{s}}_{(i)})^{\mathbf{T}}\\ \mathbf{S_{sb}} &= \frac{1}{N_T}\sum_{i=1}^{c}N_i(\mathbf{m_{s}}_{(i)}-\mathbf{m_{sy}})(\mathbf{m_{s}}_{(i)}-\mathbf{m_{sy}})^{\mathbf{T}}\end{aligned}$$

where $\mathbf{S_{sw}}$ represents the mean of within-class vector distances and $\mathbf{S_{sb}}$ represents the mean of between-class vector distances. The objective is to minimise $\mathbf{S_{sw}}$ and maximise $\mathbf{S_{sb}}$ simultaneously, that is to maximise the criterion function known as the *generalised Fisher linear discriminant function* and given by

$$\mathbf{J}(\mathbf{W_s}) = \frac{\mathbf{W_s}^T\mathbf{S_{sb}}\mathbf{W_s}}{\mathbf{W_s}^T\mathbf{S_{sw}}\mathbf{W_s}}. \tag{4.12}$$

Let $\mathbf{W_s}^*$ be the optimal solution and $\mathbf{w}^*_{\mathbf{s}(i)}$ be one of its column vectors. That is to solve the *generalised eigenvalue equation* which is given by

$$\mathbf{S_{sb}}\mathbf{w}^*_{\mathbf{s}(i)} = \lambda_{s(i)}\mathbf{S_{sw}}\mathbf{w}^*_{\mathbf{s}(i)}. \tag{4.13}$$

After Equation (4.13) is solved, we will obtain $(c-1)$ nonzero eigenvalues and their corresponding eigenvectors $[\mathbf{v_{s}}_{(1)},\ldots,\mathbf{v_{s}}_{(c-1)}]$ that create another orthogonal basis and span a $(c-1)$-dimensional canonical space. By using this basis, each point in eigenspace can be further projected to another point in this canonical space by

$$\mathbf{z_{s}}_{(i,j)} = [\mathbf{v_{s}}_{(1)},\cdots,\mathbf{v_{s}}_{(c-1)}]^{\mathbf{T}}\mathbf{y_{s}}_{(i,j)} \tag{4.14}$$

where $\mathbf{z}_{\mathbf{s}(i,j)}$ represents the new point and $[\mathbf{z}_{\mathbf{s}(i,1)}, \ldots, \mathbf{z}_{\mathbf{s}(i,N_i)}]$ is the new trajectory in canonical space. We called this orthogonal basis $[\mathbf{v}_{\mathbf{s}(1)}, \cdots, \mathbf{v}_{\mathbf{s}(c-1)}]^{\mathbf{T}}$ the *canonical space transformation matrix*. Following this analysis, different classes will be greatly separated in canonical space and this means that canonical analysis is useful to separate different classes of samples (e.g. human gait and face).

By merging equation (4.9) and equation (4.14), each template image can be directly projected into one point in the $(c-1)$-dimensional canonical space by

$$\mathbf{z}_{\mathbf{s}(i,j)} = [\mathbf{v}_{\mathbf{s}(1)}, \ldots, \mathbf{v}_{\mathbf{s}(c-1)}]^T [\mathbf{e}_{\mathbf{s}(1)}, \ldots, \mathbf{e}_{\mathbf{s}(k)}]^T \mathbf{x}_{\mathbf{s}(i,j)}. \tag{4.15}$$

4.5 Recognition

4.5.1 Spatio-temporal Correlation in Eigenspace

Spatio-temporal correlation is a particular form of template matching which is an extension of two-dimensional image correlation to three-dimensional correlation in the space and time domain, as has already been used in [37] for gait recognition. This measures the similarity between image sequences in eigenspace [37] and is described now for clarity. Let an input image sequence be represented by

$$\mathbf{x}_\mathbf{i}(t) = [x_1(t), x_2(t), \ldots, x_n(t)]^T,$$

and a reference image sequence be

$$\mathbf{y}_\mathbf{r}(t) = [y_1(t), y_2(t), \ldots, y_n(t)]^T$$

in which n is the number of pixels for each image and each element represents the pixel value at time t. Then, the spatio-temporal correlation, $Corr$, is given by

$$Corr = \sum_{t=1}^{T} \mathbf{x}_\mathbf{i}(t)^T \mathbf{y}_\mathbf{r}(t).$$

If velocity change and phase difference are considered, this equation can be rewritten by

$$Corr = \sum_{t=1}^{T} \mathbf{x}_\mathbf{i}(t)^T \mathbf{y}_\mathbf{r}(at + b)$$

where a is the velocity change and b is the phase difference, respectively. Intuitively, the assumption is that the gait is periodic and the velocity does not change during walking. So long as analysis always starts at the same place, the sequence start points need to be aligned. Therefore, only b needs to be considered.

The computational cost of correlation will be high if original images are used and it will be greatly reduced by using projected vectors in eigenspace instead. Suppose that the

basis of a k-dimensional eigenspace is $[\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]$. The input sequence and reference sequence can be projected to

$$\mathbf{x_e}(t) = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]^T \mathbf{x_i}(t)$$

and

$$\mathbf{y_e}(t) = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]^T \mathbf{y_r}(t).$$

The correlation between $\mathbf{x_i}(t)$ and $\mathbf{y_r}(t+b)$ can be written by

$$\begin{aligned} Corr &= \sum_{t=1}^{T} \mathbf{x_i}(t)^T \mathbf{y_r}(t+b) \\ &\cong \sum_{t=1}^{T} \mathbf{x_e}(t)^T [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]^T [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k] \mathbf{y_e}(t+b) \\ &= \sum_{t=1}^{T} \mathbf{x_e}(t)^T \mathbf{y_e}(t+b). \end{aligned} \tag{4.16}$$

Thus, the computational cost of template matching drops from $n \times T$ to $k \times T$ (n is the image size and k is the vector size in eigenspace) without counting the phase difference. Note that the factor of velocity change a has been eliminated from the equation by the assumption of constant velocity.

4.5.2 Accumulated Distance in Canonical Space

Different from spatio-temporal correlation [37], we propose the *accumulated distance* measure in canonical space for gait recognition. After applying EST and CST to training sequences, each class will be separated and tied to a centroid in the canonical space. The recognition confidence can be gathered by the accumulated distance to each centroid through the walking sequence.

Similar to the correlation in eigenspace, the computational cost of accumulated distance will be also greatly reduced using projected vectors in canonical space compared to using original images. The difference is that the accumulated distance is between the input sequence and the centroid of each reference sequence, rather than the correlation with the entire reference sequence. Thus, only the canonical vector of the centroid, not the entire reference sequence, needs to be retained in the training database and the size of database can be much smaller.

Suppose that the basis of a $(c-1)$-dimensional canonical space is $[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{c-1}]$ and c is the number of classes for different subjects. The input sequence and the reference sequence can be projected to $\mathbf{x_c}(t)$ and $\mathbf{y_c}(t)$ by

$$\mathbf{x_c}(t) = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{c-1}]^T [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]^T \mathbf{x_i}(t)$$

and

$$\mathbf{y_c}(t) = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{c-1}]^T [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_k]^T \mathbf{y_r}(t).$$

The centroid, $\mathbf{C_{y_r}}$, of the reference sequence, $\mathbf{y_c}(t)$, in the canonical space is given by

$$\mathbf{C_{y_r}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{y_c}(t).$$

The accumulated distance, $AccDist$, can be written by

$$AccDist = \sum_{t=1}^{T} \|\mathbf{x_c}(t) - \mathbf{C_{y_r}}\|^2$$

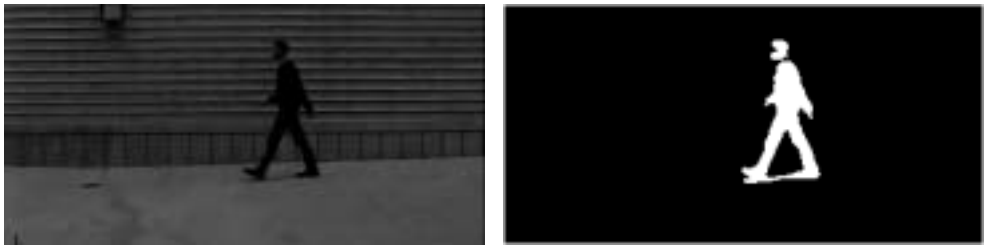
and the computational cost of accumulated distance drops from $n \times T$ to $(c-1) \times T$ (n is the image size and c is the number of classes). Compared with correlation using the original images and using projected vectors of eigenspace in Equation (4.16), one example is demonstrated here. Suppose we have the following parameters in the experiments; image size: 64×64, the number of reference subjects: $c = 6$, the number of reference frames from each subject: $n_r = 40$, the number of frames from the test subject: $n_t = 90$, the dimension of eigenspace: $k = 100$, the dimension of canonical space: $(c-1) = 6-1 = 5$. Thus, the number of pixel operations for the three approaches, using original images, using projected vectors of eigenspace and using projected vectors of canonical space, are O_o, O_e and O_c, respectively, as given by

$$\begin{aligned} O_o &= 64 * 64 * n_r * (n_t - n_r) * c \\ &= 64 * 64 * 40 * (90 - 40) * 6 \\ &= 4.9 * 10^7, \\ O_e &= 64 * 64 * k * n_t + k * n_r * (n_t - n_r) * c \\ &= 64 * 64 * k * 90 + k * 40 * (90 - 40) * 6 \\ &= 3.8 * 10^7 \end{aligned}$$

and

$$\begin{aligned} O_c &= 64 * 64 * (c-1) * n_t + (c-1) * n_t * c \\ &= 64 * 64 * 5 * 90 + 5 * 90 * 6 \\ &= 1.8 * 10^6, \end{aligned}$$

respectively. The computational cost of our approach had been reduced greatly when compared with other two approaches by factor of at least 20. The spatio-temporal correlation described in Section 4.5.1 and the accumulated distance explained in Section 4.5.2 will be applied for recognition in the eigenspace and canonical space. The results are



(a) Original image (b) Preprocessed image

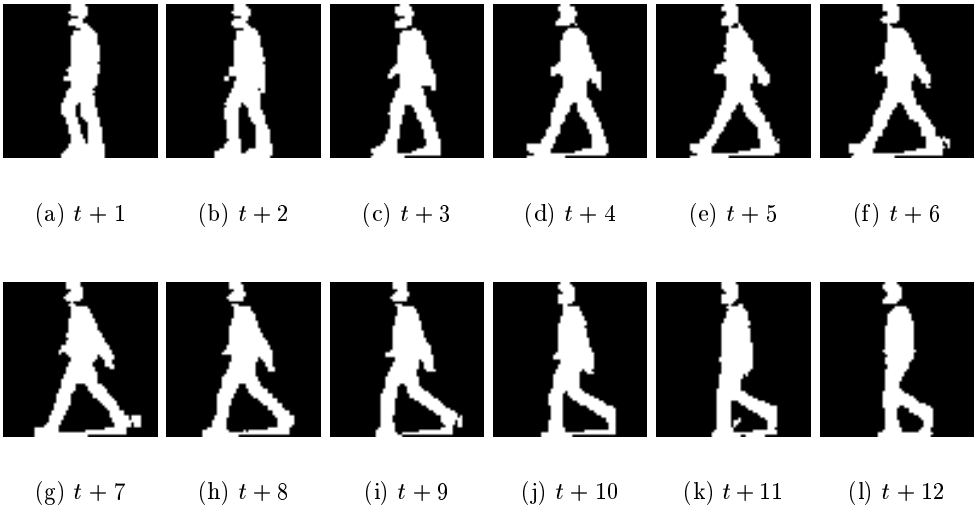
Figure 4.6: Human silhouette extraction of another subject

shown in Section 4.6.

4.6 Results of Gait Recognition

In this section, the results of gait recognition are presented. The sample human gait data came from the Visual Computing Group, University of California, San Diego [40] and had been augmented 5 subjects and 5 sequences of each to 6 subjects and 7 sequences of each. To acquire these images, a Sony Hi8 video camera was pointed towards a concrete wall in a courtyard. A number of students walked in a circular path around the camera so that only one person at a time was in the camera's field of view. The original digitised 640×480 full colour images were then translated to black and white, cropped and subsampled to a resolution of 320×160. Obtained gait data was two different sets taken at two different dates, first one is 5 subjects and 5 sequences each and second one is 6 subjects and 7 sequences of each. The second set is an extension of the first set by adding two more sequences to each of 5 original subjects and one new subject with 7 sequences. Experimental results have been conducted for the two different sets and will be shown in this section for comparison purposes.

First, spatial templates are extracted from every sequence in the preprocessing stage explained in Section 4.3. Figure 4.6 shows a sample image and a preprocessed image from a walking sequence of another subject. Sample spatial templates extracted from this sequence in temporal order are illustrated in Figure 4.7. Note that some shadow and background noise still occur after preprocessing. Since template matching is insensitive to noise [37], it would appear unlikely that those facts will affect discriminatory capability. It is not unlikely that shadow is statistically correlated with moving shape since it depends on illuminant direction or movement. It should be considered as a part of gait information by applying statistical approaches. Besides, minor background noise will be eliminated when discarding redundant components using PCA.



(a) $t+1$ (b) $t+2$ (c) $t+3$ (d) $t+4$ (e) $t+5$ (f) $t+6$

(g) $t+7$ (h) $t+8$ (i) $t+9$ (j) $t+10$ (k) $t+11$ (l) $t+12$

Figure 4.7: Sample spatial templates of another subject

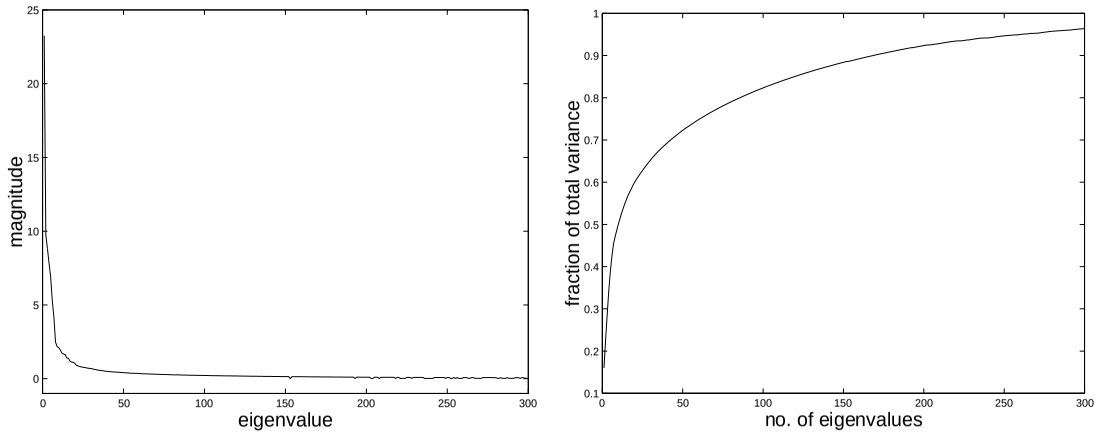
4.6.1 Training by PCA

In our experiments, one walking sequence is selected at random from each subject as the training sequence. There are 5 training sequences of 5 subjects for the first experiment and 6 training sequences of 6 subjects for the second experiment. The remaining sequences served as test sequences, there are 20 and 36 test sequences for the first and the second experiment, respectively.

After PCA, two different sets of eigenvalues and eigenvectors are calculated from the two training databases. Figure 4.8(a) shows the magnitudes of eigenvalues in the eigenspace for 5 training sequences and Figure 4.8(b) represents the accumulated variances of these eigenvalues after PCA. Figure 4.9(a) shows the magnitudes of eigenvalues in the eigenspace of 6 training sequences and Figure 4.9(b) represents the accumulated variances of these eigenvalues after PCA. Clearly, the major accumulated variance is concentrated in the first several (largest) eigenvalues.

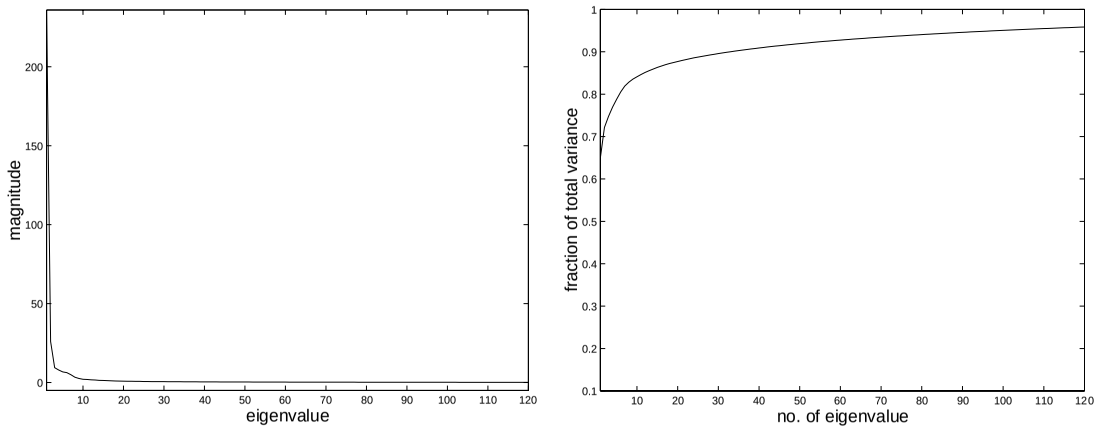
For the first training set, we choose the first 100 eigenvalues which accumulated just over 80% of the total signal energy and their corresponding eigenvectors as the eigenspace transformation matrix. As an example, Figure 4.10 shows the first 6 eigenvectors called *eigengaits* which are used to approximate each spatial template in the first experiment. For display purposes, the value in each pixel position has been rescaled from 0 to 255. For the second training set, we choose the first 110 eigenvalues which accumulated 95% of the total signal energy and their corresponding eigenvectors as the eigenspace transformation matrix. Figure 4.11 shows the first 6 eigengaits used to approximate each spatial template in the second experiment.

By using the eigenspace transformation matrix, each training spatial template is projected to the eigenspace and each training sequence became a trajectory in this space.



(a) Magnitudes (b) Accumulated variances

Figure 4.8: Eigenvalues in the eigenspace using spatial templates of 5 sequences from 5 subjects



(a) Magnitudes (b) Accumulated variances

Figure 4.9: Eigenvalues in the eigenspace using spatial templates of 6 sequences from 6 subjects

(a) $\mathbf{e}_1$ (b) $\mathbf{e}_2$ (c) $\mathbf{e}_3$ (d) $\mathbf{e}_4$ (e) $\mathbf{e}_5$ (f) $\mathbf{e}_6$

Figure 4.10: The first six eigengaits using spatial templates of 5 subjects

(a) $\mathbf{e}_1$ (b) $\mathbf{e}_2$ (c) $\mathbf{e}_3$ (d) $\mathbf{e}_4$ (e) $\mathbf{e}_5$ (f) $\mathbf{e}_6$

Figure 4.11: The first six eigengaits using spatial templates of 6 subjects

Figure 4.12(a) shows the trajectories in the eigenspace of first training set that represent 5 training gait sequences of 5 subjects after they are projected into the eigenspace. Figure 4.12(b) represents the trajectories in the eigenspace of second training set that represent 6 training gait sequences of 6 subjects after projected into the eigenspace. From those two figures, it is obvious that the trajectories which belong to different subjects overlap and their centroids are close to each other, mandating further analysis.

4.6.2 Recognition Using Spatio-temporal Correlation in Eigenspace

After training using PCA, each training sequence has been projected to the eigenspace and can be used for matching an unknown input gait sequence. After spatial template extraction, let a test gait sequence be $g(t)$, $t = 1, \ldots, T$, and brightness normalisation had been applied. This normalised template sequence is projected to the eigenspace by

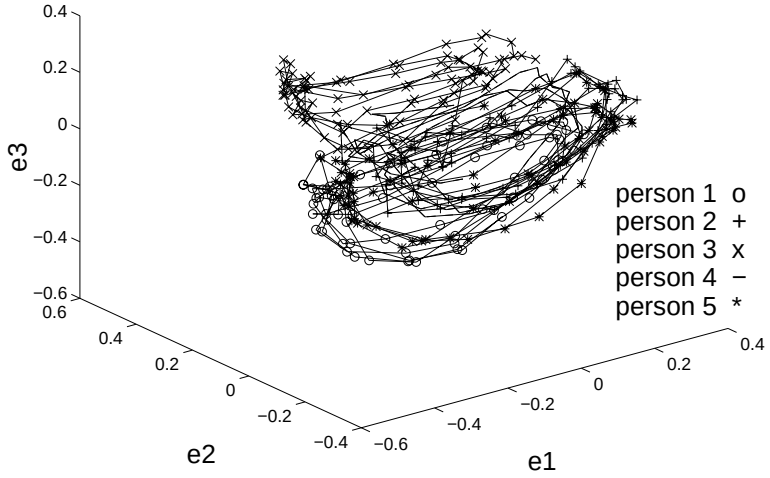
$$u(t) = [\mathbf{e}_{\mathbf{s}(1)}, \ldots, \mathbf{e}_{\mathbf{s}(k)}]^T g(t) \tag{4.17}$$

where $u(t)$ are the projected vectors in the eigenspace and $[\mathbf{e}_{\mathbf{s}(1)}, \ldots, \mathbf{e}_{\mathbf{s}(k)}]$ is the EST matrix generated from the PCA.

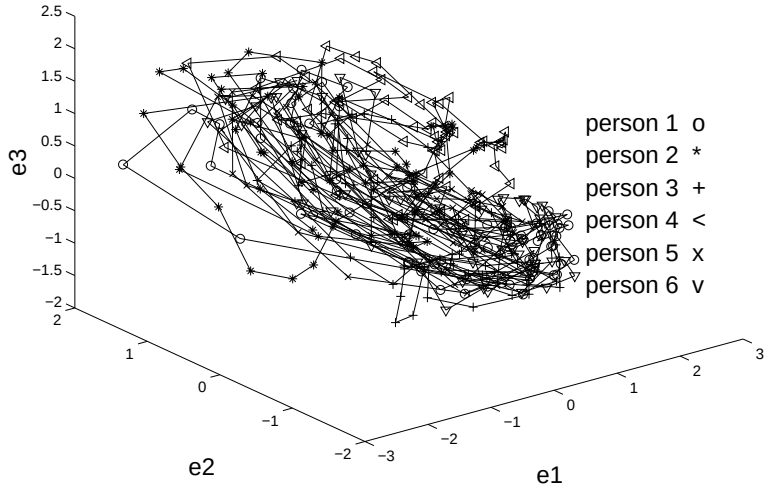
In comparison with the recognition performance of eigenspace approach [37], *spatio-temporal correlation* explained in Section 4.5.1 is applied in eigenspace for recognising a human walking sequence from a database. After the spatial templates have been converted by eigenspace transformation, the distance between the test vector sequence, $u(t)$, and training vector sequences, $\mathbf{y}_{\mathbf{s}(i,j)}$ in Equation 4.9, can be obtained by

$$d_e(i)^2 = \min_b \sum_{t=1}^{T} \|u(t) - \mathbf{y}_{\mathbf{s}(i,j)}(t+b)\|^2 \tag{4.18}$$

in which b represents the phase difference and velocity change a is not considered here by the assumption of constant velocity in gait. Based on Equation (4.16), this is equivalent



(a) Eigenspace of first training set with 5 sequences



(b) Eigenspace of second training set with 6 sequences

Figure 4.12: Trajectories of training spatial templates in two eigenspaces

to computing the maximum spatial-temporal correlation given by

$$Corr(i) = \max_b \sum_{t=1}^{T} u(t)^T \mathbf{y}_{\mathbf{s}(i,j)}(t+b).$$

The match of a test sequence $g(t)$ to a training sequence can be accomplished by choosing a value for i which has the minimum $d_e(i)^2$. The recognition rates are 100% for the first experiment with 20 test sequences and the second experiment with 36 test sequences. However, from the distribution of training sequences in Figure 4.12(a) and Figure 4.12(b), it is obvious that it becomes harder to separate different classes when the number of subjects increases and the problem of overlapped trajectories becomes worse. To overcome this problem, we propose canonical analysis for the projected vectors of training sequences in the eigenspace to separate different classes.

4.6.3 Training by CA

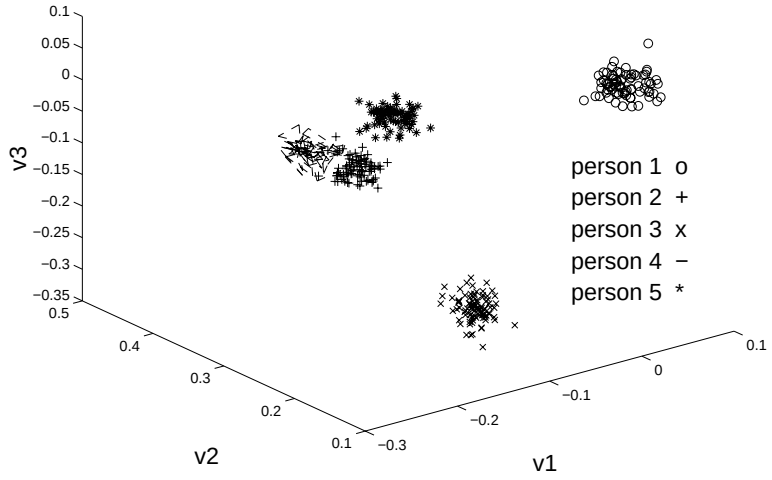
After applying canonical analysis to the projected vectors of two experiments shown in Figure 4.12(a) and Figure 4.12(b) separately, two training sets are further projected to two canonical spaces which are shown in Figure 4.13(a) and Figure 4.13(b). Figure 4.13(a) and Figure 4.13(b) show the trajectories in two canonical spaces that represent the training sequences of five subjects and six subjects after projecting their transformed vectors in the eigenspace to the canonical space. In those two figures, trajectories are widely separated into 5 different clusters in Figure 4.13(a) and 6 clusters in Figure 4.13(b). Figure 4.14(a) shows the eigenvalues in the canonical space of 5 training sequences and all of the signal energy is represented by the variance over the first 4 eigenvectors. These 4 eigenvectors are used as the CST matrix for the first experiment. Figure 4.14(b) shows the eigenvalues in the canonical space of 6 training sequences and all of the signal energy is represented by the variance over the first 5 eigenvectors. The first 5 eigenvectors are used as the CST matrix for the second experiment.

4.6.4 Recognition Using Accumulated Distance in Canonical Space

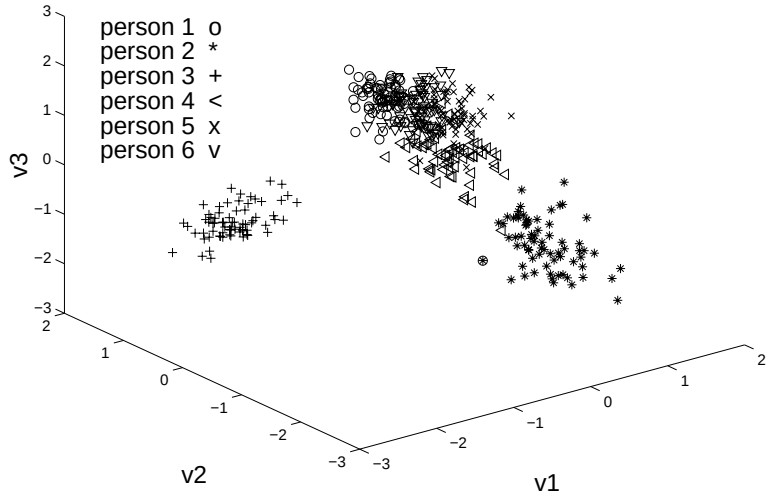
After CA, converted vectors of the test sequence, $u(t)$, in Equation (4.17), are further projected to the canonical space by

$$h(t) = [\mathbf{v}_{\mathbf{s}(1)}, \ldots, \mathbf{v}_{\mathbf{s}(c-1)}]^T u(t) \tag{4.19}$$

in which $[\mathbf{v}_{\mathbf{s}(1)}, \ldots, \mathbf{v}_{\mathbf{s}(c-1)}]$ is a $(c-1) \times k$ CST matrix generated from the CA and $h(t)$ is a vector sequence with $(c-1)$ elements in each vector. In the first experiment with 5 subjects, c equals 5 and k equals 100. For the second experiment with 6 subjects, c equals 6 and k equals 110. To recognise a human walking sequence from a database in the canonical space, we propose the *accumulated distance* to each centroid described in

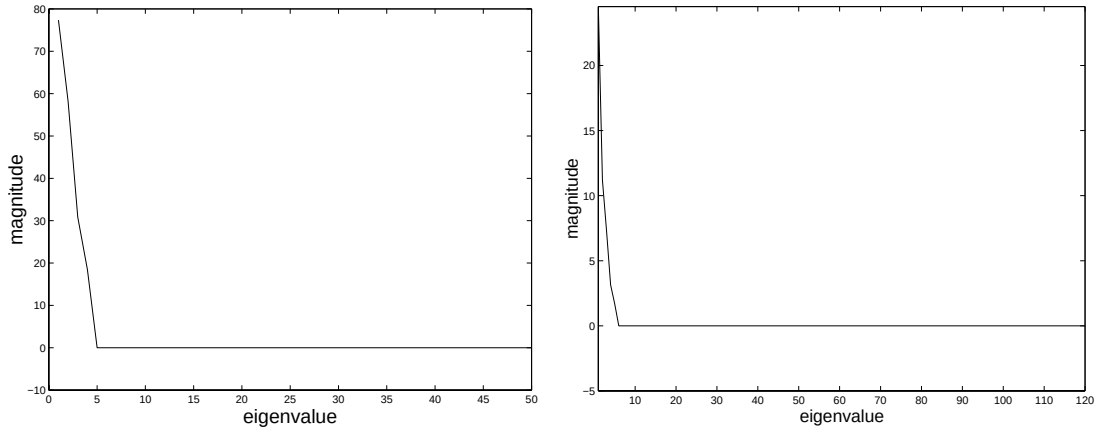


(a) Canonical space of first training set with 5 sequences



(b) Canonical space of second training set with 6 sequences

Figure 4.13: Trajectories of training spatial templates in two canonical spaces



(a) Magnitudes in the first canonical space with 5 sequences

(b) Magnitudes in the second canonical space with 6 sequences

Figure 4.14: Eigenvalues of training spatial templates in two canonical spaces

Recognition results for 20 test sequences		
method	recognition rate	class separability
EST only	100%	overlapped
EST+CST	100%	widely separated

Table 4.1: Comparison of EST and EST+CST in gait recognition using 20 test sequences of 5 subjects

Section 4.5.2. The distance between the test vector sequence, $h(t)$, in Equation (4.19), and training vector sequences, $\mathbf{z}_{\mathbf{s}(i,j)}$ in Equation (4.15), is given by

$$d_c(i)^2 = \sum_{t=1}^{T} \|h(t) - \mathbf{C_c}(i)\|^2 \tag{4.20}$$

where

$$\mathbf{C_c}(i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{z}_{\mathbf{s}(i,j)}$$

and $\mathbf{C_c}(i)$ is the centroid of class i in the canonical space. To match a test sequence $h(t)$ to a training sequence i can be accomplished by choosing the *minimum* $d_c(i)^2$.

The comparison of 2 different approaches using 20 test sequences in the first experiment is shown in Table 4.1. Clearly, the feature vectors generated by the combined approach of EST and CST yields the best recognition result, since the classes do not overlap and a high recognition rate can be achieved. By using 36 test sequences in the second experiment, the comparison of three approaches is shown in Table 4.2. Since Little and Boyd

Recognition results for 36 test sequences		
Method	Recognition rate	Class separability
Murase & Sakai's	100%	overlapped
Little & Boyd's	95.2%	N/A
EST+CST	100%	widely separated

Table 4.2: Comparison of different approaches in gait recognition using 36 test sequences of 6 subjects

	Accumulated variance of eigenvalues						
Training samples	65%	70%	75%	80%	85%	90%	95%
1 cycle	21.4%	21.4%	31.0%	47.6%	66.7%	95.2%	90.5%
2 cycles	23.8%	31.0%	31.0%	71.4%	83.3%	97.6%	100%
3 cycles	38.1%	52.4%	76.2%	83.3%	97.6%	100%	100%
4 cycles	50.0%	50.0%	83.3%	90.5%	100%	100%	100%

Table 4.3: Recognition rates using different training samples and eigenvalues

applied a different approach [40] to the same data, the best recognition rate is selected from [40] and listed in Table 4.2. Murase and Sakai's approach [37] is actually the EST method using spatio-temporal correlation. Again, the combined approach achieves the best result. Although EST method achieves the same recognition rate as the combined approach of EST and CST in Table 4.1 and Table 4.2, the proposed combined approach accomplishes better class separability which has the potential for application to a larger database.

4.6.5 The Influence of Different Training Samples and Eigenvalues

UCSD gait data with 6 subjects and 7 sequences each are used in this section. Again, one sequence from each subject is used as a training sequence and remaining 36 sequences are used for tests. In order to evaluate the recognition rates in canonical space affected by using different training samples (walking cycles) and eigenvalues, we conducted four tests which used 18, 36, 54 and 72 templates corresponding to 1, 2, 3 and 4 walking cycles from each training sequence for training. Furthermore, In each test, we choose 7 different accumulated variances ranging from 65% to 95% achieved by different numbers of eigenvalues (associated to eigenvectors) in the eigenspace.

The comparison of recognition performance is shown in Table 4.3 and Figure 4.15. In average, the results show that the best performance is accomplished by using 4 walking cycles of training samples from each subject. Actually, the recognition rate using the combined approach is already 100% by choosing 85% accumulated variance in the eigenspace. Furthermore, the recognition performance is improved by increasing

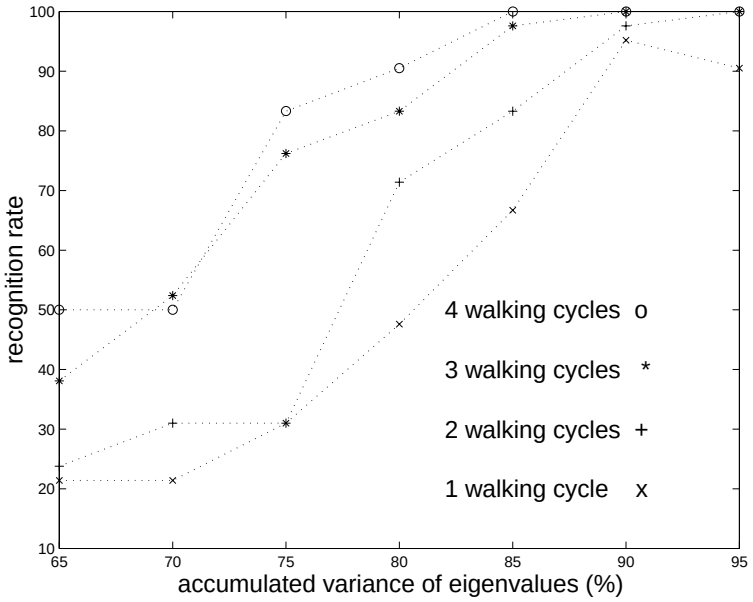


Figure 4.15: Recognition rates using different training samples and eigenvalues

the number of training samples and accumulated variances (number of eigenvectors). Therefore, results would appear to confirm sensitivity to the number of learning samples and imply that in a more extended analysis, care must be taken to include sufficient samples in the training database.

4.7 Results of Face Recognition

In order to clarify the potential advantages accrued by combining EST with CST, the approach has also been applied in face recognition. The face data came from the Olivetti Research Laboratory in Cambridge, UK. There are 10 different face images of 40 distinct subjects. The size of each image is 92×112, 8-bit grey levels. For some of the subjects, the images were taken at different times where lighting varied slightly, facial expressions altered (open/closed eyes, smiling/non-smiling) and facial adornment altered (glasses/no-glasses). All the images are taken against a dark homogeneous background and the subjects are in an upright, frontal position. Figure 4.16 shows some sample images from the face data.

The purpose of this experiment is to compare the performance of the eigenface approach [80] and our feature extraction method (EST and CST) in face recognition. In order to distinguish the effects of two approaches by different training sets, different numbers of training images are used in different experiments. We have conducted 8 experiments by using 2, 3, 4, 5, 6, 7, 8 or 9 training images from each subject. In each experiment, apart from the selected training images, the remaining images are used as the test set. Thus, the size of covariance matrix constructed by eigenvector reconstruction techniques becomes very small (e.g. 80×80 using 2 images from each subject). It is

Figure 4.16: Sample face images

more appropriate to apply PCA directly without using eigenvector reconstruction techniques. Therefore, all the 92×112 original images are subsampled to the 23×28 pixels and this results in a covariance matrix with size $(23 \times 28) \times (23 \times 28) = 644 \times 644$ which is more tractable than using original images and contains more information than using eigenvector reconstruction techniques. These reduced images are used directly by EST and the combination of EST and CST after brightness normalisation by equation (4.2). No further preprocessing, such as template extraction, is needed as in gait recognition.

Since the number of images in the entire data set is $40 \times 10 = 400$ which is smaller than the image size ($23 \times 28 = 644$), this will result in the singularity problem for Etemad and Chellappa's [101] approach which adopted CA directly to face images. Therefore, this approach cannot be tested here. Concurrently, an approach called *fisherfaces* [104] which is similar to our combined approach was applied to a larger database for testing lighting direction and facial expression in face recognition.

Now we have face patterns of 40 different subjects with 10 each in our face data. For the face recognition problem, suppose we have one test face image, $\mathbf{f}$, which is brightness normalised. This normalised face image can be projected to a point in the eigenspace by

$$\omega_{\mathbf{e}} = [\mathbf{e}_{\mathbf{f}(1)}, \ldots, \mathbf{e}_{\mathbf{f}(k)}]^T \mathbf{f}, \tag{4.21}$$

in which $\omega_{\mathbf{e}}$ is a projected vector with k elements and $[\mathbf{e}_{\mathbf{f}(1)}, \ldots, \mathbf{e}_{\mathbf{f}(k)}]$ is the eigenspace

(a) $\mathbf{e}_{\mathbf{f}(1)}$ (b) $\mathbf{e}_{\mathbf{f}(2)}$ (c) $\mathbf{e}_{\mathbf{f}(3)}$ (d) $\mathbf{e}_{\mathbf{f}(4)}$ (e) $\mathbf{e}_{\mathbf{f}(5)}$ (f) $\mathbf{e}_{\mathbf{f}(6)}$

Figure 4.17: The first six eigenfaces of 40 subjects using 8 training images each

(a) $\mathbf{v}_{\mathbf{f}(1)}$ (b) $\mathbf{v}_{\mathbf{f}(2)}$ (c) $\mathbf{v}_{\mathbf{f}(3)}$ (d) $\mathbf{v}_{\mathbf{f}(4)}$ (e) $\mathbf{v}_{\mathbf{f}(5)}$ (f) $\mathbf{v}_{\mathbf{f}(6)}$

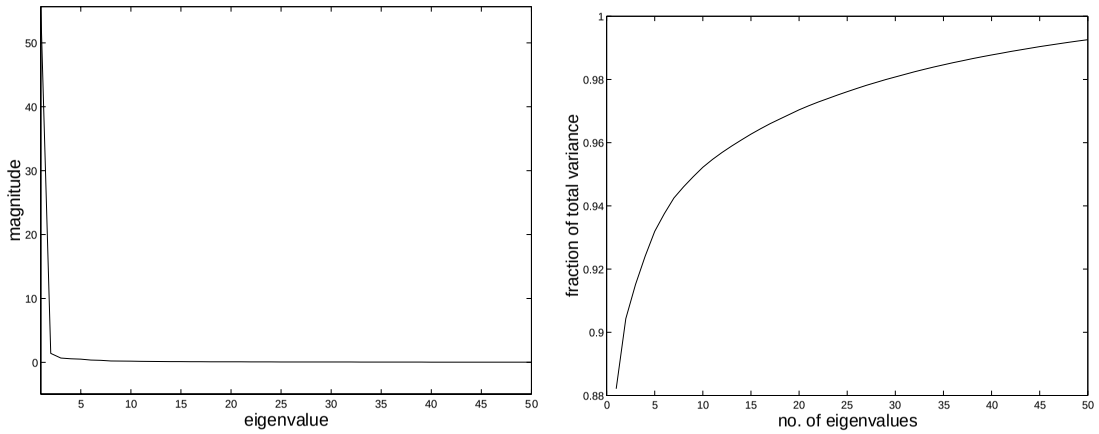
Figure 4.18: The first six canonical faces of 40 subjects using 8 training images each

transformation matrix after applying PCA to the training set. To recognise a test image from a database with 40 different subjects in the eigenspace, the *nearest-neighbour rule* in the eigenface approach [80] is used. The distance between this test vector, $\omega_{\mathbf{e}}$, and training vectors after eigenspace transformation, $\mathbf{y}_{\mathbf{f}(i,j)}$, is given by

$$fd^2_{e(i,j)} = \min_j \|\omega_{\mathbf{e}} - \mathbf{y}_{\mathbf{f}(i,j)}\|^2 \tag{4.22}$$

in which $i = 1, \cdots, 40$ and j is the number of training images from each subject. For different experiments in this section, the number j is 2, 3, 4, 5, 6, 7, 8 or 9, respectively. To match a test face $\omega_{\mathbf{e}}$ to a face in the training database can be accomplished by choosing i which minimises $fd^2_{e(i,j)}$. An equivalent measure is used for recognition in the canonical space [117] after applying CA to all training patterns in the eigenspace, except that $\omega_{\mathbf{e}}$ and $\mathbf{y}_{\mathbf{f}(i,j)}$ are replaced by $\omega_{\mathbf{c}}$ and $\mathbf{z}_{\mathbf{f}(i,j)}$ which are projected by the canonical space transformation matrix generated by CA.

Figure 4.17 shows the first 6 eigenfaces (eigenvectors) used to span the eigenspace and to approximate each face image and Figure 4.18 shows the first 6 canonical faces (eigenvectors) used to span the canonical space and to approximate each face image when choosing 8 images from each subject as the training set. From Figure 4.17 and Figure 4.18, it shows that eigenfaces capture major variations in the training set, e.g. lighting direction. However, canonical faces show the directional edges found in the training set and reduce the influence of lighting direction which are beneficial to classification. This has also been tested in the *fisherface* approach [104] using a different data set under various lighting directions. The results show that the eigenface approach [80] is suffering under



(a) Magnitudes (b) Accumulated variances

Figure 4.19: Eigenvalues in the eigenspace using 2 training faces from each of 40 subjects

the variation in lighting direction which appears to perform worse than the fisherface approach (same as our approach) in face recognition.

For all the experiments, 95% of the accumulated variance by the eigenvalues is used to choose the number of eigenvectors which are used as the EST matrix. Different numbers of eigenvectors will be selected for different training sets. Here, the experiment using 2 training patterns from each subject is taken as an example and the other experiments with different patterns can be implemented in the same way. Figure 4.19(a) shows the magnitudes of eigenvalues in the eigenspace of 80 training images after PCA. Figure 4.19(b) represents the accumulated variances of these eigenvalues. The 95% of accumulated variance is concentrated in the first 10 largest eigenvalues. Ten eigenvectors constituted the eigenspace basis in the experiment with 2 training faces each after PCA. This will result in only 10 eigenvectors in the canonical space. The eigenvalue magnitudes in the canonical space after CA are shown in Figure 4.20. Since the transformation basis of the eigenspace has only 10 eigenvectors, the maximum number of non-zero eigenvalues in canonical space can only be 10.

The recognition results of 8 experiments using 2 different approaches are shown in Table 4.4 and in Figure 4.21, the results using all data patterns are shown in in Table 4.5 and in Figure 4.22. Note that the recognition rate drops from 98.8% to 97.5% in Table 4.4 using 8 and 9 training images. The reason is that there is one mismatched sample in each test set, but the number of test samples is 80 and 40 in two sets.
Figure 4.23(a) shows the distribution of 40 subjects with 3 training images each in the eigenspace and their distribution in the canonical space is shown in Figure 4.23(b). The images belonged to the same subject are connected by lines. From those two figures, they show that class separability of 40 subjects in the canonical space is better than in the eigenspace. From the experimental results, it is shown that the combined approach of EST and CST produces a better feature set for face recognition on average, compared

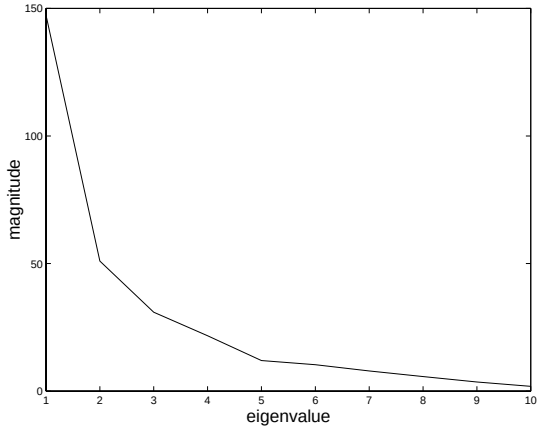


Figure 4.20: Eigenvalues in the canonical space using 2 training faces from each of 40 subjects

Face recognition results using remaining patterns								
	number of training face patterns from each subject							
method	2	3	4	5	6	7	8	9
(1)	69.4%	73.6%	80.8%	85.5%	93.8%	94.2%	95.0%	95.0%
(2)	68.8%	79.6%	86.7%	89.5%	95.6%	96.7%	98.8%	97.5%

Table 4.4: Comparison of 2 approaches using remaining patterns for face recognition ((1) and (2) represent the eigenface approach and the approach of EST+CST, respectively.)

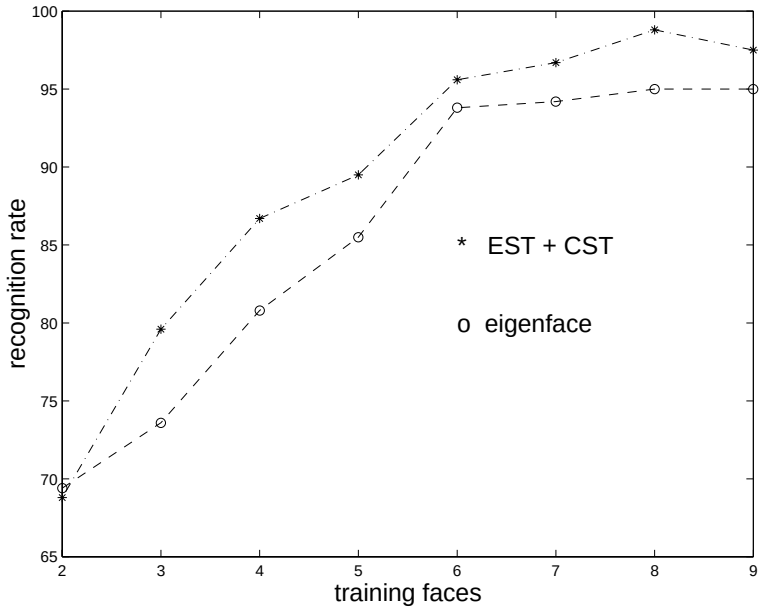


Figure 4.21: Test results of face recognition using remaining patterns

Face recognition results using all patterns								
	number of training face patterns from each subject							
method	2	3	4	5	6	7	8	9
(1)	75.5%	81.5%	88.5%	92.8%	97.5%	98.3%	99.0%	99.5%
(2)	75.0%	85.8%	92.0%	94.8%	98.3%	99.0%	99.8%	99.8%

Table 4.5: Comparison of 2 approaches using all patterns for face recognition ((1) and (2) represent the eigenface approach and the approach of EST+CST, respectively.)

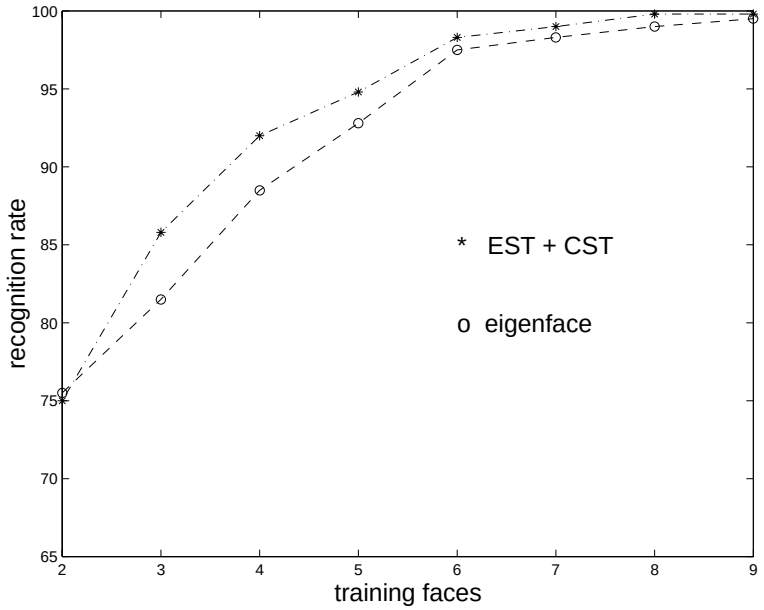


Figure 4.22: Test results of face recognition using all patterns

with the eigenface approach.

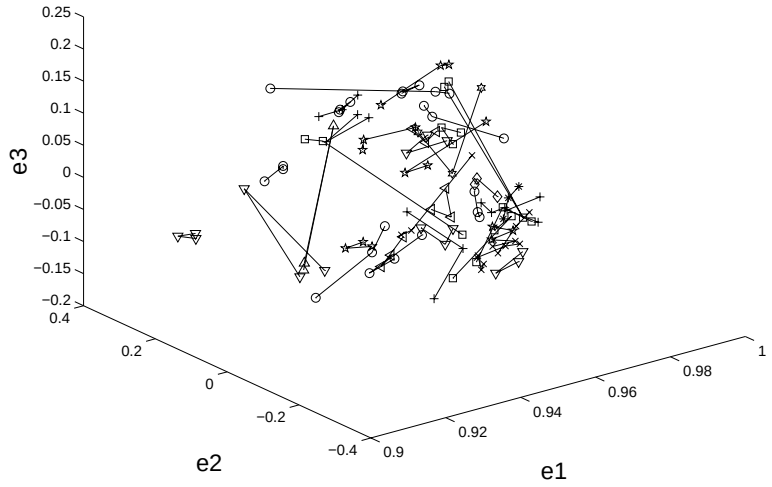
4.8 Discussions

For spatial templates of gait sequences used in this chapter, each human silhouette has been rescaled and positioned in the center of the template with size 64×64. Therefore, there is no scaling and positioning problem for gait recognition.

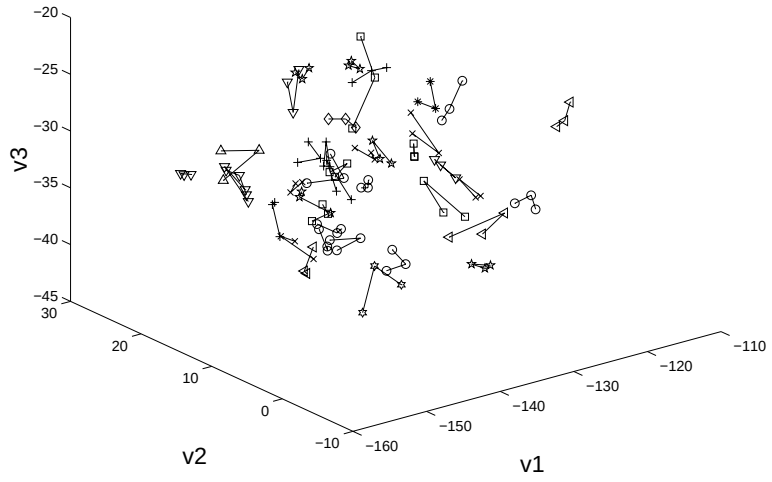
For gait recognition, we have proposed a statistical approach for feature extraction from spatial templates which can reduce dimensionality and increase separability using template matching. Promising recognition results and strong statistical evidence of variation between subjects reveal the value of template matching. Apart from the recognition performance, the computational complexity of template matching has been greatly reduced from the size of original image to the number of classes between template images. Although there are six subjects with 42 sequences in the second experiment which is larger than the first experiment which had 5 subjects and 25 sequences only, the promising recognition results still appear. However, we still need more data to evaluate the accuracy and robustness of our approach. The assumption we made about the test data is that individuals are walking laterally from right to left before a static camera. Accordingly, it appears that our approach is invariant to slight but not drastic changes of viewing angles, and we intend to investigate this further.

In the experiments of gait recognition, there are at most 6 subjects in the test data. Using template matching, gait recognition becomes a sequence matching in the database. Sequential search for the matching to each subject is used in this chapter. However, it is vital to develop an efficient searching algorithm in the database when the number of subjects increases and becomes large. Since the training database needs to be updated constantly for adding the gait sequences of new subjects, how to update the CST matrix efficiently without re-applying PCA and CA is also not a trivial issue. Eigenspace update algorithms have been proposed by researchers [122, 123]. The update algorithms of the canonical space is a similar but different problem which needs to be investigated for the purposes of computational efficiency. In this chapter, we only consider the spatial changes of human silhouettes but ignore the temporal information between silhouettes. Since gait is one kind of articulated motion, an appropriate representation to incorporate temporal information into spatial features is necessary.

For face recognition, eigenface approach suffers under variation in lighting direction and the feature extraction using the combined approach (EST and CST) achieves the better result than the eigenface (EST) approach by reducing the lighting effect. Canonical analysis is not applied alone to gait recognition in this chapter though it has the advantage for class separation. The main reason is that it is intractable to classify image data for the large size of covariance matrix (square of image size). Thus, we propose PCA to reduce the image dimension and to convert each image into a low-dimensional vector



(a) Eigenspace



(b) Canonical space

Figure 4.23: Distributions of 40 subjects with 3 images each

before applying CA for separating classes. Moreover, the singularity problem will occur in the covariance matrix if its size is larger than the number of training images.

In the other hand, applying CA directly to training images cannot be tested and compared due to the insufficient number of test images though each image size has been subsampled from 92×112 to 23×28. The performance of EST, CST and the combined approach can be compared when a larger database of face patterns is obtained in the future. Moreover, the recognition performance of statistical approaches is affected by the hair styles, viewing angles and scale changes of subjects. That is why many approaches [124, 65, 82, 125] try to normalise the face image into a standard template with only the area of eyes, the nose and the mouth before recognition. This is also need to be further investigated for the influence to our approach.

4.9 Conclusions

Gait is an emergent biometric aimed to recognise people by the way they walk. Extant approaches do not capitalise on the power of established statistical pattern recognition paradigms. In this chapter, we have proposed a gait recognition system which uses a new statistical approach for feature extraction in recognition via template matching by combining CST with EST to reduce data dimensionality and to optimise the class separability of different classes, simultaneously. Based on the hypothesis that to recognise different gait by human visual system depends on the spatial changes of the human body, we choose the method of template matching for gait recognition. Spatial templates are binary images of human silhouettes extracted from each scene image and rescaled to be image templates all with the same size which is consistent with the application of the statistical approaches of EST and CST.

This will eliminate the redundancy introduced by irrelevant background data, by scale changes during walking and by different clothes worn by the subject. Also, using spatial templates reduces the statistical problem caused by different lighting directions which occurs in face recognition using the eigenface approach. As manifest in experimental results, this improves the performance of eigenspace approach. Experimental results in this chapter show that training gait sequences are transformed into widely separated clusters in the new space. The recognition rate for test gait sequences is 100% in two experiments. In comparison with the results of EST method [37] and Little and Boyd's approach [40] independently, our new approach provides better results. The better cluster separation is achieved by canonical analysis and this might extrapolate better to larger database. However, this still needs further investigation to a larger database. Although the EST approach achieves the same recognition rate as the combined approach for gait recognition in the experimental results, better advantages of the combined approach than the EST approach are: shorter representation of each class in the database (only the centroid vector of each class needs to be stored, rather than the entire sequence),

faster matching time using accumulated distance measure in the canonical space (no problem of phase shifts), better class separation.

This combined approach - EST and CST, has been also applied to face recognition for feature extraction from images, and the results show that the combined approach achieves better recognition rates than the eigenspace approach. As such, this chapter describes a effective approach to data analysis in biometrics. Face recognition is suffering from scaling and positioning problems. The performance of statistical approaches needs to be reevaluated after solving those problems. Simultaneously, obtaining a larger database is also important for evaluation.

In solving the recognition problem of human gait which is a complicated articulated-motion, template matching shows its simplicity, accuracy and robustness using spatial templates. As such, this chapter describes a potentially effective approach for gait recognition. However, human gait includes motion information which cannot been revealed by individual spatial templates. The appropriate representation of this important information needs to be developed and included in the features for gait recognition and this will be discussed and presented in the next chapter.

Chapter 5

Incorporating Temporal Information

5.1 Introduction

In this chapter, we propose another gait recognition system [126, 127] which uses another feature - temporal templates computed from the optical flow and the combined approach - EST and CST, for feature extraction. Previously, we had proposed a statistical approach [117, 119] which combined *eigenspace transformation* (EST) with *canonical space transformation* (CST) for feature extraction from spatial templates and recognising humans by their gait successfully. The motion information of gait is actually represented by the projected trajectory in the eigenspace and the canonical space. However, human gait is one kind of articulated motion. Motion information also can be extracted from the change of human shapes between consecutive frames. Extracting motion information from gait sequences becomes a useful clue used to distinguish different subjects for recognition. There are generally two methods for extracting two-dimensional motion [128]: motion correspondence and optical flow. Motion correspondence - also called feature-based estimation [59], is concerned with the matching of characteristic tokens through time, while optical flow consists of the computation of the displacement of each pixel between frames. Due to the difficulty of finding specific features from the articulated motion of human shapes, optical flow is used here to extract a dense velocity map from human gait by considering temporal patterns as a whole instead of using motion correspondence.

Here, temporal information is incorporated from optical-flow changes between two consecutive silhouettes into temporal templates which represent distribution of velocity magnitudes in each pixel. The temporal templates are then used for gait recognition. The combined approach [117, 119] of EST and CST is used for feature extraction from templates before the recognition. Before the feature extraction by EST and CST, preprocessing is needed and applied to each gait sequence which will be converted to a

sequence of temporal templates. Temporal templates are square windows which are extracted from the optical flow computation of two consecutive scene images and rescaled to be templates all with the same size. Note that a walking subject is the only moving object in the scene. Using equally-sized templates is consistent with the statistical analysis of PCA and CA.

After preprocessing, all the template sequences from different subjects for training are analysed by principal component analysis and canonical analysis. Since temporal templates are not image templates and each pixel value represents a velocity value, brightness normalisation is not needed before analysis. After PCA and CA, a generated transformation matrix (EST + CST) can project each template sequence into the canonical space with separate clusters for different subjects. Matching a test gait sequence to a training subject is done by choosing the minimum *accumulated distance* to training centroids in the canonical space. Therefore, recognition of human gait becomes much more accurate and robust in this new space by incorporating motion information into temporal templates.

This chapter is organised as follows. Section 5.2 briefly outlines the gait recognition system using temporal templates. Section 5.3 briefly describes the basic theory of optical flow. In Section 5.4, how temporal templates are extracted from a gait sequence in the preprocessing stage is explained. The recognition performance achieved by four types of template features is evaluated in Section 5.5. Then, Section 5.6 shows the experimental results for gait recognition using temporal templates. The comparison of the EST approach, the frequency approach of Little and Boyd and our combined approach is also shown in this section. Results are discussed in Section 5.7, prior to conclusions in Section 5.8.

5.2 System Overview

The gait recognition system described in this chapter is similar to the system proposed in the previous chapter, the only difference is that temporal templates are used as features, rather than using spatial templates. By referring to Figure 4.1, the procedure of "Spatial Template Extraction" is replaced by "Temporal Template Extraction" in this chapter. After temporal templates are extracted from each gait sequence, brightness normalisation of templates is not needed.

In the training stage, all the temporal templates extracted from training gait sequences are processed by PCA and CA. The generated EST matrix, CST matrix and projected points of training sequences in the canonical space are retained in the database for the recognition of test sequences. In the test stage, test gait sequences are matched to the database with training sequences of subjects, after temporal templates are extracted and projected into the canonical space.

By using template projection, each gait sequence with motion information embedded

in each temporal template is projected to a periodic trajectory in the eigenspace and a cluster in the canonical space. Motion information is transformed from the velocity values of each template to the coefficients of each point in the eigenspace and then the canonical space.

5.3 Optical Flow Computation

Optical flow methods are very common for evaluating motion from a sequence of images. Using a sequence of images, the goal of optical flow methods is to calculate an approximation to the two-dimensional motion field, a projection of the three-dimensional velocities of surface points onto the imaging surface, from spatio-temporal patterns of image intensity. That is, to calculate the velocity of each pixel position from its motion. Consider a sequence of consecutive images where $I(x, y, t)$ denotes the image intensity of a point (x, y) at time t. Optic flow estimation is based on two constraints [129]:

1. The brightness of a particular point on the image is constant over time (the gradient constraint); and
2. Neighboring points on the image have similar velocities (the smoothness constraint).

Suppose an image point is translated a distance Δx in the horizontal direction and Δy in the vertical direction in a short time Δt. By applying the first constraint, we see that,

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t) \tag{5.1}$$

By Taylor series expansion, Equation (5.1) can be given by

$$I(x + \Delta x, y + \Delta y, t + \Delta t) = I(x, y, t) + \Delta x \frac{\partial I}{\partial x} + \Delta y \frac{\partial I}{\partial y} + \Delta t \frac{\partial I}{\partial t} + \xi \tag{5.2}$$

where ξ contains higher order terms. After substituting Equation (5.1) into Equation (5.2) and taking the limit $\Delta t \to 0$, ξ can be discarded so Equation (5.2) simplifies to

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t} = 0 \tag{5.3}$$

by retaining the first-order derivatives. This equation can be represented by

$$I_x u + I_y v + I_t = 0 \tag{5.4}$$

where $(u, v) = (\frac{dx}{dt}, \frac{dy}{dt})$ describes the horizontal and vertical velocities, $(I_x, I_y, I_t) = (\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial I}{\partial t})$ indicates partial derivatives of image brightness with respect to the spatial dimensions and time at a point (x, y). The other representation of Equation (5.4) can

be given by

$$\nabla I \cdot \mathbf{v} + I_t = 0 \tag{5.5}$$

in which $\mathbf{v} = (u, v)^T$ and $\nabla I = (I_x, I_y)$ is the spatial gradient changes of I in the x and y directions. However, the motion estimation for two unknowns, u and v, based on one equation using a single gradient constraint is ill-posed.

One more constraint can be considered is the smoothness constraint based on the assumption that neighboring points on the objects have similar velocities and the velocity field of the brightness patterns in the image varies smoothly almost everywhere [129]. One way to describe the additional constraint is to minimise the square of the magnitude of the gradient of the flow velocity. By using the smoothness constraint combined with the gradient constraint in Equation (5.4), (u, v) can be solved by minimising the quadratic error term [129] given by

$$E^2 = \int\int \{(I_x u + I_y v + I_t)^2 + \lambda^2[(u_x^2 + u_y^2) + (v_x^2 + v_y^2)]\} dxdy \tag{5.6}$$

defined over a region, where $(u_x, v_x) = (\frac{\partial u}{\partial x}, \frac{\partial v}{\partial x})$, $(u_y, v_y) = (\frac{\partial u}{\partial y}, \frac{\partial v}{\partial y})$ and λ reflects the influence of the smoothness constraint. By using two equations in (5.4) and (5.6), the two unknowns, u and v, can be solved.

Several methods have been developed by considering higher-order derivatives, different optimisation functions and characteristics of different filters. They can be divided into four classes [130]: differential methods, region-based matching, energy-based techniques and phase-based techniques. There are three processing stages common to those optical flow techniques [130]:

1. Prefiltering or smoothing with low-pass/band-pass filters in order to extract signal structure of interest and to enhance the signal-to-noise ratio,

2. the extraction of basic measurements, such as spatio-temporal derivatives or local correlation surfaces, and

3. the combination of these measurements to produce a 2-D flow field, which often involves assumptions about the smoothness of the underlying flow field.

Differential methods compute the velocity from spatio-temporal derivatives of images intensity or filtered versions of the image. Methods to compute the first order and second order derivatives were developed, although estimates from second order approaches are usually poor and sparse. In region-based matching, the velocity is defined as the shift which yields the best fit between image regions at different times by optimising some similarity or distance measures. Energy-based (or frequency-based) methods compute optical flow using the output from the energy of velocity-tuned filters in the Fourier

domain, while phase-based methods define velocity in terms of the phase behaviour of band-pass filter outputs.

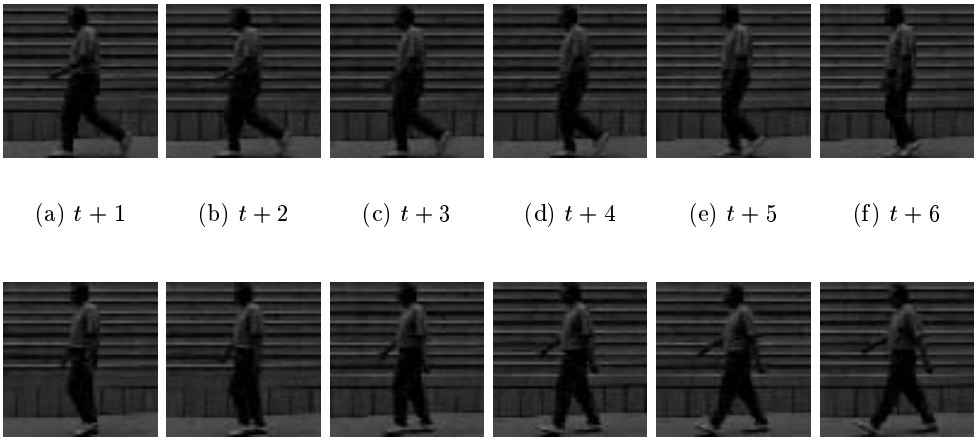
Based on the results of analysis [130], although the region-based matching techniques did not produce the most accurate velocity estimates among the techniques, as compared with the relatively large temporal duration of support used by the most successful techniques, they used only 2 or 3 frames. For gait analysis using the combined statistical approach for temporal templates, we are mainly concerned about the change of human shapes between two consecutive silhouettes, rather than the precise velocity value of each pixel. Therefore, we choose a region-based matching technique proposed by Bulthoff *et al.* [131] which uses only 2 frames to calculate the optical flow. Each temporal template which represents the change of two consecutive silhouettes is extracted from this optical flow field. This is a standard technique which has been used in [40] and given robust results for first-order flow values. Although the use of second-order flow values might also give information, they are normally more noisy than the first-order flows and not suitable for classification. Hence, they are not used here.

5.4 Temporal Template Extraction

Instead of isolating the moving figure manually, as in [40], we use the information of centroid and silhouette window from the extraction of spatial templates in the previous chapter to extract each temporal template which contains the flow within a moving window. According to the information of centroid and silhouette window in each original image, the cropped and normalised human silhouettes are displayed in Figure 5.1. The original images are shown in Figure 5.2 for comparison. The aspect ratio will be kept constant when the window size is normalised to 64×64. This shows that each human silhouette can be correctly cropped from each original image and this information can be used for cropping temporal templates after the optical flow computation between two consecutive images.

Unlike other methods, Little and Boyd [40] used dense optical flow fields, generated by minimising the sum of absolute differences between image patches [131]. However, this algorithm is sensitive to brightness change caused by reflections, shadows, and changes of illumination. Therefore, the images are firstly processed by computing the logarithm of brightness and converting the multiplicative effect of illumination change into an additive one. Secondly, each processed image is filtered by a bandpass filter (Laplacian of Gaussian) to remove the additive effects.

Basically, the algorithm searches for the displacement of each pixel among a limited set of discrete displacements by minimising the sum of absolute differences between a patch in one image and the corresponding displaced patch in the other image. The selected displacement is 5 pixels which results in a 10×10 patch (±5 pixels in each direction) in this chapter. After a best matching patch in the second image is found for each patch



(a) $t+1$ (b) $t+2$ (c) $t+3$ (d) $t+4$ (e) $t+5$ (f) $t+6$

(g) $t+7$ (h) $t+8$ (i) $t+9$ (j) $t+10$ (k) $t+11$ (l) $t+12$

Figure 5.1: Sample cropped human silhouettes from original images

in the first, the algorithm is run a second time and the roles of the the two images are switched. The results will agree for a correct match. In order to remove invalid matches, the results at each point in the first image are compared with the result at the corresponding point in the second. The second point should match the first: the sum of displacement vectors should be approximately zero. Only those matches that pass this validation test are retained. The results could be interpolated to provide sub-pixel displacements but only integral values are used here. In effect, the minimum displacement is 1.0 pixels per frame; points that are assigned non-zero displacements form a set of *moving points*. In general, optical flow techniques are computationally intensive. Based on running in a Pentium-133 PC and using the proposed region-based technique, the computation time to extract temporal templates from two consecutive 320×160 image frames takes 35 seconds for a 10×10 patch and 155 seconds for a 20×20 patch.

Three kinds of temporal templates are generated: the u-flow templates which are horizontal components of flow, v-flow templates which are vertical components of flow and $|(u,v)|$-flow templates which are the magnitudes of (u,v) as calculated from u-flow and v-flow templates. Using the first two consecutive images in Figure 5.2, Figure 5.3 shows the preprocessed image by optical flow computation for u-flow, the extracted u-flow template, the preprocessed image by optical flow computation for v-flow and the extracted v-flow template. Sample temporal templates are shown in Figures 5.4 for u-flow templates, Figures 5.5 for v-flow templates and Figures 5.6 for $|(u,v)|$-flow templates from a gait sequence, respectively. Note that each temporal template shows the temporal changes of pixels within the region of one spatial template between two consecutive images. For the u-flow templates in Figure 5.4 and v-flow in Figure 5.5, positive velocities are represented by increasing gray-level from 0 to 128 and negative velocities decreasing from 255 to 128 (associated from -1 to -128) due to the range of moving pixels

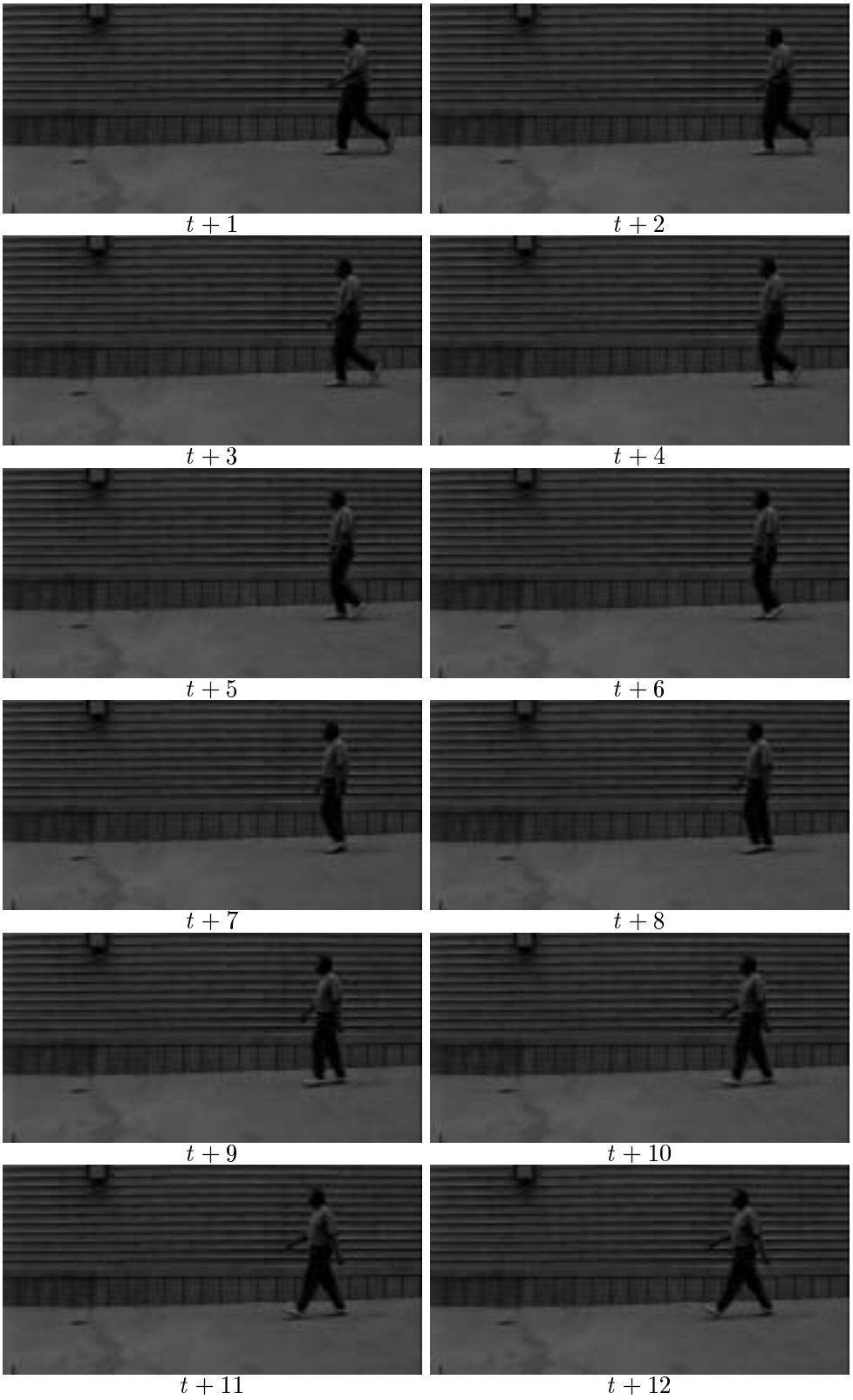


Figure 5.2: One sample sequence of original images

(a) Optical flow computation of u-flow

(b) Extracted u-flow template

(c) Optical flow computation of v-flow

(d) Extracted v-flow template

Figure 5.3: Temporal template extraction

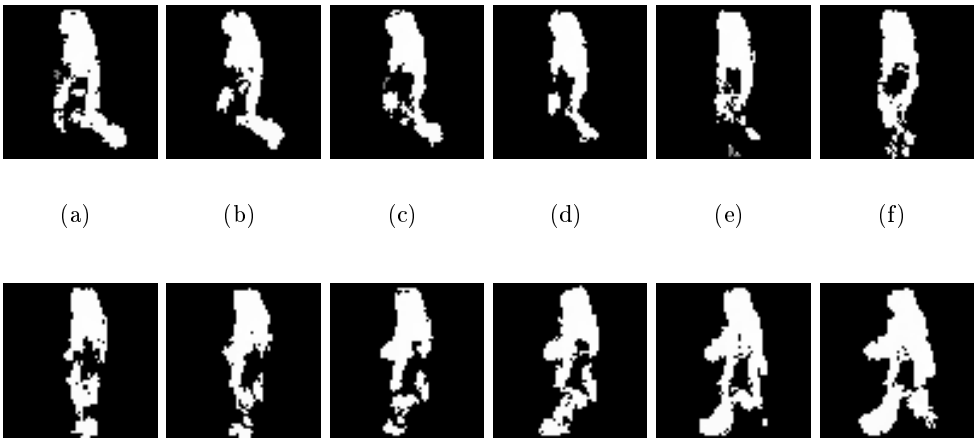
(velocities) is ± 5. For the $|(u, v)|$-flow templates in Figure 5.6, velocity magnitudes are represented from 128 to 255 (associated from 0 to 128) due to the positive values of the $|(u, v)|$ magnitudes. For display purposes, the gray-levels are associated from black (0) to white (255). Comparing Figure 5.4 with Figures 5.5, they show that the pixel movements of optical flow are less in the vertical direction than in the horizontal direction.

5.5 Performance Comparison of Different Template Features

Including the spatial templates used in the previous chapter and three kinds of temporal templates extracted in this chapter, we have four kinds of feature templates. The three kinds of temporal templates are the u-flow templates, v-flow templates and $|(u, v)|$-flow templates as extracted from the optical flow computation. By using the proposed statistical approach - EST and CST, the performance comparison in gait recognition using four feature templates individually is presented in this section.

5.5.1 Training and Recognition

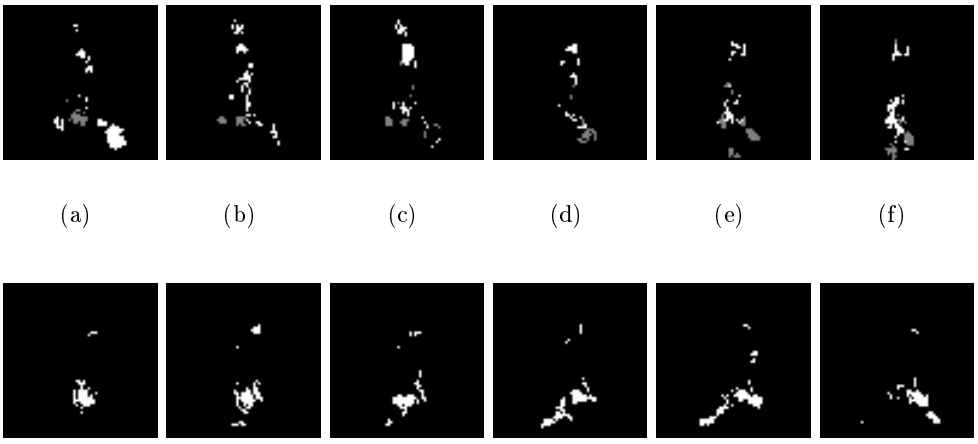
Let a test gait sequence be $g(t)$, in which $t = 1, ..., T$. Before recognition, four different kinds of templates are extracted from this test sequence and projected into individual



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k) (l)

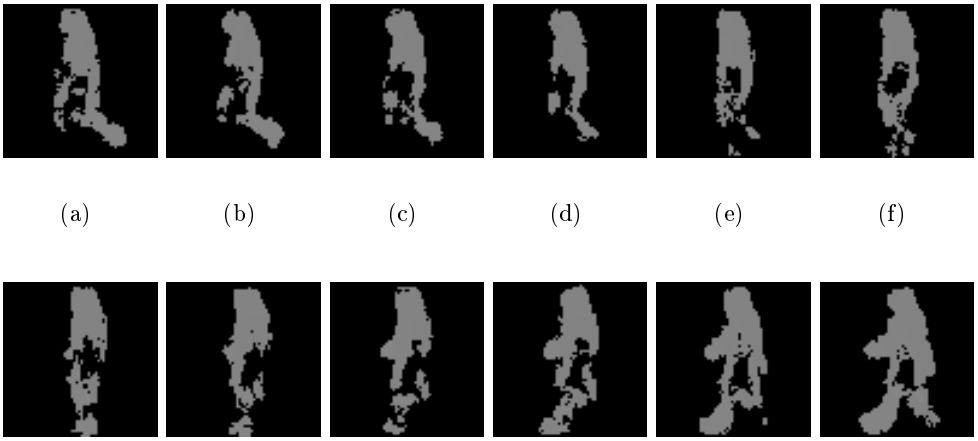
Figure 5.4: Sample u-flow templates



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k) (l)

Figure 5.5: Sample v-flow templates



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k) (l)

Figure 5.6: Sample $|(u, v)|$-flow templates

trained canonical space, given four vector sequences after projection, $h_1(t)$, $h_2(t)$, $h_3(t)$ and $h_4(t)$, representing spatial, u-flow, v-flow and $|(u, v)|$-flow templates, respectively. To recognise a human walking sequence from a trained database in each canonical space, the *accumulated distance to each centroid* is used. The accumulated distance between test vector sequences, $h_k(t)$, in which $k = 1, ..., 4$ and c centroids, $\mathbf{C}_{i,k}$, in which $i = 1, ..., c$ is

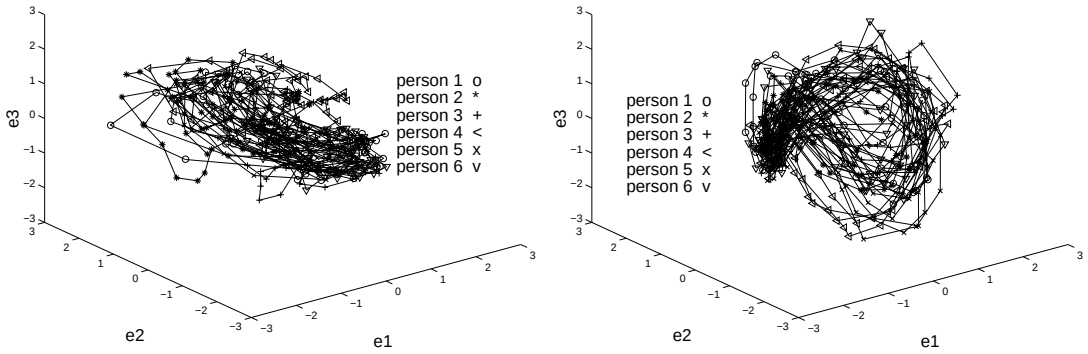
$$d^2_{i,k} = \sum_{t=1}^{T} \|h_k(t) - \mathbf{C}_{i,k}\|^2, \tag{5.7}$$

where $\mathbf{C}_{i,k}$ is the centroid of class i in canonical space k. To match a test sequence $h_k(t)$ to a training sequence i in canonical space k can be accomplished by choosing the *minimum* $d^2_{i,k}$.

5.5.2 Experimental Results

The sample human gait data which is used in the previous chapter with 6 subjects and 7 sequences of each is used for the experiments. One walking sequence is selected at random from each person as the training sequence and the remaining 36 sequences served as test sequences. Preprocessing and template extraction are applied to spatial templates, u-flow templates, v-flow templates and $|(u, v)|$-flow templates individually.

After applying PCA to the four separate training sets, four sets of training feature templates are projected into four individual eigenspaces by a generated EST. Results in Figures 5.7(a), 5.7(b), 5.7(c) and 5.7(d) show that six classes of training sequences using spatial templates, u-flow templates, v-flow templates and $|(u, v)|$-flow templates are overlapped with individual trajectories in each eigenspace. For visualisation purposes, we again show only the first three of the five possible dimensions. Linear re-scaling [109]



(a) Spatial templates

(b) u-flow templates

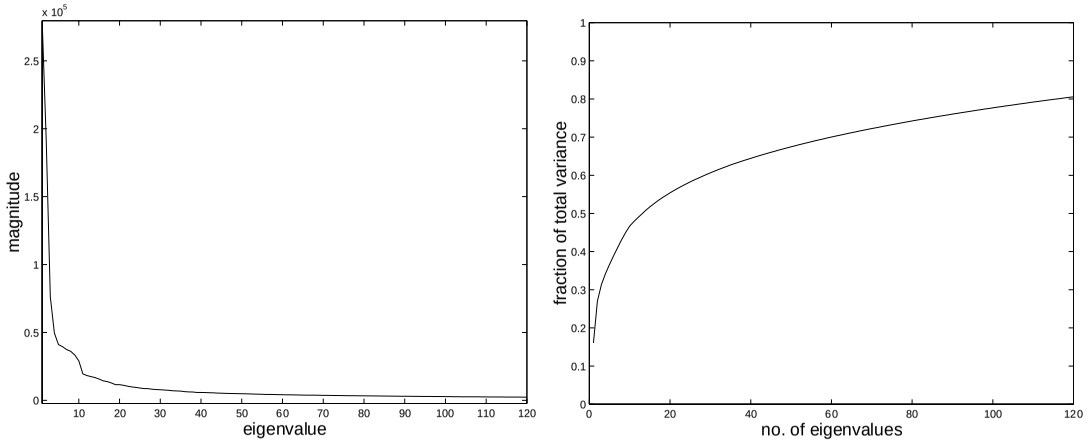
(c) v-flow templates

(d) $|(u, v)|$-flow temporal templates

Figure 5.7: Distributions of 6 training sequences in four eigenspaces

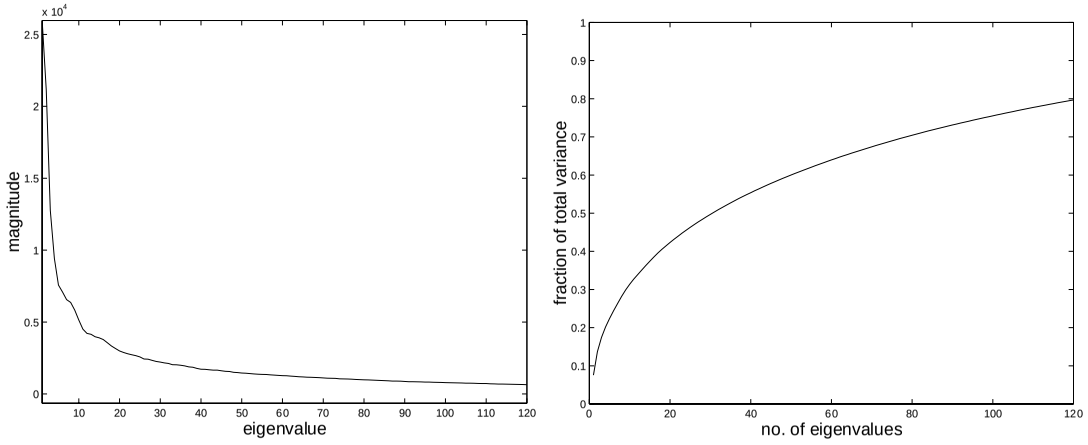
has been applied to each vector to set the average of each data set to zero and to normalise the standard deviation to unity. Figures 5.8, Figures 5.9 and Figure 5.10 show the magnitudes of eigenvalues and their accumulated variances in the eigenspace after the u-flow, the v-flow and the $|(u, v)|$-flow training templates are trained by individual PCAs. The characteristics of eigenvalues for the spatial templates have been shown in the previous chapter.

After applying individual CAs to the four projected training sets in four eigenspaces, they are further projected to four respective canonical spaces. Again, they are linear-rescaled for further comparison in recognition using accumulated distance measures for the 4 different template features in the canonical space. Results in Figures 5.11(a), 5.11(b), 5.11(c) and 5.11(d) show that the six classes of training sequences using spatial templates, u-flow templates, v-flow templates and $|(u, v)|$-flow templates are greatly separated to become clusters in each canonical space. Figure 5.12(a), Figure 5.12(b) and Figure 5.12(c) show the eigenvalues of the u-flow templates, v-flow templates and $|(u, v)|$-



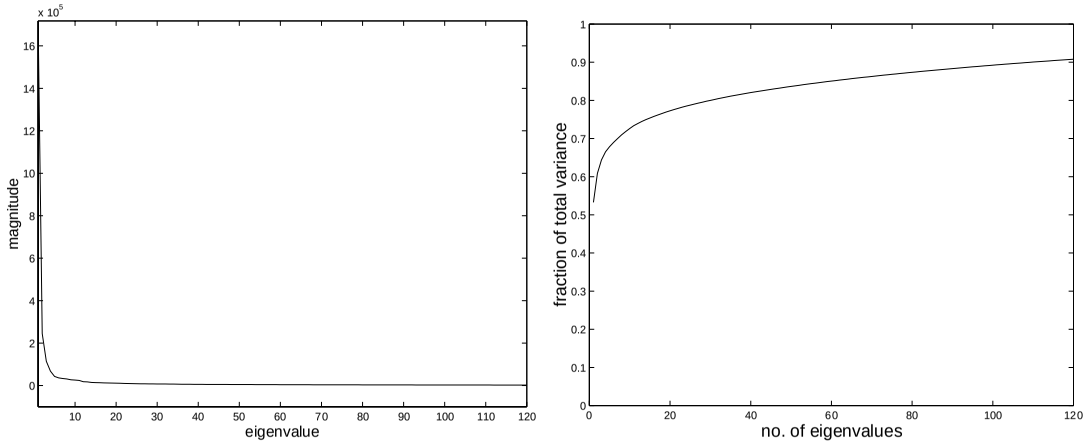
(a) Magnitudes (b) Accumulated variances

Figure 5.8: Eigenvalues in the eigenspace using u-flow templates



(a) Magnitudes (b) Accumulated variances

Figure 5.9: Eigenvalues in the eigenspace using v-flow templates

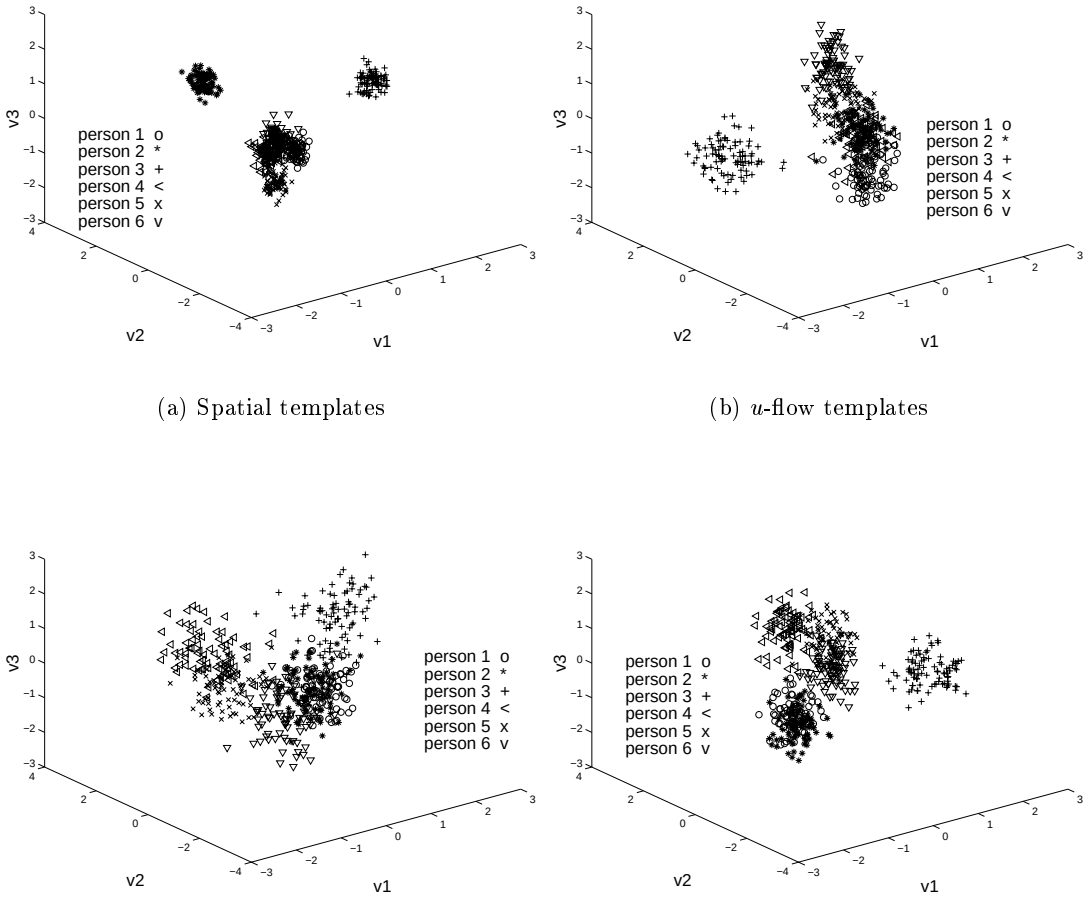


(a) Magnitudes (b) Accumulated variances

Figure 5.10: Eigenvalues in the eigenspace using $|(u, v)|$-flow templates

flow templates in the canonical space and the total variance is accumulated by the variances over the first 5 eigenvalues. Thus, the first 5 eigenvectors are used as the transformation matrix of canonical space. Actual gait recognition is achieved in this space and the size of template matching is greatly reduced from 64×64 to 5.

Figures 5.13(a) and 5.13(b) show relative accumulated distances of one training and one test sequence from subject 1. Here, the v-flow template has the lowest distances to each incorrect subject and hence the poorest discriminatory ability. Conversely, the spatial template offers the best discriminatory ability, associated with the greatest distances to incorrect subjects. Figures 5.13(c) and 5.13(d) show relative accumulated distances of two misclassified sequences, one from subject 4 and one from subject 5. Here, the distance by the v-flow template leads to confusion in classification, in Figure 5.13(c), where the fifth sequence of subject 4 can be classified as subject 1 and, in Figure 5.13(d), subjects 1 and 2 appear close to the target subject 5. Conversely, the spatial template again offers best performance, again there is little difference between the $|(u, v)|$-flow template and the u-flow template and both perform better than the v-flow template measures. The comparison of recognition performance using the combined approach of EST and CST for four different templates is shown in Table 5.1. Clearly, the feature vectors generated by the combined approach yield high recognition rates. Using template matching, the poor performance achieved by v-flow templates can be explained by the reduced information of optical flow from the extracted templates in Figure 5.5. Vertical movements of gait usually have smaller changes than horizontal movements, and thus have less discriminatory power in distinguishing different gaits. Spatial templates, u-flow templates and $|(u, v)|$-flow templates make for better performance in recognition. Although promising results have been shown here, further comparison of the three templates still needs a larger database to evaluate their performance.



(a) Spatial templates

(b) u-flow templates

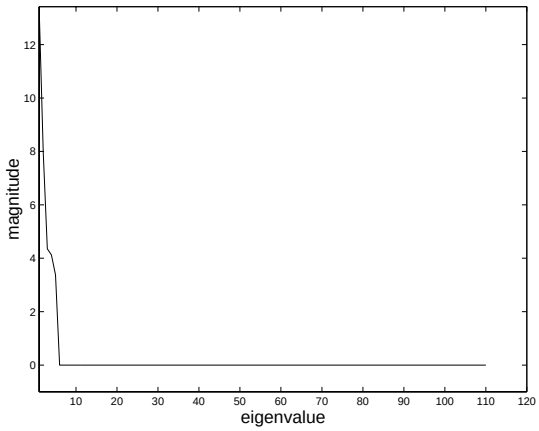
(c) v-flow templates

(d) $|(u, v)|$-flow temporal templates

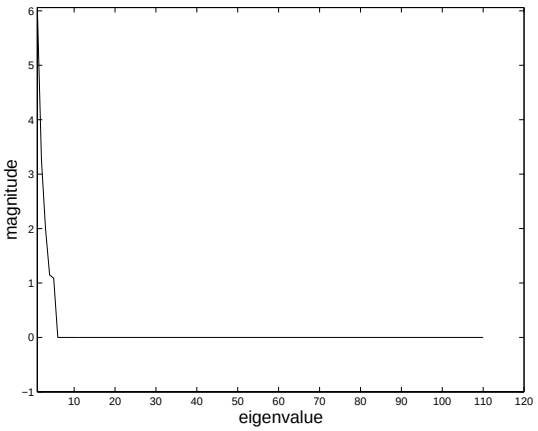
Figure 5.11: Distributions of 6 training sequences in four canonical spaces

5.6 Comparing to Other Approaches Using Temporal Templates

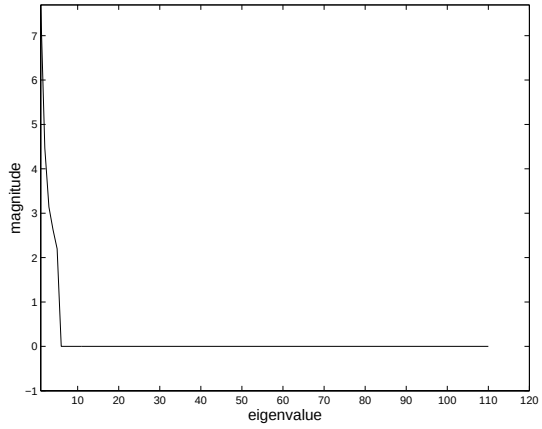
In the previous section, we have seen the experimental results of gait recognition using 4 different template features. According to the performance comparison for 3 kinds of temporal templates, although the u-flow templates achieves largely similar (or even slightly better) performance than the $|(u, v)|$-flow templates in the experiments, $|(u, v)|$-flow templates (which include the magnitude of u-flow and v-flow templates) have potentially more information than u-flow templates. Thus, $|(u, v)|$-flow templates have been selected for comparison with other approaches in gait recognition. Performance comparison using $|(u, v)|$-flow templates by our statistical approach, the eigenspace approach and Little and Boyd's approach [40] is shown in this section.



(a) u-flow templates



(b) v-flow templates



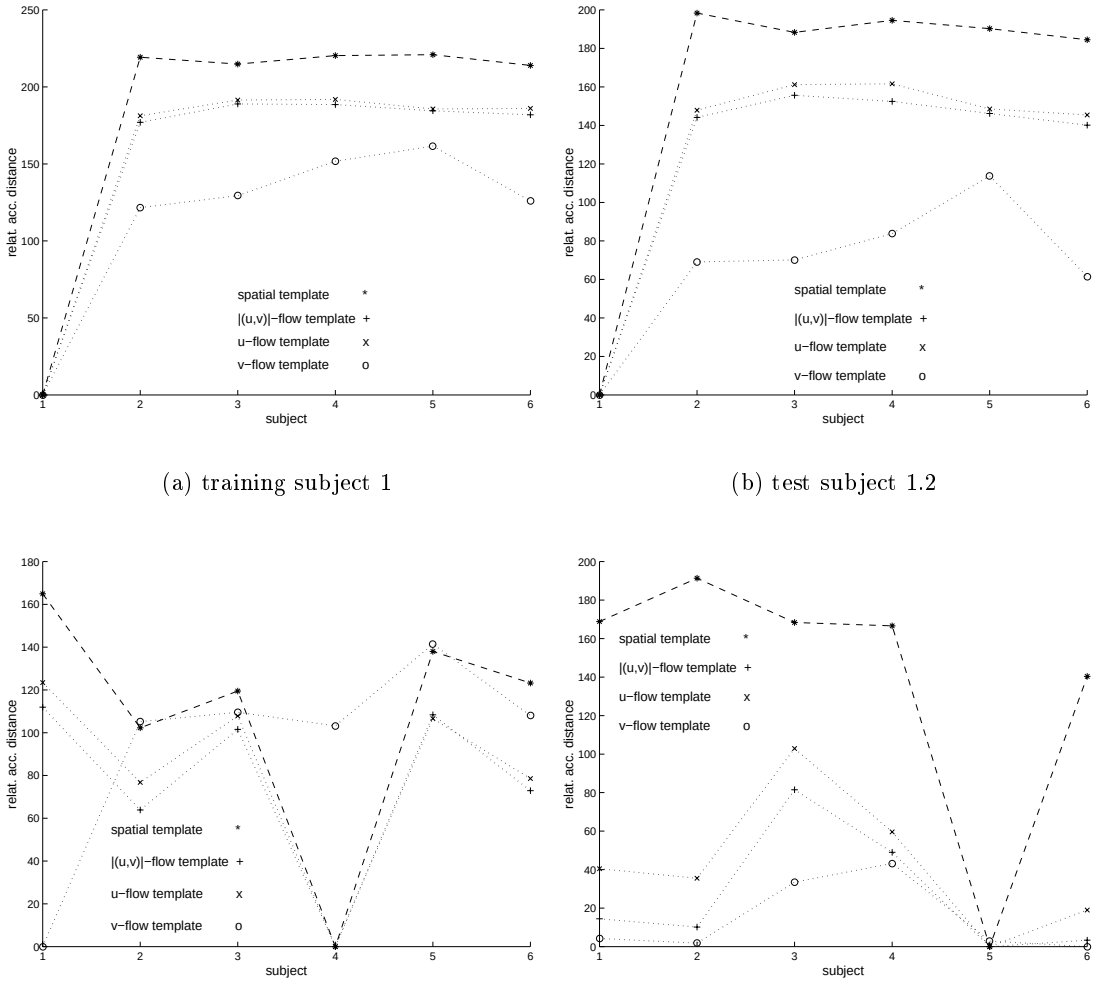
(c) $|(u, v)|$-flow templates

Figure 5.12: Eigenvalues in three canonical spaces

5.6.1 Experimental Results

The same gait data is used as in Section 5.5.2. One walking sequence is selected at random from those for each subject as the training sequence and remaining 36 sequences served as test sequences. Before training and projection, each sequence has been preprocessed and converted into a template (size 64×64) sequence as described in Section 5.4.

The characteristics of the eigenvalues for the $|(u, v)|$-flow templates have been shown in Section 5.5.2. Figure 5.10(a) shows the magnitude of each eigenvalue in the eigenspace after training templates are trained by PCA. Figure 5.10(b) shows their accumulated variances. We choose the first 110 eigenvalues which accumulate 90% of the total variance and their corresponding eigenvectors as the eigenspace transformation matrix. Thus, each temporal template can be represented by the linear combination of those 110 principal eigenvectors. Although temporal templates are not real image templates,



(a) training subject 1

(b) test subject 1.2

(c) test subject 4.5

(d) test subject 5.6

Figure 5.13: Relative accumulated distance of gait sequences

	feature used	recognition rate
(1)	spatial templates	100%
(2)	u-flow templates	100%
(3)	v-flow templates	95.2%
(4)	$\|(u,v)\|$-flow templates	100%

Table 5.1: Recognition performance using different template features

(a) $e1$ (b) $e2$ (c) $e3$ (d) $e4$ (e) $e5$ (f) $e6$

Figure 5.14: First 6 eigenvectors using temporal templates

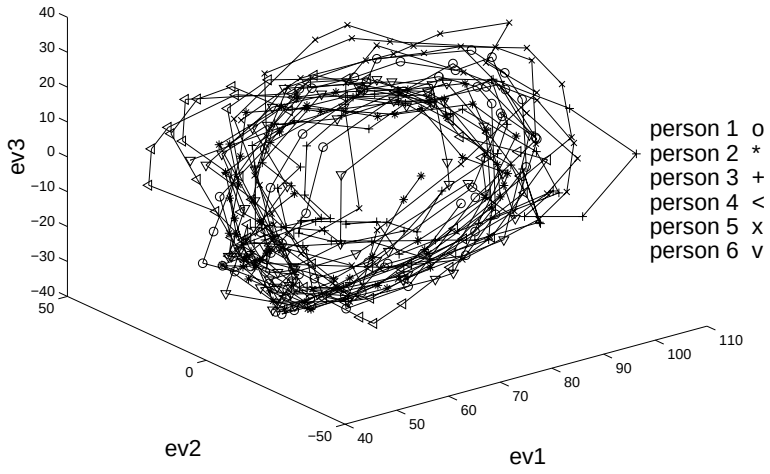
	Method	recognition rate
(1)	EST	92.7%
(2)	Little & Boyd's	95.2%
(3)	EST + CST	100%

Table 5.2: Recognition using different approaches for temporal templates

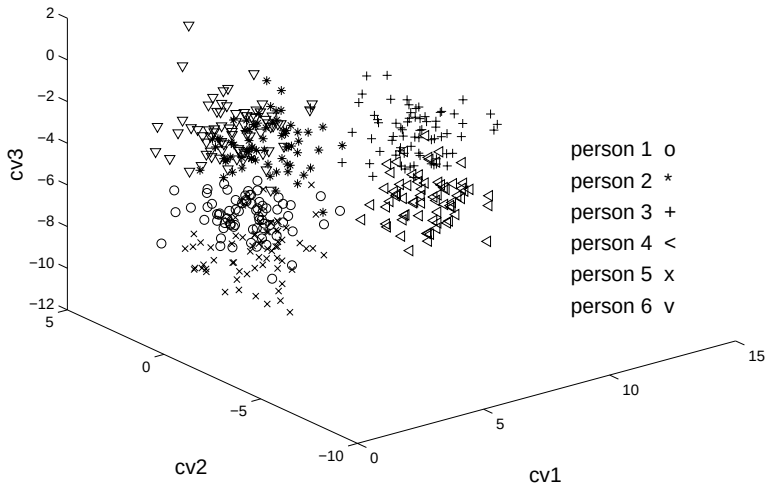
Figure 5.14 shows the first 6 eigenvectors which have been normalised to image templates for display purposes. Each $|(u, v)|$-flow template is then projected into a point in this 110-dimensional eigenspace which is much smaller than the original 64×64-dimensional image space.

Canonical analysis is further applied to these 110-dimensional vectors converted from temporal templates. Figure 5.12(c) in Section 5.5.2 shows the eigenvalues in the canonical space and the total variance is accumulated by the variances over the first 5 eigenvalues. Thus, the first 5 eigenvectors are used as the transformation matrix of canonical space and each 110-dimensional vector is further projected into a 5-dimensional canonical space. Recognition is actually achieved in this space and the operations of template matching for each pair is greatly reduced from 64×64 to 5. Figures 5.15(a) and 5.15(b) show the distribution of six training sequences in eigenspace and canonical space, respectively. For visualisation purposes, we only show the first three of five dimensions. As we can see, the class separability in canonical space is much better than in eigenspace. Note that linear-rescaling is not applied to the projected vectors in the eigenspace and the canonical space due to the comparison to other approaches only in recognition rates but not in scales as in Section 5.5.2.

In order to compare the recognition performance of EST with our approach, the *spatio-temporal correlation* [37] which is adopted in the previous chapter is used here to recognise an input gait sequence in eigenspace, although it has not been used for temporal templates before. In canonical space, we use *accumulated distance* described in the previous chapter as the distance measure to recognise different gaits. The comparison of recognition performance using three different approaches is shown in Table 5.2. Clearly, the feature vectors generated by the combination of EST and CST yield the best recognition rate among those three approaches. Although the same gait data is used as in



(a) eigenspace



(b) canonical space

Figure 5.15: Distributions in two subspaces

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	21.4%	45.2%	69.0%	83.3%	85.7%	83.3%	85.7%
2 cycles	19.0%	78.6%	90.5%	95.2%	95.2%	92.9%	95.2%
3 cycles	14.3%	81.0%	90.5%	95.2%	97.6%	97.6%	97.6%
4 cycles	69.0%	90.5%	95.2%	97.6%	97.6%	100%	100%

Table 5.3: Recognition rates using different training samples and eigenvalues

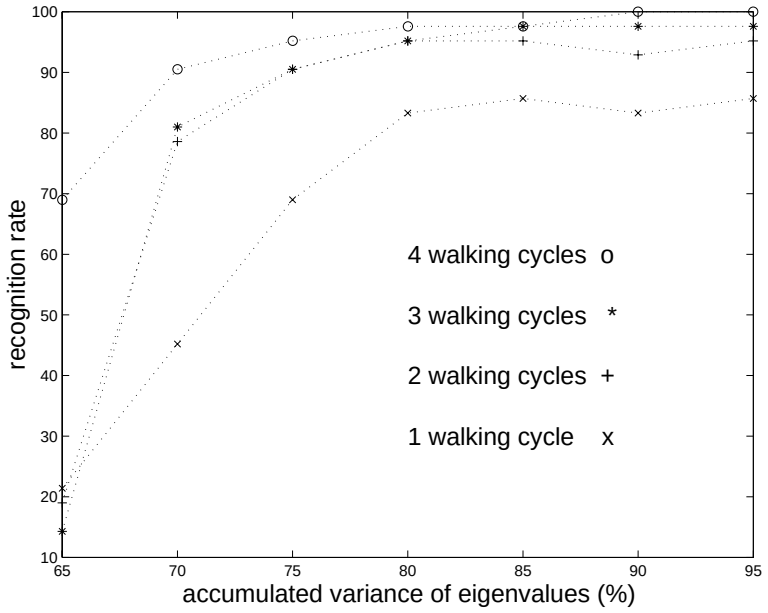


Figure 5.16: Recognition rates using different training samples and eigenvalues

previous chapter, the recognition performance using only EST for temporal templates in the eigenspace is no longer as good as using spatial templates as features.

5.6.2 The Influence of Different Training Samples and Eigenvalues

For the gait recognition using $|(u, v)|$-flow templates in Section 5.6.1, we choose the first 110 eigenvalues which accumulate 90% of total variance in eigenspace and their corresponding eigenvectors as the eigenspace transformation matrix. In order to evaluate the recognition effects in canonical space using different training samples (walking cycles) and eigenvalues, here we conducted four tests which used 18, 36, 54 and 72 templates corresponding to 1, 2, 3 and 4 walking cycles from each training sequence for training. Furthermore, In each test, we choose 7 different accumulated variances ranging from 65% to 95% achieved by different numbers of eigenvalues (associated to eigenvectors). The comparison of recognition performance is shown in Table 5.3 and Figure 5.16. The results reveal that the best performance is accomplished by using over 90% accumulated variance of eigenvalues and 4 walking cycles of training samples from each subject. Fur-

thermore, the recognition performance is improved by increasing the number of training samples and accumulated variances (number of eigenvectors). Thus, results would appear to confirm sensitivity to the number of training samples and implies that in a more extended analysis, care must be taken to include sufficient samples in the training database.

5.7 Discussions

Although gait patterns may be different depending on age, sex, clothing, etc., here we only consider the gait as a discriminatory feature between different subjects rather than considering any effect introduced by the subject. Such factors will doubtless be of concern in later work, but our concern here is more the basic nature of gait, as described by standard measures. In this chapter, we have proposed one new feature - temporal templates, for gait recognition. Unlike spatial templates which are human silhouettes extracted from the gait sequence and invariant to different illumination conditions, temporal templates are extracted from the computation of optical flow which suffers under variation in lighting direction. Moreover, different algorithms will generate different flow values.

Optical flow is an approximation to the two-dimensional flow field from image intensities. Accurate and dense measurements are difficult to achieve. There are four major limitations for accurate optical flow analysis [59]:

1. Optical flow is very sensitive to noise due to its dependence on spatio-temporal gradients;
2. Optical flow requires that motion be smooth and small;
3. Optical flow requires that motion varies continuously over the image; and
4. Optical flow is affected by object occlusion and choice of initial or boundary conditions.

One problem with optical flow in general, is that it is susceptible to the aperture problem [128] which, in some conditions, only allows the precise computation of the normal flow, i.e. the component parallel to the gradient. Despite this, it has successfully been used as a source of motion information and the approximation of the change of human shapes between two frames has achieved an appropriate representation and performed promising results in gait recognition using statistical approaches. Naturally, the flow technique is sensitive to changes in frame rate and in speed of the subject. However, this is not a issue here where the sequences are of exactly the same type, but for other data the technique would need to be modified via reformulation to compensate for these factors.

When comparing the recognition performance of different approaches, the feature of $|(u, v)|$-flow templates is selected. Using the same feature, recognition results achieved by our statistical approach, the eigenspace approach and Little and Boyd's approach [40] show that the combination of EST and CST accomplishes the best performance. This shows that temporal templates do provide useful motion information for gait recognition and the new approach still has promising discriminatory power for temporal templates, not just for spatial templates.

Using the combined statistical approach for feature extraction, experimental results show that spatial templates, u-flow templates and $|(u, v)|$-flow templates are better features than v-flow templates for gait recognition. This can be explained by the statement that vertical movements of gait usually have smaller changes than horizontal movements, thus have less discriminatory power in distinguishing different gaits. The feature of spatial templates achieves the best result when compared to the other three temporal templates, this means that the motion information extracted from optical flow computation is useful but not fully satisfied for gait recognition. It is necessary to compute the optical flow using stable algorithms under a well constrained environment.

5.8 Conclusions

In this chapter, a gait recognition system using temporal templates as features has been presented. The statistical approach proposed in the previous chapter which combined EST with CST has been used for feature extraction. The accumulated distance has been used for recognition in the canonical space. By inheriting the advantages of dimension reduction and class separability, the presented gait recognition system reveals promising recognition performance using temporal templates which have incorporated the temporal motion information between two consecutive image frames by the computation of optical flow. Three kinds of temporal templates are generated: the u-flow templates (the horizontal components of flow); v-flow templates (the vertical components of flow) and $|(u, v)|$-flow templates (the magnitudes of (u, v) as calculated from the u-flow and v-flow templates). Apart from spatial templates used in previous chapter, those three new features provide motion information from each gait sequence by integrating temporal information into each template. Thus, the motion information is actually included into the template sequence and can be analysed by the new statistical approach.

Two experimental results have been shown in this chapter. At first, the $|(u, v)|$-flow templates which combine the information of u-flow templates and v-flow templates have been selected as features for gait recognition. By comparing with the eigenspace approach and the Little and Boyd's approach, the proposed system still achieves the best result in gait recognition. Secondly, the analysis and comparison of recognition performance for each individual feature has been also discussed in this chapter and the results show that the spatial templates, the horizontal u-flow templates and the magnitude $|(u, v)|$-flow

templates show better discriminatory power than the vertical v-flow templates in gait recognition.

Although promising results have been achieved by the proposed statistical approach using the spatial and temporal templates as features individually in gait recognition, those two feature templates either lack the temporal information in the spatial templates or the spatial information in the temporal templates. How to integrate spatial and temporal information becomes a nontrivial issue. Motivated by the extended feature vectors [98] which suggest that orthogonal feature sets (uncorrelated to each other) should be chosen to reduce the variance of a final match measure, extended feature vectors which combine the spatial and temporal templates are proposed as features for gait recognition in next chapter.

Chapter 6

Combining Spatial with Temporal Information

6.1 Introduction

In this chapter, we propose a gait recognition system [132, 133] using the spatial-temporal feature which combines the spatial and temporal information from the projected vectors of spatial templates and temporal templates in the canonical space after EST and CST. Recognition is achieved in the extended canonical space using the accumulated distance as the metric. Using template matching, promising performance in gait recognition has been shown in the previous two chapters by using spatial and temporal information separately. However, spatial and temporal templates have either only spatial or temporal information of human gait though they have achieved promising results for gait recognition. Motivated by the extended features for face recognition [98], extended vectors which combine spatial and temporal information would appear to have potential for gait recognition. Therefore, the hypothesis used in this chapter is that recognising people by their gait could be more robust and accurate if the spatial and the temporal information can be integrated together. Based on the spatial and temporal information provided by the projected vectors of spatial and temporal templates in each individual canonical space, how to fuse these two information into one single feature becomes the next issue.

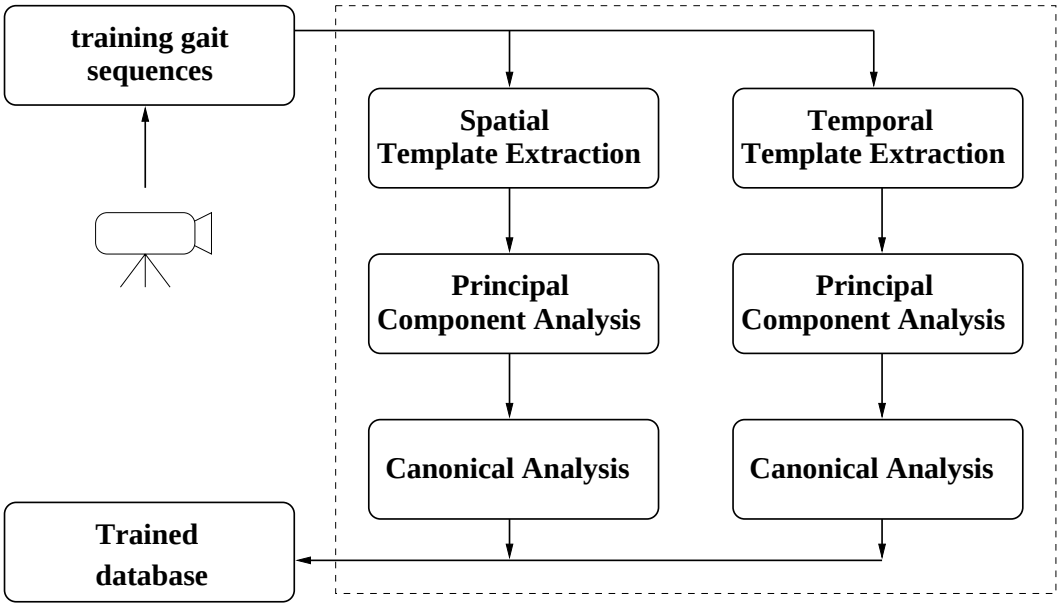
Since the spatial information has been included in each spatial template and the temporal information of motion between two consecutive spatial templates is embedded in each temporal template, the integration can be done by simply concatenating the projected vectors of each spatial and temporal template in the canonical space together at each time instant. However, the scales of two canonical spaces for spatial templates and temporal templates are different after their individual EST and CST. Thus, linear rescaling [109] is needed to set the average of each data set to zero and to normalise the standard deviation to unity before the concatenation. Although Boyle [92] has proposed an approach to compute optimal scaling factors of individual augmented components for

balancing their contributions in the eigenspace, it seems natural to directly concatenate spatial and temporal features together, after the linear re-scaling. In order to maximise the performance of gait recognition, the analysis for the contributions of individual gait features in the canonical space needs to be further investigated. By incorporating spatial and temporal information into extended feature vectors in canonical space, gait recognition can potentially become more robust and accurate than using any single feature alone. The recognition scheme is based on low-level features of motion, and the recognition or tracking of specific parts of the subject is not required.

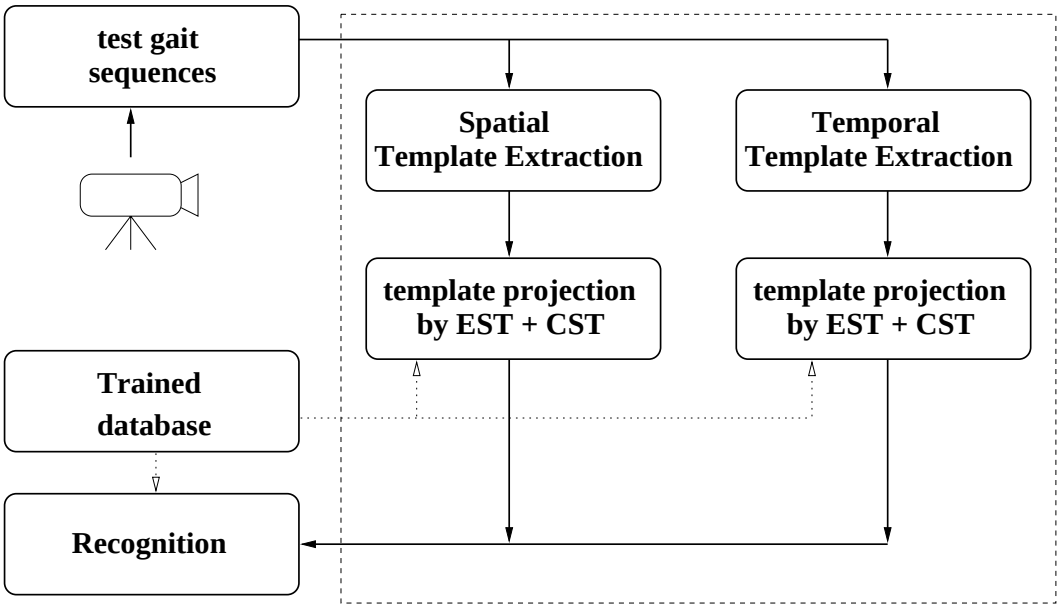
This chapter is organised as follows. Section 6.2 briefly describes the gait recognition system using extended features. How to extract spatial and temporal templates from each sequence is explained in Section 6.3. In Section 6.4, the corresponding methods of feature extraction for spatial and temporal templates and how to integrate them in the canonical space are described. Section 6.5 presents the strategy for recognising test sequences in the extended space after spatial and temporal templates are projected into individual canonical space and combined as extended vectors. Using the UCSD gait data, the recognition performance achieved by the extended features and the comparison with the performance using spatial and temporal templates are evaluated in Section 6.6. Using new gait data, the recognition results for different feature templates are evaluated in Section 6.7. The combined data set of UCSD and SOTON is evaluated for the extended feature vectors in Section 6.8. Results are discussed in Section 6.9, prior to conclusions in Section 6.10.

6.2 System Overview

Since the recognition system proposed here uses extended features which integrate the feature vectors projected from spatial templates and temporal templates by EST and CST, this system is actually the combination of two independent systems described in the previous two chapters for spatial and temporal templates. The training diagram of the combined system is depicted in Figure 6.1(a) and the test diagram is shown in Figure 6.1(b). In the training diagram, the input to the dashed box is the training set of gait sequences belonged to different subjects. After the parallel processing to two training processes for extracted spatial and temporal templates, the projected feature vectors in two independent canonical spaces are concatenated into extended vectors which constitute an extended canonical space. The generated transformation matrixes (by PCA and CA) and the extended vectors which represent gait sequences of different subjects are retained in the trained database for further transformation and recognition of test gait sequences. In the test diagram, the test gait sequences are preprocessed by template extraction and projection. The projected vectors (by EST and CST) of spatial and temporal templates are concatenated into extended vectors before recognition. Then, those extended vectors are matched to the trained database according to



(a) Training diagram



(b) Test diagram

Figure 6.1: Block diagrams of training and test for extended features

the accumulated distance measure.

6.3 Feature Template Extraction

Template extraction of spatial and temporal templates has been described in the previous two chapters. Since two different sets of temporal templates are generated and used in this chapter, only the extraction of temporal templates is briefly explained here.

For the extraction of temporal templates, Little and Boyd's [40] technique, as based on the algorithm of Bulthoff *et al.* [131], is used to generate optical flow fields between two consecutive frames. Instead of isolating the moving figure manually, as in [40], we use the information of centroid and silhouette window from the extraction of spatial templates to extract each temporal area from the optical flow diagram generated by two consecutive images. Then each temporal area is fitted in a 64×64 temporal template by normalising its position and size with constant aspect ratio. Basically, the algorithm searches for the displacement of each pixel among a limited set of discrete displacements by minimising the sum of absolute differences between a patch in one image and the corresponding displaced patch in the other image. Apart from the selected displacement of 5 which results in a 10×10 patch in the previous chapter, in this chapter, the displacement 10 which results in a 20×20 patch is also selected to generate a second set of temporal templates. This will introduce some outliers into each temporal template and can test the robustness of the proposed systems for noise flows.

Four kinds of templates are extracted which are spatial templates, u-flow templates, v-flow templates and $|(u, v)|$-flow templates, respectively. In this chapter, the $|(u, v)|$-flow templates are used to represent temporal gait information. Sample temporal templates are shown in Figure 6.2 using displacement 10 for UCSD data. As shown in Figure 6.2, the obtained temporal templates have more motion than that due to human motion (i.e. are contaminated more by noise than the temporal templates in the previous chapter). Temporal templates generated by displacement 5 have been shown in the previous chapter. Temporal information is incorporated from optical-flow changes between two consecutive spatial templates into temporal templates which represent the distribution of velocity magnitudes in each pixel and are combined with spatial templates as extended features for gait recognition in the extended space.

Intuitively, recognising humans by gait depends on how the silhouette of individual subjects changes, either spatially or temporally. According to this hypothesis, here we use an extended feature set which incorporates spatial and temporal information from spatial and temporal templates. Before training and recognition, each gait sequence is converted into two template sequences, spatial templates and temporal templates at the preprocessing stage. This is the first stage of dimensionality reduction. Each template has been aligned to a fixed-size window for further processing.

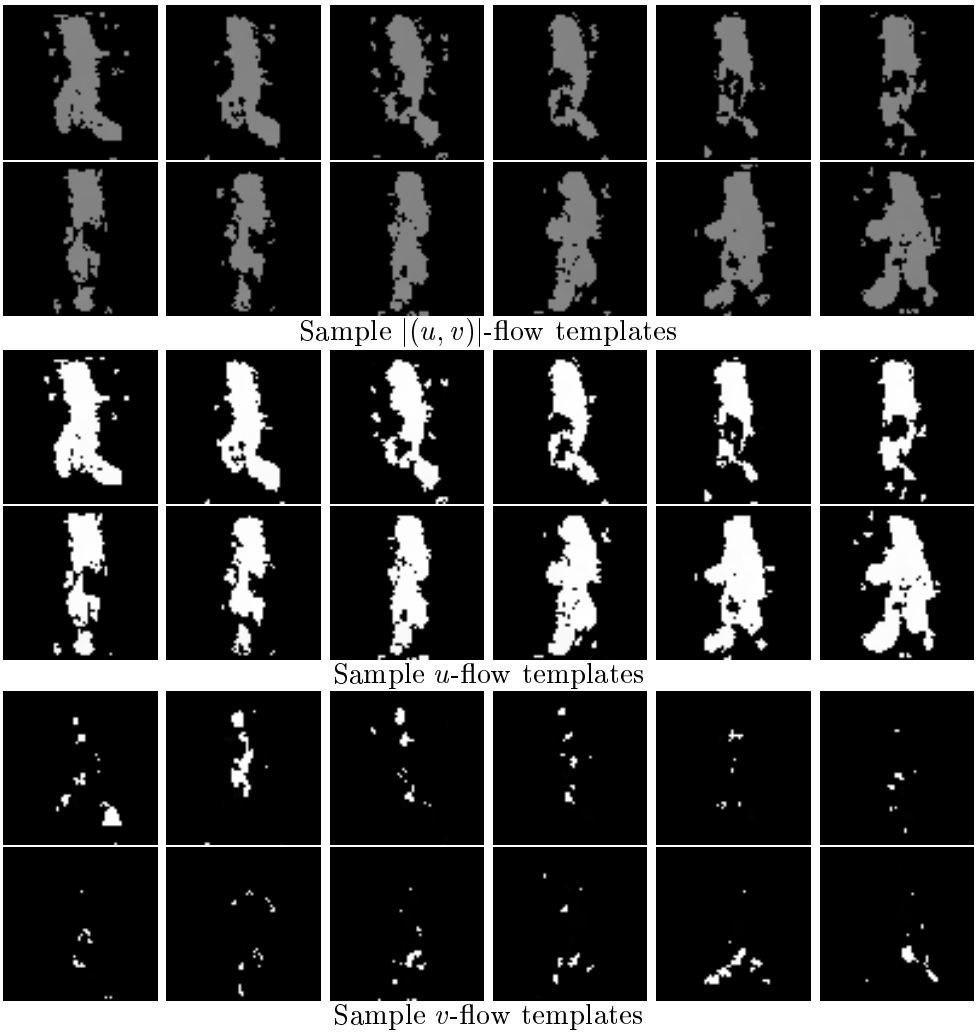


Figure 6.2: Sample temporal templates from a gait sequence using a 20×20 patch

6.4 Transformation, Training and Integration

This is the second stage of dimensionality reduction. In order to reduce data dimensionality and to optimise the class separability of different templates simultaneously, we use EST combined with CST for feature extraction. After the extraction of spatial and temporal templates from gait sequences, all feature templates from training sequences are used for training by EST and CST. Given c classes for training, each class represents a template sequence of a single subject with spatial or temporal templates. Each subject has a sequence with N_i, $i = 1, \cdots, c$ templates and $N_T = N_1 + N_2 + \cdots + N_c$ is the total number of templates from all training sequences. After brightness normalisation of each spatial template, the training set with spatial templates is represented by

$$[\mathbf{x}_{\mathbf{s}(1,1)}, \cdots, \mathbf{x}_{\mathbf{s}(1,N_1)}, \mathbf{x}_{\mathbf{s}(2,1)}, \cdots, \mathbf{x}_{\mathbf{s}(c,N_c)}] \tag{6.1}$$

where each sample $\mathbf{x}_{\mathbf{s}(i,j)}$ is a spatial template with n pixels. The training set with temporal templates is given by

$$[\mathbf{x}_{\mathbf{t}(1,1)}, \cdots, \mathbf{x}_{\mathbf{t}(1,N_1)}, \mathbf{x}_{\mathbf{t}(2,1)}, \cdots, \mathbf{x}_{\mathbf{t}(c,N_c)}] \tag{6.2}$$

where each sample $\mathbf{x}_{\mathbf{t}(i,j)}$ is a temporal template with n pixels. Note that brightness normalisation is not needed for temporal templates, as described in the previous chapter.

After PCA, these two training sets are projected into their individual eigenspaces by

$$\begin{aligned} \mathbf{y}_{\mathbf{s}(i,j)} &= [\mathbf{e}_{\mathbf{s}(1)}, \cdots, \mathbf{e}_{\mathbf{s}(k)}]^T \mathbf{x}_{\mathbf{s}(i,j)} \\ \mathbf{y}_{\mathbf{t}(i,j)} &= [\mathbf{e}_{\mathbf{t}(1)}, \cdots, \mathbf{e}_{\mathbf{t}(k)}]^T \mathbf{x}_{\mathbf{t}(i,j)} \end{aligned}$$

where $i = 1, \ldots, c$ and $j = 1, \ldots, N_c$ and individual canonical spaces after CA by

$$\begin{aligned} \mathbf{z}_{\mathbf{s}(i,j)} &= [\mathbf{v}_{\mathbf{s}(1)}, \cdots, \mathbf{v}_{\mathbf{s}(c-1)}]^T \mathbf{y}_{\mathbf{s}(i,j)} \\ \mathbf{z}_{\mathbf{t}(i,j)} &= [\mathbf{v}_{\mathbf{t}(1)}, \cdots, \mathbf{v}_{\mathbf{t}(c-1)}]^T \mathbf{y}_{\mathbf{t}(i,j)} \end{aligned}$$

After the training of spatial and temporal templates by PCA and CA, the projected vectors, $\mathbf{z}_{\mathbf{s}(i,j)}$ and $\mathbf{z}_{\mathbf{t}(i,j)}$, in two canonical spaces can be directly obtained by

$$\mathbf{z}_{\mathbf{s}(i,j)} = [\mathbf{v}_{\mathbf{s}(1)}, \cdots, \mathbf{v}_{\mathbf{s}(c-1)}]^T [\mathbf{e}_{\mathbf{s}(1)}, \cdots, \mathbf{e}_{\mathbf{s}(k)}]^T \mathbf{x}_{\mathbf{s}(i,j)} \tag{6.3}$$

$$\mathbf{z}_{\mathbf{t}(i,j)} = [\mathbf{v}_{\mathbf{t}(1)}, \cdots, \mathbf{v}_{\mathbf{t}(c-1)}]^T [\mathbf{e}_{\mathbf{t}(1)}, \cdots, \mathbf{e}_{\mathbf{t}(k)}]^T \mathbf{x}_{\mathbf{t}(i,j)} \tag{6.4}$$

and concatenated to the extended feature vectors, $\mathbf{z}_{\mathbf{e}(i,j)}$, by

$$\mathbf{z}_{\mathbf{e}(i,j)} = [\tilde{\mathbf{z}}_{\mathbf{s}(i,j)}, \tilde{\mathbf{z}}_{\mathbf{t}(i,j)}] \tag{6.5}$$

where $\mathbf{z}_{\mathbf{s}(i,j)}$ and $\mathbf{z}_{\mathbf{t}(i,j)}$ are linear re-scaled to $\tilde{\mathbf{z}}_{\mathbf{s}(i,j)}$ and $\tilde{\mathbf{z}}_{\mathbf{t}(i,j)}$. Before concatenation, linear re-scaling [109] has been applied to each vector to set the average of each data set to zero and to normalise the standard deviation to unity using the mean and variance of training sequences. By calculating the mean $\bar{\mathbf{z}}_{\mathbf{s}}$ and variance σ_s^2 with respect to the training set of spatial templates, the linear re-scaling to $\mathbf{z}_{\mathbf{s}(i,j)}$ can be given by

$$\bar{\mathbf{z}}_{\mathbf{s}} = \frac{1}{N_T} \sum_{i=1}^{c} \sum_{j=1}^{N_i} \mathbf{z}_{\mathbf{s}(i,j)} \tag{6.6}$$

$$\sigma_s^2 = \frac{1}{N_T - 1} \sum_{i=1}^{c} \sum_{j=1}^{N_i} (\mathbf{z}_{\mathbf{s}(i,j)} - \bar{\mathbf{z}}_{\mathbf{s}})^2 \tag{6.7}$$

$$\tilde{\mathbf{z}}_{\mathbf{s}(i,j)} = \frac{\mathbf{z}_{\mathbf{s}(i,j)} - \bar{\mathbf{z}}_{\mathbf{s}}}{\sigma_s} \tag{6.8}$$

and $\mathbf{z}_{\mathbf{t}(i,j)}$ can be processed in the same way.

The point, $\mathbf{z}_{\mathbf{e}(i,j)}$, in the extended canonical space integrates the spatial and temporal

information into a single vector with class information. The *centroid* of each training sequence in the extended canonical space is given by

$$\mathbf{C_e}(i) = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{z_e}_{(i,j)} \tag{6.9}$$

and used to calculate accumulated distances for gait recognition.

6.5 Recognition

Let a test gait sequence be $g(t)$, in which the time index $t = 1, \ldots, T$. Before recognition, spatial templates and temporal templates are extracted from this test sequence and projected into individual trained canonical spaces by Equation (6.3) and Equation (6.4), given two vector sequences after projection

$$\begin{aligned} \mathbf{h_s}(t) &= (v_{s,1}(t), \ldots, v_{s,c-1}(t)) \\ \mathbf{h_t}(t) &= (v_{t,1}(t), \ldots, v_{t,c-1}(t)), \end{aligned}$$

in which $\mathbf{h_s}(t)$ and $\mathbf{h_t}(t)$ represent feature vectors of spatial template and temporal templates. They can further be combined into an extended vector sequence $\mathbf{h_e}(t)$ which is given by

$$\mathbf{h_e}(t) = (v_{s,1}(t), v_{t,1}(t), \ldots, v_{s,c-1}(t), v_{t,c-1}(t)).$$

Before concatenation, linear re-scaling has been applied to each vector by using the mean and variance calculated in training. New centroids in the extended space constituted by training spatial and temporal templates can be obtained in the same way by

$$\mathbf{C_e}(i) = (u_{s,1}(i), u_{t,1}(i), \ldots, u_{s,c-1}(i), u_{t,c-1}(i)), \tag{6.10}$$

where $\mathbf{C_s}(i) = (u_{s,1}(i), \ldots, u_{s,c-1}(i))$ and $\mathbf{C_t}(i) = (u_{t,1}(i), \ldots, u_{t,c-1}(i))$ are training centroids of spatial and temporal templates in each canonical space.

To recognise a human walking sequence from a trained database in canonical spaces of spatial templates, temporal templates and extended features, the distance measure of *accumulated distance to each centroid* in each canonical space is used. The matched

class i to the test sequence $g(t)$ can be given by

$$G_s = \arg\min_i \sum_{t=1}^{T} \|\mathbf{h_s}(t) - \mathbf{C_s}(i)\|^2 \quad (6.11)$$

$$G_t = \arg\min_i \sum_{t=1}^{T} \|\mathbf{h_t}(t) - \mathbf{C_t}(i)\|^2 \quad (6.12)$$

$$G_e = \arg\min_i \sum_{t=1}^{T} \|\mathbf{h_e}(t) - \mathbf{C_e}(i)\|^2 \quad (6.13)$$

where $\mathbf{C_s}(i)$, $\mathbf{C_t}(i)$ and $\mathbf{C_e}(i)$ are the centroids of class $i, i = 1, \ldots, c$ in canonical spaces of spatial, temporal and extended features, respectively. G_s, G_t and G_e are matched classes with minimum accumulated distances achieved by $g(t)$ in three individual canonical spaces. Experimental results to evaluate the gait recognition in the extended canonical space are presented in the following sections.

6.6 Evaluation Results Using UCSD Data

This section presents the experimental results in gait recognition by UCSD gait data used in the previous chapter. Temporal templates (10×10 patch) in the previous chapter are used in Section 6.6.1 and Section 6.6.2. For comparison purposes, the recognition results using new temporal templates (20×20 patch) are shown in Section 6.6.3.

6.6.1 Number Selection of Training Templates

The test gait data used in this section is the same as used in the previous two chapters with 6 subjects and 7 sequences of each. One walking sequence is selected from each subject as a training sequence and the remaining 36 sequences served as test sequences. In order to equalise probability, each training sequence has been cut to the same length, here we choose 70 consecutive images which covers about 4 walking cycles as shown in Figure 6.3(a). Human gait is one kind of periodic motion [134, 40, 39], especially for walking laterally. This can be shown by the changes of first component from each projected feature vector of template sequences in the eigenspace. Figure 6.3(a) shows the time series of the first component from six training vector sequences and their corresponding frequency responses are shown in Figure 6.3(b). Here, we use the least-squares linear prediction (LSLP) method [135] to calculate the spectrum. This can avoid the inaccurate frequency estimates of short sample sinusoidal data by applying maximum entropy spectrum estimation. Figure 6.3(b) shows that gait is really a periodic motion with each training subject having a similar walking frequency in the main component value. In average, Figure 6.3(a) shows that the walking period of each training subject is 4 walking cycles with 70 frames. Therefore, the number of templates used for training should be selected for at least one walking cycle from each subject. This is the minimum

number to fulfill the requirement of statistical analysis.

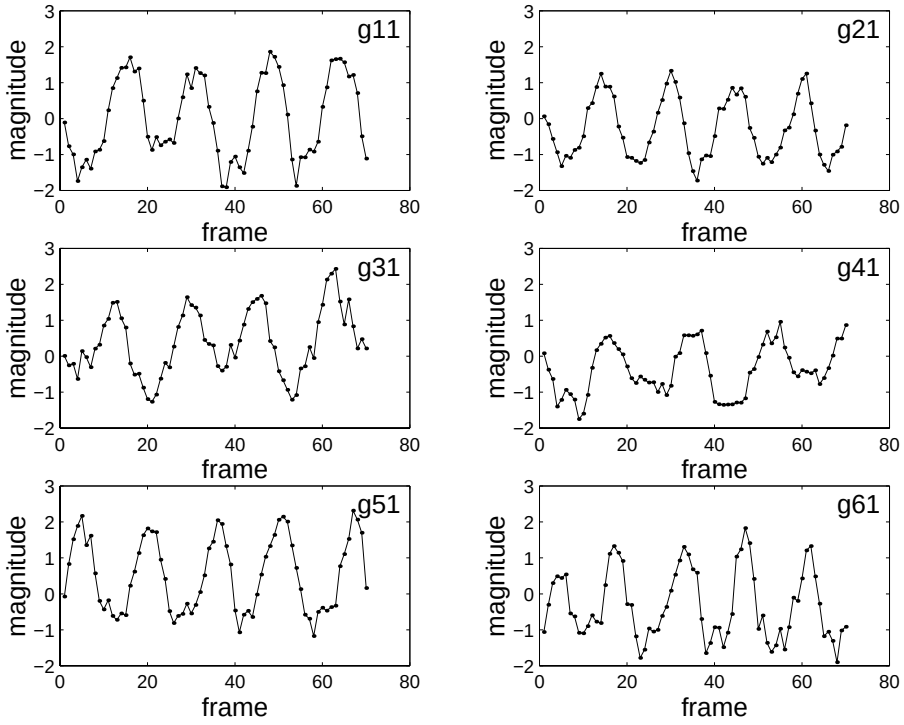
6.6.2 Performance of Extended Features

In this section, extended features which concatenate the feature vectors from spatial and temporal templates are used for recognition. Before concatenation, linear re-scaling has been applied to each canonical vector to set the average of each data set to zero and to normalise the standard deviation to unity, according to the mean and the variance of training template sequences. Figures 6.4(a) and 6.4(b) show the distributions of six training sequences in two individual canonical spaces after linear re-scaling. For visualisation purposes, we only show the first three of five dimensions.

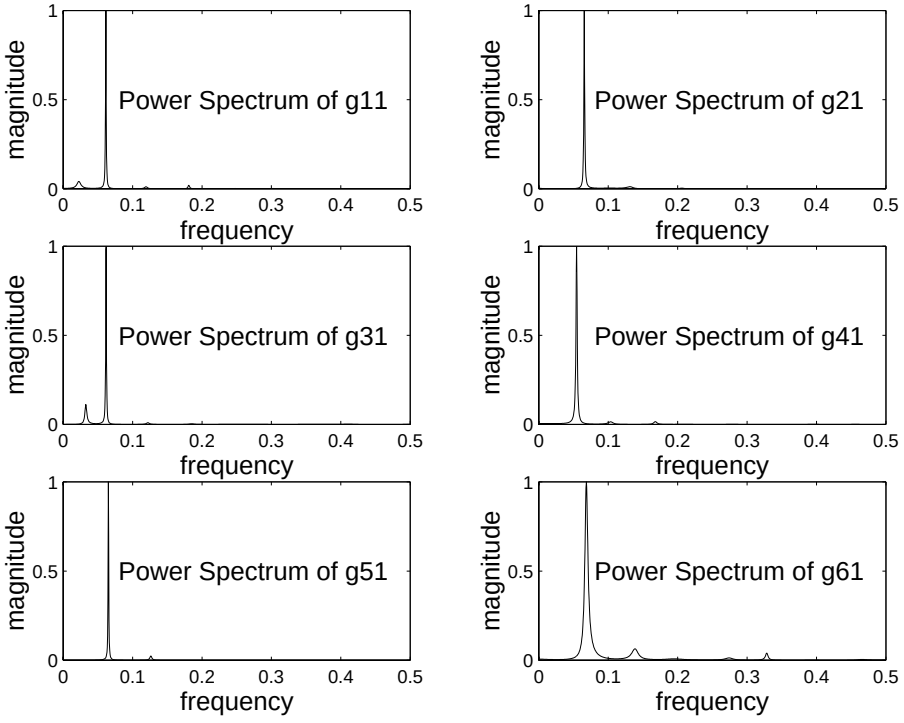
In the extended canonical space, we use Equation (6.13) as the distance measure to recognise different gaits. Equation (6.11) and Equation (6.12) are applied in individual canonical spaces of spatial and temporal templates. Recognition rates achieved by three different features are all 100% for 42 test gait sequences using 90% accumulated variance in the eigenspace. Note that the temporal templates are generated by using the 10×10 patch size during the optical flow computation. The comparison of recognition performance and distance measures of 42 sequences using three different features are shown in Figure 6.5. In order to have a qualitative measure to evaluate the recognition performace, here we use the difference between the accumulated distances of second to first best matched centroid. The feature with higher difference value achieves a better first match and recognition performance. Each point in Figure 6.5 shows the difference of accumulated distances between minimum and second minimum matches using 3 different features. Negative values represent misclassifications. Different peak values accrue from unequal length of gait sequences. Although the three different features achieve the same recognition rates (100%), the extended feature vectors generated by the combination of spatial and temporal templates yield the best performance among the three features (consistent with the larger value of relative accumulated distance).

From the obtained results, it is shown that the proposed approach has three advantages: (1) the recognition scheme is based on low-level features of motion, and the feature extraction of specific parts of the subject is not required; (2) EST and CST has been used to reduce data dimensionality and to optimise the class separability of different classes simultaneously, largely reducing the computation time and improving the performance of the eigenspace approach; and (3) the use of extended features greatly increases the robustness and accuracy of recognition.

However, the recognition performance is degraded when an improper patch size is selected to calculate the optical flow. This is drawn in the next section.

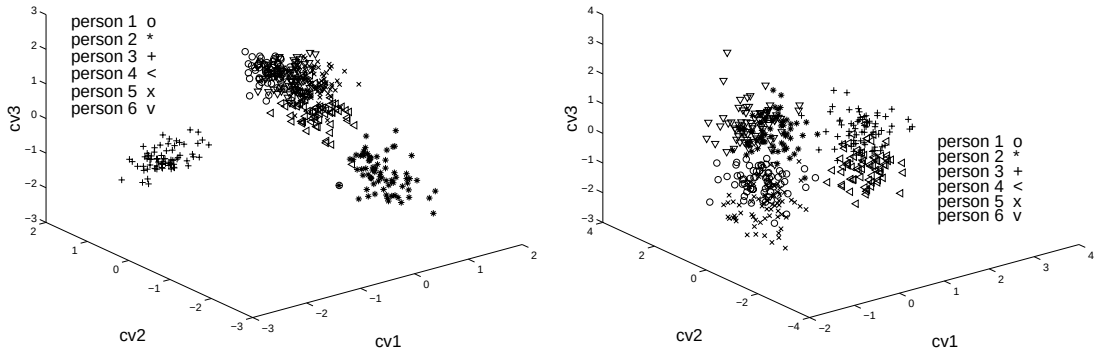


(a) Time domain



(b) Frequency domain

Figure 6.3: Periodicity analysis of one gait sequence



(a) Spatial templates (b) Temporal templates

Figure 6.4: Distributions in two canonical spaces

6.6.3 Results Using New Temporal Templates

In this section, for comparison purposes, new temporal templates (20×20 patch size) of u-flow templates, v-flow templates and $|(u, v)|$-flow templates are used individually for gait recognition. The recognition results in canonical space using different training samples (walking cycles) and eigenvalues are evaluated. Four tests which used 18, 36, 54 and 72 templates corresponding to 1, 2, 3 and 4 walking cycles from each training sequence for training are conducted. Furthermore, in each test, we choose 7 different accumulated variances in the eigenspace ranging from 65% to 95% achieved by different numbers of eigenvalues (associated to eigenvectors). The distribution of training temporal templates in canonical space using 72 templates from each of 6 subjects is shown in Figure 6.6. The comparisons of recognition performance using $|(u, v)|$-flow templates, u-flow and v-flow are shown in Table 6.1, Table 6.3 and Table 6.5, respectively. Their corresponding figures are shown in Figure 6.7(a), Figure 6.8(a) and Figure 6.9(a). In order to compare with the performance achieved by the temporal templates (10×10 patch size) used in the previous chapter, the recognition performance using $|(u, v)|$-flow templates, u-flow and v-flow are shown in Table 6.2, Table 6.4 and Table 6.6, respectively. The corresponding figures are shown in Figure 6.7(b), Figure 6.8(b) and Figure 6.9(b). The results reveal that the recognition performance of temporal templates using the 20×20 patch is worse than using the 10×10 patch. The noise introduced by using a larger patch size in computing the optical flow affects the statistical analysis and furthermore, degrades the recognition rates.

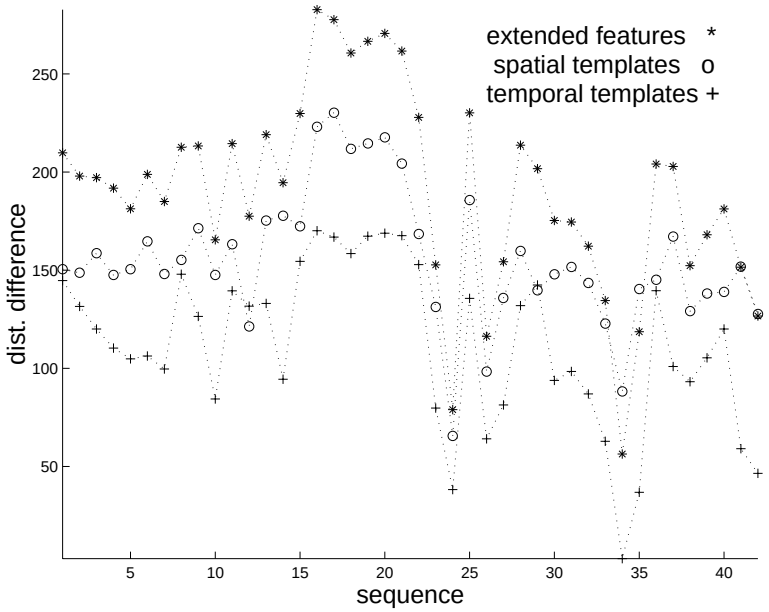


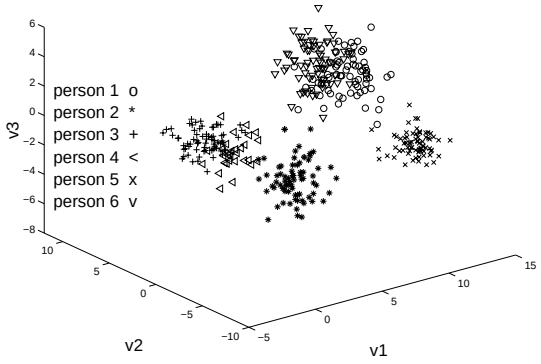
Figure 6.5: Distance measures of 42 sequences

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	33.3%	50.0%	50.0%	59.5%	61.9%	54.8%	50.0%
2 cycles	42.9%	76.2%	76.2%	76.2%	78.6%	76.2%	78.6%
3 cycles	66.7%	78.6%	78.6%	78.6%	78.6%	78.6%	78.6%
4 cycles	66.7%	73.8%	78.6%	78.6%	78.6%	78.6%	78.6%

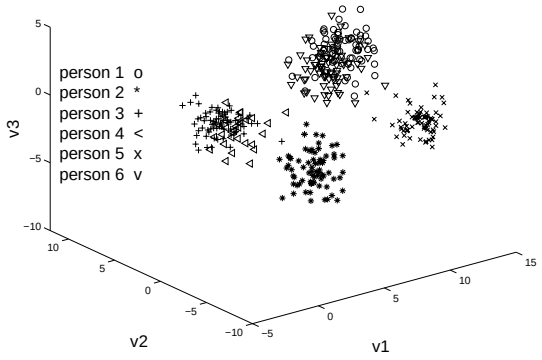
Table 6.1: Recognition rates of different training samples and eigenvalues ($|(u, v)|$-flow templates are from UCSD data using a 20×20 patch)

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	21.4%	45.2%	69.0%	83.3%	85.7%	83.3%	85.7%
2 cycles	19.0%	78.6%	90.5%	95.2%	95.2%	92.9%	95.2%
3 cycles	14.3%	81.0%	90.5%	95.2%	97.6%	97.6%	97.6%
4 cycles	69.0%	90.5%	95.2%	97.6%	97.6%	100%	100%

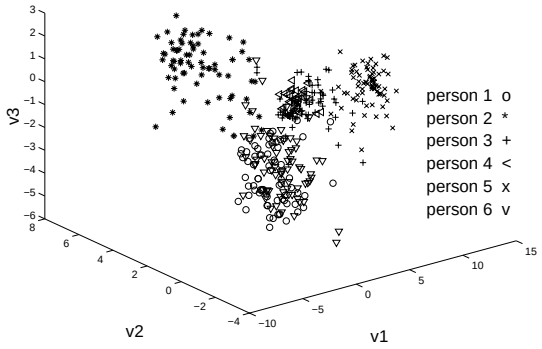
Table 6.2: Recognition rates of different training samples and eigenvalues ($|(u, v)|$-flow templates are from UCSD data using a 10×10 patch)



(a) $|(u, v)|$-flow templates



(b) u-flow templates



(c) v-flow templates

Figure 6.6: Distribution of temporal templates from UCSD data in the canonical space using a 20×20 patch

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	40.5%	45.2%	61.9%	66.7%	64.3%	64.3%	59.5%
2 cycles	40.5%	69.0%	76.2%	76.2%	76.2%	78.6%	78.6%
3 cycles	54.8%	71.4%	76.2%	78.6%	78.6%	78.6%	78.6%
4 cycles	64.3%	66.7%	76.2%	78.6%	78.6%	78.6%	78.6%

Table 6.3: Recognition rates of different training samples and eigenvalues (u-flow templates are from UCSD data using a 20×20 patch)

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	23.8%	45.2%	78.6%	88.1%	92.9%	85.7%	88.1%
2 cycles	16.7%	69.0%	85.7%	90.5%	95.2%	95.2%	97.6%
3 cycles	21.4%	76.2%	92.9%	92.9%	97.6%	97.6%	97.6%
4 cycles	16.7%	90.5%	95.2%	95.2%	97.6%	100%	100%

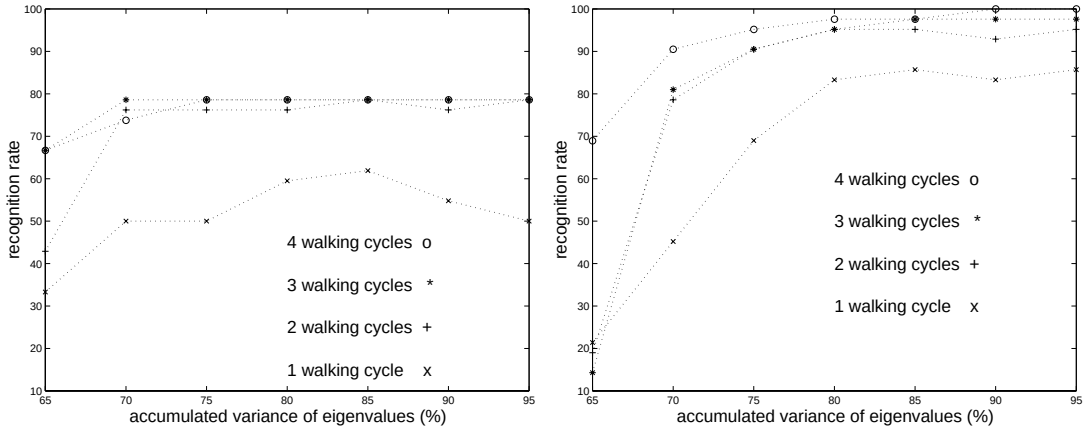
Table 6.4: Recognition rates of different training samples and eigenvalues (u-flow templates are from UCSD data using a 10×10 patch)

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	16.7%	19.0%	21.4%	19.0%	19.0%	21.4%	23.8%
2 cycles	28.6%	33.3%	35.7%	35.7%	31.0%	31.0%	28.6%
3 cycles	52.4%	52.4%	57.1%	57.1%	59.5%	54.8%	54.8%
4 cycles	64.3%	69.0%	66.7%	64.3%	64.3%	66.7%	66.7%

Table 6.5: Recognition rates of different training samples and eigenvalues (v-flow templates are from UCSD data using a 20×20 patch)

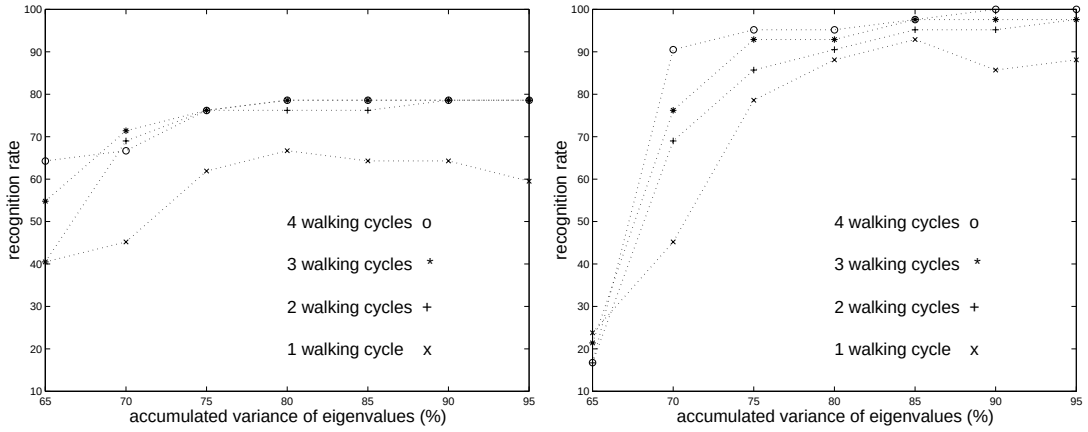
Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	42.9%	38.1%	35.7%	33.3%	26.2%	23.8%	31.0%
2 cycles	85.7%	88.1%	90.5%	85.7%	83.3%	83.3%	78.6%
3 cycles	88.1%	90.5%	90.5%	90.5%	90.5%	90.5%	88.1%
4 cycles	90.5%	90.5%	90.5%	90.5%	88.1%	88.1%	88.1%

Table 6.6: Recognition rates of different training samples and eigenvalues (v-flow templates are from UCSD data using a 10×10 patch)



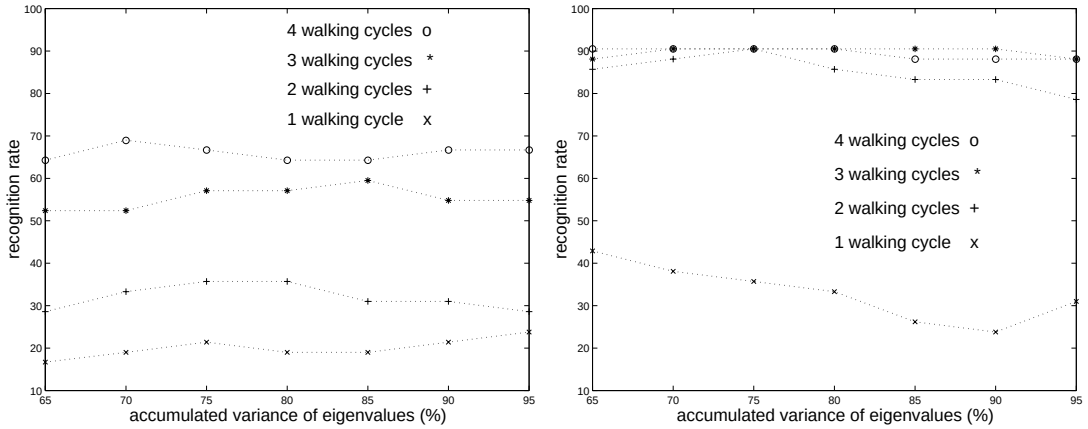
(a) 20×20 patch (b) 10×10 patch

Figure 6.7: Recognition rates of different training samples and eigenvalues using $|(u, v)|$-flow templates from UCSD data



(a) 20×20 patch (b) 10×10 patch

Figure 6.8: Recognition rates of different training samples and eigenvalues using u-flow templates from UCSD data



(a) 20×20 patch (b) 10×10 patch

Figure 6.9: Recognition rates of different training samples and eigenvalues using v-flow templates from UCSD data

6.7 Evaluation Results Using SOTON Data

In this section, we use a new set of gait data - SOTON data, which was taken at the University of Southampton, UK, in June 1998. The SOTON data has 6 subjects and 4 sequences each. For these 4 sequences from each subject, two of them are sequences with striped trousers. This can be used to test the influences in recognition performance using temporal templates achieved by different clothing. The differences between UCSD and SOTON data are:

- Environment: UCSD data was taken outdoors without lighting control and with complex background, but SOTON data was taken indoors with lighting control and a plain background.
- Frame rate: UCSD data are 30 frame/sec, SOTON data are 25 frame/sec.
- Subject clothes: Each subject in UCSD data wore the same clothes during video taping, but each subject in SOTON data wore different clothes in each set (with or without striped trousers).

Those three factors can be used to test the robustness of temporal templates which are computed from the optical flow method. Two sample sequences of one subject from the original images are shown in Figure 6.10 and Figure 6.11, one with and one without striped trousers. Sample spatial templates without and with striped trousers are shown in Figures 6.12. Using the 10×10 patch size, sample temporal templates without and with striped trousers are shown in Figures 6.13 and Figures 6.14. Using the 20×20 patch size, sample temporal templates without and with striped trousers are shown in Figures 6.15 and Figures 6.16.

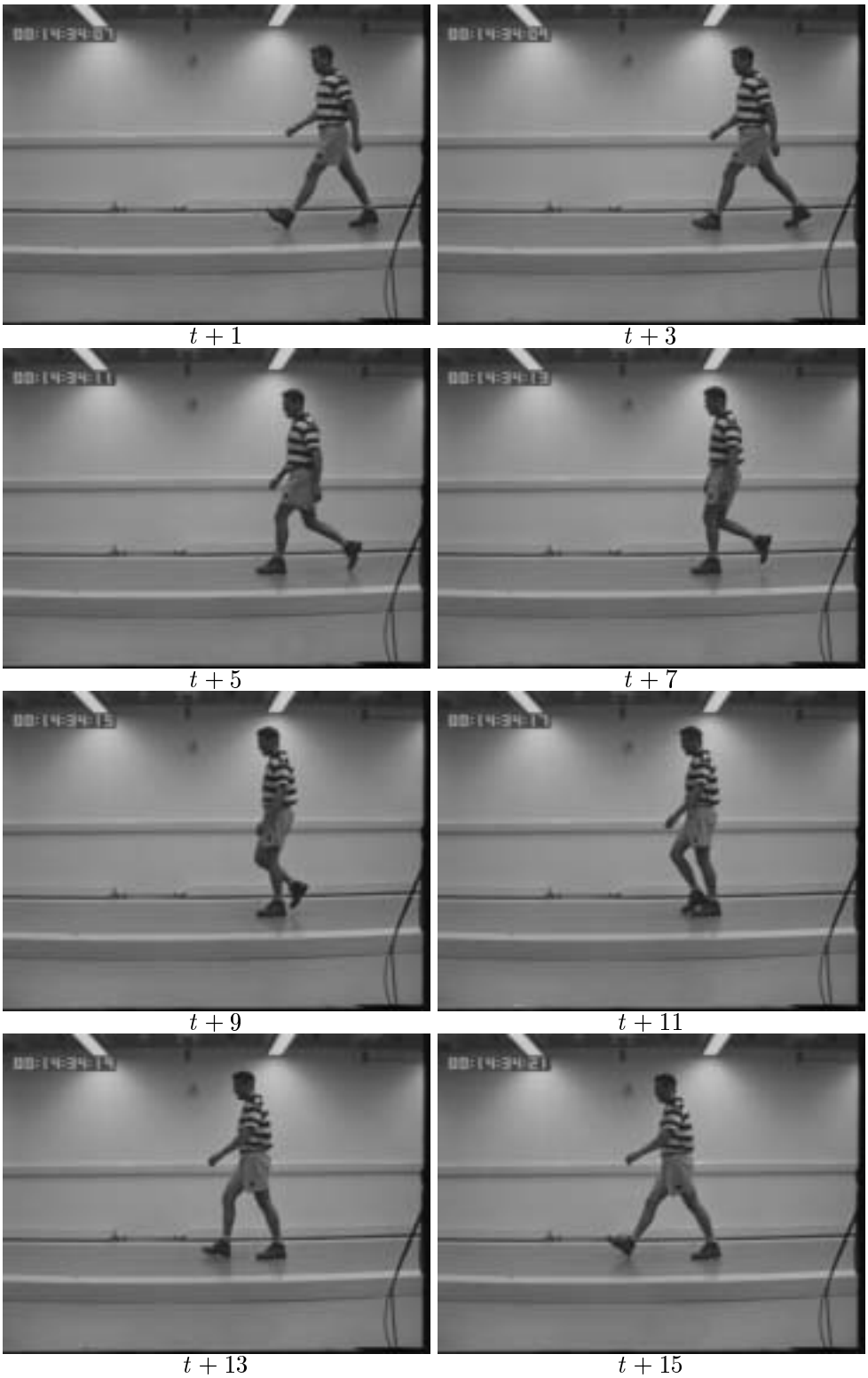


Figure 6.10: One sample sequence of original images from one subject of SOTON data without striped trousers

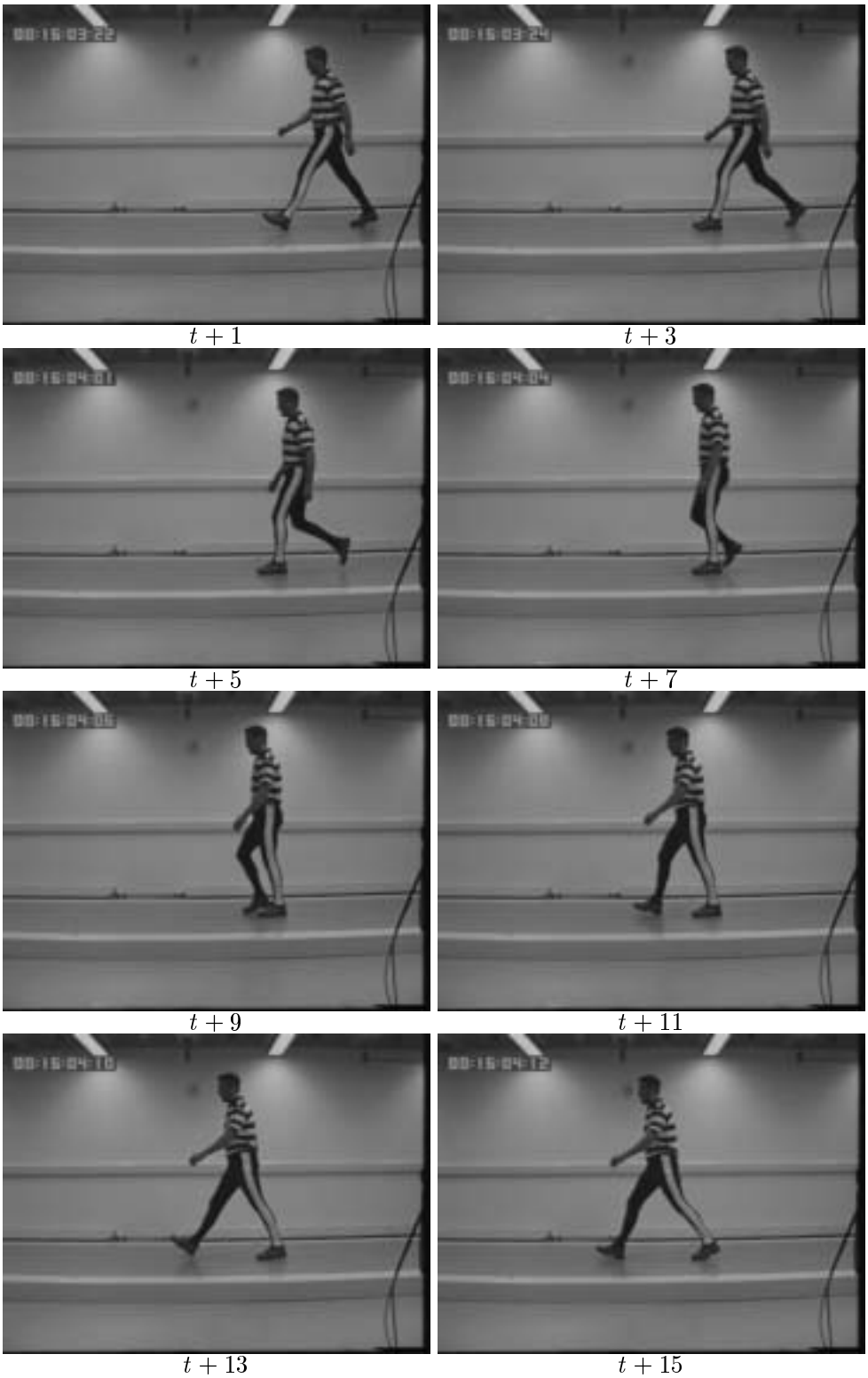


Figure 6.11: One sample sequence of original images from one subject of SOTON data with striped trousers

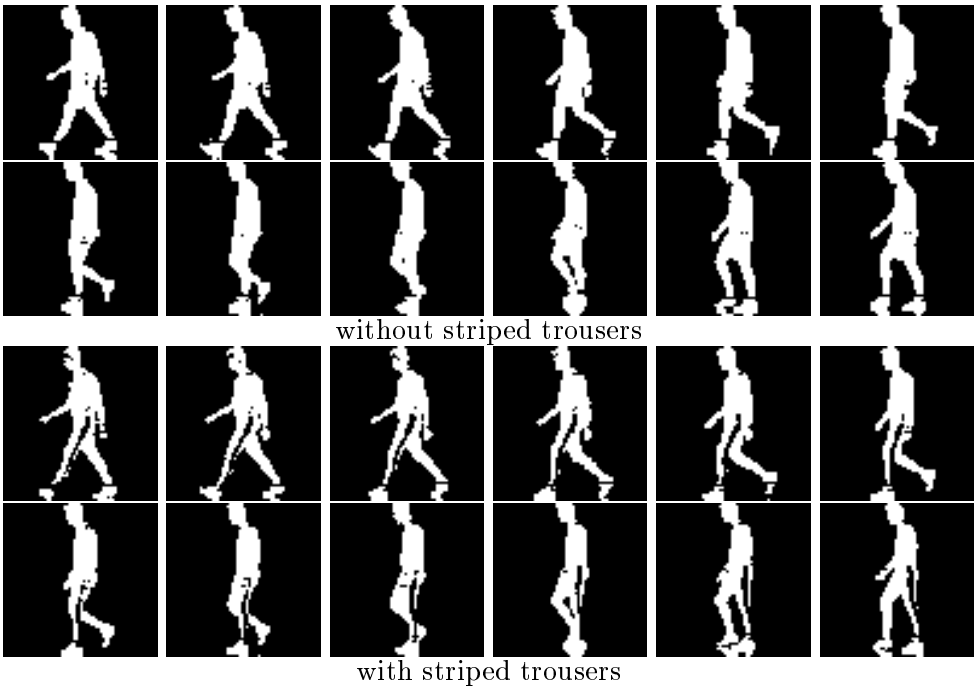


Figure 6.12: Sample spatial templates from a gait sequence of SOTON data

In this section, spatial templates, u-flow templates, v-flow templates and $|(u, v)|$-flow templates are used individually for gait recognition. The gait sequences of subjects with and without striped trousers are not isolated from each subject. Thus, each subject has 4 sequences including two situations of different trousers. Total number are 24 gait sequences for 6 subjects. Since the shortest length in all gait sequences is 40 frames covering about 2 walking cycles, 2 walking cycles are selected as the length of each training sequence. Therefore, for each type of feature templates, two tests which used 19 and 38 templates (corresponding to 1 and 2 walking cycles) from each training sequence are conducted. The training sequences are taken from 6 subjects with one each at random. The recognition effects in canonical space using different training samples (walking cycles) and eigenvalues are evaluated. Furthermore, in each test, we choose 7 different accumulated variances in the eigenspace ranging from 65% to 95% achieved by different numbers of eigenvalues (associated to eigenvectors).

After PCA and CA, the distribution of training spatial templates in canonical space using 38 templates from each of 6 subjects is shown in Figure 6.17. The distribution of training $|(u, v)|$-flow templates in canonical space using two patch sizes of 10×10 and 20×20 is shown in Figure 6.18. The distribution of training u-flow templates in canonical space using two patch sizes of 10×10 and 20×20 is shown in Figure 6.19. The distribution of training v-flow templates in canonical space using two patch sizes of 10×10 and 20×20 is shown in Figure 6.20.

Recognition rates achieved by different training sets of spatial templates are presented in Table 6.7 and Figure 6.21. Using the 20×20 patch size, recognition rates achieved by using $|(u, v)|$-flow templates, u-flow and v-flow are shown in Table 6.8, Table 6.10

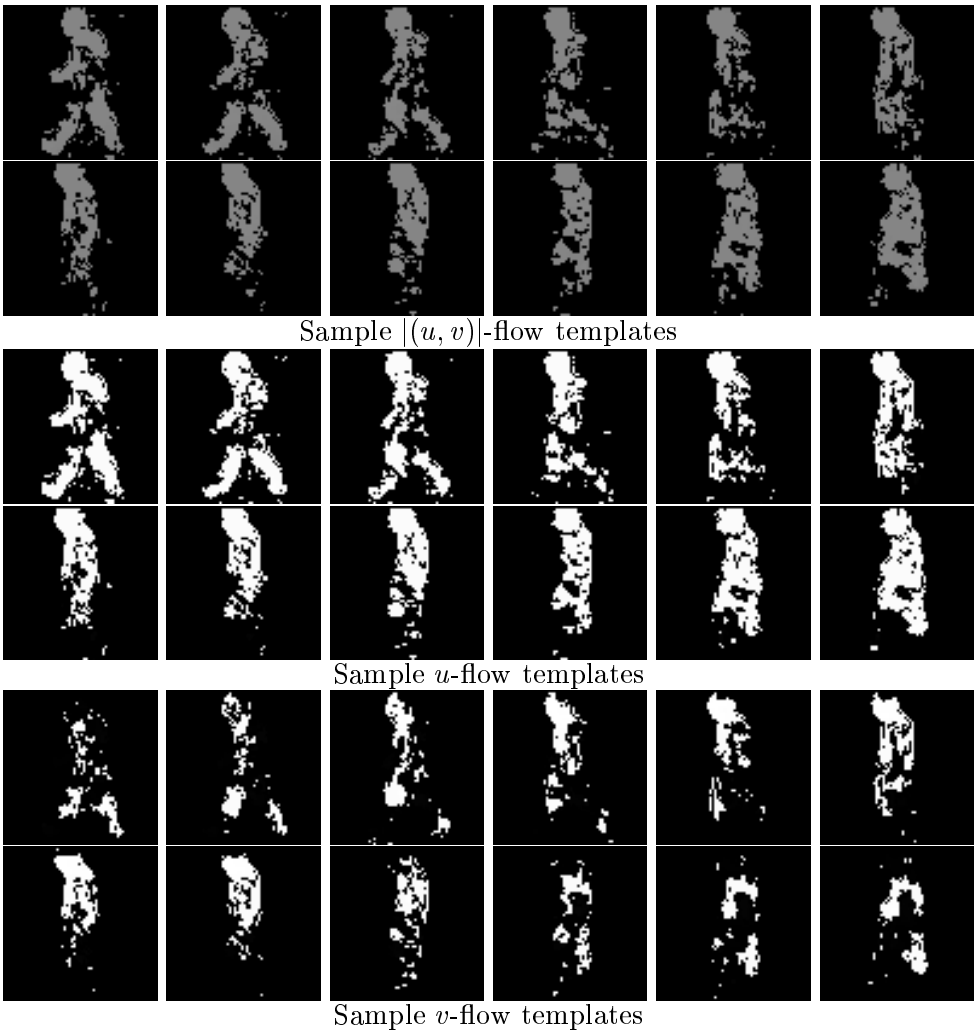


Figure 6.13: Sample temporal templates from a SOTON subject without striped trousers using a 10×10 patch

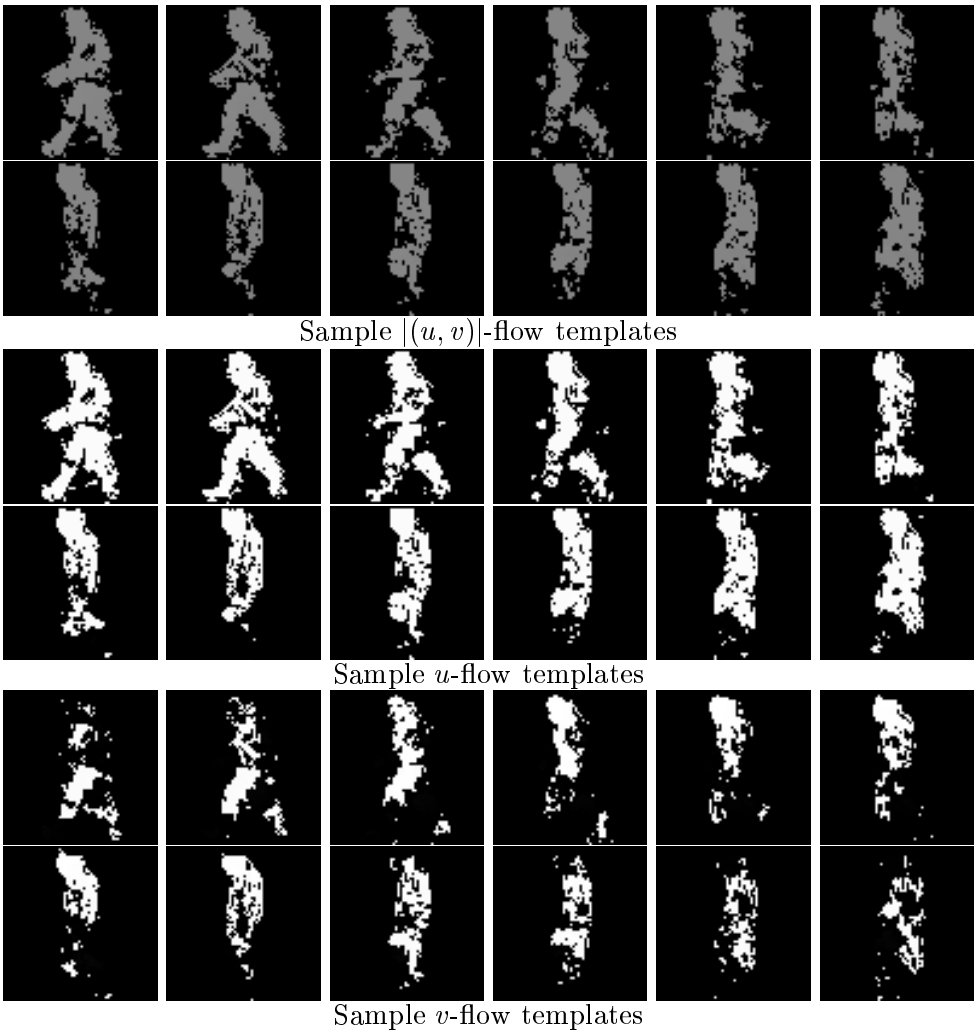


Figure 6.14: Sample temporal templates from a SOTON subject with striped trousers using a 10×10 patch

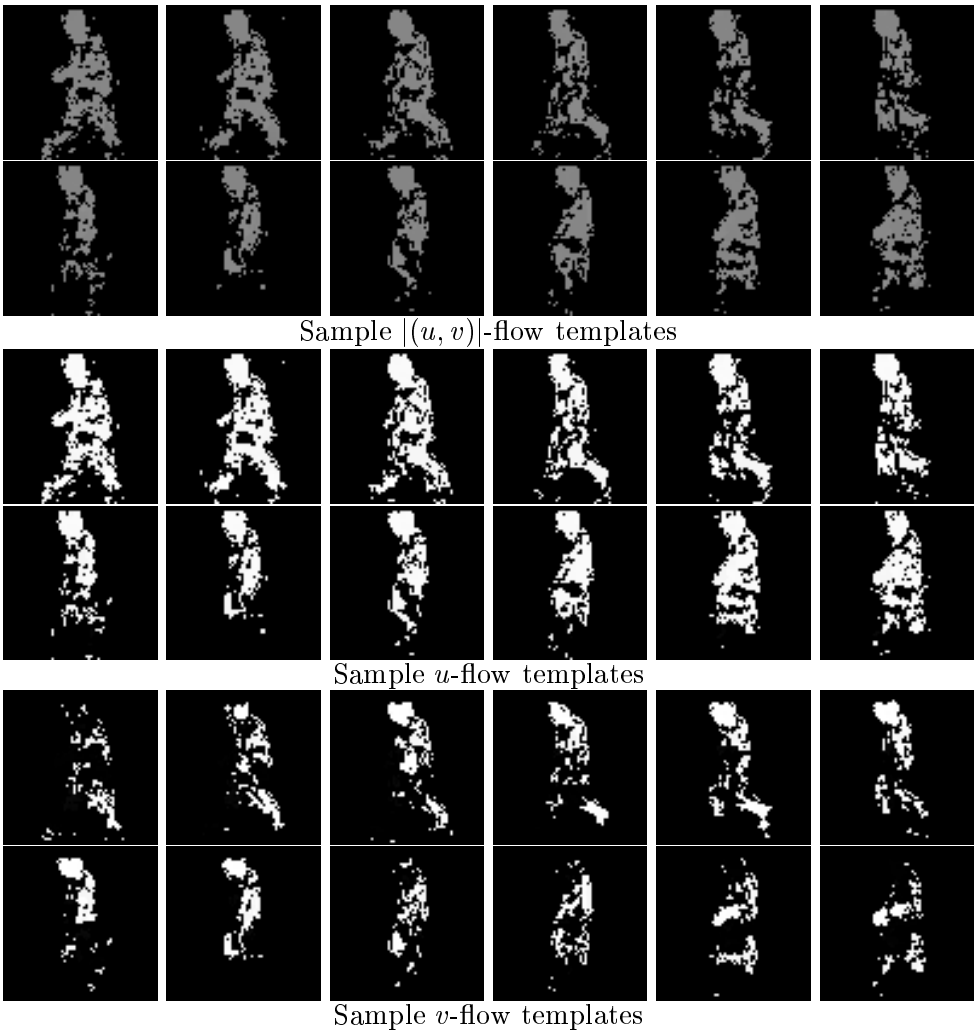


Figure 6.15: Sample temporal templates from a SOTON subject without striped trousers using a 20×20 patch

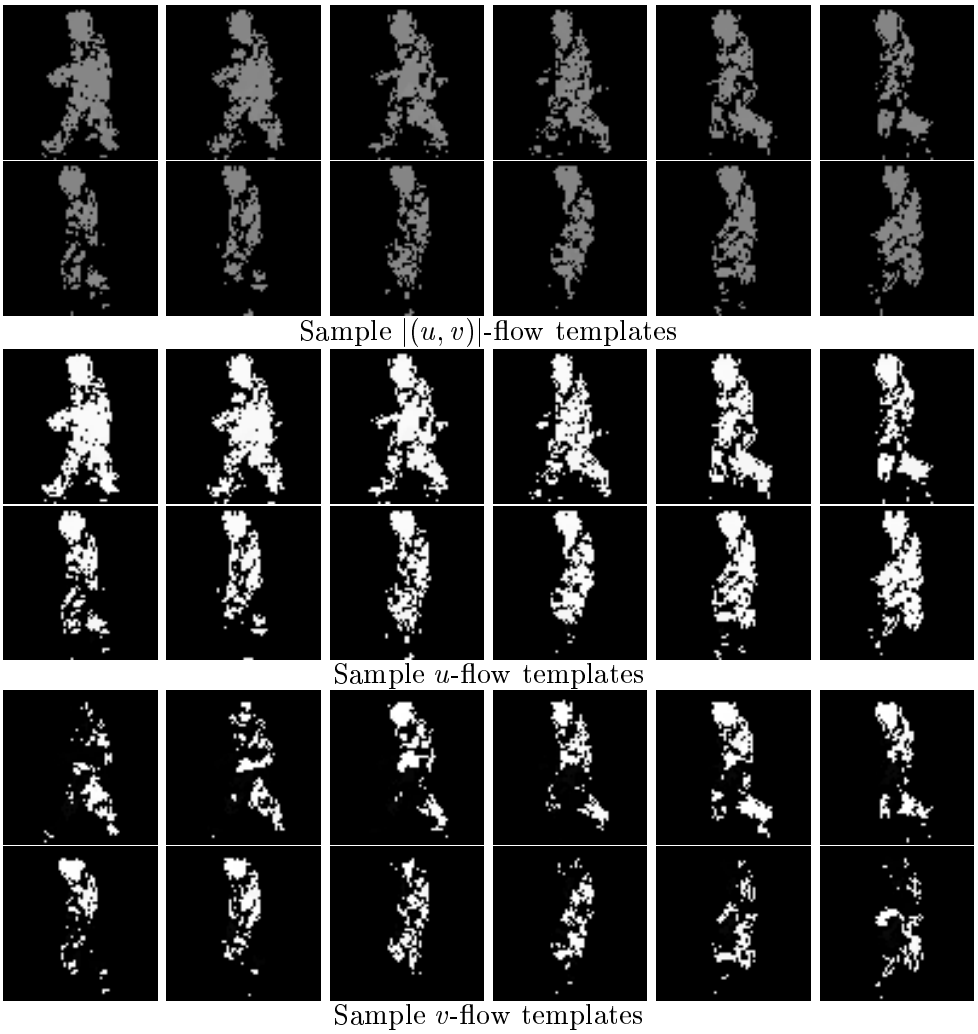


Figure 6.16: Sample temporal templates from a SOTON subject with striped trousers using a 20×20 patch

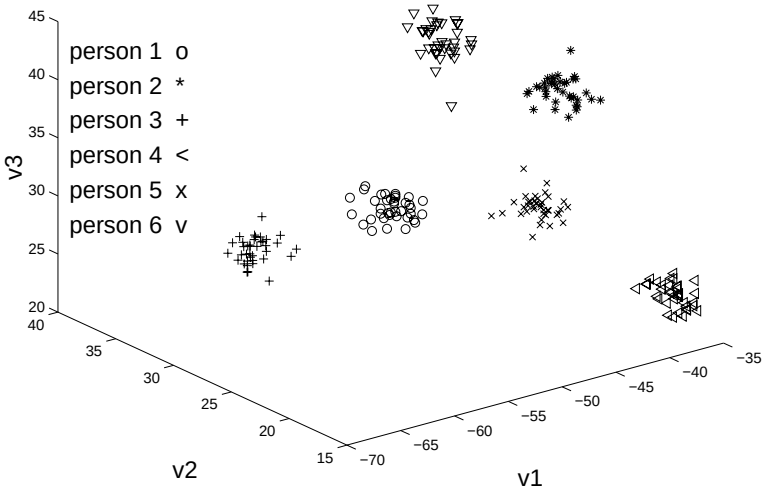
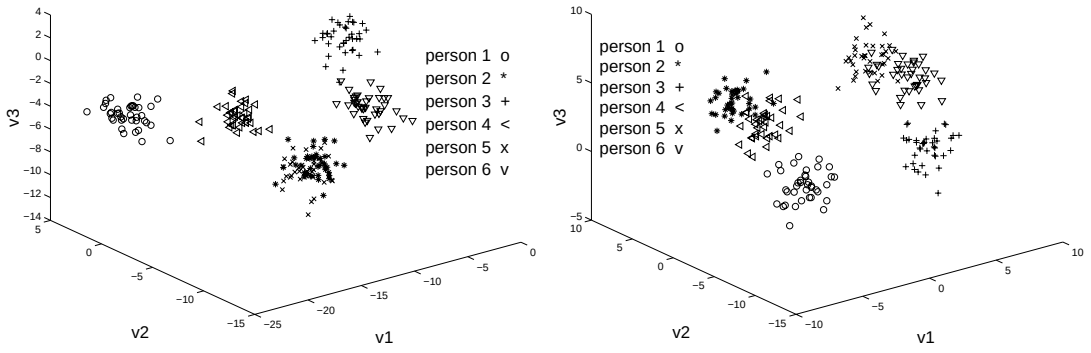
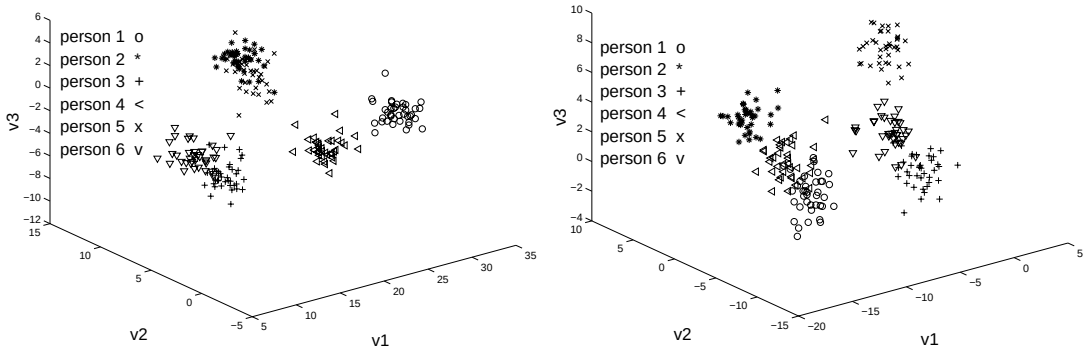


Figure 6.17: Distribution of training spatial templates from SOTON data in the canonical space



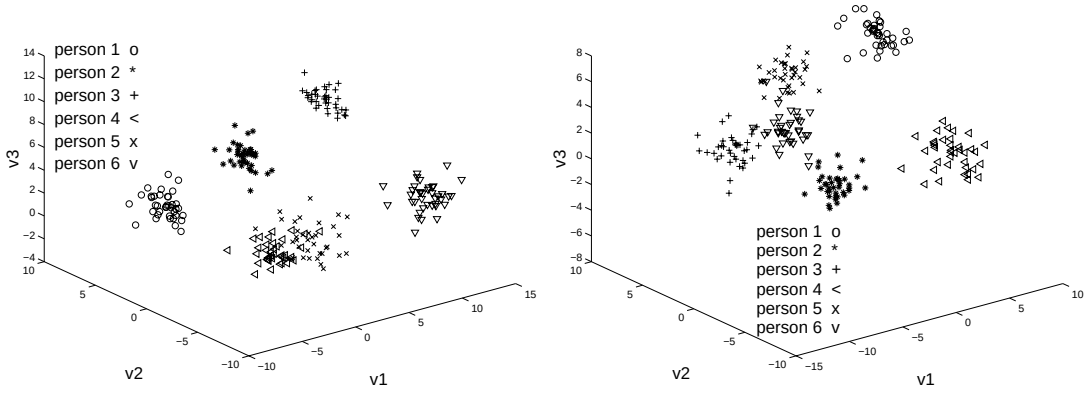
(a) 20×20 patch (b) 10×10 patch

Figure 6.18: Distribution of training $|(u, v)|$-flow templates from SOTON data in the canonical space



(a) 20×20 patch (b) 10×10 patch

Figure 6.19: Distribution of training u-flow templates from SOTON data in the canonical space



(a) 20×20 patch (b) 10×10 patch

Figure 6.20: Distribution of training v-flow templates from SOTON data in the canonical space

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	37.5%	41.7%	45.8%	70.8%	87.5%	95.8%	95.8%
2 cycles	41.7%	50.0%	54.2%	66.7%	83.3%	87.5%	95.8%

Table 6.7: Recognition rates of different training samples and eigenvalues (spatial templates are from SOTON data)

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	79.2%	87.5%	87.5%	87.5%	91.7%	91.7%	70.8%
2 cycles	91.7%	95.8%	91.7%	91.7%	91.7%	95.8%	91.7%

Table 6.8: Recognition rates of different training samples and eigenvalues ($|(u, v)|$-flow templates are from SOTON data using a 20×20 patch)

and Table 6.12, respectively. Using the 10×10 patch size, recognition rates achieved by using $|(u, v)|$-flow templates, u-flow and v-flow are shown in Table 6.9, Table 6.11 and Table 6.13, respectively. Their corresponding figures are shown in Figure 6.22, Figure 6.23 and Figure 6.24.

Again, the results show that the recognition rates achieved by temporal templates using the 20×20 patch are generally less than using the 10×10 patch. The performance achieved by using training samples of 2 walking cycles is better than 1 walking cycle. It appears that recognition performance is improved by increasing the number of training samples. Note that the recognition rates achieved by using v-flow templates (10×10 patch size) in Table 6.13 and Figure 6.24(b) are not consistent with the previous results. By using 2 walking cycles and different accumulated variances, v-flow templates can all achieve 100% recognition rates. This needs to be investigated in the future work. For spatial templates, the best recognition rate achieved is 95.8% using 95% accumulated variance in the eigenspace. This shows that the performance is not affected by the sequences with the same subject wearing different clothes. The noise flows introduced by using a larger patch size to compute optical flow do affect the statistical analysis and

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	12.5%	66.7%	70.8%	75.0%	83.3%	70.8%	79.2%
2 cycles	50.0%	75.0%	91.7%	100%	100%	91.7%	100%

Table 6.9: Recognition rates of different training samples and eigenvalues ($|(u, v)|$-flow templates are from SOTON data using a 10×10 patch)

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	87.5%	83.3%	83.3%	91.7%	87.5%	83.3%	75.0%
2 cycles	87.5%	95.8%	95.8%	95.8%	95.8%	95.8%	95.8%

Table 6.10: Recognition rates of different training samples and eigenvalues (u-flow templates are from SOTON data using a 20×20 patch)

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	12.5%	54.2%	79.2%	83.3%	87.5%	91.7%	83.3%
2 cycles	12.5%	83.3%	100%	100%	100%	100%	100%

Table 6.11: Recognition rates of different training samples and eigenvalues (u-flow templates are from SOTON data using a 10×10 patch)

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	79.2%	75.0%	79.2%	79.2%	75.0%	62.5%	62.5%
2 cycles	83.3%	83.3%	83.3%	83.3%	87.5%	87.5%	91.7%

Table 6.12: Recognition rates of different training samples and eigenvalues (v-flow templates are from SOTON data using a 20×20 patch)

Training samples	Accumulated variance of eigenvalues						
	65%	70%	75%	80%	85%	90%	95%
1 cycle	79.2%	87.5%	87.5%	79.2%	79.2%	79.2%	87.5%
2 cycles	100%	100%	100%	100%	100%	100%	100%

Table 6.13: Recognition rates of different training samples and eigenvalues (v-flow templates are from SOTON data using a 10×10 patch)

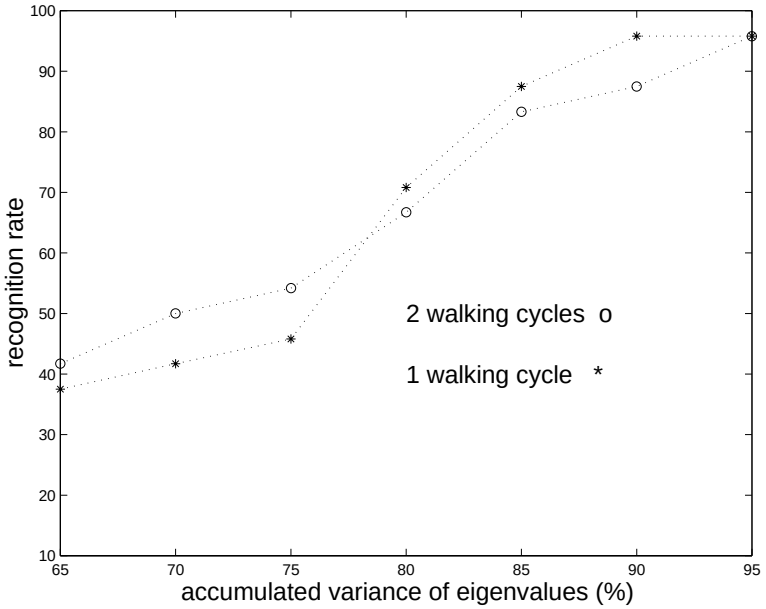
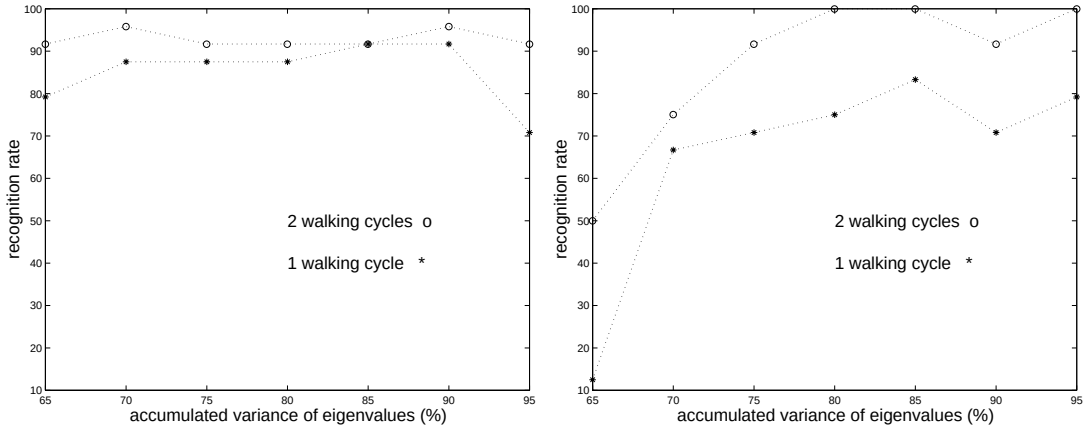
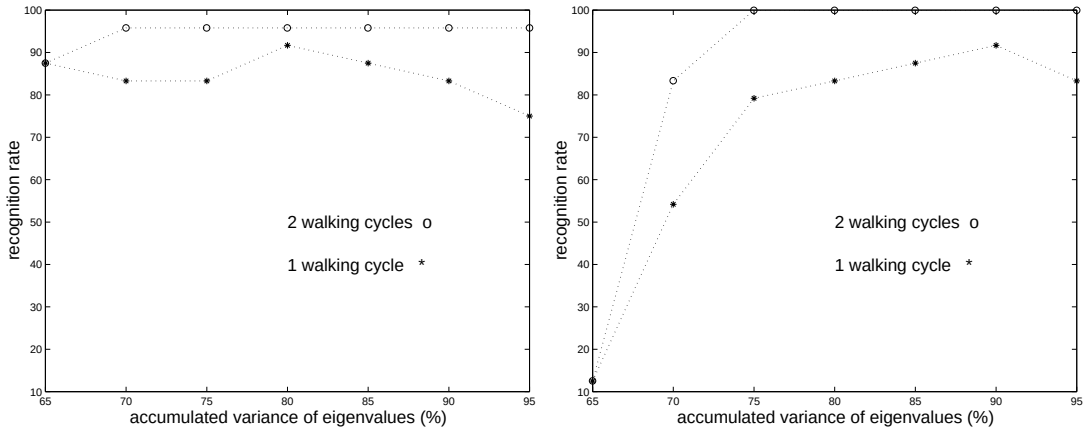


Figure 6.21: Recognition rates of different training samples and eigenvalues (spatial templates are from SOTON data)



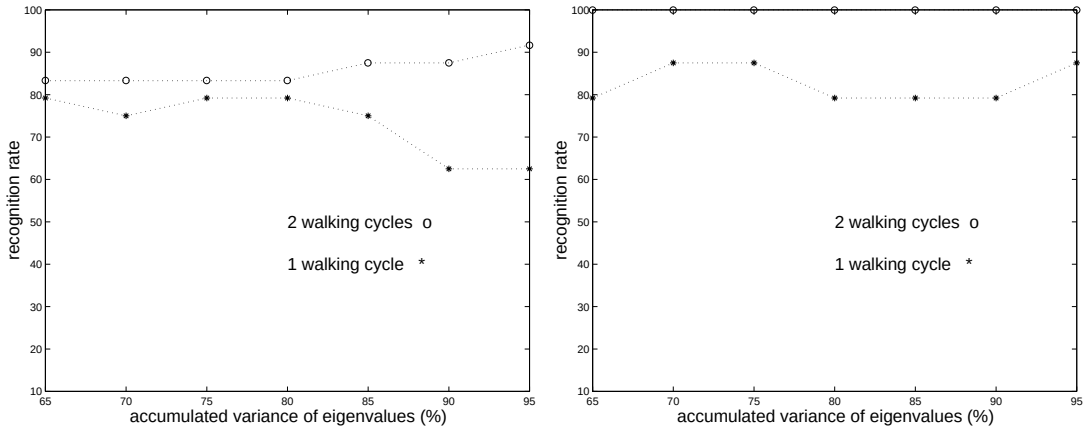
(a) 20×20 patch (b) 10×10 patch

Figure 6.22: Recognition rates of different training samples and eigenvalues ($|(u, v)|$-flow templates are from SOTON data)



(a) 20×20 patch (b) 10×10 patch

Figure 6.23: Recognition rates of different training samples and eigenvalues (u-flow templates are from SOTON data)



(a) 20×20 patch (b) 10×10 patch

Figure 6.24: Recognition rates of different training samples and eigenvalues (v-flow templates are from SOTON data)

furthermore, degrade the recognition rates. Since the average best performance for different data sets is accomplished by using 95% accumulated variance in the eigenspace, this is used in the next section for a combined data set.

6.8 Evaluations of Extended Features Using UCSD and SOTON Data

In this section, the performance of proposed extended features is evaluated in a large data set. Meanwhile, two patch sizes, 20×20 and 10×10, used to extract temporal templates are compared for their effects in gait recognition. The UCSD data is augmented with the SOTON data to have a large data set with 12 subjects and 66 sequences in total. One sequence from each of the 12 subjects is selected at random as the training sequence and 38 frames (2 walking cycles) are used as training samples. Thus, there are 12 sequences with 38 frames each in the training set. The remaining 54 sequences are used as the test set. The combination of EST and CST is used to extract the feature vectors from spatial templates and $|(u, v)|$-flow templates individually. After feature extraction, their feature vectors in the canonical space are integrated into extended vectors for recognition in the extended canonical space. The 95% accumulated variance is used in the eigenspace of spatial templates and $|(u, v)|$-flow templates to choose the number of eigenvectors for the EST matrix. Basically, two experiments are conducted. The first one is the integration of spatial templates and $|(u, v)|$-flow templates using the 20×20 patch size, while $|(u, v)|$-flow templates using the 10×10 patch size are used in the second experiment. Note that linear re-scaling is applied to the feature vectors of spatial templates and $|(u, v)|$-flow templates individually before the integration.

After the training of spatial templates by PCA and CA, the training templates are projected into the canonical space. The magnitudes of the eigenvalues after CA and the distribution of 12 training sequences in the canonical space are shown in Figure 6.27. The 11 eigenvectors associated with nonzero eigenvalues are used as the CST matrix. Figure 6.25 shows the first 12 eigenvectors (eigengaits) which span the eigenspace and Figure 6.26 presents the 11 nonzero eigenvectors which span the canonical space.

The distributions of the training $|(u, v)|$-flow templates with the 20×20 patch size and with the 10×10 patch size are shown in Figure 6.28. It appears that the class separation in Figure 6.28(b) is better than in Figure 6.28(a). The training results after PCA and CA using the $|(u, v)|$-flow templates with the 10×10 patch size are also shown here. Figure 6.29 shows the first 12 eigenvectors which span the eigenspace and Figure 6.30 presents the 11 nonzero eigenvectors which span the canonical space of $|(u, v)|$-flow templates.

After the linear re-scaling and the integration of feature vectors, the extended features are used for gait recognition. The recognition rates achieved by the 5 different features to the test set of 54 gait sequences are listed in Table 6.14. As shown in the table, the

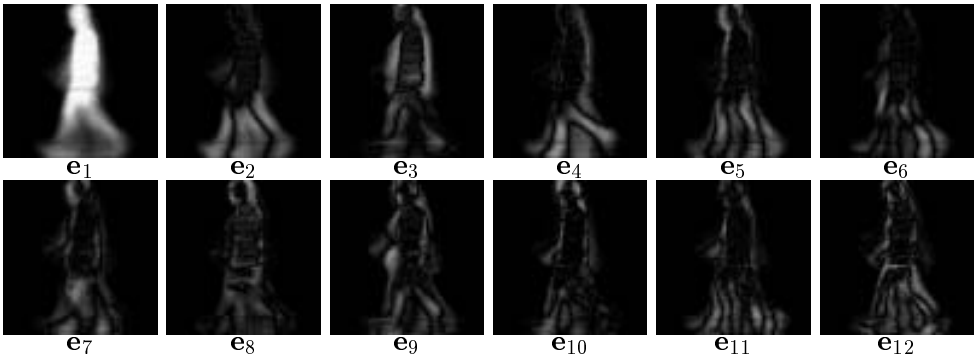


Figure 6.25: First 12 eigenvectors which span the eigenspace of spatial templates from 12 subjects

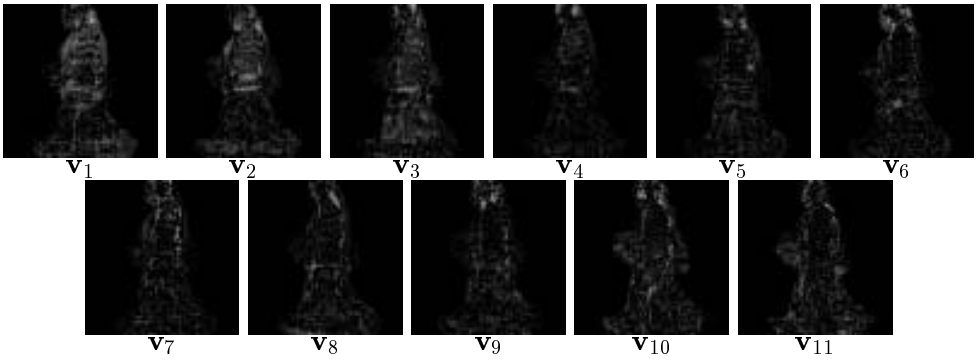
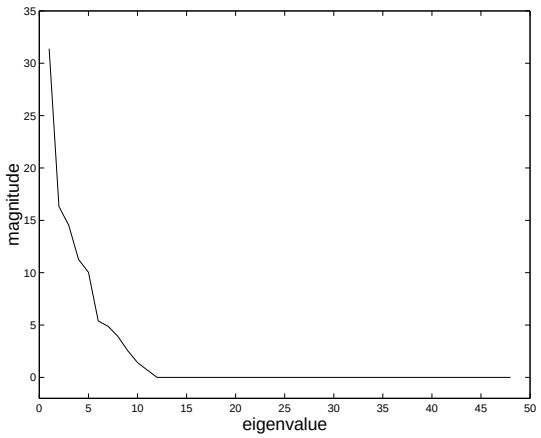
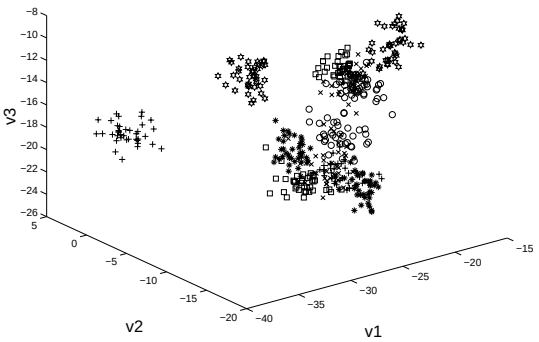


Figure 6.26: 11 nonzero eigenvectors which span the canonical space of spatial templates from 12 subjects

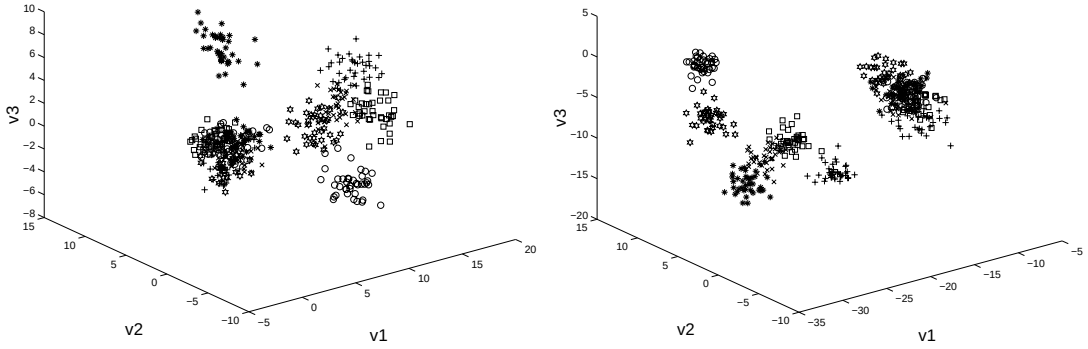


(a) Magnitudes of eigenvalues after CA



(b) Distribution of training templates

Figure 6.27: Characteristics of training spatial templates in the canonical space of the augmented data set



(a) 20×20 patch (b) 10×10 patch

Figure 6.28: The distribution of training $|(u, v)|$-flow templates in the canonical space using two patch sizes

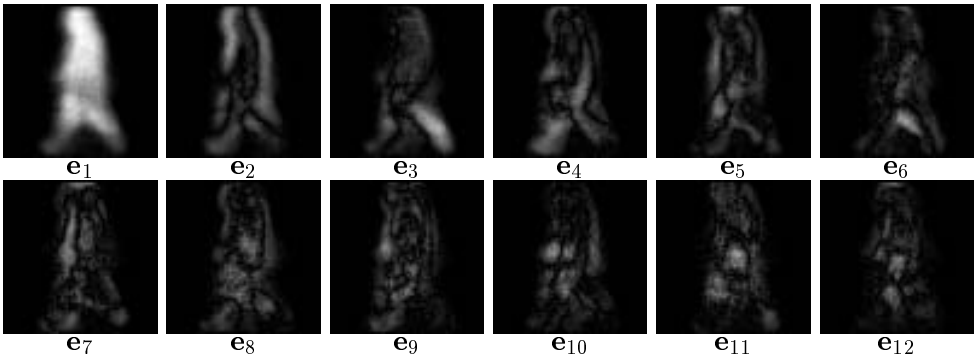


Figure 6.29: First 12 eigenvectors which span the eigenspace of $|(u, v)|$-flow templates from 12 subjects using a 10×10 patch

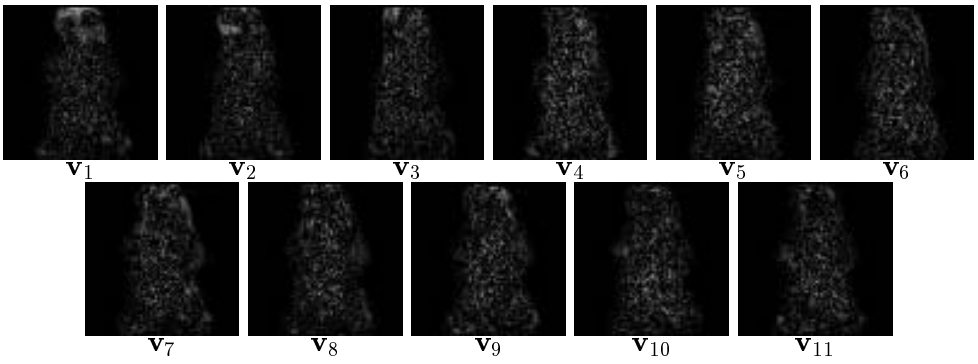


Figure 6.30: 11 nonzero eigenvectors which span the canonical space of $|(u, v)|$-flow templates from 12 subjects using a 10×10 patch

	feature used	recognition rate
(1)	spatial templates	98.2%
(2)	$\|(u,v)\|$-flow templates (10×10 patch)	100%
(3)	$\|(u,v)\|$-flow templates (20×20 patch)	75.9%
(4)	extended features of (1) + (2)	100%
(5)	extended features of (1) + (3)	96.3%

Table 6.14: Comparison of extended features using template features

$|(u,v)|$-flow templates with the 20×20 patch size have the worst performance. Again, this suggests that the noise incurred by using larger patch size degrades the recognition performance. Recognition rates of the augmented data set still show promising results by using the features from spatial templates, $|(u,v)|$-flow templates with the 10×10 patch size and 2 extended features. This shows the robustness of our proposed gait recognition systems using spatial templates, temporal templates and extended features.

In order to perform qualitative analysis of different features in Table 6.14, the relative measure of distance accumulation in Section 6.6.2 is used. The comparison of recognition performance and distance measures of 54 sequences using three different features is shown in Figure 6.31 and Figure 6.32. Figure 6.31 and Figure 6.32 show the differences of accumulated distance between minimum and second minimum matches using 3 different features. Negative values represent misclassifications. Different peak values accrue from the unequal length of the gait sequences. In Figure 6.31, the extended features achieve nearly the same performance as spatial templates and temporal templates have the worst performance incurred by noise. The noise also degrades the performance of extended features when compared with Figure 6.32. Although the same recognition rate is achieved by using $|(u,v)|$-flow templates (10×10 patch) and extended features of (1)+(2) in Table 6.14, Figure 6.32 shows that the extended features perform better than the $|(u,v)|$-flow templates from the qualitative analysis. Extended features appear to have more discriminatory ability than using spatial or temporal templates alone. This shows that the proposed extended features which incorporate the spatial and temporal information do have more information for discriminating different gaits.

6.9 Discussions

In this chapter, the extended features incorporating spatial and temporal information are proposed for gait recognition. Using the extended features and the combination of EST and CST for feature extraction, experimental results show that using the extended features performs better than the performance of using spatial or temporal templates alone. This appears that the fusion of feature vectors do have more information for classification purposes. However, the performance is deteriorated if the temporal templates

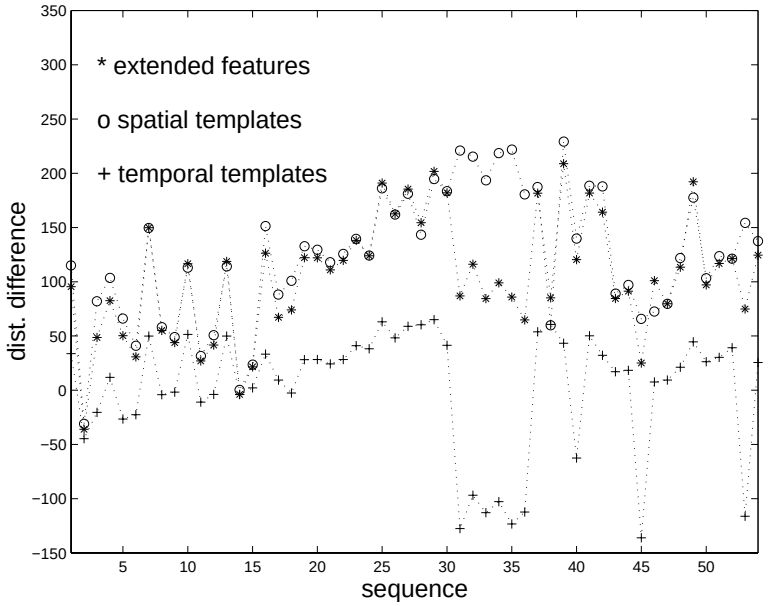


Figure 6.31: Distance measures of 54 sequences using extended features, spatial templates and $|(u, v)|$-flow templates using a 20×20 patch

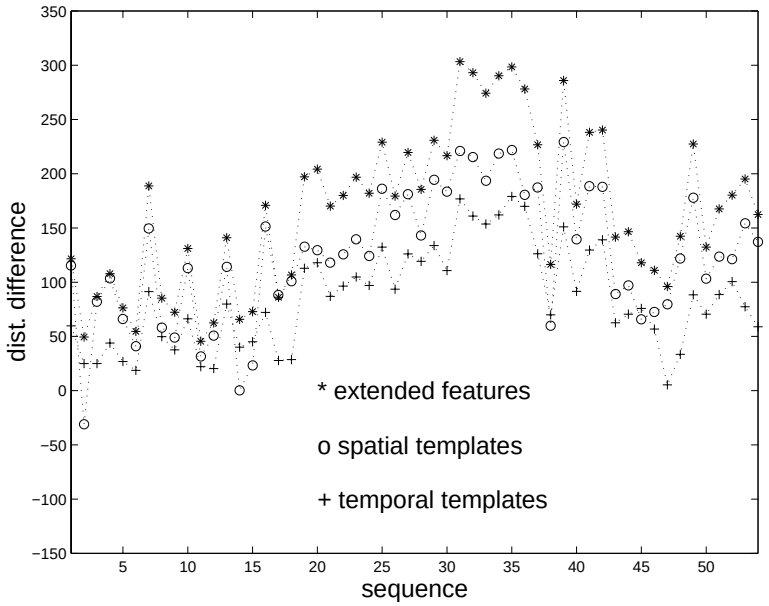


Figure 6.32: Distance measures of 54 sequences using extended features, spatial templates and $|(u, v)|$-flow templates using a 10×10 patch

with the 20×20 patch size are used. This appears that using region-based methods to calculate optical flow need to be taken care of different region sizes. From the experimental results, it appears that choosing a larger patch size incurs more noise flows and degrades the recognition performance. How to select an optimal region size for different applications is not a trivial problem. The performance comparison using the patch size of 20×20 and 10×10 to extract temporal templates is shown in this chapter by using UCSD and SOTON gait data. The size of 10×10 achieves better performance than 20×20 in gait recognition. This can be realised by the analysis in Section 6.6.3 and Section 6.7. Using temporal templates for gait recognition, the influence of recognition results using different patch sizes need to be further investigated.

Two data sets are used for recognition in this chapter, they are 42 sequences with 6 subjects from the UCSD data and 24 sequences with 6 subjects from the SOTON data. They are taken in different conditions, one outdoors and one indoors. The differences have been described in Section 6.7. Actually, temporal templates are extracted from the optical flow computation which can be affected by the different clothes worn by the subjects, frame rate change, variant lighting and scale changes. Those problems do not happen in the extraction of spatial templates which only depends on the used segmentation method. Since lighting changing in outdoor scenes is more variant than indoor scenes, extracted temporal templates are expected to be different. This can be seen by the extracted temporal templates from UCSD data and SOTON data when the patch size of 20×20 is used. In Figure 6.2, extracted temporal templates from UCSD data include more noise flows than the temporal templates in Figure 6.16 where the same patch size is used. According to the recognition results shown in Section 6.7, they appear to be more variant than the results from the UCSD data. This can be affected by two factors: two different clothes worn by one subject; and insufficient training samples which are only 2 walking cycles.

The augmented data set by UCSD and SOTON data is used for recognition in Section 6.8. According to the qualitative analysis in Figure 6.32, the performance of extended features is better than using the spatial or temporal templates alone. Although the $|(u, v)|$-flow templates with the 10×10 patch size achieve the same performance as the extended features, the relative distance measure shows that extended features are qualitatively better. Again, in supporting the analysis in Section 6.6.2, extended features do have better discriminant power than individual features even when the augmented data set is used.

6.10 Conclusions

In this chapter, a gait recognition system using the extended features which incorporate spatial and temporal information and the combination of EST and CST for feature extraction is proposed. The statistical methods of EST and CST are used for extracting

the feature vectors from spatial and temporal templates separately. These two different feature vectors in the canonical space are concatenated into single vectors for the gait recognition in the extended canonical space. This new feature incorporates spatial and temporal information into each extended vector. Each single point in the extended space carries spatial and temporal information and each trajectory in the extended space represents the movement of each gait sequence.

Two data sets (UCSD and SOTON) and their augmented set have been tested for gait recognition using the proposed system. Experimental results show that the proposed extended features achieve better recognition performance than spatial templates and temporal templates presented in the previous two chapters. This shows that by including spatial with temporal information can increase the robustness of gait recognition using statistical techniques. Different experiments using various patch sizes to extract temporal templates are performed. Depending on different situations and patch sizes to extract temporal templates, recognition performace is affected. Using the larger patch size - 20×20, the recognition rates are affected and degraded when compared to using the smaller patch size - 10×10. The effects of temporal templates extracted by different patch sizes need to be further investigated. The requirements are that the template extraction need to extract the real motion information and suppress noise flows.

Although promising results have been shown here, further evaluation on a larger database is still needed. Future work will also concentrate on looking for more precise and robust features.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

To address automatic gait recognition, this thesis has presented a statistical approach which combines EST with CST. Based on this combined approach for feature extraction, gait recognition can be achieved by using the proposed template features. The features used are: spatial templates, temporal templates and extended features. Before applying statistical analysis, each gait sequence is converted into a template sequence. Each template sequence is consecutively projected by two transformations, EST and CST, into the eigenspace and then, into the canonical space. The recognition is accomplished by selecting the minimum accumulated distance of unknown gait sequences to training centroids in the canonical space.

The EST matrix is generated by applying Principal Component Analysis (PCA) to training template sequences of different subjects in the database. Taking the advantage of PCA in dimensionality reduction, each template sequence is projected from a high-dimensional image space into a low-dimensional eigenspace. A gait sequence is represented by a trajectory in the new space. The CST matrix is produced by applying Canonical Analysis (CA) to the projected vectors of the training template sequences in the eigenspace. Each template sequence in the eigenspace is further projected into a canonical space. Since CA can optimise the class separability by maximising between-class and minimising within-class variations, training template sequences of different subjects in the database are tied to individual clusters in the canonical space. Training centroids calculated from the clusters can be stored in the database and used for recognition. This greatly reduces the size of the database by retaining the centroids, rather than the entire gait sequences.

A gait recognition system has been developed which uses spatial templates as features. The motion information of gait is actually represented by the projected trajectory in the eigenspace and the canonical space. The extraction of spatial templates is achieved by subtracting the scene image with a walking subject from the background image. In

order to increase robustness in segmentation, region growing is further applied after subtraction. The objective of gait recognition is to recognise people by the way they walk. Thus, the main concern is the changes of human shape without regard to the clothing or to differing background. Therefore, spatial templates which are binary images of human silhouettes extracted from each scene image are proposed as features. Each human silhouette has been rescaled and positioned in the center of the template. Therefore, there is no scaling and positioning problem. This eliminates the redundancy introduced by irrelevant background data, by scale changes during walking and by different clothing worn by the subject. Furthermore, using the templates with the same size is consistent with application of statistical approaches.

In comparison with the results of two other approaches independently, the proposed combined approach appears to provide improved results in gait recognition. The combined approach - EST and CST, has been also applied to face recognition for feature extraction, and the results show that the combined approach achieves better performance than the eigenspace approach in automatic face recognition. However, face recognition suffers from scaling and positioning problems. The performance of statistical approaches needs to be re-evaluated after solving those problems.

In Chapter 5, temporal templates were used as features for gait recognition. For the extraction of temporal templates, a region-based matching technique which uses two frames to calculate the optical flow is used. Each temporal template, which represents the change of two consecutive silhouettes, is extracted from this optical flow field. Temporal information is incorporated from optical-flow changes between two consecutive silhouettes into temporal templates which represent distribution of velocity magnitudes in each pixel. Temporal templates are square windows extracted from the optical flow computation of two consecutive scene images which have been rescaled to be templates with the same size. Three types of temporal templates are generated: the u-flow templates (the horizontal components of flow); v-flow templates (the vertical components of flow) and $|(u, v)|$-flow templates (the magnitudes of (u, v) as calculated from the u-flow and v-flow templates).

The $|(u, v)|$-flow templates combine the information of u-flow templates and v-flow templates and have been selected as features for gait recognition. By comparing with earlier approaches, the new system still achieves attractive performance in gait recognition. Furthermore, the comparison of recognition performance achieved by each individual feature template shows that the spatial templates, the horizontal u-flow templates and the magnitude $|(u, v)|$-flow templates have better discriminatory power than the vertical v-flow templates. However, unlike spatial templates which are invariant to different illumination conditions, temporal templates are extracted from optical flow and suffer under variation in lighting direction. Moreover, different algorithms can result in different flow values.

Chapter 6 concerned extended features which combine the spatial and temporal in-

formation. By incorporating spatial and temporal information in canonical space, gait recognition can potentially become more robust and accurate than when using any single feature alone. Since the spatial information has been included in each spatial template and the temporal motion information between two consecutive silhouettes is embedded in each temporal template, the combination is accomplished by simply concatenating the projected vectors of each spatial and temporal template in the canonical space together at each time instant. However, the scales of two canonical spaces for spatial and temporal templates differ after their individual EST and CST. Thus, linear re-scaling was applied to set the average of each data set to be zero and to normalise the standard deviation to unity before the concatenation.

Experimental results showed that using extended features can achieve a better performance than by using the spatial templates or temporal templates alone. This shows that combining spatial with temporal information can increase the robustness of gait recognition using statistical techniques. However, since the region-based matching technique is used in optical flow computation to extract the temporal templates, the selected region size affects the quality of extracted temporal templates. The recognition rates are degraded using the larger region size: 20×20 when compared to using the small region size: 10×10. Therefore, the effects of different region sizes in the extraction of temporal templates need to be further investigated. The requirements are that the real motion information is extracted whilst suppressing noise.

Although the approaches proposed in this thesis have provided impressive results in gait recognition, a larger database is needed to further evaluate their robustness. Meanwhile, their potential improvements and extensions also need to be developed. The remainder of this chapter describes future areas for research.

7.2 Future Work

Gait research is still in an exploratory phase, rather than at an established one. Accordingly, gait extraction and recognition offer a rich avenue of research opportunity. These opportunities exist not only in development and extension of basic technique, but also in application and as a potential contributor to multi-biometric systems.

For gait, techniques are required to isolate moving articulated objects. Based on the assumption of static background and single walking subject, the technique of simple subtraction used for the extraction of spatial templates needs to be improved when the scene's environment changes in lighting and motion. Better segmentation techniques should be used. The techniques could be aimed at extracting the generic shape of the human body. As such, this requires extraction of arbitrary moving articulated shapes, as required for human motion analysis and recognition. To extract temporal templates, a region-based technique is used to calculate the optical flow field between two image frames. The recognition performance degrades if an improper region size chosen, which

incurs noise. Apart from evaluating the effects accrued by different region sizes, future directions should look for more robust methods to calculate the optical flow field.

Currently, after normalisation, extended features concatenate feature vectors directly from spatial and temporal templates in the canonical space. The data fusion of different features is a general problem in pattern recognition. Evaluating the importance of each feature becomes necessary before fusion. Therefore, developing algorithms to calculate the weighting of individual features is unavoidable for further fusion. The significance will be revealed especially when more features are extracted from the gait sequences and fused. To fuse different data sets from disparate domains and maximise their information contents in the eigenspace, Boyle [92] has proposed an approach to compute optimal scaling factors of individual augmented components for balancing their contributions. In order to maximise the performance of gait recognition, the analysis for the contributions of individual gait features in the canonical space needs to be further investigated. Moreover, two systems which integrate multi-biometric traits for human identification have been proposed, one [136] uses voice and faces and another one combines faces and fingerprints [137]. The integration of gait signatures (with motion information) with other biometric features also provides potentials in human identification.

In the proposed system for gait recognition, the training process using the combined approach of EST and CST consumes most of the computation time. This becomes more and more computationally expensive when the database is getting larger after more training subjects have been added in. It is unrealistic to perform the training each time when new training subjects appear and the entire updated database is used. To solve the problems in pattern recognition using only the eigenspace representation, eigenspace update algorithms [122, 123] have been developed. Instead of training the new database constituted by old and new subjects using PCA, only the information from new subjects needs to be updated. Thus, using only the new subjects, the training is merely to modify the old EST matrix which is computationally feasible. Since there are two matrices, EST and CST matrix, involved in the training of proposed systems for gait recognition, to update their combined matrix becomes more complicated. Based on these eigenspace update algorithms, algorithms to update the combination of EST and CST matrix need to be developed.

The combination of PCA and CA is a linear projection method aimed to map the original data into a separable feature space, according to the results of traditional statistical analysis. Also, nonlinear statistical techniques for data projection and classification, such as neural networks, have been proposed in recent years [138, 139, 140, 141, 142]. After the training of network models by learning algorithms, neural networks use a nonlinear transformation to map data from an input space to a new feature space where the classification is achieved. A new architecture of neural networks, *Support Vector Machines* (SVMs) [143], select *support vectors* from a subset of training patterns to separate the boundaries between classes according to specific kernel functions. This is

suitable for pattern classification. After the dimension reduction performed by PCA, the nonlinear analysis for the data distribution of feature templates in the eigenspace by neural networks needs to be further investigated and evaluated for their performance in gait recognition.

Since the experiments conducted in this thesis have used a limited number of subjects, only sequential search is used to find the best match from the database. As the database grows larger when more subjects are included, sequential search becomes more time consuming and is less attractive for searching the best match from the database. Therefore, more efficient and quicker search algorithms need to be developed.

The objective of medical research in gait has been to classify the components of gait for the treatment of pathologically abnormal patients. In order to process gait data quickly and identify the functional deficiencies of a patient, classification methods are needed to characterise a patient's gait and direct clinical analysis to the movement abnormalities. The problem of classifying gait disorders is a problem of mapping a multivariate temporal pattern to a most likely known disorder. Based on the new approaches for identifying individuals, the analysis of gait patterns from different classes of abnormal patients to identify certain disorder can be investigated aimed to assist the clinical treatment. Also, the new approaches can be analysed for the possibility of recognising different motions performed by humans doing sports.

The gait patterns used in this thesis only consider one subject walking laterally and parallel to the image plane. To achieve gait recognition in the real world, these constraints should be relaxed. Gait patterns with subjects walking at different view angles should be extracted and analysed. The problem of feature extraction from image scenes involved in multiple walking subjects becomes complicated, especially when occlusion happens. To solve this problem, object tracking and detection methods in three-dimensional image space need to be developed.

The tracking of human walking can be achieved by tracking individual parts of the human body. The change in joint angles corresponding to gait motion can be extracted by some model-based approaches mentioned in Chapter 2. In this thesis, only template features which reveal the motion of whole human body have been considered, the change of joint angles in each body part during walking also plays an important role in human gait. Therefore, the statistical analysis of those features can also be analysed for recognition purposes.

As with all biometrics, gait research will benefit from an established database for purposes of development, preferably with a separate database for test purposes and hopefully with the stringency of the FERET test. Clearly any database will need to include variation in factors that can affect the perception of gait. These include variety in clothing, and in footwear and with subjects carrying common articles such as handbags or shopping. Also, we will require subjects walking with a wide variety of trajectories relative to the camera together with normal views as used in preliminary studies. Such a

database will allow establishment of the properties and limits of signatures derived from gait. As such, they will provide an estimate of the confidence that can be associated with the use of gait to buttress other biometric measures. As stated earlier, gait may be evident where other biometrics can be assessed with limited precision, or are obscured. Clearly, gait can be used to buttress other biometrics, but a good database can only serve to evidence the uniqueness or otherwise of automatic gait measurement. There is great scope for future research effort, both in application and development. Clearly, gait would benefit from an established database on which to assess new developments. These developments could be improvements in recognition procedure or in technique. As such, future work will establish more precisely the results that can be achieved by this new biometric.

List of Publications

Journal Papers:

1. **P.S. Huang, C.J. Harris, and M.S. Nixon** "Recognising Humans by Gait via Parametric Canonical Space", *Journal of Artificial Intelligence in Engineering, Special Issue on EIS'98* (accepted for publication).

2. **P.S. Huang, C.J. Harris, and M.S. Nixon** "Human Gait Recognition in Canonical Space Using Temporal Templates", *IEE Proceedings - Vision, Image and Signal Processing* (accepted for publication).

3. **P.S. Huang, C.J. Harris, and M.S. Nixon** "Automatic Gait Recognition via Statistical Approaches for Extended Template Features", *IEEE Transactions on Pattern Analysis and Machine Intelligence* (submitted).

Book Chapters:

1. **M.S. Nixon, J.N. Carter, D. Cunado, P.S. Huang and S.V. Stevenage** "Automatic gait recognition", in *BIOMETRICS - Personal Identification in Networked Society*, A. Jain, R. Bolle and S. Pankanti, Eds., chapter 11, pp. 231-249. Kluwer Academic Publishers, January 1999.

Conference Papers:

1. **P.S. Huang, C.J. Harris, and M.S. Nixon** "Recognising Humans by Gait via Parametric Canonical Space", in the *Proceedings of ICSC International Symposium on Engineering of Intelligence Systems*, Tenerife, Spain, vol. 3, pp. 384-389, February 1998.

2. **P.S. Huang, C.J. Harris, and M.S. Nixon** "Canonical Space representation for Recognizing Humans by Gait and Face", in the *Proceedings of IEEE Southwest Symposium on Image Analysis and Interpretation*, Tucson, Arizona, USA, pp. 180-185, April 1998.

3. **P.S. Huang, C.J. Harris, and M.S. Nixon** "Comparing Different Template Features for Recognizing People by Their Gait", in the *Proceedings of BMVA Ninth British Machine Vision Conference*, Southampton, UK, vol. 2, pp. 639-648, September 1998.

4. **P.S. Huang, C.J. Harris, and M.S. Nixon** "A Statistical Approach for Recognizing Humans by Gait Using Spatial-Temporal Templates", in the *Proceedings of IEEE International Conference on Image Processing*, Chicago, Illinois, USA, vol. 3, pp. 178-182, October 1998.

5. **P.S. Huang, C.J. Harris, and M.S. Nixon** "Visual Surveillance and Tracking of Humans by Face and Gait Recognition", in the *Proceedings of 7th IFAC International Symposium on Artificial Intelligence in Real-Time Control*, Grand Canyon National Park, Arizona, USA, 43-44 (Extended Version on CD), October 1998. (Best paper award in Visualization and Imaging session)

6. **P.S. Huang, C.J. Harris, and M.S. Nixon** "Recognizing Humans by Gait Using a Statistical Approach for Temporal Templates", in the *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, La Jolla, California, USA, vol. 5, pp. 4556-4561, October 1998.

7. **M.S. Nixon, J.N. Carter, D. Cunado, P.S. Huang, J.M. Nash and S.V. Stevenage** "Automatic gait recognition", in the *Proceedings of IEE Colloquium on Motion Analysis and Tracking*, London, UK, pp. 3/1-3/6, May 1999.

References

[1] M.S. Nixon, J.N. Carter, D. Cunado, P.S. Huang, and S.V. Stevenage, "Automatic gait recognition", in *BIOMETRICS - Personal Identification in Networked Society*, A. Jain, R. Bolle, and S. Pankanti, Eds., chapter 11, pp. 231–249. Kluwer Academic Publishers, January 1999.

[2] R.B. Davis and P.A. DeLuca, "Clinical gait analysis - current methods and future directions", in *Human Motion Analysis*, G.F. Harris and P.A. Smith, Eds., chapter 2, pp. 17–42. IEEE Press, 1997.

[3] M.P. Murray, A.B. Drought, and R.C. Kory, "Walking patterns of normal men", *Journal of Bone Joint Surgery*, vol. 46-A, no. 2, pp. 335–360, 1964.

[4] M.P. Murray, "Gait as a total pattern of movement", *American Journal of Physical Medicine*, vol. 46, no. 1, pp. 290–332, 1967.

[5] A-L Kairento and G. Hellen, "Biomechanical analysis of walking", *Biomechanics*, vol. 14, no. 10, pp. 671–678, 1981.

[6] K.R. Kaufman and D.H. Sutherland, "Future trends in human motion analysis", in *Human Motion Analysis*, G.F. Harris and P.A. Smith, Eds., chapter 11, pp. 187–215. IEEE Press, 1997.

[7] G.F. Harris and P.A. Smith, Eds., *Human Motion Analysis*, IEEE Press, 1997.

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

[9] C. Cedras and M. Shah, "A survey of motion analysis from moving light displays", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, June 1994, pp. 214–221.

[10] R.F. Rashid, "Towards a system for the interpretation of moving light displays", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 6, pp. 574–581, 1980.

[11] W.H. Dittrich, "Action categories and the perception of biological motion", *Perception*, vol. 22, pp. 15–22, 1993.

[12] G.P. Binham, R.C. Shmidt, and L.D. Rosenblum, "Dynamics and the orientation of kinematic forms for visual event recognition", *Journal of Experimental Psychology: Human Perception and Performance*, vol. 21, no. 6, pp. 1473–1493, 1995.

[13] L.T. Kozlowski and J.E. Cutting, "Recognising the sex of a walker from a dynamic point light display", *Perception and Psychophysics*, vol. 21, pp. 575–580, 1977.

[14] S. Runenson and G. Frykholm, "Kinematic specification of dynamics as an informational basis for person-and-action perception: expectation, gender recognition and deceptive intention", *Journal of Experimental Psychology: General*, vol. 112, pp. 585–615, 1983.

[15] G. Mather and L. Murdock, "Gender discrimination in biological motion displays based on dynamic cues", *Proceedings of Royal Society London*, vol. B:258, pp. 273–279, 1994.

[16] J.E. Cutting and D.R. Proffitt, "Gait perception as an example of how we perceive events", in *Intersensory Perception and Sensory Integration*, R.D. Walk and H.L. Pich, Eds., chapter 8, pp. 249–273. Plenum Press, London UK, November 1981.

[17] 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.

[18] J.E. Cutting, D.R. Proffitt, and L.T. Kozlowski, "A biochemical invariant for gait recognition", *Journal of Experimental Psychology: Human Perception and Performance*, vol. 4, pp. 357–372, 1978.

[19] S.V. Stevenage, M.S. Nixon, and K. Vince, "Visual analysis of gait as a cue to identity", *Applied Cognitive Psychology*, 1999, at press.

[20] D. Marr and H.K. Nishihara, "Representation and recognition of the spatial organization of three-dimensional shapes", *Proceedings of Royal Society London*, vol. B:200, pp. 269–294, 1978.

[21] D. Hogg, "Model-based vision: a program to see a walking person", *Image and Vision Computing*, pp. 5–20, 1983.

[22] D. Hogg, *Interpreting Images of a Known Moving Object*, PhD thesis, University of Sussex, Brighton, UK, 1984.

[23] K. Rohr, "Incremental recognition of pedestrians from image sequences", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, New York, FL, USA, June 1993, pp. 8–13.

[24] K. Rohr, "Towards model-based recognition of human movements in image sequences", *CVGIP: Image Understanding*, vol. 59, no. 1, pp. 94–115, 1994.

[25] K. Akita, "Image sequence analysis of real world human motion", *Patttern Recognition*, vol. 17, no. 1, pp. 73–83, 1984.

[26] H-J Lee and Z. Chen, "Determination of 3D human body postures from a single view", *Computer Vision, Graphics, and Image Processing*, vol. 30, pp. 148–168, 1985.

[27] Z. Chen and H-J Lee, "Knowledge-guided visual perception of 3-D human gait from a single image sequence", *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 2, pp. 336–342, 1992.

[28] Y. Guo, G. Xu, and S. Tsuji, "Understanding human motion patterns", in *Proceedings of 12th International Conference on Pattern Recognition*, Jerusalem, Israel, October 1994, vol. 2, pp. 325–329.

[29] Y. Guo, G. Xu, and S. Tsuji, "Tracking human body motion based on a stick figure model", *Journal of Visual Communication and Image Representation*, vol. 5, no. 1, pp. 1–9, March 1994.

[30] A.G. Bharatkumar, K.E. Daigle, M.G. Pandy, Q. Cai, and J.K. Aggarwal, "Lower limb kinematics of human walking with the medial axis transformation", in *Proceedings of IEEE Nonrigid and Articulated Motion Workshop*, TX, USA, 1994, pp. 70–76.

[31] A. Azarbayejani, C. Wren, and A. Pentland, "Real-time 3-D tracking of the human body", in *Proceedings of IMAGE'COM 96*, Bordeaux, France, May 1996, pp. 911–916.

[32] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: Real-time tracking of the human body", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 780–785, 1997.

[33] J. O'Rourke and N.I. Badler, "Model-based image analysis of human motion using contraint propagation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, no. 6, pp. 522–536, 1980.

[34] S. Kurakake and R. Nevatia, "Description and tracking of moving articulated objects", *Systems and Computers in Japan*, vol. 25, no. 8, pp. 16–26, 1994.

[35] R. Polana and R. Nelson, "Detecting activities", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, New York, NY, USA, June 1993, pp. 2–7.

[36] J. Little and J. Boyd, "Describing motion for recognition", in *Proceedings of International Symposium on Computer Vision*, Coral Gables, FL, USA, November 1995, pp. 235–240.

[37] H. Murase and R. Sakai, "Moving object recognition in eigenspace representation: gait analysis and lip reading", *Pattern Recognition Letters*, vol. 17, pp. 155–162, 1996.

[38] L. Cambell and A. Bobick, "Recognition of human body motion using phase space constraints", in *Proceedings of Fifth International Conference on Computer Vision*, Cambridge, MA, USA, 1995, pp. 624–630.

[39] R. Polana and R. Nelson, "Detection and recognition of periodic, nonrigid motion", *International Journal of Computer Vision*, vol. 23, no. 3, pp. 261–282, 1997.

[40] J. Little and J. Boyd, "Recognizing people by their gait: the shape of motion", *Videre*, vol. 1, no. 2, pp. 1–32, 1998.

[41] J.W. Davis and A.F. Bobick, "The representation and recognition of human movement using temporal templates", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, June 1997, pp. 928–934.

[42] S.A. Niyogi and E.H. Adelson, "Analysis and recognizing walking figures in XYT", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, June 1994, pp. 469–474.

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

[44] N. Murphy, N. Byrne, and K. O'Leary, "Long sequence analysis of human motion using eigenvector decomposition", *SPIE:Intelligent Robots and Computer Vision XII*, vol. 2056, pp. 400–410, 1993.

[45] M.J. Black, Y. Yacoob, A.D. Jepson, and D.J. Fleet, "Learning parameterized models of image motion", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Puerto Rico, June 1997, pp. 561–567.

[46] M. Assereto, G. Figari, and A. Tesei, "Robust approach to tracking human motion in real scenes", *Electronics Letters*, vol. 30, no. 24, pp. 2013–2014, 1994.

[47] S. Wachter and H.-H. Nagel, "Tracking of persons in monocular image sequences", in *Proceedings of IEEE Nonrigid and Articulated Motion Workshop*, San Juan, Puerto Rico, June 1997, vol. 3, pp. 2–9.

[48] J.C. Cheng and J.M.F. Moura, "Tracking human walking in dynamic scenes", in *Proceedings of IEEE International Conference on Image Processing*, Santa Barbara, CA, USA, October 1997, pp. 137–140.

[49] I.A. Kakadiaris and D. Metaxas, "Model-based estimation of 3D human motion with occlusion based on active multi-viewpoint selection", in *Proceedings of IEEE*

Conference on Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, June 1996, pp. 81–87.

[50] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Training models of shape from sets of examples", in *Proceedings of Third British Machine Vision Conference*, Leeds, UK, September 1992, BMVA, pp. 9–18.

[51] A. Baumberg and D. Hogg, "Learning flexible models from image sequences", in *Proceedings of European Conference on Computer Vision*, May 1994, pp. 299–308.

[52] A.M. Baumberg and D. Hogg, "A efficient method for contour tracking using active shape models", in *Proceedings of The Workshop on Motion of Non-Rigid and Articulated Objects*, TX, USA, November 1994, pp. 194–198.

[53] L-Q Xu and D.C. Hogg, "Neural networks in human motion tracking - an experimental study", *Image and Vision Computing*, vol. 15, pp. 607–615, 1997.

[54] A. Baumberg and D. Hogg, "Generating spatiotemporal models from examples", *Image and Vision Computing*, vol. 14, pp. 525–532, 1996.

[55] A. Pentland and B. Horowitz, "Recovery of non-rigid motion and structure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 7, pp. 730–742, 1991.

[56] J. Denzler and H. Niemann, "Real-time pedestrian tracking in natural scenes", in *Proceedings of 7th International Conference on Computer Analysis of Images and Pattterns*, Kiel, September 1997, pp. 42–49.

[57] S.X. Ju, M.J. Black, and Y. Yacoob, "Cardboard people: A parameterized model of articulated motion", in *Proceedings of 2nd International Conference on Automatic Face and Gesture Recognition*, Killington, Vermont, October 1996, pp. 38–44.

[58] K.J. Bradshaw, I.D. Reid, and D.M. Murray, "The active recovery of 3D motion trajectories and their use in prediction", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 3, pp. 219–233, 1997.

[59] J.K. Aggarwal and N. Nandakumar, "On the computation of motion from sequences of image - a review", *Proceedings of the IEEE*, vol. 76, no. 8, pp. 917–935, 1988.

[60] T.S. Huang and A.N. Netravali, "Motion and structure from feature correspondences: a review", *Proceedings of the IEEE*, vol. 82, no. 2, pp. 252–267, 1994.

[61] M.K. Leung and Y-H Yang, "Human body motion segmentation in a complex scene", *Pattern Recognition*, vol. 20, no. 1, pp. 55–64, 1987.

[62] A. Shio and J. Sklansky, "Segmentation of people in motion", in *Proceedings of IEEE Workshop on Visual Motion*. IEEE, 1991, pp. 325–332.

[63] D. Meyer, J. Denzler, and H. Niemann, "Model based extraction of articulated objects in image sequences for gait analysis", in *Proceedings of International Conference on Image Processing*, Santa Barbara, CA, USA, October 1997, IEEE, vol. 3, pp. 78–81.

[64] M. Oren, C.P. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio, "Pedestrian detection using wavelet templates", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Puerto Rico, June 1997, pp. 193–199.

[65] K-K Sung and T. Poggio, "Example-based learning for view-based human face detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 39–51, 1998.

[66] L.R. Rabiner and B.H. Juang, "An introduction to Hidden Markov Models", *IEEE ASSP Magazine*, pp. 4–16, January 1986.

[67] L.R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition", *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[68] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden markov model", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1992, pp. 379–385.

[69] M. Kirby and L. Sirovich, "Application of the Karhunen-Loève procedure for the characterization of human faces", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, 1990.

[70] M.J. Black, Y. Yacoob, and S.X. Ju, "Recognizing human motion using parameterized models of optical flow", in *Motion-Based Recognition*, M. Shah and R. Jain, Eds., chapter 11, pp. 245–269. Kluwer Academic Publishers, Boston, 1997.

[71] G.I. Chiou and J-N Hwang, "Lipreading from color video", *IEEE Transactions on Image Processing*, vol. 6, no. 8, pp. 1192–1195, 1997.

[72] G. Gioftsos and D.W. Grieve, "The use of neural networks to recognize patterns of human movements: gait patterns", *Clinical Biomechanics*, vol. 10, no. 4, pp. 179–183, 1995.

[73] H. Ushida, T. Yamaguchi, and T. Takagi, "Human-motion recognition via a fuzzy associative memory system", *Systems and Computers in Japan*, vol. 26, no. 8, pp. 90–104, 1995.

[74] T. Yamaguchi, K. Goto, and T. Takagi, "Two-degree-of-freedom fuzzy model using associate memories and its applications", *Information Sciences*, vol. 71, pp. 65–97, 1993.

[75] D. Cunado, M.S. Nixon, and J.N. Carter, "Using gait as a biometric, via phase-weighted magnitude spectra", in *First International Conference, AVBPA'97*, Crans-Montana, Switzerland, March 1997, pp. 95–102.

[76] H.M. Lakany and G.M. Hayes, "An algorithm for recognising walkers", in *First International Conference, AVBPA'97*, Crans-Montana, Switzerland, March 1997, pp. 111–117.

[77] C. Bregler, "Learning and recognizing human dynamics in video sequences", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, June 1997, pp. 568–574.

[78] D. Meyer, "Human gait classification based on Hidden Markov Models", in *3D Image Analysis and Synthesis'97*, December 1997, pp. 139–146.

[79] D. Meyer, J. Pösl, and H. Niemann, "Gait classification with HMMs for trajectories of body parts extracted by mixture densities", in *Proceedings of Ninth British Machine Vision Conference*, Southampton, UK, September 1998, BMVA, pp. 459–468.

[80] M. Turk and A. Pentland, "Eigenfaces for recognition", *Journal of Cognitive Neuroscience*, vol. 3, pp. 71–86, 1991.

[81] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition", in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, June 1994, pp. 84–91.

[82] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, 1997.

[83] H. Murase and S.K. Nayar, "Learning and recognition of 3D objects from appearance", in *Proceedings of IEEE Workshop on Qualitative Vision*, New York, June 1993, IEEE, pp. 39–50.

[84] H. Murase and S.K. Nayar, "Illumination planning for object recognition using parametric eigenspaces", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 12, pp. 1219–1227, 1994.

[85] H. Murase and S.K. Nayar, "Visual learning and recognition of 3-D objects from appearance", *International Journal of Computer Vision*, vol. 14, pp. 5–24, 1995.

[86] H. Murase and S.K. Nayar, "Three-dimensional object recognition from appearance - parametric eigenspace method", *Systems and Computers in Japan*, vol. 26, no. 8, pp. 45–54, 1995.

[87] H. Murase and S.K. Nayar, "Detection of 3D objects in cluttered scenes using hierarchical eigenspace", *Pattern Recognition Letters*, vol. 18, pp. 375–384, 1997.

[88] S.K. Nayar, H. Murase, and S.A. Nene, "Learning, positioning, and tracking visual appearance", in *Proceedings of IEEE International Conference on Robotics and Automation*, San Diego, May 1994, IEEE, pp. 3237–3244.

[89] J.J. Weng, "On comprehensive visual learning", in *Proceedings of NSF/ARPA Workshop on Performance vs. Methodology in Computer Vision*, Seattle, WA, USA, June 1994, pp. 152–166.

[90] S.K. Nayar, S.A. Nene, and H. Murase, "Subspace methods for robot vision", *IEEE Transactions on Robotics and Automation*, vol. 12, no. 5, pp. 750–758, 1996.

[91] K. Ohba and K. Ikeuchi, "Detectability, uniqueness, and reliability of eigen windows for stable verification of partially occluded objects", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 9, pp. 1043–1048, 1997.

[92] R.D. Boyle, "Scaling additional contributions to principal components analysis", *Pattern Recognition*, vol. 31, no. 12, pp. 2047–2053, 1998.

[93] Y. Moses, Y. Adini, and S. Ullman, "Face recognition: The problem of compensating for changes in illumination direction", in *Proceedings of European Conference on Computer Vision*, 1994, pp. 286–296.

[94] W. Bledsoe, "Man machine facial recognition", Tech. Rep. RPI-22, Panoramic Research Inc., Palo Alto, CA, USA, 1966.

[95] A. Samal and P. Iyengar, "Automatic recognition and analysis of human faces and facial expressions: A survey", *Pattern Recognition*, vol. 25, pp. 65–77, 1992.

[96] R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey", *Proceedings of the IEEE*, vol. 83, no. 5, pp. 705–740, 1995.

[97] R. Brunelli and T. Poggio, "Face recognition: Features versus templates", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 10, pp. 1042–1052, 1993.

[98] X. Jia and M.S. Nixon, "Extending the feature vector for automatic face recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 12, pp. 1167–1176, 1995.

[99] D. Valentin, H. Abdi, A.J. O'Toole, and G.W. Cottrell, "Connectionist models of face processing: A survey", *Pattern Recognition*, vol. 27, pp. 1209–1230, 1994.

[100] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images", in *First International Conference, AVBPA'97*, Crans-Montana, Switzerland, March 1997, pp. 127–142.

[101] K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images", *Journal of the Optical Society of America - Series A*, vol. 14, no. 8, pp. 1724–1733, 1997.

[102] T-M Tu, C-H Chen, J-L Wu, and C-I Chang, "A fast two-stage classification method for high-dimensional remote sensing data", *IEEE Transactions Geoscience and Remote Sensing*, vol. 36, no. 1, pp. 182–191, January 1998.

[103] D.L. Swets and J.J. Weng, "Using discriminant eigenfeatures for image retrieval", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, 1996.

[104] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[105] B.F. Merembeck and B.J. Turner, "Directed canonical analysis and the performance of the classifiers under its associated linear transformation", *IEEE Transactions on Geoscience and Remote Sensing*, vol. GE-18, no. 2, pp. 190–196, 1980.

[106] C-C T. Chen and D.A. Landgrebe, "A spectral feature design system for the hiris/modis era", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 27, no. 6, pp. 681–686, 1989.

[107] R.A. Fisher, "The statistical utilization of multiple measurements", *Annals of Eugenics*, vol. 8, pp. 376–386, 1938.

[108] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, second edition, 1990.

[109] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, 1996.

[110] B.H. Bharucha and T.T. Kadota, "On the representation of continuous parameter processes by a sequence of random variables", *IEEE Transactions on Information Theory*, vol. IT-16, no. 2, pp. 139–141, 1970.

[111] A.K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall, 1989.

[112] H. Murakami and V. Kumar, "Efficient calculation of primary images from a set of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, no. 5, pp. 511–515, 1982.

[113] H. Murase and M. Lindenbaum, "Partial eigenvalue decomposition of large images using spatial temporal adaptive method", *IEEE Transactions on Image Processing*, vol. 4, no. 5, pp. 620–629, 1995.

[114] N. Ahmed, T. Natarjan, and K.R. Rao, "Discrete cosine transform", *IEEE Transactions on Computers*, pp. 90–93, January 1974.

[115] John A. Richards, *Remote Sensing Digital Image Analysis: An Introduction*, Springer-Verlag, second edition, 1993.

[116] P.S. Huang, C.J. Harris, and M.S. Nixon, "Recognising humans by gait via parametric canonical space", in *Proceedings of International Symposium on Engineering of Intelligent Systems*, Tenerife, Spain, February 1998, ICSC, vol. 3, pp. 384–389.

[117] P.S. Huang, C.J. Harris, and M.S. Nixon, "Canonical space representation for recognizing humans by gait and face", in *Proceedings of Southwest Symposium on Image Analysis and Interpretation*, Tucson, Arizona, USA, April 1998, IEEE, pp. 180–185.

[118] P.S. Huang, C.J. Harris, and M.S. Nixon, "Visual surveillance and tracking of humans by face and gait recognition", in *Proceedings of International Symposium on Artificial Intelligence in Real-Time Control*, Arizona, USA, October 1998, IFAC, pp. 43–44(Extended Version on CD).

[119] P.S. Huang, C.J. Harris, and M.S. Nixon, "Recognising humans by gait via parametric canonical space", *Journal of Artificial Intelligence in Engineering, Special Issue on EIS'98*, 1999, to be published.

[120] M-P Dubuisson and A.K. Jain, "Contour extraction of moving objects in complex outdoor scenes", *International Journal of Computer Vision*, vol. 14, no. 6, pp. 83–105, 1995.

[121] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis and Machine Vision*, Chapman and Hall Compution, first edition, 1993.

[122] S. Chandrasekaran, B.S. Manjunath, Y.F. Wang, J. Winkeler, and H. Zhang, "An eigenspace update algorithm for image analysis", *Graphical Models and Image Processing*, vol. 59, no. 5, pp. 321–332, 1997.

[123] P.M. Hall, D. Marshall, and R.R. Martin, "Incremental eigenanalysis for classification", in *Proceedings of Ninth British Machine Vision Conference*, Southampton, UK, September 1998, BMVA, pp. 286–295.

[124] H.A. Rowley, S. Baluja, and T. Kanade, "Neural network-based face detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, pp. 23–38, 1998.

[125] B. Moghaddam and A. Pentland, "Probabilistic matching for face recognition", in *Proceedings of Southwest Symposium on Image Analysis and Interpretation*, Tucson, Arizona, USA, April 1998, IEEE, pp. 186–191.

[126] P.S. Huang, C.J. Harris, and M.S. Nixon, "Recognizing humans by gait using a statistical approach for temporal templates", in *Proceedings of International Conference on Systems, Man, and Cybernetics*, La Jolla, California, USA, October 1998, IEEE, vol. 5, pp. 4556–4561.

[127] P.S. Huang, C.J. Harris, and M.S. Nixon, "Human gait recognition in canonical space using temporal templates", *IEE Proceedings - Vision, Image and Signal Processing*, 1999, to be published.

[128] C. Cedras and M. Shah, "Motion-based recognition: a survey", *Image and Vision Computing*, vol. 13, no. 2, pp. 129–155, 1995.

[129] B.K.P. Horn and B.G. Schunck, "Determining optical flow", *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

[130] J.L. Barron, D.J. Fleet, and S.S. Beauchemin, "Performance of optical flow techniques", *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

[131] H. Bulthoff, J. Little, and T. Poggio, "A parallel algorithm for real-time computation of optical flow", *Nature*, vol. 337, pp. 549–553, February 1989.

[132] P.S. Huang, C.J. Harris, and M.S. Nixon, "Comparing different template features for recognizing people by their gait", in *Proceedings of Ninth British Machine Vision Conference*, Southampton, UK, September 1998, BMVA, vol. 2, pp. 639–648.

[133] P.S. Huang, C.J. Harris, and M.S. Nixon, "A statistical approach for recognizing humans by gait using spatial-temporal templates", in *Proceedings of International Conference on Image Processing*, Chicago, Illinois, USA, October 1998, IEEE, vol. 3, pp. 178–182.

[134] F. Liu and R.W. Picard, "Detecting and segmenting periodic motion", Tech. Rep. 400, M.I.T. Media Laboratory Perceptual Computing Section, Cambridge, MA 02139, USA, 1996.

[135] I. Barrodale and R.E. Erickson, "Algorithms for least-squares linear prediction and maximum entropy spectral analysis", *Geophysics*, vol. 45, no. 3, pp. 420–432, March 1980.

[136] R. Brunelli and D. Falavigna, "Personal identification using multiple cues", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955–966, 1995.

[137] L. Hong and A. Jain, "Integrating faces and fingerprints for personal identification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1295–1307, 1998.

[138] P. Baldi and K. Hornik, "Neural networks and principal component analysis: Learning from examples without local minima", *Neural Networks*, vol. 2, pp. 53–58, 1989.

[139] A.R. Webb and D. Lowe, "The optimised internal representation of multilayer classifier networks performs nonlinear discriminant analysis", *Neural Networks*, vol. 3, pp. 367–375, 1990.

[140] P. Gallinari, S. Thiria, F. Badran, and F. Fogelman-Soulie, "On the relations between discriminant analysis and multilayer perceptrons", *Neural Networks*, vol. 4, pp. 349–360, 1991.

[141] S.Y. Kong and J.S. Taur, "Decision-based neural networks with signal/image classification applications", *IEEE Transactions on Neural Networks*, vol. 6, no. 1, pp. 170–181, 1995.

[142] J. Mao and A.K. Jain, "Artificial neural networks for feature extraction and multivariate data projection", *IEEE Transactions on Neural Networks*, vol. 6, no. 2, pp. 296–317, 1995.

[143] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.