UNIVERSITY OF SOUTHAMPTON

Model-Based Gait Extraction and Recognition

by

Ziheng Zhou

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

May 2007

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Ziheng Zhou

Extracting full-body motion from monocular video sequences for gait recognition is an important and difficult problem. Very often, the motion will be highly articulated and have complex changing boundaries and images may suffer from high level correlated and random noise from the real world. Moreover, the large variations of the appearances of walking people caused by, for instance, carrying objects or wearing clothing, make the problem even more complicated. In this thesis, we propose a consistent and easily extensible Bayesian framework for the gait extraction problem using strong prior knowledge. This knowledge is imposed by a single two-dimensional articulated model having both time-invariant (static) and time-variant (dynamic) parameters. The model is easily extended to handle the variations of body shapes. To exploit the dynamics of human walk, we use a hidden Markov model to detect the phases of images in a walking cycle. The PDF projection theorem is introduced to learn the observation probability distributions accurately. We build a strong prior model from the statistics of the parameters of the articulated model, which are learned from noise-free indoor training data. The system parameters are first bootstrapped from a small amount of data and then refined by the Bayesian updating. We demonstrate our approach on both high-quality indoor and noisy outdoor video data, as well as high-quality data with synthetic noise and occlusions added, and walkers with rucksacks, skirts and trench coats. Results are quantified in terms of the chamfer distance and average pixel error between automatically extracted body points and corresponding points hand-labelled. No one part of the system is novel in itself, but the overall framework makes it feasible to extract gait from very much poorer quality image sequences than hitherto. We devise a simple gait recognition algorithm based on the extracted model parameters obtained by the Bayesian framework. The gait signature consists of the static parameter and the amplitude and phase information measured by fitting Fourier series to the trajectories of the dynamic parameters. The algorithm is tested on both indoor and outdoor data and its performance is compared with the standard baseline algorithm. We have achieved a much higher recognition rate on the outdoor noisy data, once again proving the robustness of the framework against real-world noise.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my appreciation and gratitude to my supervisors Dr. Adam Prügel-Bennett and Prof. Robert I. Damper for their insightful advice, continuous technical support and helpful discussions. Thanks also to Prof. Mark S. Nixon and Dr. John N. Carter for allowing me to use the Southampton HiD gait database in my work.

As well as the people mentioned above, I would like to thank my parents who have been so supportive to my study at Southampton. Their priceless advice and understanding helped me get through many of those hard times in my life. In addition, I would like to say a big 'thank-you' to my fiancée, Shanshan Zheng, who shared the sweetness and bitterness of my life in the past five years. It is her love that has given me strength and confidence to overcome any difficulty on the way toward my PhD. Finally, thanks to all my friends, especially Yisong Xiao and Letu Yang, who brought me so much happiness and made me really enjoy the life at Southampton.

Chapter 1

Introduction

1.1 Problem statement

Understanding human motion in video sequences is one of the most challenging tasks for computer vision. Because of the complexity of the human body structure, the motion is highly articulated, which means the targeted object has a constantly changing boundary and frequently causes self-occlusions. In general, it is difficult to deal with arbitrary human motion unless we have knowledge about the variations of the motion. This thesis focuses on the extraction of a universal type of human motion: walking motion. This work can benefit many applications including analysing human gait for medical, computer-animation, or biometric purposes.

From early psychological experiments (Johansson 1975) we know that humans have the ability to identify people by their gait. Recent research on human gait also indicates the potential of using gait as a new biometric. Although many gait recognition systems have been built and achieved a high recognition rate, the current technology is far from being used in practice. The major problem is that most of the systems were designed and tested on some particular gait database produced in a well-controlled environment (e.g., in a indoor laboratory). The image quality is high and the background is relatively simple. Most gait information can therefore be obtained through segmentation but such segmentation can be hardly achieved from video sequences taken by a normal surveillance system. Recently, efforts have been made to build a gait database in which walkers are filmed under real-world conditions, for example, the Southampton large HiD gait database (Grant et al. 2004). However, the quality of images is still high, which makes clues such as the skin colour available for gait extraction (Zhang, Collins, and Liu 2004). An ideal system should handle not only the real-world noise, but the dynamic nature of human gait. For example, the appearance of a walking individual in images can be changed significantly by wearing different clothing or carrying objects. Any practical system should be able to cope with such variations.

Our goal is to develop a general framework capable of dealing with the uncertainties inherent in the problem of human gait extraction. We want to explore the limit of the kind of information we can obtain from the sort of difficult data which can be observed in practice and to see the capability of the information to identify walking people. The image sequences used in this work are from the Southampton HiD gait database. However, there is no intention to use the high quality of the images in this database to facilitate the motion extraction. Instead of using the images directly, we performed simple background subtraction at first and then used the silhouettes as input. Some examples of the silhouettes can be seen in Appendix A.

To achieve our objective, we have adopted a Bayesian approach exploiting strong prior information about how humans walk. By *strong*, we mean very basic, almost inviolable knowledge such as the fact that all humans have a head and two legs, with each leg jointed at the knee. This strong prior information is imposed by a two-dimensional articulated model of a walker, which can be easily extended, for example, to allow for the walker carrying a bag, or wearing a coat or skirt. At this stage, we consider only walkers moving perpendicular to the camera as this is typical of the current state of the art, and because we have a large database collected under these conditions. As well as exploiting prior information about body articulation, knowledge of the characteristic, dynamic movement of the body was built into a hidden Markov model (HMM). The adopted framework also allows us to learn the statistics of normal walkers from high-quality video images. In Bayesian language, we can use good data to obtain a posterior for the model parameters; this posterior can then be used as a prior when presented with noisy data. Such 'bootstrapping' via Bayesian updating prevents us having to obtain extensive statistics of human walkers manually, which might otherwise make this approach prohibitively expensive. Because we aim to extract the fullest information about articulated motion that the image quality allows, rather than to track walkers in real time, it is highly advantageous to process the whole sequence of images. In this way, and combined with the use of prior information, we can obtain a global solution that, for instance, copes with extreme noise, occlusions, etc.

1.2 Contributions

The main contribution of this work is to introduce a consistent Bayesian framework for addressing the important and well-studied problem of human gait extraction from video sequences. This framework allows us to integrate strong prior knowledge with data-driven learning and thereby to produce results comparable with the best reported in the literature, handling noise and occlusion especially well. Each component in the framework is itself quite simple. The novelty comes from combining them together to solve a complicated problem within a Bayesian framework. More specifically, the contributions are as follows.

- Providing a systematic way for the construction of a Bayesian framework for the problem of fitting a parametric model to observations in noisy image data. The likelihood is built from the robust chamfer distances and the prior is defined as a multivariate Gaussian using the statistics of the parameters of the model learned from noise-free training data. The bootstrapping which involves manual work is relatively cheap and system parameters can be refined by the Bayesian updating recursively.

- Building carefully an HMM for modelling the dynamics of human walk. The HMM structure is well-designed, using tied-states to represent the true transition probabilities. The probability density function (PDF) projection theorem is used to learn the high-dimensional observation probability distributions accurately. Using this relatively new theorem in the gait-related problem is novel and proven to be useful in constructing the HMM.

- The overall framework is flexible and easily extensible. The components in the framework are independent of each other so that we can easily modify one of them without breaking the whole framework. The extensibility has been demonstrated in the experiments where we extend the basic articulated model to cope with walkers carrying a rucksack, wearing a long skirt, or wearing a trench coat. For all three cases, the only cost is adding one or two static parameters.

- Various experiments have been designed to test the system. The framework was first tested on the clean data with artificial noise added. The chamfer distances were used to quantify the results. The fitting results for the outdoor noisy sequences were then shown visually. We marked manually five key joints on the original images and measured the errors between the marked joints and the ones recovered by the model. To prove further the accuracy, we built a gait recognition system that uses the extracted parameters to identify walkers. We have implemented the baseline algorithm for gait recognition of Sarkar et al. (2005) on the same data. The comparison of performance of our algorithm with baseline has proven the robustness of the framework against random noise and occlusions.

Produced publications:

- Zhou, Z., A. Prügel-Bennett and R. I. Damper (2006). A Bayesian framework for extracting human gait using strong prior knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28(11)*, pp.1738-1752.

- Zhou, Z., R. I. Damper and A. Prügel-Bennett (2006) Model Selection within a Bayesian approach to extraction of walking motion. *IEEE Conference on Computer Vision and Pattern Recognition Workshop* (CVPRW' 06), p.44.

1.3 Outline of Thesis

The rest of this thesis is organised as follows. In Chapter 2, we give a review of the previous research related to our work. We first describe the approaches aiming to extract general human body motion. After that, we focus on the extraction of the motion of walking people. Finally, we review some gait recognition systems.

Chapter 3 describes the data used in our work as well as the normalisation procedure. As mentioned previously, we are using the Southampton HiD gait database. The database consists of three kinds of data: indoor, outdoor and supplemental image sequences. Some examples are given for each of them. We then describe the normalisation of the silhouettes obtained from background subtraction. The purpose of the normalisation is to locate the walking individual in images, cropping a sub-image where he or she is centred and resizing the sub-image to a proper size. This is done by a tracking algorithm using the generalised Hough transform and dynamic programming.

Chapter 4, which is the main chapter of this thesis, gives the details of the whole Bayesian framework for the human gait extraction. At the beginning of this chapter, we justify the reason why we adopt a Bayesian framework for this problem. We then describe sequentially the way we exploit our prior knowledge of human walk in building the framework. An articulated model is constructed to represent a walker and a hidden Markov model used to model the periodicity of human gait. The PDF projection theorem is used to learn the observation probability distributions, which is one of the novelties in this thesis. We then give the definition of the likelihood and prior which are based on the chamfer distances and some statistics of the model parameters respectively. We introduce the bootstrapping and the Bayesian updating component to learn the system parameters in an efficient way. In the end, we extend the framework to cope with walkers with different body configurations.

Chapter 5 shows the experimental results to demonstrate the power of the framework. Three experiments have been designed to test the system. We first test the system on the artificial noisy image sequences. Two kinds of noise, salt and pepper noise and occlusion, are added to the noise-free indoor data. We show the system performance on the images with various levels of noise added in terms of the chamfer distance. Simulations are designed to test the effect of possible poor normalisation. The sequences used in the simulations are generated from the indoor sequences by perturbing the positions of silhouettes and adding artificial noise. After that, the framework is tested on the real outdoor noisy sequences. The results are quantified by the errors between the hand-marked joints and the joints obtained by model fitting. Finally, we test the extensibility of the system on the supplemental data. We modified the model to cope with a walker carrying a rucksack, wearing a long skirt or wearing a trench coat. We show some encouraging preliminary results from this experiment.

Chapter 6 describes a system that uses the gait information extracted by the Bayesian framework to identify walking people. The overall recognition algorithm is quite simple. For each given sequence, a feature vector is constructed for classification. The vector is built from the static parameters that describe the body size of a walker, and the dynamic features which are the phases and amplitudes obtained by fitting Fourier series to the extracted joint-angle trajectories. Each element in the vector is then normalised and weighted by its F-statistic. The classification is done by a simple nearest-mean classifier. The competitive baseline algorithm is implemented to make a comparison with our system.

In the final chapter, we conclude our work and propose some possible future work.

Chapter 2

Related Work

2.1 Human Body Motion Extraction

Understanding human body motion in image sequences is one of most challenging tasks in computer vision. The motion itself is highly articulated and not easy to predict. Usually, it is recorded in images and the loss of the depth information will cause the singularity problem when recovering motion in 3D. Moreover, the imperfect image segmentation encountered in practice makes the problem even more difficult. In this section, we review the previous work on this problem. Good surreys can be found in Aggarwal and Cai (1999),Gavrila (1999),Moeslund and Granum (2001) and Wang et al. (2003).

To model the non-rigidity of the human body, various models have been constructed in the previous work. Articulated models are commonly used to represent the full or partial human body. In general, such a model consists of two components: a representation of the skeletal structure and a representation for body appearance. The former is comprised of a collection of segments connecting joints on the human body. Around the segments are the geometries approximating the articulated body parts. The shapes used for a 2D model include circles, rectangles, trapezoids and ellipses (Ju et al. 1996; Ning et al. 2002; Lan and Huttenlocher 2004; Wagg and Nixon 2004a). Most of the 3D models use shapes from a class of geometric primitives, super-quadrics (Barr 1984), to model body parts (O'Rourke and Badler 1980; Hogg 1983; Rehg and Kanade 1994; Rehg and Kanade 1995; Bregler and Malik 1998; Stenger et al. 2001). These include a large number of shapes such as cylinders, spheres, ellipsoids and hyper-rectangles. A combination of some of these primitives can usually result in satisfactory modelling of a human body. More accurate models can be achieved by deformations on the super-quadrics (Gavrila and Davis 1996; Kakadiaris and Metaxas 2000). In addition, parametric contour models have been also used to model the articulated human body. Examples of using contour models constructed from B-splines to extract the motion of walking people and moving hands can be found in Baumberg and Hogg (1994) and Isard and Blake (1998) respectively.

Having had a suitable model, we attempt to find a set of parameters (state) that fit the model best to the object in the images. Many previous motion extraction systems did the fitting within a probabilistic framework. Very often, the fitting is done by measuring the posterior state distribution. The state space is usually high-dimensional and it is not straightforward to compute this distribution. We tackle this problem by decomposing the posterior according to the Bayes' rule, which states that the posterior is proportional to the multiplication of the likelihood and prior. The likelihood can be computed directly from the current frame while the prior is propagated from the state distribution learned for the previous frame.

Based on the assumption that the state is a Gaussian random variable, the Kalman filter (Kalman 1960) provides a framework for propagating this random variable recursively and has been used in some previous work (Baumberg and Hogg 1994; Blake et al. 1993; Kakadiaris and Metaxas 2000). Note that the original Kalman filter can only describe a linear system. For a nonlinear system, the random variable is propagated through the first-order linearisation of the nonlinear state estimation function. In this case, the Kalman filter is called the extended Kalman filter (EKF). Julier and Uhlmann (1997) introduced the unscented Kalman filter to improve the accuracy of the linearisation from first-order to third-order. Stenger et al. (2001) utilised this technique in their hand tracking system. The unimodal state density assumed by the Kalman filter is sometimes not suitable in situations where the true state distribution is multi-modal (e.g., the background is cluttered or the object is occluded). To propagate the true state distribution, Isard and Blake (1998) proposed their *condensation* algorithm. The state distribution was represented and propagated by a set of samples (particles) over time. The propagation was done by two steps: building priors from the state distribution learned at the previous time using the learned dynamic models and measuring likelihoods of the current observation given the particles. The state density was then updated using Bayes' rule. A drawback of such non-parametric methods is that they require a large number of samples to be maintained to describe the distribution accurately enough, especially in a high-dimensional space. Cham and Rehg (1999), however, modelled the multimodal distribution by a piecewise Gaussian representation. Only the peaks of the posterior distribution were sampled and propagated to the next frame to generate a prior density, which greatly reduced the number of samples.

Instead of inferring a probabilistic model, the inverse kinematics technique provides an alternative to recover human motion from image sequences using a parametric 3D articulated model (Yamamoto and Koshikawa 1991; Rehg and Kanade 1994; Rehg and Kanade 1995; Bregler and Malik 1998). In such a system, the state space is mapped onto the image space by a nonlinear measurement function that generates the 2D projection of the 3D model and calculates the error between the projected model and images. The inverse kinematics approach inverts this mapping to measure changes of the parameters of the model which minimise the error. The inverting involves the linearisation of the

measurement function which is described by the function's Jacobian matrix and is done by a gradient-based optimisation procedure. Minimising directly the error between the synthesised model and image features is another way to find the optimal set of parameters of the model (Hogg 1983; Ohya and Kishino 1994; Gavrila and Davis 1996). Gavrila and Davis measured the error as the chamfer distance between the projected model and the object in images. They used a decomposition approach and a best-first technique to search the optimal state. Ohya and Kishino carried out the optimisation using the genetic algorithm.

2.2 Gait Extraction System

Having described the work on general human motion extraction, we focus on a particular kind of human motion: walking motion. Extracting the motion of walking people still remains a challenge in computer vision. The motion is highly articulated, which causes complex changing boundaries and serious self-occlusion. Furthermore, walking in a cluttered environment and the large variations of body appearance make the problem even more difficult. However, we have a lot of prior knowledge of how human beings walk (spatial and temporal information of movement of walking people), which can be used to compensate for the complexity of this problem. Most of the previous work can be divided into two categories in terms of the way the motion is extracted and recovered.

Fitting an articulated model is a common method to capture the motion of walking people. Hogg (1983) was among the first who used an articulated model to track walking people. His model consisted of 14 cylinders and the edge projection of the model was compared to the edge features of the walker. Based on Hogg's work, Rohr (1994) incorporated a Kalman filter to predict and smooth the model parameters in each frame. Such a linear dynamic system requires relatively noise-free image input and it is difficult to reach high accuracy. Ju et al. (1996) presented a cardboard person model, a set of connected planar patches, to track limbs of a walker. Rather than using edges, they exploited optical flows. They showed the tracking results for a person walking on a treadmill from different view points. However, the lack of ability to handle self-occlusion and the dependence on the optical flows limit its use in real cluttered situations. To model the sequential changes of the posterior distribution of the model parameters, Ning et al. (2002) implemented the condensation algorithm. The positions of the limbs were captured from indoor noise-free image sequences. Recently, Lan and Huttenlocher (2004) developed a tracking system aiming to handle self-occlusion and changes in view angles. They introduced a pictorial structure spatial model (similar to the cardboard person model) and a hidden Markov model (HMM) was used to model the temporal movement of walkers with each state exemplified by a pictorial structure model. The observation probabilities were approximated by sampling from the posterior distribution. Tracking was done by finding a maximum a posteriori (MAP) state sequence. We use a similar idea

in our HMM definition, that is, using walking models rather than selecting images from the sequence as the exemplars to compute the observation probabilities. They showed visually the tracking results on a multi-viewpoint sequence, but no quantitative results about the accuracy of the fitting were given. For purpose of extracting articulated motion accurately rather than purely tracking, we give an MAP solution for each frame in an image sequence. Generally, such a high-dimensional optimisation problem is intractable. But using the results of HMMs as the initialisation and a strong prior model, we can greatly reduce the search space and achieve good fitting results. We demonstrate the accuracy of the fitting both visually and quantitatively in Chapter 5.

Tracking contours of walking people is the other widely used method for gait extraction. In Baumberg and Hogg (1994), active shape models built from closed B-splines were used to track the contours of walking people. They performed principal component analysis on the training data to reduce the dimension of the state space to be searched. A Kalman filter was used to carry out the tracking overtime. Exemplar-based methods (Toyama and Blake 2002) have proven to be efficient for tracking human motion and gesture recognition. The exemplars are acquired directly from the training data and the similarities between exemplars and images are modelled by a probability distribution in an image-distance metric space. Although the approaches are computationally efficient, they only track contours roughly which makes it difficult to recover high-level articulated structure of human motion. Recently, two systems have been built to extract the contours of walkers for higher-level gait analysis (e.g., gait recognition). Wagg and Nixon (2004a) constructed a geometric shape model for a walking individual. They found the parameters that determined the shape sizes from a global temporal accumulation of the input image sequence. Medical data of the variations of the joint angles were used to find the limb positions. Finally, they used an active shape model to force the edges of the model to match the contours of walkers. Zhang et al. (2004) constructed a sophisticated deformable body contour model, which was controlled by landmarks. These landmarks were then modelled by a Bayesian network and sampled sequentially using sequential Monte Carlo on a single image frame. The major drawback of using such contour models is that all possible variations of body appearance need to be well represented in the training data. For walkers with significant appearance changes, the models have to be re-constructed and the shape priors have to be re-learned. In contrast, assuming that people will not change their gaits significantly when their appearance changes greatly (e.g., through carrying a rucksack or wearing a trench coat), our system allows us to cope with the variations easily without learning a new prior. This is illustrated in Section 5.6.

We have described two classes of methods for human gait extraction: one is fitting an articulated model and the other is tracking body contours of walking people. The former exploits prior knowledge of the human body. The motion is captured by finding the optimal set of parameters that minimise the error between the synthesised model and the objects in images. The posture can be naturally recovered by the model parameters.

The advantage of using an articulated model is that we can integrate our prior knowledge of human walking easily into the model to compensate for the potential noise and uncertainties. For example, we can control the parameters of the model to vary in a constrained region to prevent the model being distracted by the noise resulting in an unrealistic posture for a walker. Or the images can be corrupted so that image features representing the body parts are disconnected. An articulated model can compensate well for such discontinuity. There is a tradeoff between the complexity of the model and the system performance. The model should be constructed to take into account the quality of the images. An over-complicated model would introduce a high-dimensional state space which makes the fitting difficult and computationally expensive. The contour-tracking systems are, in general, more computationally efficient since there is usually no need to solve a high-dimensional optimisation problem. The disadvantage is that it is image-quality sensitive. Some systems learn the spatial variations of points on contours to compensate for the noise when sampling. Another disadvantage is that it is not straightforward to obtain articulated motion from the extracted contours. However, for an appearance-based gait recognition system, the contours can be used directly as the gait features.

2.3 Human Gait Analysis

Human gait has been increasingly studied because of its potential as a new biometric, that is, using gait to identify walking people. In comparison with other biometrics such as the face and fingerprint, human gait can be captured passively at a distance by any surveillance system, which makes it difficult for the moving subjects to camouflage their real gait. In the rest of this section, we first introduce a few established gait databases which have significantly contributed to the current research on the gait recognition. After that, we describe some early psychological experiments that have largely motivated this research. Finally, we review some of the reported gait recognition systems.

2.3.1 Human Gait Database

A few human gait databases have been built for the purpose of using gait as a biometric around the world. They are the Southampton *human identification at a distance* (HiD) database (Shutler et al. 2002), CMU HiD database (Gross and Shi 2001), UMD HiD database (Chalidabhongse et al. 2001), MIT HiD database (Lee 2003) and USF HiD database (Sarkar et al. 2005).

The Southampton HiD database contains 4824 video sequences from 115 people viewed from the side, walking at normal speed under both indoor laboratory and outdoor real-world environments. High-quality silhouettes have been extracted from the indoor data

through chroma-key techniques. The outdoor data are much more challenging because of the real-world noise caused, for instance, by the natural illumination, shadows and moving objects in the background. The database also contains indoor images of walkers carrying bags, rucksacks, wearing clothing such as long skirts or trenchcoats. This work uses subset of this database and more details of this database can be found in the following chapter.

The CMU HiD database contains sequences of 25 subjects walking on treadmills. For each subject, six sequences were taken simultaneously from different viewpoints. Subjects were asked to perform four types of walk: fast walk, slow walk, walking on an inclined surface and slow walk carrying a ball.

There are 55 subjects included in the UMD HiD database. Each subject walked a T-shaped pattern in a parking place and was filmed by two surveillance cameras from two orthogonal viewpoints. As a result, we can see the frontal, back, left and right sides of the walker in each sequence.

The data in the MIT HiD database are gathered from 24 subjects under an indoor laboratory condition. Subjects were asked to wear different clothing ranging from sweaters and long pants to T-shirts and skirts during the data collection sessions on different days. All video sequences were recorded by one camera placed perpendicular to the predetermined walking direction.

The USF HiD database (also called the gait challenging database) contains 122 subjects with 1870 outdoor video sequences. Sarkar et al. had each subject walk multiple times counterclockwise around each of two similar sized and shaped elliptical courses. Factors that could influence human gait were considered in this database including walking surface, viewpoints, shoes, carrying conditions and time.

2.3.2 MLD Experiments

Using gait as a biometric is largely motivated by Johansson's early psychophysical experiments (Johansson 1975). His experiments with moving light displays (MLD) attached to body joints of an actor showed that human observers could almost instantly recognise human motion patterns directly from several moving dots without any structural information, since they were not connected. After that, more experiments using MLDs were designed and reported. Cutting and Kozlowski (1977) showed that the gender of a person could be recognised from the moving dots. Moreover, Cutting et al. (1978) demonstrated from their experiments that people could even identify the gait of their friends by watching the moving dots. From the psychological point of view, there are two theories explaining the experiments mentioned above (Cëdras and Shah 1995). In the first, people recover a certain structure from the MLD type stimuli and then use the structure for recognition. The second theory states that the motion information

is directly used without recovering the structure. In terms of whether the structural information is used or not, there are also two classes of methods in the computer-vision domain for gait recognition: model-based and model-free.

2.3.3 Model-based Methods

Model-based methods attempt to recover a particular body structure from image sequences and use the structural information for recognition. The capture of such information usually involves searching a motion model. The model could be simply a stick-figure representing the skeletal structure or shapes connected articulatedly representing the whole or partial body of walking people.

Johnson and Bobick (2001) gave an example of using a 2D stick-figure to model a walking person. The whole body was modelled by four points that marked the positions of the head, pelvis and both feet. They used magnetic sensors to locate the four points on subjects walking in their laboratory environment. Four distances were calculated from the model (the vertical distance between the head and foot, the distance between the head and pelvis, the distance between the pelvis and foot, and the distance between the left foot and right foot) at the maximal separation point of the feet during the double-support phase of the gait cycle. Feature vectors were built from the distances for personal identification. Tanawongsuwan and Bobick (2001) also used sensors to find joints on a walking person in their experiment. Instead of using static information (distances between the detected points), we recovered the joint angles and used the angle trajectories to identify walkers. Dynamic time warping (DTW) was implemented to align and compare each pair of trajectories.

Cunado et al. (2003) built a simple pendulum model for the motion of the upper legs. An evidence-gathering process, namely the velocity Hough transform, was implemented to capture the model from image sequences automatically. The sequences were filmed in the indoor laboratory environment and subjects were asked to wear special clothing to help find the midline of the thigh. Fourier analysis was performed on the extracted thigh angles to get the amplitudes and phases of the harmonics of the first few orders. These numbers were then used to recognise walkers. Extending the above work, Yam et al. (2004) constructed a model for the motion of both upper and lower legs. The model is a stick-figure connecting three joints: the pelvis, knee and ankle. They used temporal template matching to extract the model from images and then performed Fourier analysis to built the gait signature. The system was tested to identify subjects walking or running on a treadmill by their gait. In Yoo et al. (2002), a stick-figure for the full-body motion of a walker was constructed for gait recognition. The model was found using the edges detected in images and the recovered joint-angle trajectories were then input to a back-propagation neural network for recognition.

Shape models have also been used in model-based gait recognition methods. In general, they include some geometric shapes representing the body parts of a walker. These shapes are connected to simulate the body structure. In Bissacco et al. (2001), a shape model consisting of elliptical texture patches, which were connected to form a kinematic chain, was formed to model a human skeleton. A linear dynamic system was identified from the joint-angle trajectories recovered from the model. They computed distances between sequences using the system parameters to recognise three kinds of gait: walking, running and going up and down a staircase. Bhanu and Han (2002) constructed a 3D shape model for the full-body motion. For each video sequence, they fitted the model to four selected key frames (silhouettes) and then used the extracted stationary parameters, which controlled the sizes of the body parts, for recognition.

More recent work using shape models can be found in Wang et al. (2003) and Wagg and Nixon (2004a). In the former, a sophisticated 2D articulated model was built for each of the subjects involved in their experiments. The model was fitted to images through a probabilistic framework: the condensation algorithm. They also carried out the shape analysis described in Boyd (2001) on silhouettes subtracted from video sequences. The extracted joint-angle trajectories were normalised by DTW and used for identification. Wagg and Nixon presented a model-based gait extracted system guided by the biomechanical analysis of walking people. They represented the head and torso with two ellipses and outlined legs with parallel lines. These shapes were extracted from a global temporal accumulation of the given image sequence in a coarse-to-fine way.

2.3.4 Model-free Methods

Model-free methods focus on using the motion information directly without recovering any human body structure. Very often, such information is the monochromatic silhouettes or motion flow extracted from video sequences. Various gait features have been calculated from the motion information.

One way to do that is to measure the scalar information (i.e., width, height, centroid, etc.) of silhouettes or flows. A typical example is the pioneering work done by Little and Boyd (1998). They computed the dense optical flow from video sequences and measured various scalar features of the flow. They then calculated phase features from the sequences of the scalars and used the feature vectors to recognise walking people. Lee and Grimson (2002) presented a similar gait-recognition system, which was tested on a much larger gait database. Instead of using optical flow, they calculated scalar features directly from silhouettes. Each silhouette in which a walking person had been centred was divided into seven regions. An ellipse was fitted to the pixels in each region. They then found the dominant walking frequency and computed first-order Fourier features (amplitudes and phases) from the time series formed by the scalars describing those ellipses. The identification of walkers was performed by classifying the feature vectors.

BenAbdelkader and Davis (2002) built a system that detected objects carried by walkers. They first extracted silhouettes from video sequences and found the bounding box for the foreground pixels. They then defined five regions in a boundary box and computed the maximal width of the pixels in each region. The width signals were used to detect people carrying objects. In Kale et al. (2004), the widths of a silhouette at each row were chosen as the features for recognition. An HMM was built to and trained for each walker in the database to classify the feature vectors. In a testing sequence, they computed the likelihoods of the sequence given each of the HMMs. The larger the corresponding likelihood was, the more likely the walker appeared in the sequence.

Generating a spatiotemporal pattern from an image sequence for the purpose of recognition is another important way of using the motion information without a model. Early work can be found in Niyogi and Adelson (1994). They stacked frames of an image sequence to form a cube (XYT image cube). The pattern (XT-slice) was obtained by slicing the cube at a particular place on the Y-axis. They found that the XT-slice of a walker at the ankle contained the braided walking pattern. The pattern was then used to track and recognise people in images. Cutler and Davis (2000) introduced self-similarity plots to detect and identify periodic motion in video sequences. Each row or column in the plot corresponds to a frame in the sequence and the values in the plot gave the correlations between each pair of frames. They showed that periodic motion had its own characteristic patterns in the self-similarity plots. Moreover, the experiments for walking/running humans and walking/running dogs suggested that different periodic motions have different spatiotemporal patterns. Later on, the self-similarity plots were used to identify walking individuals in video sequences (BenAbdelkader et al. 2002). Principal component analysis was applied to reduce the dimensionality of the calculated self-similarity plots. Personal identification was then done by a k-nearest neighbour classifier. Bobick and Davis (2001) introduced an appearance based approach to recognise different human movements. They created temporal templates consisting of motion-energy images and motion-history images for various movements. The Hu moments were computed as discriminatory features and the Mahalanobis distances between feature vectors were used to recognise different motion.

There is also some work using motion information other than the methods described in the previous paragraph. Boyd (2001) developed a system that measured the phase information at each pixel through a phase-locked loop. He showed that the patterns generated by the output phases had certain discriminatory capabilities to distinguish individuals by their gait. Shutler and Nixon (2001) computed the Zernike velocity moments to describe the motion of a walking person. Hayfron-Acquah et al. (2003) analysed the symmetry of walking motion using the generalised symmetry operator. The operator was performed both on silhouettes and optical flows. They defined the gait signature as the average of the symmetry maps and quantified the similarities using the Euclidean distance between the Fourier descriptions of each pair of signatures.

Recently, Sarkar et al. (2005) introduced a very simple baseline algorithm for human gait classification and tested it on a set of challenging data. The algorithm performed recognition by computing the temporal correlations of silhouettes.

2.3.5 Discussion

We have described the two categories of methods (model-based and model-free) for gait-related problems. We can see that the model-based approaches involve a shape model representing the skeletal structure of the human body. Very often, the parameter space of the model is constrained based on our knowledge or assumption of the walk. We then search the set of parameters that make the shape model best fit the walkers in the images. Such a mechanism gives a model-based system the advantage that we can build our prior knowledge about human gait into the system naturally, which potentially makes the system robust to outside noise. However, the fitting of a sophisticated model could be computationally inefficient and sometimes bring difficulty in finding optima in a high-dimensional space. Using model-free methods, on the other hand, has an obvious advantage, that is, the gait features can be computed from images efficiently. The disadvantage is that most of the methods are appearance-based. Such a characteristic means that the systems could be very sensitive to the quality of the image sequences, which makes the techniques unreliable for real-world applications.

2.4 Summary

In this chapter, we have reviewed some of the previous work related with the topic of this thesis. In Section 2.1, we described approaches for the extraction of general human body motion. Some of them obtain the motion by computing the posterior probabilities of the model parameters given an observation in the state space. We discussed the two common frameworks for the propagation of the probability distributions over time: the Kalman filter and the particle filter. Other methods involve minimising the error between the synthesised model and the observed object, which can be computed by a nonlinear measurement function. The optimal state can be either obtained by the inverse kinematics that approximates the gradient information of the measurement function by linearisation or by searching directly in the state space. In Section 2.2, we described the systems extracting a particular type of human motion, i.e., walking. Articulated models and contour models are both used to capture the motion of walking people. We gave some examples of using these two kinds of models in this section. Discussion of the advantages and disadvantages of using the models was given in the end of the section. We then gave an overview of the existing gait recognition technology. We first described some of the established gait databases and then the early MLD experiments which

motivated greatly the current research. Gait systems were classified as model-based and model-free, and described in detail afterwards.

Chapter 3

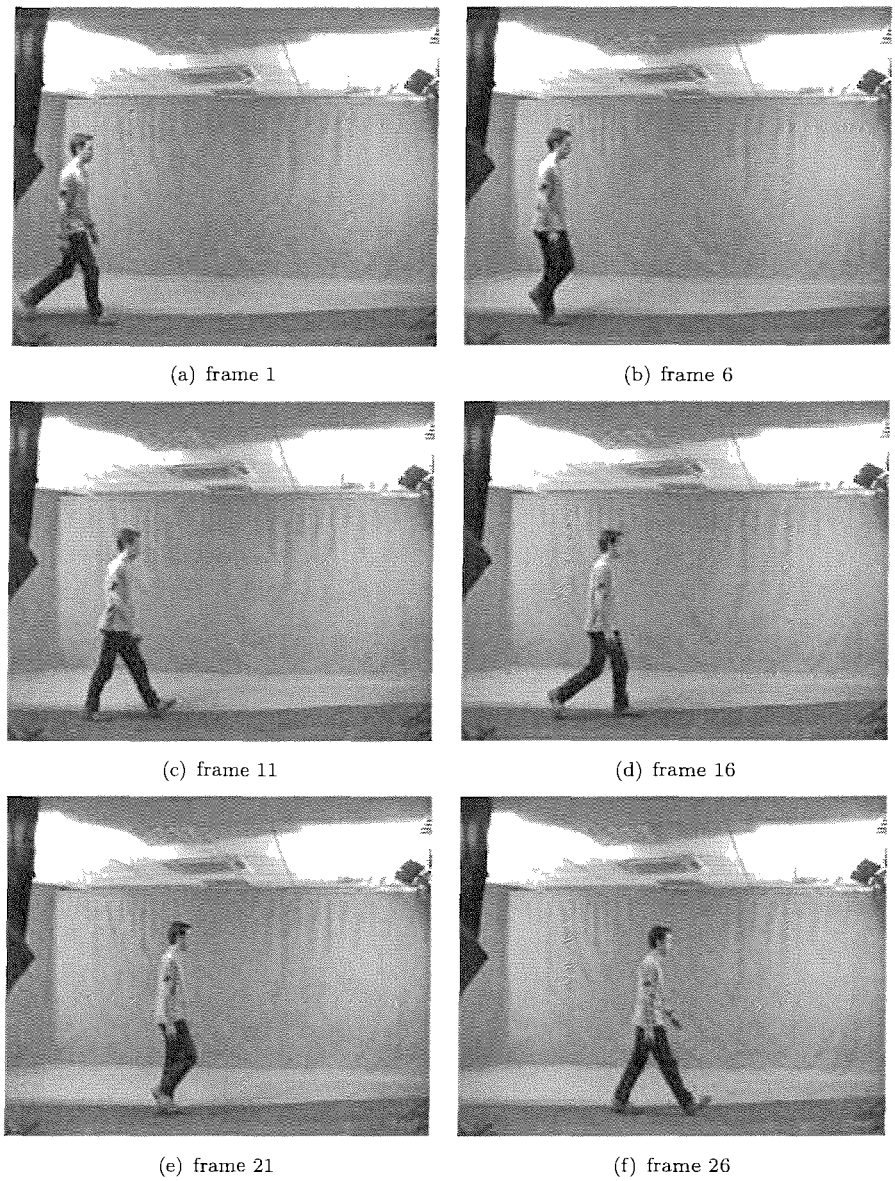
Data Description and Preprocessing

This research has used a subset of the Southampton *human identification at a distance* (HiD) database (Shutler et al. 2002). It consists of both high-quality (indoor) data and lower quality outdoor data, representative of a real application. There is also supplemental data of some of the walkers carrying bags, wearing coats, skirts, etc. Our concern is to devise a system that is capable of extracting gait information from image sequences at least as challenging as the outdoor and supplemental data, exploiting the high-quality indoor data for initial learning only. Rather than taking raw color images as input, our algorithms are designed to work with simple extracted silhouettes. All silhouettes are normalised to ease the following procedure of extracting gait.

3.1 The Southampton HiD Database

The Southampton HiD database has been designed and built to provide a large multi-purpose dataset enabling the investigation of gait as biometric as well as other sequence-based vision applications. It contains sequences of just over 100 walkers, viewed from the side and filmed at 25 frames per second. There are different kinds of data stored in the database including: indoor, outdoor and supplemental image sequences.

Indoor Sequences: These were filmed under laboratory conditions with a high-quality camera, controlled lighting and a constant green background to facilitate silhouette extraction. These data are clearly not representative of a real application scenario. For our purposes, they are treated as initial training data for learning typical body shape and motion parameters, and their variations. Figure 3.1 shows a sample sequence filmed indoor.



(a) frame 1 (b) frame 6

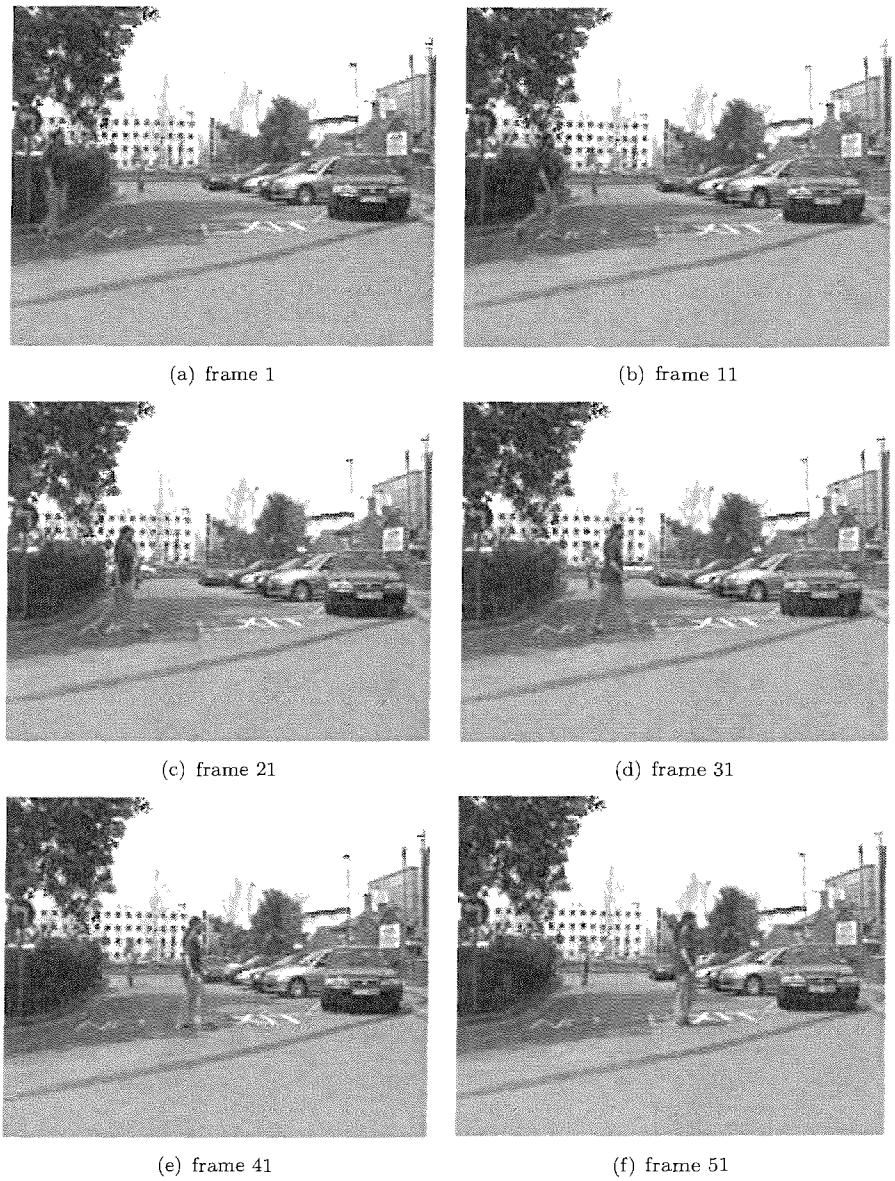
(c) frame 11 (d) frame 16

(e) frame 21 (f) frame 26

FIGURE 3.1: An example of indoor video sequences. The sequence is labelled by its walkerID = 001 and sequenceID = 01R in the indoor image data.

Outdoor Sequences: To test the potential of our Bayesian framework on data more representative of a practical application, we used the outdoor image sequences in the HiD database. These images are affected by changes in illumination, motion of trees, passers-by and cars, and ambiguous colour contrasts between the walker and background. Examples are shown in Figure 3.2.

Supplemental Data: The database contains supplemental images of walkers carrying bags, rucksacks, wearing clothing such as long skirts or trenchcoats which obscure the legs, etc. Some of these have been used to test the Bayesian framework. Although collected under laboratory conditions, these represent difficult data, which stretch the



(a) frame 1 (b) frame 11

(c) frame 21 (d) frame 31

(e) frame 41 (f) frame 51

FIGURE 3.2: An example of outdoor video sequences. The sequence is labelled by its walkerID = 004 and sequenceID = 00R in the outdoor image data.

methods developed in a different way from the outdoor data. Examples of such data are given in Figure 3.3. The top row shows walkers carrying a rucksack and the middle and bottom rows are images of walkers wearing, respectively, a long skirt and a trenchcoat, where the motion of limbs was severely occluded.

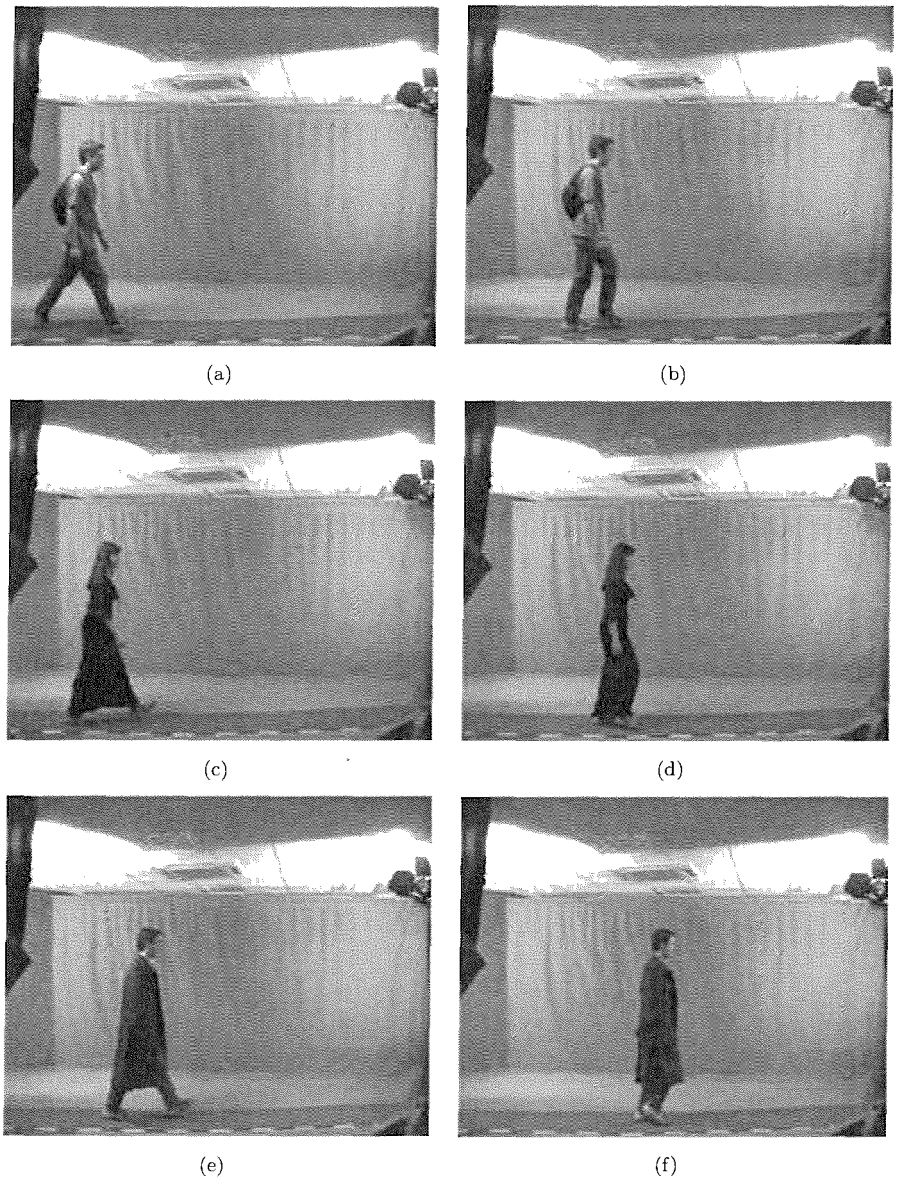


FIGURE 3.3: Examples of supplemental image data: rucksack (a, b), long skirt (c, d), and trench coat (e, f).

3.2 Silhouette Extraction

For indoor image data, high-quality silhouettes were obtained using a chroma-key technique. Figure 3.4 shows examples of the silhouettes extracted from the indoor image data. It can be seen that the foreground pixels (walkers) were extracted nearly perfectly from the pure coloured backdrop and the effect of shadows was significantly reduced.

Silhouettes for the outdoor images were produced by background subtraction (Grant et al. 2004). Figure 3.5 shows typical examples. It can bee seen that the quality of silhouettes largely depends on the environments when filming walkers. For example, in

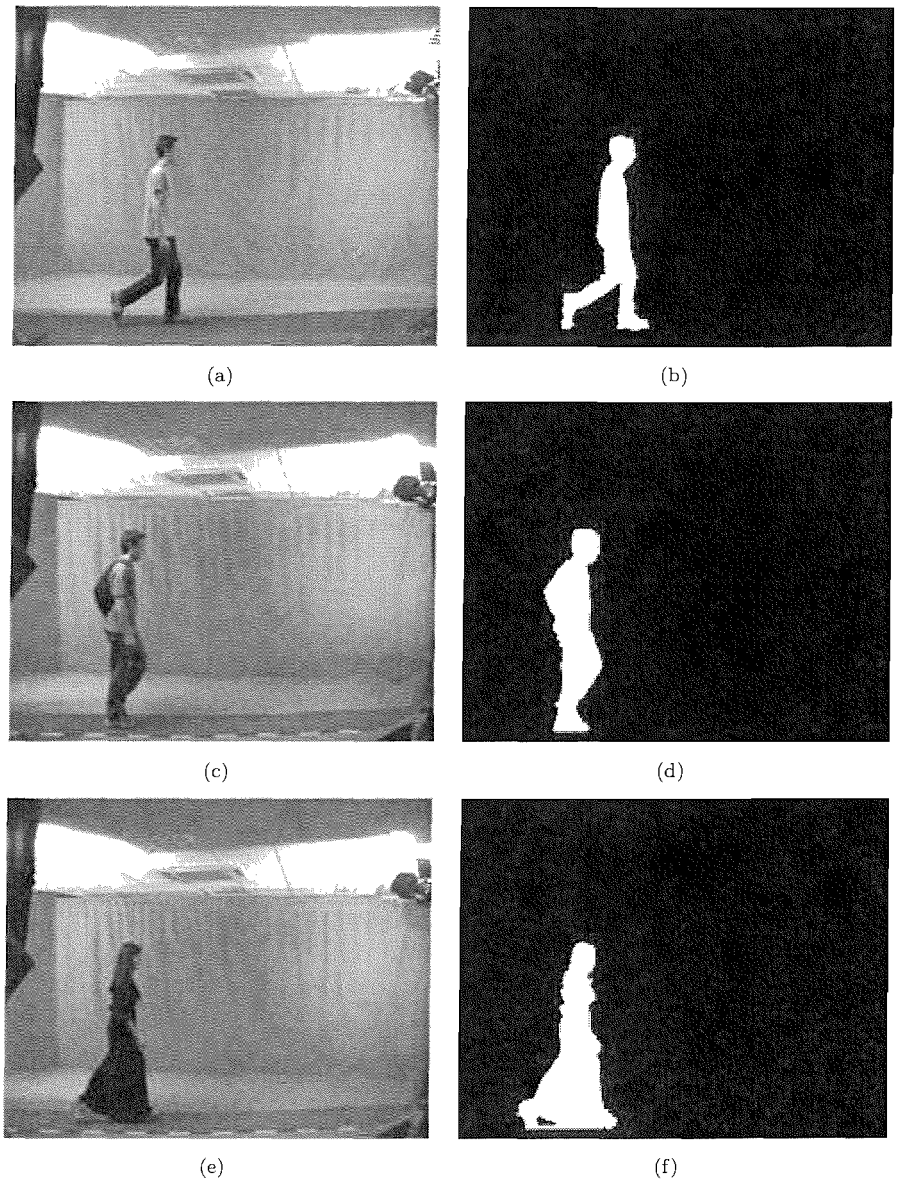


FIGURE 3.4: Examples of silhouettes extracted from indoor image data. The left column shows the raw colour images (a, c, e) and the right column the extracted silhouettes (b, d, f).

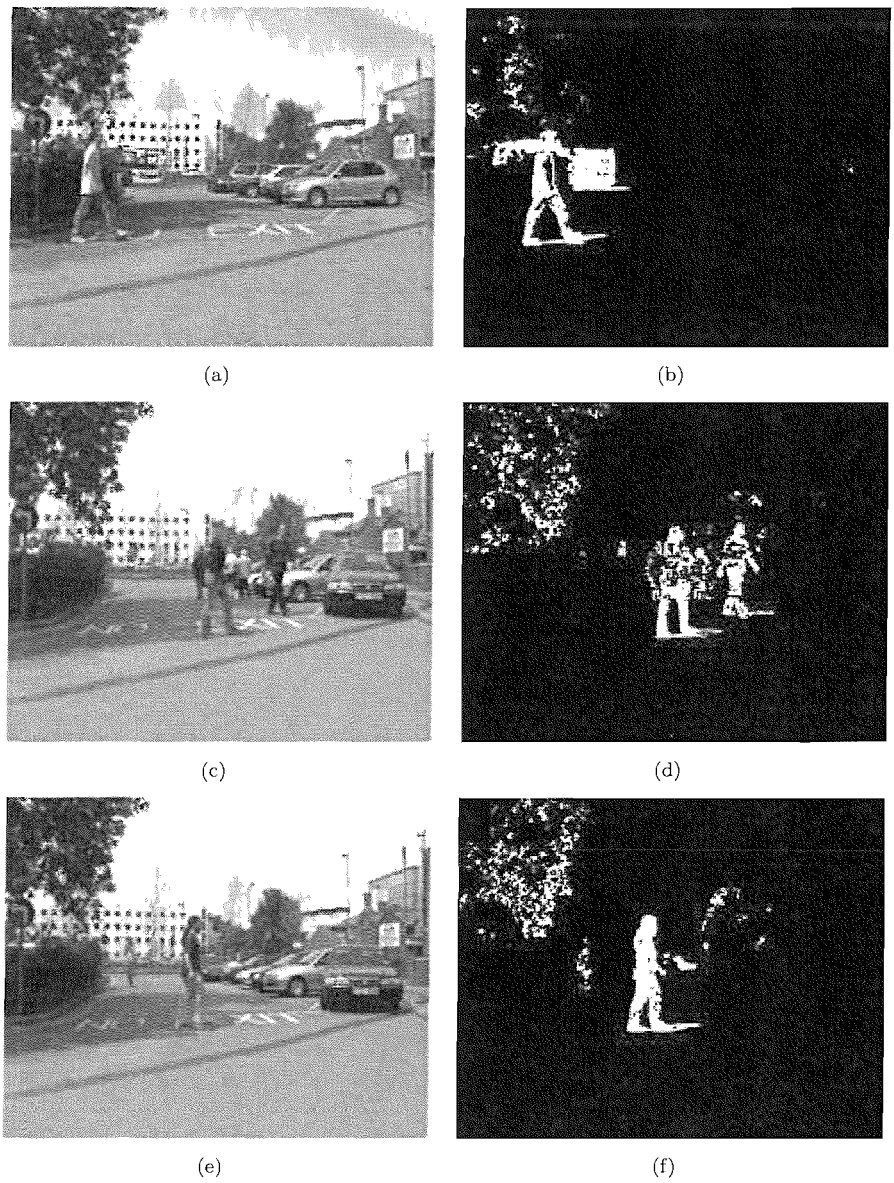


FIGURE 3.5: Examples of silhouettes extracted from outdoor image data. The left column shows the raw colour images (a, c, e) and the right column the extracted silhouettes (b, d, f).

Figure 3.5(a), a passing bus caused a big block of noise in 3.5(b) while the relatively less cluttered background in Figure 3.5(e) resulted in a better silhouette obtained in Figure 3.5(f). Note that in Grant et al. (2004), further operations were carried out to clean and repair the silhouettes after background subtraction. To explore the potential of the Bayesian framework with challenging data, no such post-processing was implemented in our work.

3.3 Silhouette Normalisation

To simplify the subsequent processing of gait extraction, silhouettes are normalised (i.e., centered in each image) before being input into the framework. Although normalisation could be simply and straightforward for the indoor data, the noise inherent in the outdoor data dictates the use of a relatively more sophisticated approach. For this purpose, we use an evidence-based tracking algorithm described by Lappas et al. (2002), who extended the dynamic Hough transform (GHT) to detect arbitrary shapes undergoing arbitrary affine motion. Details of the GHT can be found in Appendix B. In this section, we first describe their method briefly and then give examples of the normalised silhouettes for outdoor data.

The tracking algorithm processes the whole image sequence globally and the optimal object trajectory is found by maximising its associated energy. For a sequence with length T, an object trajectory consists of a set of points $\{(x_t, y_t)\ 1 \leq t \leq T\}$ where (x_t, y_t) locates the object to be tracked in frame t. The speed V_t and orientation ϕ_t are computed for the point (x_t, y_t) in frame t as:

$$V_t = \sqrt{(x_{t-1} - x_t)^2 + (y_{t-1} - y_t)^2} \tag{3.1}$$

$$\phi_t = \arctan\left[\frac{(y_{t-1} - y_t)}{(x_{t-1} - x_t)}\right] \tag{3.2}$$

where (x_{t-1}, y_{t-1}) is the location in the previous frame $t-1$. The energy function E_{traj} of an object trajectory consists of two terms, the Hough energy E_{Hough} and the motion energy E_{motion}, and can be expressed as:

$$E_{\text{traj}} = w_1 E_{\text{Hough}} - w_2 E_{\text{motion}} \tag{3.3}$$

where w_1 and w_2 are weights that can be adjusted to vary the relative importance of each term. In their original paper, there is a third term called deformation energy standing for the smoothness of the changes in scale and rotation. Since walkers in the HiD database were viewed and filmed from the side, we assumed no deformation energy here and therefore omitted the third term. The Hough energy is defined as:

$$E_{\text{Hough}} = \sum_{t=1}^{T} p_t \tag{3.4}$$

where p_t is simply the peak value at (x_t, y_t) obtained after fitting a template to frame t using the generalised Hough transform (GHT) (Ballard 1981). This term forces the

trajectory to pass through the points with maximum structure evidence. The motion energy represents the elasticity and rigidity of the trajectory, and has the form:

$$E_{\text{motion}} = \sum_{t=2}^{T-1} |V_{t-1} - V_t| + \sum_{t=2}^{T-1} |\phi_{t-1} - \phi_t| \tag{3.5}$$

where the first term penalises the large changes in speed and the second penalises the large changes in direction. The optimisation is achieved using temporal dynamic programming.

In our work, the template used for the Hough transform is the very rough contour of the upper body (head and torso) formed by averaging across walkers. To generate the template, we chose 5 indoor sequences for each of the 50 walkers in our gait database. We positioned the walking subjects in the centre of the images by calculating the centroid of silhouettes. We averaged all the silhouettes and the template was obtained by the edge detection on the mean silhouette. A limited number of templates with different sizes were generated and the optimal trajectory found for each. The overall best trajectory tells us the locations of the walker in the sequence, and the size of the optimal template is used to normalise the silhouettes. Figure 3.6 illustrates the normalisation for an outdoor sequence. Fig. 3.6(a) shows the average template; the sequence depicted in Fig. 3.6(b) shows the position found by the tracking algorithm with the optimal template superimposed. (Note that the polarity of the silhouette has been inverted to show the superimposed template more clearly.) As seen, the walker is reasonably located even in the presence of a passing bus. Finally, Fig. 3.6(c) shows the bounding box obtained by re-centering the walker. Because walkers in the indoor images are large (in terms of number of pixels) relative to the outdoor images, they were reduced to fit in a (70×70) bounding box. The outdoor images were simply cropped to (120×120) pixels. Images can be smaller in the former case as (70×70) images were found adequate for bootstrapping Bayesian learning, which is the main purpose of these data.

To improve the computational efficiency of the tracking algorithm for noisy outdoor data as in Figure 3.4 , we introduce some heuristic spatial constraints to define a reasonably small search window in each frame for GHT, so as to search for the position of the walking subject in a small restricted area instead of the whole image area. In the remainder of this section, we give the details of how we impose these constraints.

Horizontal Constraints: Based on the fact that subjects walked normally when being filmed in the gait database, we assume that the walking speed of a walker remains approximately unchanged. The pixel-wise speed is then computed by dividing the width of an image by the number of frames of a sequence. Therefore, we can know roughly the horizontal position of a walker in a particular frame (by "roughly", we mean within an interval of 40-pixel width).

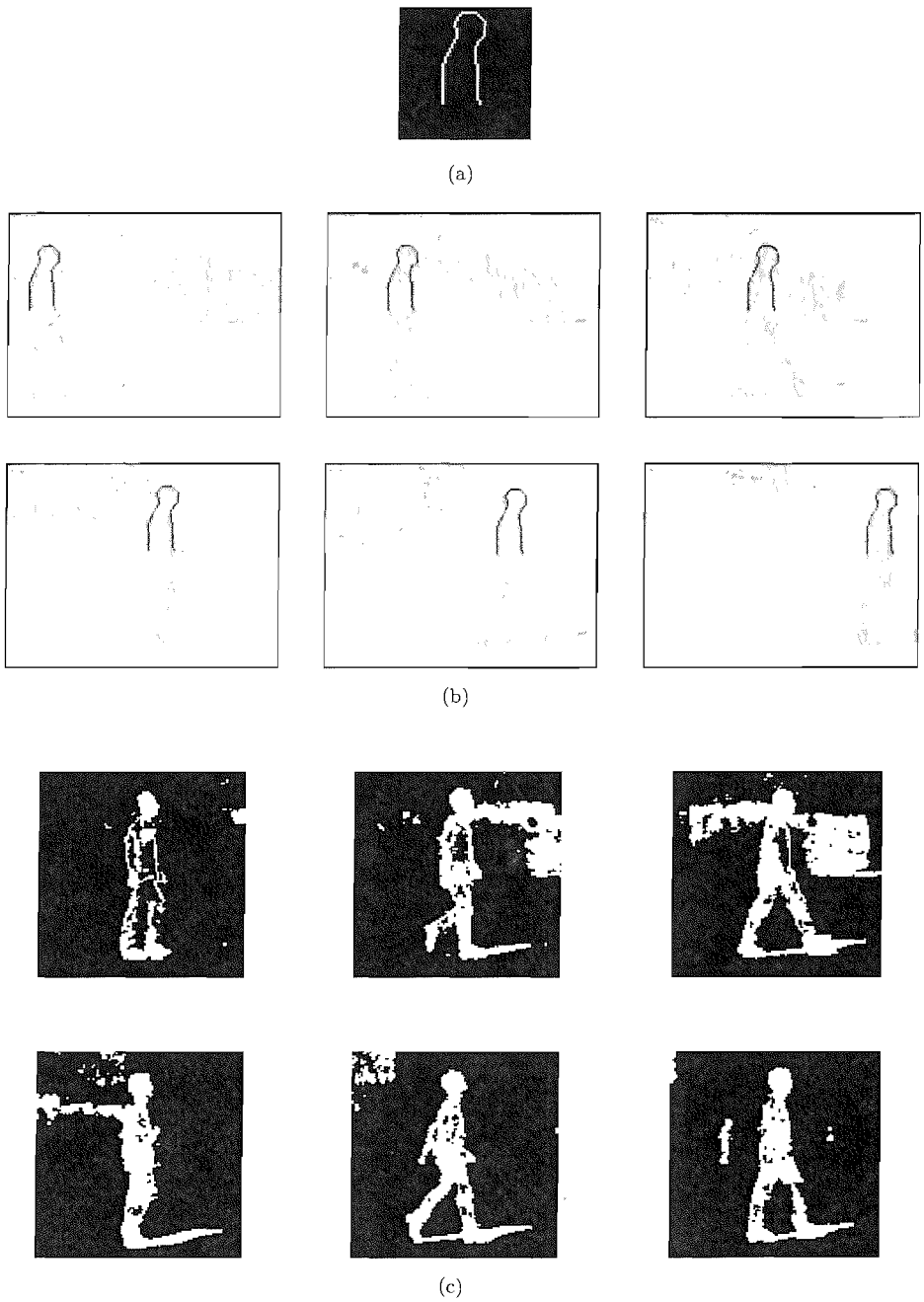


FIGURE 3.6: Example of normalisation of an outdoor image sequence: (a) shows the average template somewhat enlarged relative to the other sequences; (b) shows a typical silhouette sequence with the best-fit position of this template superimposed; (c) shows the final (120 × 120) pixel bounding box obtained.

Vertical Constraints: To limit the search window on the perpendicular direction, we try to locate the head of the walking subject in a vertical interval. For silhouettes with a clean background, it is rather simple to do so by scanning each row of the silhouettes and counting the number of foreground pixels in a row from top to bottom. The position of the head can be located at a point where the number changes significantly. However, for noisy silhouettes, we can not implement the above method because many foreground pixels might be just noise. Also we might miss some pixels belonging to the walker in the silhouettes.

To reduce the extent of noise, we use the difference between two consecutive frames to detect the location of the head. Given sequence $\boldsymbol{X} = \{\boldsymbol{x}_t\}_{t=1}^{T}$, for two consecutive frames $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$, the different image $\hat{\boldsymbol{x}}_t$ is defined as the difference between $\boldsymbol{x}_{t-1}$ and $\boldsymbol{x}_t$. Figures 3.7(a) and 3.7(b) show two consecutive frames. Their difference image $\hat{\boldsymbol{x}}_{36}$ is given by Figure 3.7(c). We can see that the noise at the background is significantly reduced and the motion of the walking subject is captured by the pixels different in the two silhouettes.

Since we have already had the horizontal constraints, we only have to consider the pixels in a small horizontal interval. We then chop a sub-image $\hat{\boldsymbol{x}}_t^*$ with 40-pixel width, which is the length of the corresponding horizontal search window, from the difference image $\hat{\boldsymbol{x}}_t$. Figure 3.7(d) shows the chopped sub-image $\hat{\boldsymbol{x}}_{36}^*$ from the difference image on the left. In the sub-image, we can see that the majority of the pixels describe the motion of the walking subject. We then count the number of pixels at each row for $\hat{\boldsymbol{x}}_t^*$ to see if we can judge the head position from these numbers. Figure 3.7(e) shows the histogram of the number of pixels at each row for $\hat{\boldsymbol{x}}_{36}^*$. It is clear that the value changes dramatically at the head position. Since the vertical position of the head of the walker does not change much in a sequence, we count the number of pixels over the whole sequence, that is, add the numbers from all $\{\hat{\boldsymbol{x}}_t^*\}_{t=1}^{T}$. We denote the total number at row r as y_r. Figure 3.8 shows the histogram for these numbers computed for the sample sequence in Figure 3.7. The shape of the histogram in Figure 3.8 is similar to that shown in Figure 3.7(e), but much smoother. We believe that it is more robust to use the total number of different pixels at each row to detect the head position.

To find the jump in the histogram, we fit a sigmoid curve f_{sig}. The curve is defined as:

$$f_{\text{sig}}(x; a, b, c) = \frac{a}{1 + \exp(-b(x + c))} \tag{3.6}$$

where a, b and c are the parameters controlling the shape of the curve. We try to find a set of parameters (a^*, b^*, c^*) which make the sigmoid curve best-fit the histogram. This is done by maximising sum of squares $\mathcal{E}(a, b, c) = \sum_{r=1}^{R} \left(y_r - f_{\text{sig}}(r, a, b, c)\right)^2$ where r is the row number and y_r is the number of pixels at row r. In 3.8 the best fitted sigmoid curve is shown by the solid line and we can see that it fits the histogram well. The

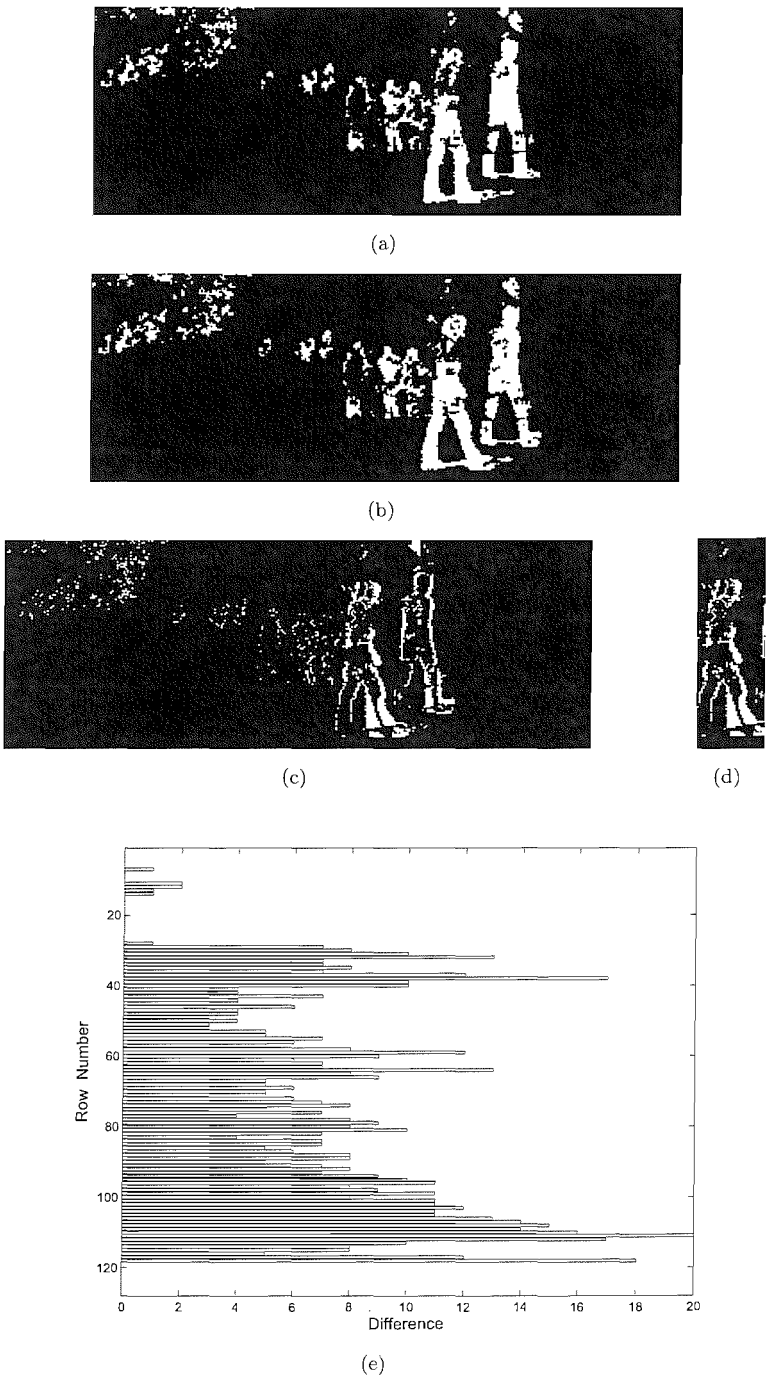


FIGURE 3.7: (a) Frame 35, (b) Frame 36, (c) difference image $\hat{\boldsymbol{x}}_{36}$ between (a) and (b), (d) chopped sub-image $\hat{\boldsymbol{x}}^*_{36}$ from $\hat{\boldsymbol{x}}_{36}$ and (e) histogram of the number of pixels at each row of $\hat{\boldsymbol{x}}^*_{36}$. The sequence is labelled in the gait database as: Walker ID = 002 and Sequence ID = 01R.

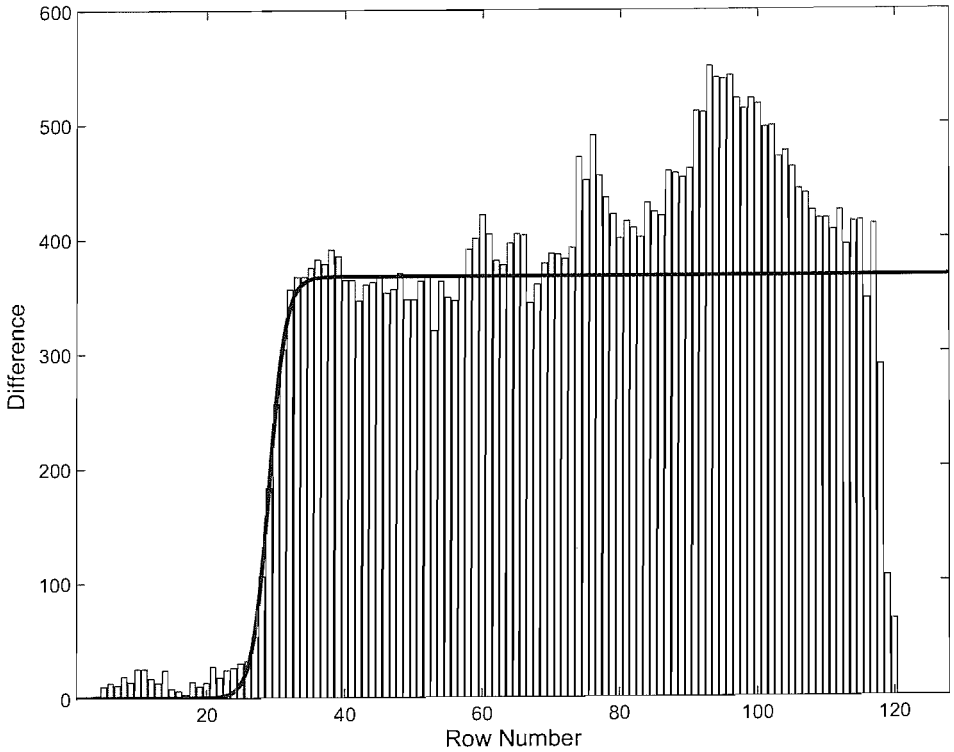
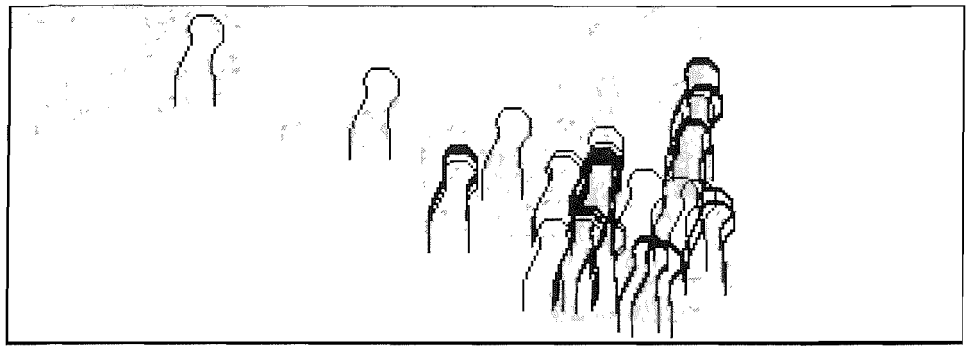


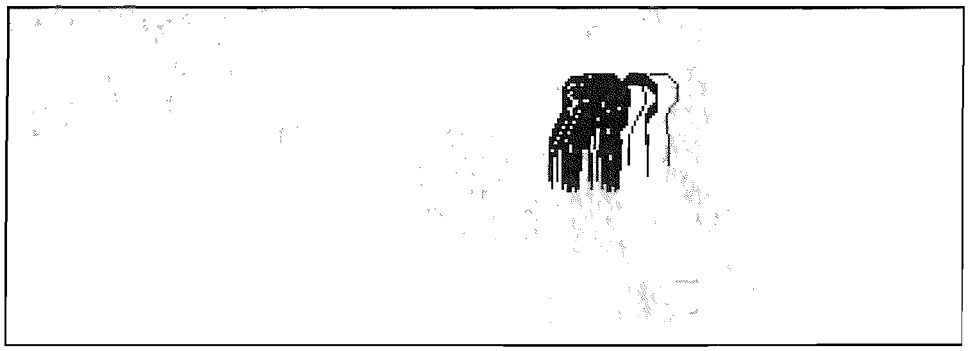
FIGURE 3.8: Histogram of the total number of different pixels for each row over the whole sequence and the best fitted sigmoid curve. The sequence is labelled in the gait database as: Walker ID = 002 and Sequence ID = 01R.

vertical interval is then defined as $\left[\text{round}\left(-c^* - \frac{b^*}{2}\right) - 10, \text{round}\left(-c^* - \frac{b^*}{2}\right) + 10\right]$ where function round() converts a decimal to the closet integer.

The object tracking algorithm in Lappas et al. (2002) requires the implementation of the GHT to frames in a sequence to provide candidates of positions of the object. In our work, the first $N = 50$ positions with the largest peak values are chosen as candidates from which the optimal position is determined for each frame. Figure 3.9 shows a sample frame fitted by the template using the GHT. In Figure 3.9(a), the template was fitted without any spatial constraint. The upper body of the walker was very noisy and not he was properly located by the template. In comparison with the results in Figure 3.9(b) using constraints the spatial constraints significantly improved the fitting performance. We chose 20 sequences randomly from total 500 outdoor sequences and checked the normalised silhouettes by eye. For all of the 20 sequences, the walkers were properly positioned in the images.



(a)



(b)

FIGURE 3.9: Results of fitting the upper-body template to a sample silhouette using Hough transform. The foreground pixels are displayed in gray while the template in black. The first $N = 50$ positions with largest peak value are shown in both images. (a) Without spatial constraints. (b) With spatial constraints.

3.4 Summary

In this chapter, we have described the image data used in our work and discussed how we pre-process the images. The Southampton Hid database provides us various kinds of gait sequences including high-quality indoor data, lower quality outdoor data and challenging supplemental data. The indoor data from which high-quality silhouettes can be obtained by a chroma-key technique are used as the training data in our Bayesian framework to learn the priors. On the contrary, the outdoor sequences containing real-world noise are used to test the ability of the framework to handle noise. To prove our system's virtues of consistency and flexibility, the framework will be adapted to the supplemental data that include unusual gait data, such as walkers carrying bags, wearing long skirts and trenchcoats, etc., with only minor changes.

All silhouettes extracted from raw colour images are normalised before being input to the framework. We have implemented the tracking algorithm described in Lappas et al. (2002) to locate the upper body of the walker in images. The utilisation of the generalised Hough transform and dynamic programming makes the algorithm computational efficient. To improve the efficiency further, we introduce some spatial

constraints to make the normalisation more robust. In the next chapter, we will describe the whole Bayesian framework, which is the key part in this thesis, in detail. It will be seen that the framework encodes successfully our prior knowledge of human walking to cope with the real encountered noise in a Bayesian formulism. Moreover, each component of the framework is relatively dependent from others, making it flexible and extensible for various situations.

Chapter 4

A Bayesian Framework for Gait Extraction

Extracting human gait from real-world images poses a severe challenge for a computer vision system. The motion itself is highly articulated, which means the walking object have complex changing boundaries. In real-world situations, the motion will occur in a cluttered scene making segmentation difficult and ambiguous. Moreover, the large variations of the body appearance of walkers make the problem even more difficult. For example, carrying a rucksack changes the shape of the upper body of a walker dramatically and wearing a long skirt causes severe occlusions of legs. However, human gait is also a domain that has been well studied and therefore contains strong prior knowledge. The complexity of the problem demands a system with strong capability of handling uncertainty. Meanwhile, it should be able to incorporate existing knowledge of human gait to improve extraction. A Bayesian framework fits the requirements well. Such a framework allows us to combine observations (what we can see) and prior knowledge (what we know) systematically and to model the uncertainties in a probabilistic way.

4.1 Overview

The problem we try to cope with is to extract human gait from image sequences with real-world noise or walkers with different body configurations. We decompose this complex problem into sub-problems and each of the sub-problems is solved by some simple component. Figure 4.1 shows the components and the way they are combined to extract human gait from image sequences. It can be seen that the framework has a simple structure and the components themselves are quite simple and based on some well-established techniques. We regard this simplicity as a virtue, since the components can be easily extended and modified to cope with various data (e.g., images with walkers

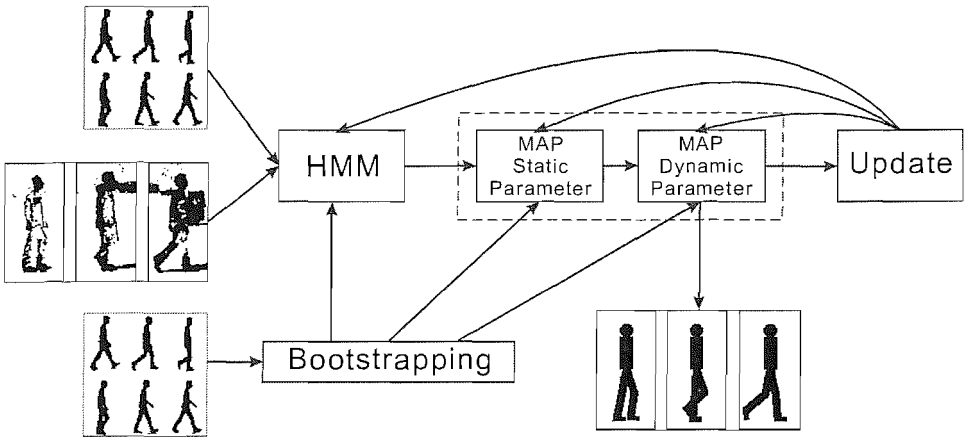
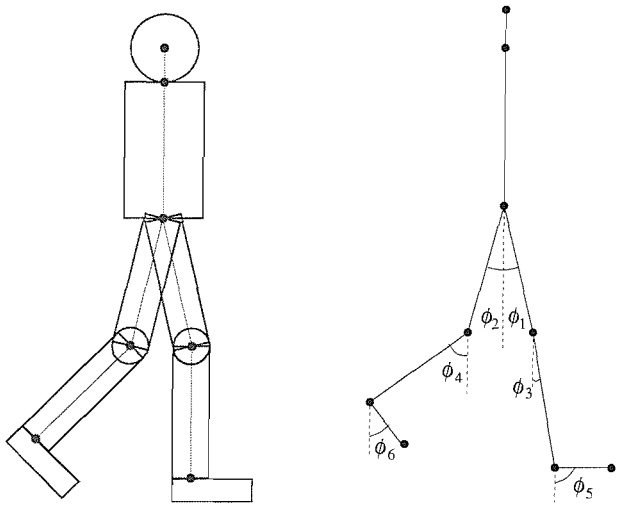


FIGURE 4.1: Framework structure showing how human gait if extracted from image sequences. There are four components within the framework: 1) a hidden Markov model learning the phases of images in the gait cycel; 2) a maximum a posteriori component optimising the parameters (including static and dynamic parameters) of a model fitted to walkers in images; 3) a Bayesian updating component refining system parameters; and 4) a component bootstrapping system parameters.

carrying a rucksack or wearing a long skirt etc). In this work, we give an example of how to build a consistent, extensible and pricinpled Bayesian framework. The key contribution here is to illustrate the advantages of deploying an appropriate mix of strong prior knowledge and simple but powerful learning methodologies in just the way that the Bayesian framework allows.

To define strong prior knowledge of shapes and movement of humans, we build a single articulated model. This is described in Section 4.2. We try to fit the model to walking individuals in the images and extract their gait by the parameters of the model. Since the variations of the pose of a walker are large, it makes fitting the model difficult as there is a large region in the parameter space corresponding to feasible walker models. To ease this problem, we use a hidden Markov model to learn the phase of images in the gait cycle. The new PDF projection theorem (see Appendix D) is novelly used to learn observation probability distributions in this gait system. Sections 4.3 and refsec:Framework-pdf describe the HMM component and the implementation of the PDF projection theorem in the framework. After the HMM decoding, we try to solve a maximum a posteriori problem. In Section 4.5, we define the posterior probability to be maximised. There are two kinds of parameters of the articulated model: the static parameters that define the sizes of the body parts and the dynamic parameters controlling the post of the model. They are optimised separately using different strategies (see Section 4.6). A bootstrapping component is constructed to estimate the system parameters from a small amount of indoor training data. Section 4.7 describes the two components. In the last section, we extend the framework to handle different body configurations.



(a) (b)

static parameters	dynamic parameters (angle)
head radius τ_1	front thigh ϕ_1
torso width τ_2	back thigh ϕ_2
torso length τ_3	front calf ϕ_3
leg width τ_4	back calf ϕ_4
thigh length τ_5	front foot ϕ_5
calf length τ_6	back foot ϕ_6

(c)

FIGURE 4.2: The basic articulated model of a walker: (a) shows the body parts; (b) defines the various joint angles; and (c) lists the model's static and dynamic parameters. The arms are omitted in an attempt to match the complexity of the model appropriately to the available data.

4.2 Articulated Model

A cornerstone of our approach is the exploitation of strong prior knowledge of human walkers and walking. The most basic level at which this knowledge is introduced is the articulated walker model. Both two-dimensional and three-dimensional articulated models have been used to model walking people in some applications (e.g., Hogg 1983; Rohr 1994; Ju et al. 1996; Ning et al. 2002; Lan and Huttenlocher 2004). Most of the shapes having been used to build an articulated model are surveyed in Gavrila (2000). Instead of a sophisticated three-dimensional model, we build a very simple two-dimensional articulated model using three circles and seven rectangles for walkers viewed from the side. Figure 4.2 shows the model used and lists the parameters that control it. The basic model has 12 parameters which divide into two groups: those determining the sizes of the body parts which remain constant for all images in the sequence; and the angles between the body parts, which vary from frame to frame. We refer to these as *static* and *dynamic* variables respectively.

Clearly, the model is only a crude approximation to a real walker. No account is taken of perspective, parts of the body such as the neck, arms and hands are missing, foot lengths and widths are fixed across all sequences, and there is no distinction made between the left and right sides of the body. We rather distinguish between the front and back legs alone. Thus, our definition of a gait cycle is half the length of the one defined in Murray (1967). These simplifications reduce computational complexity and, more importantly, reflect our view of matching model complexity to the available data. For example, there is unlikely to be sufficient information in the outdoor images (see examples shown in the previous chapter) to be able to fit details such as the arms. (We have explores the use of arms in our articulated model, but results were no better than those reported later in this thesis.) Furthermore, we avoid possible ambiguities in fitting the model arising from having to determine which leg is which.

Note that the basic model can be easily extended to cope with large variations of the appearance of a walker caused by carrying bags or wearing some special clothes. In our work later reported in this thesis, the model is generalised by the addition of a rucksack, a long skirt, and a trenchcoat respectively. The rucksack is represented by a half ellipse, a long skirt by filling the gaps between two legs, and a trapezoid is added to stand for a trenchcoat. The overall cost is no more than two static parameters added and no other changes for the framework.

Gait information is extracted by finding the best set of model parameters to fit any given silhouette. In the Bayesian framework, this means determining the likelihood of the image given the model. To achieve this, we generate from the model a silhouette of the appropriate size. This 'model silhouette' is then matched against the observed data silhouette. In the following, we denote the set of parameters of the articulated model as $\boldsymbol{\theta}$ and the model silhouette as $\mathcal{I}(\boldsymbol{\theta})$.

4.3 Locating Phase in the Gait Cycle

As human walking can be considered approximately periodic (Murray 1967; Cunado et al. 2003), we can think of the dynamic parameters of the articulated model (i.e., joint angles) as a strong function of the phase within the gait cycle. The Bayesian framework exploits this information by finding which part of the gait cycle an image comes from. To automate this, we use a hidden Markov model (HMM) since it provides a natural framework of processing sequential stochastic data.

In the rest of this section, we will describe the issues of how we design the HMM fitted in our gait-extraction problem. In Section 4.3.1, we give a brief introduction to HMMs. The construction of the HMM is described in Section 4.3.2 including the whole HMM architecture and the learning of the transition probabilities and initial probabilities. Obtaining the observation probability distributions is not straightforward

in this case. We describe how to use the new PDF projection theorem to measure these high-dimensional unknown distributions in the next section.

4.3.1 An Introduction to Hidden Markov Models

A hidden Markov model describes a special Markov process whose random variables are not observable (hidden), but can only be observed through another stochastic process which produces a sequence of observations. See Rabiner (1989) for more detailed description of HMMs. Before describing the components of an HMM, we give the definition of a Markov chain. Given a discrete-time stochastic process $Y = \{Y_t : t = 1, 2, \dots\}$ where t represents the time instance, we use S to denote the state space of Y which contains all possible values of any element in Y. The stochastic process Y is said to be a Markov chain if the distribution of the random variable at time t, Y_t, given all the past of the process depends only on the immediate past, Y_{t-1}. That is the following equation holds:

$$P(Y_t = s_i | Y_{t-1} = s_j, Y_{t-2} = s_k, \dots) = P(Y_t = s_i | Y_{t-1} = s_j) \tag{4.1}$$

where states $s_i, s_j, s_k, \dots$ are in S. The subscripts satisfies $1 \leq i, j, k \leq N_S$ where N_S is the size of S. The probability $P(Y_t = s_i | Y_{t-1} = s_j)$ is called the transition probability and denoted by a_{ij}.

A hidden Markov model is an extension of a Markov chain. It consists of three main components: the initial probability distribution π, the transition probability distribution A, and the observation probability distribution B. A brief explanation of each of these components is given as follows.

1. Initial probability distribution π: The initial probabilities $\pi = \{\pi_i : 1 \leq i \leq N_S\}$ are the probabilities of the hidden stochastic process Y beginning in each state s_i in the state space S.

2. Transition probability distribution A: The distribution A is comprised of all possible transition probabilities $\{a_{ij} : 1 \leq i, j \leq N_S\}$. Therefore, there are $N_S \times N_S$ entries in A. For two states s_i and s_j, if there is no connection between them, a_{ij} is set to zero.

3. Observation probability distribution B: The observation probabilities connect the hidden process with the observed process. In the case of discrete observations, the distribution in state s_j is a table listing the probabilities of all distinct observations observed when the hidden process is in s_j. For continuous observations, we need to estimate the probability density of the observations in each of the states in S.

The complete set of components of an HMM λ can be expressed as $\lambda = (A, B, \pi)$. The problem to be solved here is that given an observation sequence $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$ and an HMM λ, how to find out a state sequence $\boldsymbol{Q} = (s_{q_1}, s_{q_2}, \ldots, s_{q_T})$ that maximises the likelihood $P(\boldsymbol{Q}|\boldsymbol{X}, \lambda)$:

$$\begin{aligned} P(\boldsymbol{Q}|\boldsymbol{X}, \lambda) &= \frac{P(\boldsymbol{X}|\boldsymbol{Q}, \lambda)P(\boldsymbol{Q}|\lambda)}{P(\boldsymbol{X}|\lambda)} \\ &= c\left(\prod_{t=2}^{T} p(\boldsymbol{x}_t|q_t)\right)\left(\prod_{t=2}^{T} a_{q_{t-1}q_t}\right)\pi_{q_1}. \end{aligned} \tag{4.2}$$

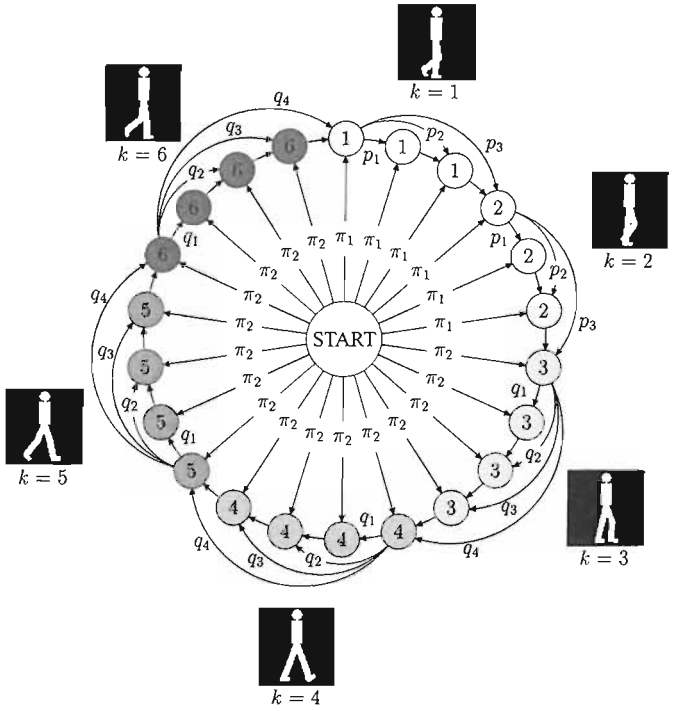
Here, c is a normalisation constant and $p(\boldsymbol{x}_t|q_t)$ is the probability of observation $\boldsymbol{x}_t$ given state s_{q_t} and λ. We omitted λ and replace s_{q_t} by q_t to simply the above equation. The standard Viterbi algorithm can solve the problem (see Rabiner 1989 for details) efficiently and is implemented in my work.

We exploit our prior knowledge of human walking by locating the phase of an image in the gait cycle, or in other words, which part of the gait cycle an image comes from. The input image sequence can be viewed as the observations. Our task is to find a state sequence which is most likely to generate the observed image sequence. One of the challenges to build up the HMM is to measure the observation probability distributions. Since the probability is a conditional probability of an image, the distributions are in the high-dimensional image space. It is very difficult to measure them directly because we do not have enough data to represent the true distributions. Alternatively, we use the relatively new PDF projection theorem to estimate such distributions.

4.3.2 Constructing the Hidden Markov Model

We hand-craft carefully a cyclic HMM to model the gait cycle. The structure of the HMM is shown in Figure 4.3. The gait cycle is divided into $K = 6$ sections and each section is modelled by some states with the same grey colour in the figure. These states are tied together to share the same initial probability π_k and observation probability distribution $p(\boldsymbol{x}|k)$. Here $\boldsymbol{x}$ is an observed image and k is the section number.

Given the capability of modelling sequential stochastic data, HMMs have been used in previous gait-related work (e.g., Meyer et al. 1998; Lee et al. 2003; Sundaresan et al. 2003; Lan and Huttenlocher 2004). A major structural difference between the HMMs they used and the one shown in Figure 4.3(a) is that they represented each section of the gait cycle by a single state with a self-transition and we used tied-states here. We believe that tied-states capture the dynamics of human walking more precisely. If we assume that a section is modelled by a single state with self-transition probability a, then the



(a)

probabilities	values
π_1	0.033
π_2	0.050
p_1	0.475
p_2	0.475
p_3	0.050
q_1	0.050
q_2	0.530
q_3	0.370
q_4	0.050

(b)

FIGURE 4.3: Hand-crafted hidden Markov model used to locate images within the gait cycle: (a) shows the architecture of the HMM with sections labelled by an image of the corresponding prototype; (b) lists the values of transition probabilities.

probability of n consecutive frames observed in the section is given by an exponential function $p(n)$:

$$p(n) = a^{n-1}(1-a) \tag{4.3}$$

which reaches a maximum when $n = 1$. That probability function tells us that the most likely number of images in a section is always equal to one, which is apparently not true for the real situations. In contrast, the tied-state structure provides us a more accurate way to reflect the real transition distribution. For example, to model section $k = 3$ (the section numbers are given inside the circles standing for states), we use four tied states. The transition probabilities $\{q_i\}_{i=1}^{4}$ are listed in Figure 4.3(b). Observing two or three consecutive frames is most likely for this section.

There are two sections only having three tied states in the HMM while others have four. Originally, the walking cycle was divided into $K = 5$ sections, each having four tied states, to give a 20 state HMM. The intention was that each section was occupied for approximately the same time. Four states per section were chosen because this is a reasonable upper limit on the number of frames per section. Skip transitions will therefore model cycles of less than 20 frames. The largest skips (transitions p_3 and

q_4) are included to model possible missing frames, although this did not occur in the gait data. Subsequently, it was realised that for one particular section of the cycle, the variation in walker pose was very large as a consequence of rapid limb movement in this phase. This occurred at the bottom of the leg swing, where the dynamic movement is high. To remedy this, we split this section into two to give $K = 6$ sections with the split pair represented by three HMM states. The transition probabilities were adjusted to cater for this.

The transition probabilities listed in Figure 4.3(b) are learned from a set of indoor image sequences. They are selected by choosing three sequences from each of seven random walkers in the gait database. Instead of the classical Baum-Welch learning algorithm, we used a rather simple method to learn the transition probabilities as follows:

- Hand-labelling all the selected image sequences according to our prior knowledge so that for each frame we know the particular section the image comes from.
- Counting the frequency of n consecutive frames occurring in the same section ($n = 1, 2, 3, 4$). The transition probabilities are then computed according to these frequencies. The counting process was done separately for the sections having 3 tied states and those having 4 tied states.

The initial probabilities of states in the same section are assigned equally. For section k, the probability π^k is computed as:

$$\pi^k = \frac{\text{length of section } k}{(\text{length of a gait cycle}) \times (\text{number of states in section } k)} \tag{4.4}$$

Since the two short sections (sections 1 and 2) are half the length of other sections, the initial probabilities of the states in these two sections are different from others as listed in Figure 4.3(b).

Note that we have not given any explanation of the silhouettes in Figure 4.3(a). These silhouettes correspond to the mean walker models of the sections and are to be used to measure the observation probability distributions.

4.4 Modelling Observation Probability Distribution

Having discussed the structure of the HMM, the transition probabilities and the initial probabilities, we need to define the observation probabilities to complete the HMM. Estimating the probability of an image (an observation) belonging to (i.e., being emitted by) a particular state is non-trivial. We use the probability density function (PDF) projection theorem (Baggenstoss 2003) to derive the observation probability densities.

This theorem provides a general framework to project PDFs from a low-dimensional feature space back to the raw data space which is usually high-dimensional so as to avoid the curse of dimensionality. In Minka (2004) and Thayananthan et al. (2004), they discussed a scheme of estimating the likelihood PDFs for template-based (or exemplar-based) matching using the PDF projection theorem.

4.4.1 The PDF Projection Theorem

Let $\boldsymbol{x}$ be the data points in the raw data space. We define features $\boldsymbol{z}$ by $\boldsymbol{z} = T(\boldsymbol{x})$ where $T()$ is a many-to-one mapping from the raw data space to a feature space. Given a fixed reference hypothesis H_0 with known PDFs $p(\boldsymbol{x}|H_0)$ and $p(\boldsymbol{z}|H_0)$, the PDF projection theorem says that function $p(\boldsymbol{x})$ defined by:

$$p(\boldsymbol{x}) = \frac{p(\boldsymbol{x}|H_0)}{p(\boldsymbol{z}|H_0)} p(\boldsymbol{z}), \tag{4.5}$$

is a PDF in the raw data space where $p(\boldsymbol{z})$ is a PDF in the feature space. Very often (i.e, in the case of template matching), we want to estimate the observation density under the hypothesis H_k, $p(\boldsymbol{x}|H_k)$. Let $\boldsymbol{z}_k$. be the features extracted in a way associated with H_k. Baggenstoss (2003) states that if $\boldsymbol{z}_k$ is a sufficient statistic for H_k versus H_0, we can estimate $p(\boldsymbol{x}|H_k)$ using Equation 4.5, that is:

$$p(\boldsymbol{x}|H_k) = \frac{p(\boldsymbol{x}|H_0)}{p(\boldsymbol{z}_k|H_0)} p(\boldsymbol{z}_k|H_k). \tag{4.6}$$

Note that it is difficult to establish the sufficiency in practice. Baggenstoss discussed this issue. It is said that the sufficiency is not essential in practice. Equation 4.6 provides us a way to approximate the real PDF. However, we can achieve near optimal performance by carefully choosing the features that contain enough information related to the raw data. In this case, the raw data space is the image space. Appendix D gives more details about the PDF projection theorem.

4.4.2 H_k, z_k and H_0

To use the PDF projection theorem, we have to define hypothesis H_k feature $\boldsymbol{z}_k$ and reference hypothesis H_0. Moreover, we model the PDFs in the feature space explicitly using gamma distributions so that the projected PDFs in the raw data space can be derived analytically.

Hypothesis H_k: In our case, hypothesis H_k states that an image $\boldsymbol{x}$ comes from section k. The observation PDF of a state in that section can then be written as $p(\boldsymbol{x}|H_k)$.

Feature z_k: We define the feature as an image distance between the observed image and an exemplar associated with section k. All exemplars are the silhouettes shown in Figure 4.3(a). Since the image distance is a scalar, we use z_k instead of $\boldsymbol{z}_k$ to denote the feature. Each of the exemplar is a model silhouette that embodies the average articulated-model parameters extracted from its associated section of the walk. The average parameters are denoted by $\{\boldsymbol{\theta}_k\}_{k=1}^K$. Given image $\boldsymbol{x}$ feature z_k can then be computed as:

$$z_k = \rho\Big(\boldsymbol{x}, \mathcal{I}(\boldsymbol{\theta}_k)\Big) \tag{4.7}$$

where ρ is a function measuring the similarity between two images. To quantify the similarity, we use the chamfer distance, whose power has been proven in object detection and tracking (Borgefors 1988; Gavrila 2000; Toyama and Blake 2002). There are of course other image distances, e.g., the one based on the number of pixels having same or different values in two silhouettes (Ju et al. 1996; Sarkar et al. 2005). However, they are not suitable to be used on the outdoor data which have poor silhouette quality.

We denote the chamfer distance between silhouettes $\boldsymbol{I}_1$ and $\boldsymbol{I}_2$ by $\rho(\boldsymbol{I}_1, \boldsymbol{I}_2)$. Given their edge-point sets $U = \{\boldsymbol{u}_n\}_{n=1}^N$ and $V = \{\boldsymbol{v}_m\}_{m=1}^M$, the chamfer distance from U to V is defined as:

$$\rho(U, V) = \frac{1}{N} \sum_{\boldsymbol{u}_n \in U} \min_{\boldsymbol{v}_m \in V} \|\boldsymbol{u}_n - \boldsymbol{v}_m\|. \tag{4.8}$$

In this work, the chamfer distance is computed efficiently using the chamfer transform. The real (silhouette) images and the prototype images are converted to edge images using the Sobel edge detector. The real edge images then serve as reference; they are chamfer transformed. We use a (3×3) mask with a (3,4)/3 distance measure to approximate a Euclidean distance using integer arithmetic as described by Borgefors (1988). More details of chamfer distances can be found in Appendix C.

Reference Hypothesis H_0: The reference hypothesis H_0 should be carefully chosen so as to make the estimations of $p(\boldsymbol{x}|H_0)$ and $p(\boldsymbol{z}|H_0)$ tractable. We define H_0 as that an image comes from the walking cycle. We also assume that the probability of an image x belonging to the walking cycle, $p(\boldsymbol{x}|H_0$, is a constant. Since the reference hypothesis H_0 is the union of all the hypotheses $\{H_k\}_{k=1}^K$:

$$H_0 = \bigcup_{k=1}^{K} H_k, \tag{4.9}$$

We can measure $p(z_k|H_0)$ by:

$$p(z_k|H_0) = \sum_{j=1}^{K} p(z_k|H_j)p(H_j). \tag{4.10}$$

The prior $p(H_j)$ can be thought as the probability of being in the jth section and is simply assigned by the proportion of the length of that section in the gait cycle. Substituting Equation 4.10 to 4.6, we have the way to approximate the probability of an image $\boldsymbol{x}$ coming from the kth section:

$$p(\boldsymbol{x}|H_k) \approx p(\boldsymbol{x}|H_0)\frac{p(z_k|H_k)}{\sum_{j=1}^{K} p(z_k|H_j)p(H_j)} \tag{4.11}$$

where $p(z_k|H_j)$ is the distribution of the chamfer distances between the exemplar for the kth section and the images from the jth section. It has to be noticed that Equation 4.11 is a way to approximate the true probability distribution since we can not prove the statistical sufficiency of z_k.

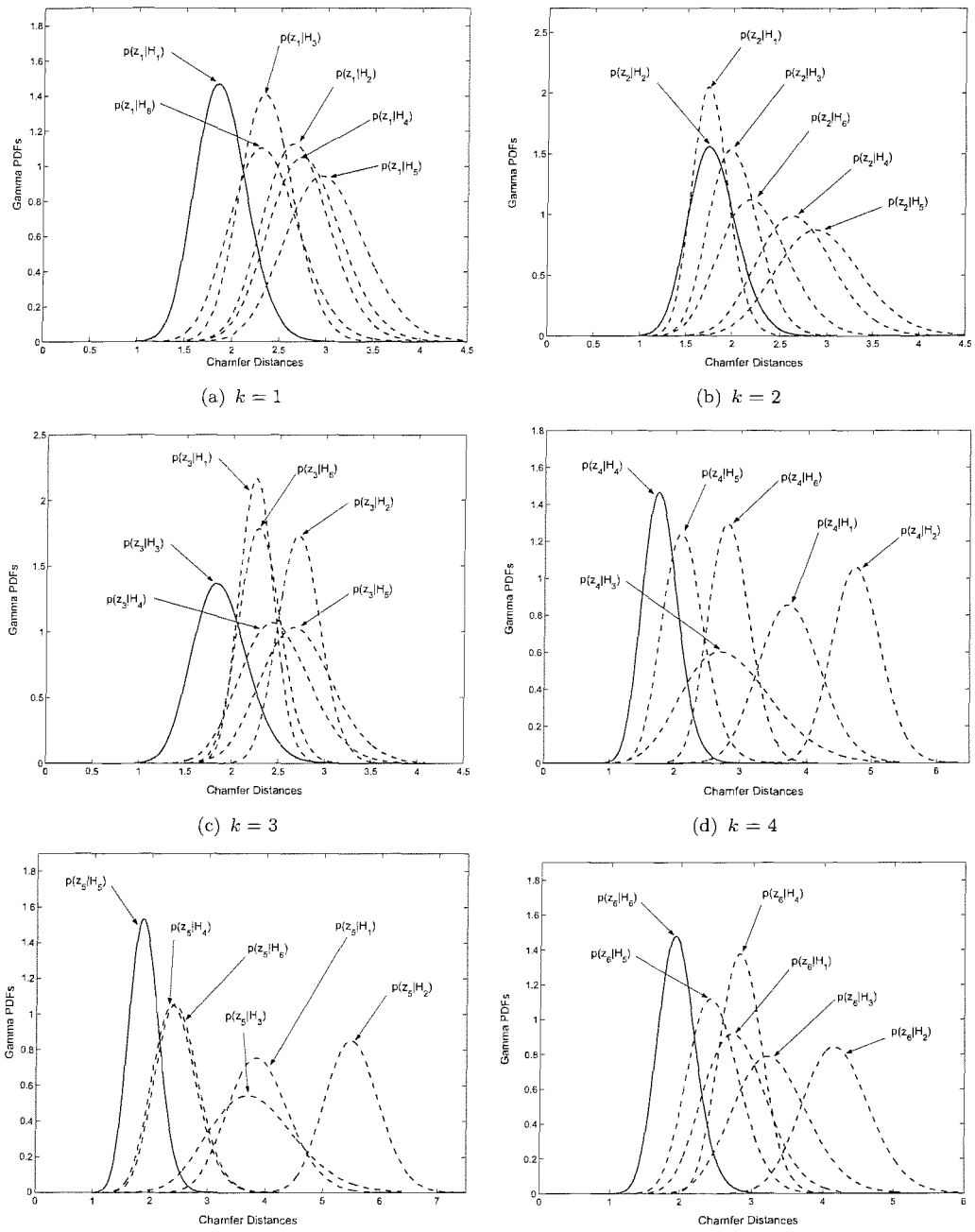
4.4.3 Gamma Representation

We model the probability distribution of the chamfer distances between the exemplar of the kth section and the observed images coming from the jth section of the walk by a gamma distribution:

$$p(z) = \frac{b^a z^{a-1} \mathrm{e}^{-bz}}{\Gamma(a)}. \tag{4.12}$$

Thus, to use the PDF projection theorem we need to know the parameters a and b for each of the gamma distributions describing the spread of chamfer matches between the images in the jth section of the walk and the prototype silhouette for section k—giving a total of K^2 gamma distributions. Figure 4.4 shows the learned gamma distributions for the indoor image data and Figure 4.5 illustrates how the computed gamma distributions fit the empirical chamfer distances. The ones for outdoor noisy image sequences are shown in Figure 4.6. As seen, the probability distributions learned from the outdoor data are wider and more overlapped by each other than those from the indoor data, which is coherent with the level of noise in the images.

The parameters of the gamma distributions will be different for the high-quality indoor data and the noisy outdoor data. We therefore need to perform an initial calibration for each database used. Given a set of labelled images, we can perform this calibration by first calculating the chamfer distances between the images in each section of the walk and each of the prototypes and then finding the parameters of the gamma distributions that maximise the likelihood of the chamfer distance values. For the indoor data, we used

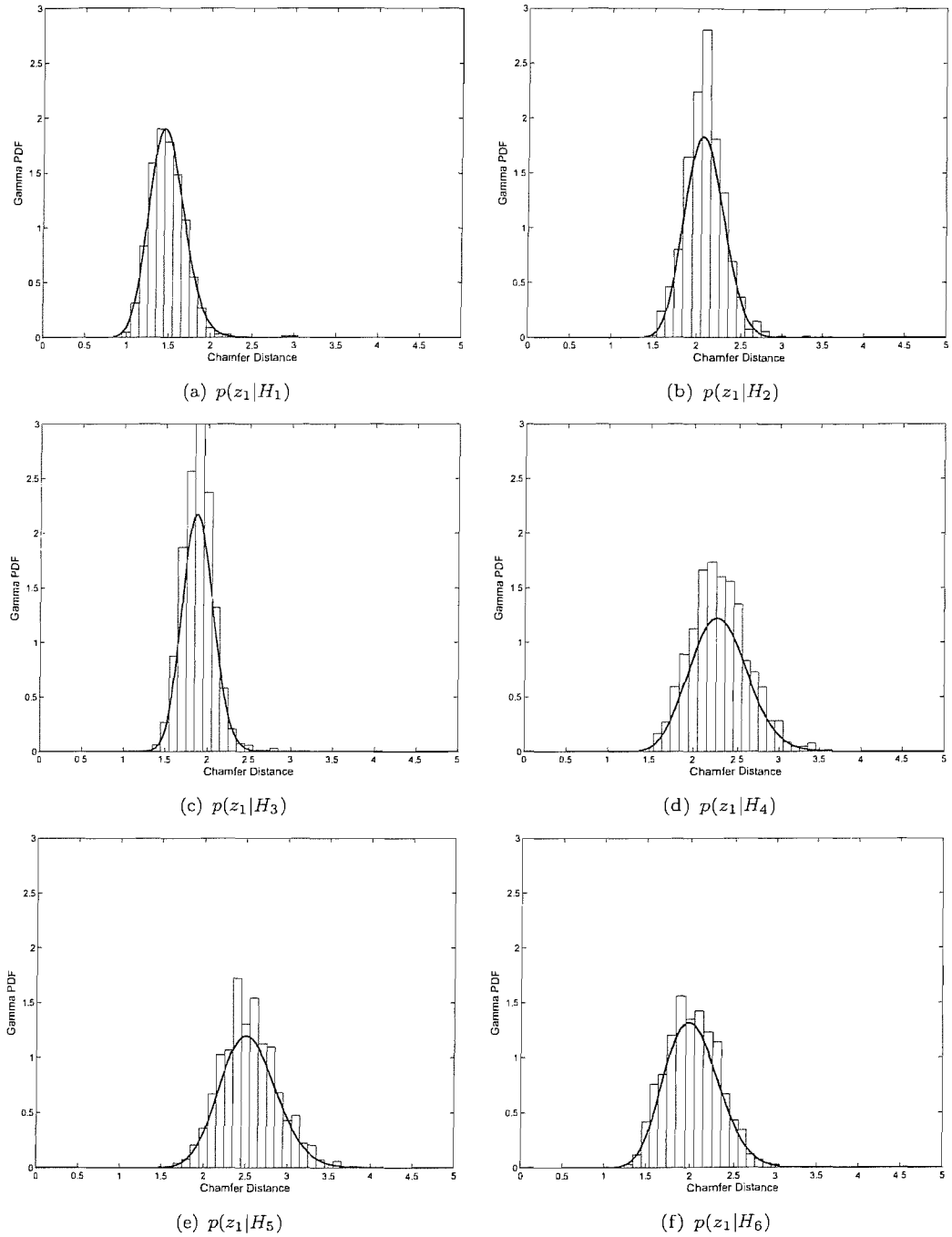


(a) $k = 1$ (b) $k = 2$

(c) $k = 3$ (d) $k = 4$

(e) $k = 5$ (f) $k = 6$

FIGURE 4.4: Gamma distributions learned for indoor data in order to derive the observation probability densities using the PDF projection theorem. PDF $p(z_k|H_j)$ gives the distribution of the chamfer distances between all images coming from section j and the prototype silhouette corresponding to section k where $1 \leq k, j \leq K$.

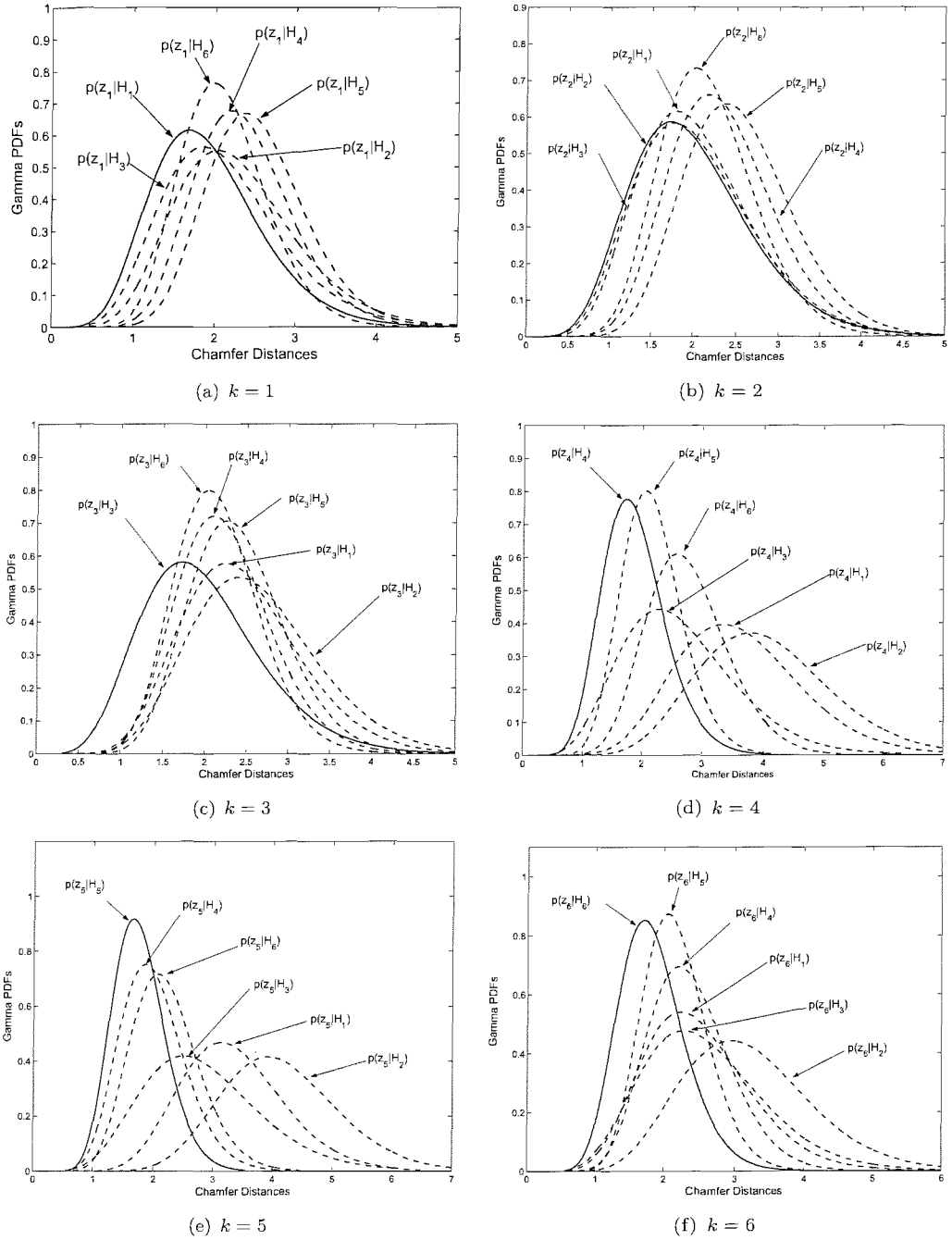


(a) $p(z_1|H_1)$ (b) $p(z_1|H_2)$

(c) $p(z_1|H_3)$ (d) $p(z_1|H_4)$

(e) $p(z_1|H_5)$ (f) $p(z_1|H_6)$

FIGURE 4.5: Examples showing how the gamma distributions fit the empirical chamfer distances calculated from the indoor data.



(a) $k = 1$ (b) $k = 2$

(c) $k = 3$ (d) $k = 4$

(e) $k = 5$ (f) $k = 6$

FIGURE 4.6: Gamma distributions learned for outdoor noisy data in order to derive the observation probability densities using the PDF projection theorem. PDF $p(z_k|H_j)$ gives the distribution of the chamfer distances between all images coming from section j and the prototype silhouette corresponding to section k where $1 \leq k, j \leq K$.

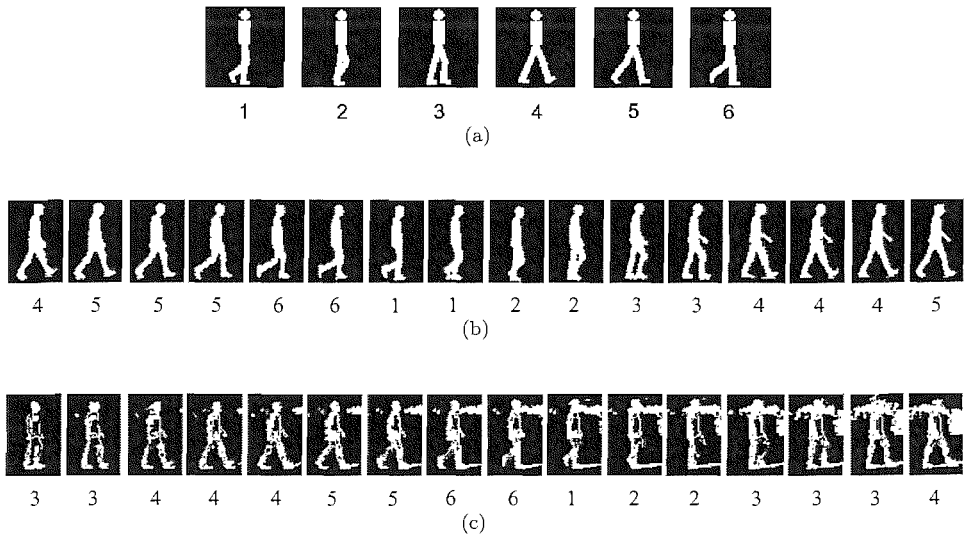


FIGURE 4.7: Labelling the phase of the gait cycle using the HMM: (a) shows the $K = 6$ mean models used as prototypes. Typical labellings produced by the HMM are shown for (b) an indoor sequence and (c) an outdoor sequence.

the same training data as used to compute the average prototype parameters; for the outdoor later, we hand-labelled a similar number of images. Typical results are shown in Figure 4.7, where Figure 4.7(a) shows the prototype silhouettes and Figure 4.7(b) and 4.7(c) show labelling for illustrative indoor and outdoor sequences respectively.

4.5 Posterior Probability for Model Parameters

Having labelled each image according to its section in the walking cycle, we are in a position to find the parameters of the articulated model which best fit each of the images. The static parameters describing sizes of the body parts and the dynamic parameters describing the angles of the limbs are treated differently. The static parameters are assumed to remain constant over all frames in a sequence. We thus accumulate evidence for these values from a large number of images. The dynamic parameters are optimised on a frame-by-frame basis. In both cases, we maximise a posterior probability for the parameters. For the dynamic parameters, this is the posterior given a particular image, while for the static parameters, it is the posterior given a sequence of images. The posterior for the sequence is the product of the posteriors for each of the single images.

The posterior probability of the parameters $\boldsymbol{\theta}$, given an image from section k and a model $\mathcal{M}_k$, can be written as:

$$p(\boldsymbol{\theta}|\boldsymbol{x}, \mathcal{M}_k) \propto p(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M}_k)\, p(\boldsymbol{\theta}|\mathcal{M}_k) \tag{4.13}$$

where $p(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M}_k)$ is the likelihood of the image given the parameters and $p(\boldsymbol{\theta}|\mathcal{M}_k)$ is the prior for the parameters. The constant of proportionality is independent of the model so does not influence the maximum a posteriori parameters. We cannot use the pdf projection theorem for calculating the likelihood, $p(\boldsymbol{x}|\boldsymbol{\theta})$ of an image, $\boldsymbol{x}$, given a set of model parameters, $\boldsymbol{\theta}$, because we now have a continuum of models. Instead we follow the conventional (maximum entropy) assumption that the likelihood is exponentially distributed in the chamfer distance:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) \propto \mathrm{e}^{-b\,\rho(\boldsymbol{x},\mathcal{I}(\boldsymbol{\theta}))} \tag{4.14}$$

where b is a Lagrange multiplier to be determined empirically. We make the additional assumption that the number of images with chamfer distance $\rho(\boldsymbol{x}, \mathcal{I}(\boldsymbol{\theta}))$ equal to r grows as a polynomial r^{a-1}. That is, the distribution of chamfer distances is given by:

$$p\Big(\rho(\boldsymbol{x}, \mathcal{I}(\boldsymbol{\theta})) = r\Big|\boldsymbol{\theta}\Big) = \int p(\boldsymbol{x}|\boldsymbol{\theta})\,\delta\Big(\rho(\boldsymbol{x}, \mathcal{I}(\boldsymbol{\theta})) - r\Big)\,\mathrm{d}\boldsymbol{x} \propto r^{a-1}\mathrm{e}^{-br} \tag{4.15}$$

i.e., a gamma distribution. Empirically, this distribution fits the data well as shown in Figure 4.8, in which the theoretical distribution is compared to a histogram of values obtained from selected calibration data (3606 chamfer distances), chosen on the basis of good visual fit. The parameters a and b can be found by fitting empirical data using maximum likelihood. This way to determine the exponent b is identical to that of Toyama and Blake (2002), although we give a slightly different (and we believe more direct) motivation.

We assume a Gaussian prior for the parameters:

$$p(\boldsymbol{\theta}|\mathcal{M}_k) \propto \exp\Big(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_k)^{\mathrm{T}}\boldsymbol{C}_k^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_k)\Big) \tag{4.16}$$

where $\boldsymbol{\theta}_k$ and $\boldsymbol{C}_k$ are the averages and covariances for the parameters in section k of the walk. The average parameters are the same as those of our prototypes described in the previous subsection. The covariance matrix is learned using Bayesian updating described in Section 4.7. In the next subsection, we discuss the practical details of finding the parameters that maximise the posterior probability.

4.6 Optimising Parameters

The posterior probability is a non-linear function of the parameters $\boldsymbol{\theta}$, which may have many local maxima. From Equation 4.13 and 4.14, we know that evaluating this function involves computing the chamfer distances. Such a distance is non-differentiable, which

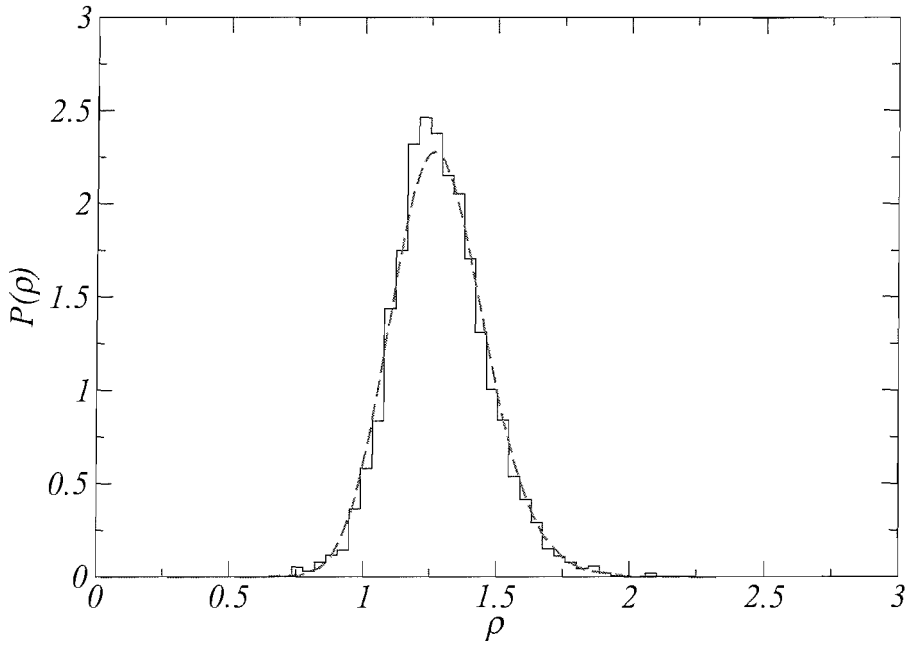


FIGURE 4.8: Distribution of empirical chamfer distance values for 3606 data points selected as calibration data and fitted gamma distribution. As can be seen, the fit is excellent.

makes the standard gradient-based optimisation methods not suitable here. To find the best-fit parameters, we use a standard multi-dimensional continuous optimisation algorithm (Powell 1964) to maximise the log-posterior. It is an iterative method and at each iteration, we define carefully N directions in the search space so that each pair of them are conjugate. Here N is the space dimension. We start by picking up a direction and search the maximum on the line defined by the direction. We then move to the maximum and start another search along the next direction. This process is carried out until we have done the line search for all the directions. The optimisation process stops when it reaches a local maximum.

During the optimisation, we need to compute the likelihood between the images and a 'silhouette model'. Pre-computing the chamfer transform for all the images speeds up this computation considerably. The chamfer transform is also used by the HMM in computing the likelihood of the image coming from a particular section of the walker. The quality of the solution found, as well as the time taken to perform the optimisation, depends on the initial values of the parameters.

4.6.1 Optimising Static Parameters

We start by finding the optimal static parameters over the sequence as they are time-invariant in a sequence. Here we denote the static parameters as τ. Knowing which part of the gait cycle images come from, we can easily cluster all images of a given

sequence into K image groups, $\{G_k\}_{k=1}^{K}$, according to the output of the HMM. We denote the images in G_k as $\{\boldsymbol{x}_{n_k}\}_{n_k=1}^{N_k}$ where N_k is the number of images in this group. In this subsection, we introduce an iterative learning algorithm that can not only find the reasonable values for the static parameters, but learn the initial values $\boldsymbol{\phi}_k$ of the dynamic parameters for the images in group G_k.

We use $\hat{\boldsymbol{\theta}}_k$ to denote the set of model parameters including $\boldsymbol{\tau}$ and $\boldsymbol{\phi}_k$. Base on Bayes' theory, the posterior $p(\hat{\boldsymbol{\theta}}_k|G_k, \mathcal{M}_k)$ can be computed as:

$$\begin{aligned} p(\hat{\boldsymbol{\theta}}_k|G_k, \mathcal{M}_k) &\propto p(\hat{\boldsymbol{\theta}}_k|\mathcal{M}_k)p(G_k|\hat{\boldsymbol{\theta}}_k, \mathcal{M}_k) \\ &\propto p(\hat{\boldsymbol{\theta}}_k|\mathcal{M}_k) \prod_{n_k=1}^{N_k} p(\boldsymbol{x}_{n_k}|\hat{\boldsymbol{\theta}}_k, \mathcal{M}_k). \end{aligned} \tag{4.17}$$

Here to simplify the computation, we assume that the images in G_k are mutually independent. Having Equation 4.17, we can define another posterior probability over the whole sequence:

$$p(\hat{\boldsymbol{\Theta}}|\boldsymbol{G}, \boldsymbol{\mathcal{M}}) = \prod_{k=1}^{K} p(\hat{\boldsymbol{\theta}}_k|G_k, \mathcal{M}_k) \tag{4.18}$$

$$\begin{aligned} \hat{\boldsymbol{\Theta}} &= \{\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_K\} \\ \boldsymbol{G} &= \{G_1, \ldots, G_K\} \\ \boldsymbol{\mathcal{M}} &= \{\mathcal{M}_1, \ldots, \mathcal{M}_K\}. \end{aligned} \tag{4.19}$$

Again for simplicity of computation, $\{\hat{\boldsymbol{\theta}}_k\}_{k=1}^{K}$ are assumed to be conditionally independent with respect to G_k and $\mathcal{M}_k$. Figure 4.9 shows the iterative learning algorithm for the static parameters. We use the average parameters $\{\boldsymbol{\theta}_k\}_{k=1}^{K}$ to initialise $\boldsymbol{\tau}$ and $\{\boldsymbol{\phi}_k\}_{k=1}^{K}$. Note that these parameters are also used in the HMM to generate the prototype silhouettes. In each iteration, we first optimise $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_K$ one-by-one using the static parameters obtained in the previous iteration. After that, we update the static parameters by maximising Equation 4.19 using the latest learned $\{\boldsymbol{\phi}_k\}_{k=1}^{K}$. The algorithm stops if the posterior probability $p(\hat{\boldsymbol{\Theta}}|\boldsymbol{G}, \boldsymbol{\mathcal{M}})$ is no larger than the one in the last iteration. Having determined the static parameters and the initial values of the dynamic parameters, we can optimise the dynamic parameters for each frame in a sequence.

Input:

- Image groups $G_1, \ldots, G_K$ created according to the output of the HMM.
- Average articulated-model parameters $\theta_1, \ldots, \theta_K$ for the sections of the walk.

Initialise:

- Set τ equal to the static parameters of the average parameters.
- Set ϕ equal to the dynamic parameters of θ_k $(k = 1, \ldots, K)$.

Do

for $k = 1, 2, \ldots, K$ **do**
Calculate $\phi_k \leftarrow \arg\max_{\phi} p(\tau, \phi | G_k, \mathcal{M}_k)$.
end
Calculate $\tau \leftarrow \arg\max_{\tau} p(\tau, \phi \,|\, \boldsymbol{G}, \boldsymbol{\mathcal{M}})$.

Until $p(\tau, \phi \,|\, \boldsymbol{G}, \boldsymbol{\mathcal{M}})$ **stops increasing.**

Output:

- τ
- ϕ_k $(k = 1, \ldots, K)$

Figure 4.9: Algorithm for learning the static parameters iteratively. In each iteration, we first optimise K sets of dynamic parameters for each of the K image groups. We then optimise the static parameters based on the latest learned parameters. The algorithm halts when there is no increment for the posterior over the whole sequence.

4.6.2 Optimising Dynamic Parameters

The dynamic parameters are extracted on an image-by-image basis. We optimise the dynamic parameters twice, from two different starting positions in the parameter space. The first initialisation are the initial values $\{\phi_k\}_{k=1}^{K}$ obtained from the learning algorithm described in the previous subsection. The second one is generated by a linear prediction using the parameters learned in the previous time steps. If we denote the optimal parameters found in frame $t-1$ and t as $\theta(t-1)$ and $\theta(t)$ respectively, the prediction of the dynamic parameters at time $t+1$, $\tilde{\theta}(t+1)$, can be computed as:

$$\tilde{\theta}(t+1) = \theta(t) + \big(\theta(t) - \theta(t-1)\big). \tag{4.20}$$

As we do the optimisation twice from two starting points, it ends up with two sets of dynamic parameters. We choose the one corresponding to the higher maximum posterior probability.

Input:

- Image sequence $\mathcal{X} \leftarrow \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$.
- Number sequence $Q \leftarrow \{q_1, q_2, \ldots, q_T\}$ output by the HMM. Each number q_t $(1 \leq q_t \leq T)$ tells which section image $\boldsymbol{x}_t$ comes from.
- Learned static parameters $\boldsymbol{\tau}$.
- Initial values learned in Figure 4.9, $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\}$.

for $t = 1, 2$ **do**
  Set initialisation $\tilde{\boldsymbol{\theta}}_t \leftarrow \left[\boldsymbol{\tau}^{\mathrm{T}}, \boldsymbol{\phi}_{q_t}^{\mathrm{T}}\right]^{\mathrm{T}}$.
  Compute the optimal parameters $\boldsymbol{\theta}_t$ by maximising the posterior probability $p(\boldsymbol{\theta}|\boldsymbol{x}_t, \mathcal{M})$ with respect to the dynamic parameters.
end
for $t = 3, 4, \ldots, T$ **do**
  Set initialisation $\tilde{\boldsymbol{\theta}}_t^1 \leftarrow \left[\boldsymbol{\tau}^{\mathrm{T}}, \boldsymbol{\phi}_{q_t}^{\mathrm{T}}\right]^{\mathrm{T}}$.
  Compute the optimal parameters $\boldsymbol{\theta}_t^1$ by maximising the posterior probability $p(\boldsymbol{\theta}|\boldsymbol{x}_t, \mathcal{M})$ with respect to the dynamic parameters.
  Set initialisation $\tilde{\boldsymbol{\theta}}_t^2 \leftarrow \boldsymbol{\theta}_{t-1} + (\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_{t-2})$.
  Compute the optimal parameters $\boldsymbol{\theta}_t^2$ by maximising the posterior probability $p(\boldsymbol{\theta}|\boldsymbol{x}_t, \mathcal{M})$ with respect to the dynamic parameters.
  if $p(\boldsymbol{\theta}_t^1|\boldsymbol{x}_t, \mathcal{M}) > p(\boldsymbol{\theta}_t^2|\boldsymbol{x}_t, \mathcal{M})$ **then**
    $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t^1$
  else
    $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t^2$
  end
end

Output: Articulated-model parameters $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_T\}$.

Figure 4.10: Algorithm for learning the dynamic parameters. Besides the initial values learned in Figure 4.9, we use a linear prediction to generate another initialisation for the optimisation. We choose the fit with larger posterior.

In principle, we can refine our estimates for the static and dynamic parameters iteratively. However, in practice, we found that after a single iteration our estimates for the model parameters were adequate and the improvements obtained by further iterations were insignificant.

4.7 Bootstrapping and Updating

To automate our gait-extraction framework, we have to learn the statistics of the parameters of the articulated model, namely, the average parameters $\{\boldsymbol{\theta}_k\}_{k=1}^K$ and the

covariance matrices $\{\boldsymbol{C}_k\}_{k=1}^K$. Measuring these statistics accurately demands a large number of data points in the model-parameter space. Each of these points represent the parameters optimised from a given image. It could be expensive to get enough points by marking images manually. To avoid the tedious labour, we introduce a bootstrapping process to learn the statistics from clean indoor image data. After the system has been automated, we use the Bayesian updating to refine the statistics. The learned statistics are then used to build the strong priors against noise in images.

Figure 4.11 shows the details of the bootstrapping process. We do the following operations in the bootstrapping:

- We chose 21 image sequences from 7 random walkers in the indoor database; each of them had 3 sequences. We then performed normalisation on the silhouettes and labelled them manually to know which section each frame came from.
- The hand-labelled section numbers were used to calculate the transition probabilities for the HMM.
- Given some initial guess of the average model parameters, we used the same optimisation strategies described in Section 4.6 to find a best-fit model for each image except that we maximised the likelihood $p(\boldsymbol{x}|\boldsymbol{\theta}, \mathcal{M})$ rather than the posterior $p(\boldsymbol{\theta}|\boldsymbol{x}, \mathcal{M})$. Some hard constraints were imposed to prevent the model behaving unrealistically. For instance, we limited the radius of the head to be between 3 and 5 pixels.
- We computed the average parameters $\{\boldsymbol{\theta}_k\}_{k=1}^K$ and the covariance matrices $\{\boldsymbol{C}_k\}_{k=1}^K$ from the learned model parameters.
- Using the average parameters, we learned the gamma distributions so that the likelihood of an image given a state in the HMM could be measured by the PDF projection theorem.
- We then built the statistics of the parameters into the prior model to make the measure of posteriors available. After that, the system was fully automated.

Figure 4.12 shows the complete updating process. Once a new normalised image sequence is input into the framework, we first label it by the HMM. The transition probabilities are updated using the output section numbers by the HMM. For each frame, we learned a set of parameters by maximising the posterior. These parameters are used to update the average parameters $\{\boldsymbol{\theta}_k\}_{k=1}^K$ and covariance matrices $\{\boldsymbol{C}_k\}_{k=1}^K$. To do that, we need to remember the number of the sets of parameters, N, that have been used to learn the statistics. For a new frame from the kth section of the walk, we denote the extracted parameters by $\boldsymbol{\theta}^*$. The updated average parameters $\boldsymbol{\theta}_k^{\text{new}}$ can be computed as:

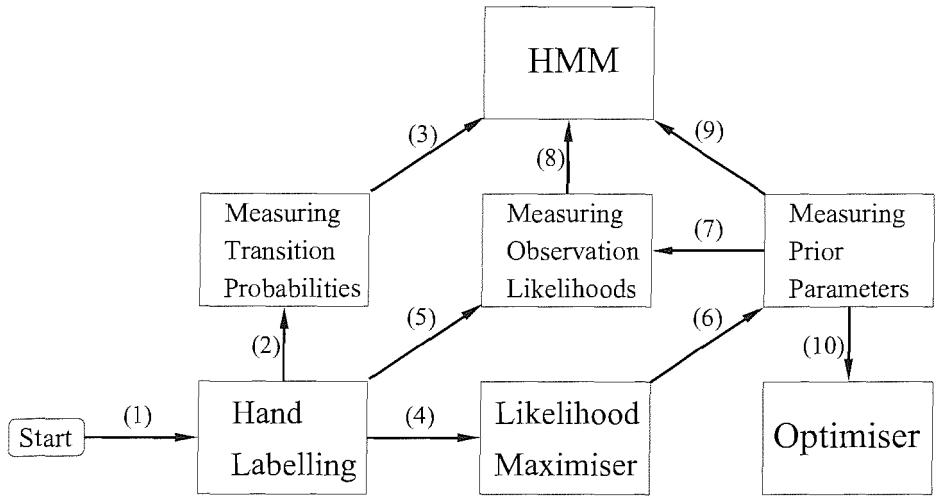


FIGURE 4.11: A flowchart describing the bootstrapping process for the Bayesian framework. The numbers in the figure are explained as follows: (1) 21 normalised gait sequences from 3 random chosen walkers; (2, 5) hand-labelled section numbers for the sequences; (3) learned transition probability distributions for the HMM; (4) those normalised sequences together with the hand-labelled section number; (6) extracted the articulated-model parameters; (7, 9) average parameters $\{\theta_k\}_{k=1}^K$; (8) observation probability densities for states in the HMM; and (10) average parameters $\{\theta_k\}_{k=1}^K$ and covariance matrices $\{\boldsymbol{C}_k\}_{k=1}^K$

$$\theta_k^{\text{new}} = \frac{N}{N+1}\theta_k^{\text{old}} + \frac{1}{N+1}\theta^* \tag{4.21}$$

We calculated the updated covariance matrix $\boldsymbol{C}_k^{\text{new}}$ as:

$$\begin{aligned}
\boldsymbol{C}_k^{\text{new}} &= \boldsymbol{R}_k^{\text{new}} - \theta_k^{\text{new}}(\theta_k^{\text{new}})^{\mathrm{T}} \\
&= \frac{N}{N+1}\boldsymbol{R}_k^{\text{old}} + \frac{1}{N+1}\theta^*(\theta^*)^{\mathrm{T}} - \theta_k^{\text{new}}(\theta_k^{\text{new}})^{\mathrm{T}} \\
&= \frac{N}{N+1}\left(\boldsymbol{C}_k^{\text{old}} + \theta_k^{\text{old}}(\theta_k^{\text{old}})^{\mathrm{T}}\right) + \frac{1}{N+1}\theta^*(\theta^*)^{\mathrm{T}} \\
&\quad -\theta_k^{\text{new}}(\theta_k^{\text{new}})^{\mathrm{T}}
\end{aligned} \tag{4.22}$$

where $\boldsymbol{R}$ represents the correlation of the parameters. The new statistics are then used to update the priors of the Bayesian framework. The observation probability densities are also updated using the new average parameters.

Note that for outdoor noisy data, we use the same method to choose 21 sequences and label them manually to learn the corresponding observation probability densities. These densities are updated as more sequences are labelled automatically by the HMM. Only the densities are learned and updated since we have already obtained the strong priors from clean data.

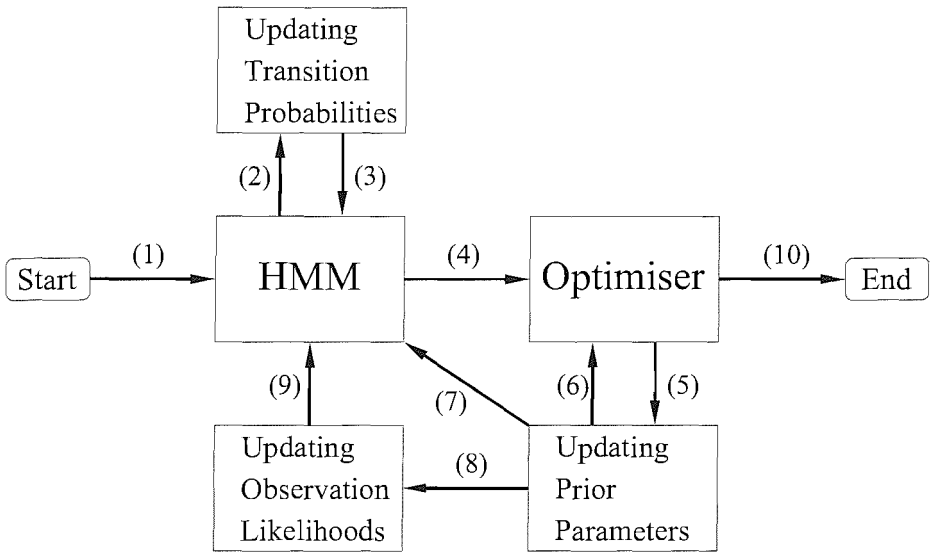


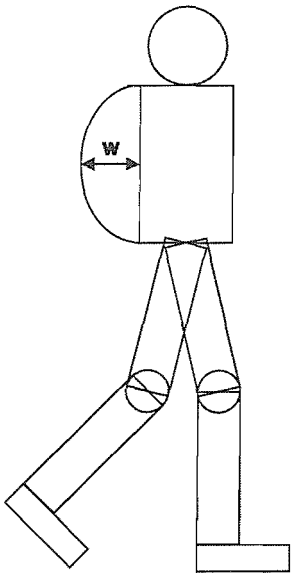
FIGURE 4.12: A flowchart describing the process of gait extraction and Bayesian updating. The numbers in the figure are explained as follows: (1) an input normalised gait sequence; (2) Automatically learned section numbers for images in sequence; (3) updated the transition probability distributions; (4) the normalised sequence together with the section numbered output by the HMM; (5, 10) extracted articulated-model parameters; (6) updated average parameters $\{\boldsymbol{\theta}_k\}_{k=1}^K$ and covariance matrices $\{\boldsymbol{C}_k\}_{k=1}^K$; (7, 8) updated average parameters $\{\boldsymbol{\theta}_k\}_{k=1}^K$; and (9) updated state observation probability densities.

A severe problem with gait extraction arises when the body appearance is changed by, for example, a carried rucksack or the limbs are obscured by, for example, an overcoat or carried briefcase. Because our approach combines a powerful statistical approach with a simple articulated model of a walker, it offers a straightforward way to cope with this situation by extending the walker model. We illustrate this with three examples in which the model is generalised by the addition of a rucksack, a long skirt, or a trenchcoat.

4.7.1 Framework Extension

The Bayesian framework is the combination of some simple components and therefore, can be easily extended. Here we show this advantage by extending the framework to handle a server problem inherent in gait extraction, that is the body-appearance changes or the motion occlusions caused by carrying objects or wearing special clothes. Because our approach combines a powerful statistical approach with a simple articulated model of a walker, it offers a straightforward way to cope with this situation by extending the articulated model. We illustrate this with three examples in which the model is generalised by the addition of a rucksack, a long skirt, or a trenchcoat.

The basic articulated model (see Section 4.2) is simply modified to cater for each of the three cases. Figures 4.13, 4.14 and 4.15 show the extended models for walkers with a rucksack, a long skirt or a trenchcoat respectively. A rucksack is represented by a half



(a)

	Added Parameters
1	Rucksack Width w

(b)

FIGURE 4.13: The extended articulated model for a walker carrying a rucksack: (a) shows the body parts; and (b) lists the only static parameter added to control the half ellipse standing for a rucksack.

ellipse, a long skirt by filling the gaps between the legs, and a trapezoid is added to stand for a trenchcoat. A maximum of two more static parameters is added to the basic model for the extra body appearance. Note that the parameters of the basic articulated model are still used to control the behaviour of the model so that we can still use the statistical priors learned from the indoor data to constrain the parameters in optimisation.

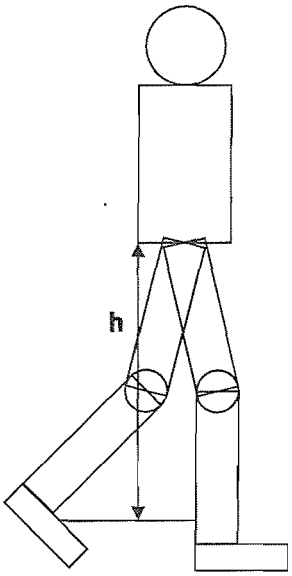
We denote the parameters of the extended model as $\boldsymbol{\theta}'$. Following the definition in Equation 4.14, given an image $\boldsymbol{x}$ from the kth section of the gait cycle and the new model $\mathcal{M}'_k$ we write the likelihood probability as:

$$p(\boldsymbol{x}|\boldsymbol{\theta}', \mathcal{M}'_k) \propto \mathrm{e}^{-b\,\rho(\boldsymbol{x}, \mathcal{I}(\boldsymbol{\theta}'))}. \tag{4.23}$$

We still use a multi-variate Gaussian as the prior probability:

$$p(\boldsymbol{\theta}'|\mathcal{M}'_k) \propto \exp\Big(-\frac{1}{2}(\boldsymbol{\theta}' - \boldsymbol{\theta}'_k)^{\mathrm{T}}(\boldsymbol{C}'_k)^{-1}(\boldsymbol{\theta}' - \boldsymbol{\theta}'_k)\Big) \tag{4.24}$$

where $\boldsymbol{\theta}'_k$ and $\boldsymbol{C}'_k$ are the new average parameters and covariance for the extended model. We denote the added static parameters by $\boldsymbol{\tau}'$ and their means and covariance by $\boldsymbol{\theta}_{\boldsymbol{\tau}'}$ and $\boldsymbol{C}_{\boldsymbol{\tau}'}$. The average parameters $\boldsymbol{\theta}'_k$ can be expressed as:

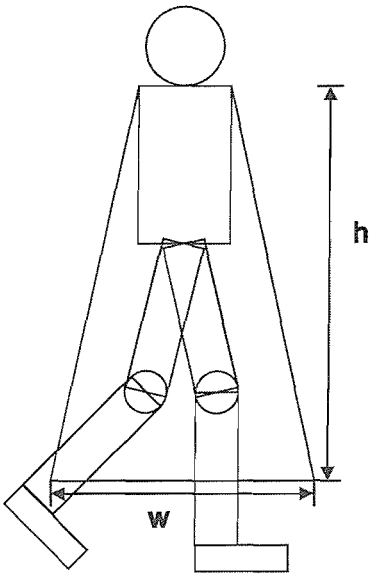


(a)

	Added Parameters
1	Skirt Length h

(b)

FIGURE 4.14: The extended articulated model for a walker wearing a long skirt: (a) shows the body parts; and (b) lists the only static parameter controlling the length of the skirt which is represented by filling the gaps between two legs.



(a)

	Added Parameters
1	Trenchcoat Width w
2	Trenchcoat Height h

(b)

FIGURE 4.15: The extended articulated model for a walker wearing a trenchcoat: (a) shows the body parts; and (b) lists the two static parameters added to control the trapezoid standing for the trenchcoat.

$$\theta_k^{'} = \begin{pmatrix} \tau^{'} \\ \theta_k \end{pmatrix} \tag{4.25}$$

where θ_k are the average parameters of the basic articulated model in section k. We build the covariance $C_k^{'}$ simply from the previously learned covariance C_k and the covariance of the added parameters $C_{\tau'}$:

$$C_k^{'} = \begin{pmatrix} C_{\tau'} & 0 \\ 0 & C_k \end{pmatrix} \tag{4.26}$$

where $\mathbf{0}$ stands for matrices with all elements equal to zero. This is based on the assumption that the parameters of the basic model are tightly correlated between each other since they describe the body structure, but less correlated with the added parameters. With this definition, we only need to learn the statistics of the added static parameters from the training data, which significantly eases the learning process.

A simplified bootstrapping process is carried out for each of the three cases (rucksack, skirt and trenchcoat) as follows:

- Select 10 walkers randomly and choose 1 sequence for each of them.
- Hand-label these 10 sequences and measure the added static parameters for each of the sequences.
- Compute the average parameters $\{\theta_k^{'}\}_{k=1}^{K}$ and the covariance $\{C_k^{'}\}_{k=1}^{K}$ using Equation 4.25 and 4.26.
- Measure the observation probability densities for the HMM using $\{\theta_k^{'}\}_{k=1}^{K}$.

As more sequences input into the framework, the statistics ($\{\theta_k^{'}\}_{k=1}^{K}$ and $\{C_k^{'}\}_{k=1}^{K}$) and the observation likelihoods can be updated through the process described in Figure 4.12.

4.8 Summary

We have discussed the whole Bayesian framework in this chapter. Our prior knowledge including the knowledge of human body structure and the knowledge of the walk has been naturally built into the framework in terms of a two-dimensional articulated model and a hidden Markov model (HMM) that detects which part of the gait cycle an image comes from. The articulated model is simply comprised by basic geometric shape (circles and rectangles) and controlled by 12 parameters (6 static parameters and 6 dynamic parameters). The simplicity of the model stems from the idea of matching model

complexity to the available data, which will be noisy image data encountered in real-world conditions.

In this work, we divide the gait cycle into a few sections and each of them is modelled by tied-states rather than single states with self-transitions for more accurate modelling of transition probabilities. To measure the observation likelihoods for the HMM, we exploit the PDF projection theorem. The probability densities in high-dimensional image space are projected from an image-distance metric space without the curse of dimensionality. Here, the distance metric is the chamfer distance because of its robustness and computational efficiency.

The posterior probability is computed using Bayes' rule. The parameter in the exponent is obtained empirically from the assumed gamma distribution of the chamfer distances between the images and their best-fit models. The prior probabilities are defined as a multivariate Gaussian using the means and covariance of the parameters of the models extracted for each section. The static and dynamic parameters are optimised using different strategies. The static parameters are measured by an iterative algorithm upon the whole sequence since they are invariant for images in the same sequence. After that, we extract the dynamic parameters frame by frame.

We then described the bootstrapping and Bayesian-updating processes. Initially, we measure all the parameters, such as the transition probabilities, observation probability densities and the statistics of the parameters by labelling several clean indoor sequences manually. After that, we make the whole system automatic. Once new sequences are input, these parameters can be updated to be more accurate. This prevents doing a large amount of manual work that is very expensive.

We have claim the consistency and flexibility of our gait-extracting framework. The components within the framework are relatively independent which means changes of any module will not influence very much the others. For noisy outdoor images, we only need to learn the observation likelihoods which are different to those of the clean data. We also give an example of how easily the framework to be extended to cope with some difficult problems, e.g., to extract gait information for the walkers carrying a rucksack or wearing a long skirt or a trenchcoat. By adding one or two static parameters, we can achieve the tasks easily within the framework. Being able to work on such variations of image data is an important advantage of our system over other gait systems.

Chapter 5

Experiments and Results

We have tested our method on various sequences in the Southampton HiD database (Shutler et al. 2002), not only the indoor, outdoor and supplemental data as outlined in Chapter 3 but also some artificially modified data. The modifications tested are the addition of synthetic 'salt and pepper' noise and gait extraction in the presence of occluding bars. We believe it is very important to derive quantitative figures of merit for our results. In the case of high-quality images (i.e., indoor data), it is sufficient to use the chamfer distance for this purpose, since the extracted silhouettes are of a high fidelity. To see how possible poor normalisation could affect the gait extraction, we have generated some simulated sequences from indoor data by perturbing the positions of silhouettes and adding artificial noise as well. We have tested our system on the simulated data and used the chamfer distance to quantify results. For outdoor data, however, the chamfer distance is unreliable. We have, therefore, established (approximate) ground truth by hand labelling body points in a selection of images. The hand labels are 'unseen' in the automatic extraction. This approximate ground truth is then used to calculate an average pixel error per body point.

5.1 Indoor Data

Figure 5.1(a) illustrates three typical extracted model sequences superimposed on the corresponding high-quality indoor data. To enable the reader to judge the significance of the chamfer distance as an error measure, their respective values are shown below each image. As can be seen, most errors are attributable to the simplifications implicit in the model; for example, the walker on the bottom row has a pony-tail and long tee-shirt which are not well modelled by the rectangle-plus-circle representation of the torso and head. Nonetheless, it is clear that for the most part, the model fitting is very good with an average chamfer distance error of about 1 pixel. (Recall that the 'model silhouette'

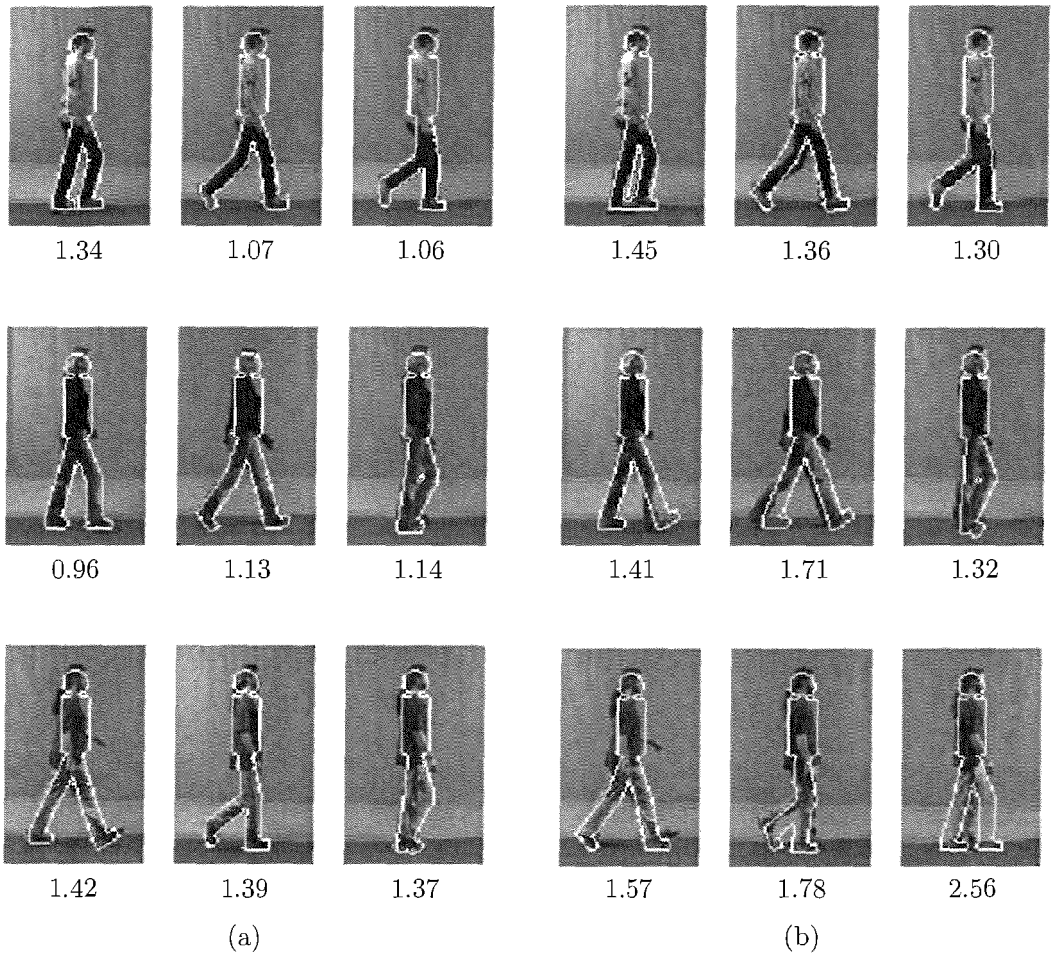
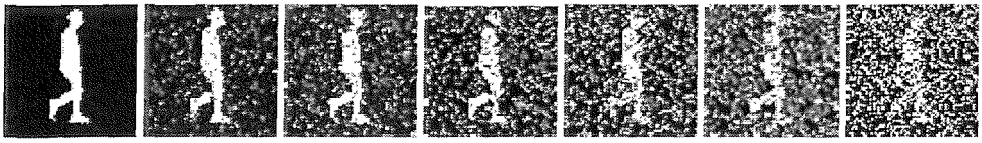


FIGURE 5.1: Examples of extracted models overlaid on their original images. The models shown in (a) were found on noise-free data whereas those in (b) were found after 50% salt and pepper noise—not shown here—had been added to the silhouettes. The numbers below each image are the calculated chamfer distances between the fitted models and the noise-free silhouettes.

was edge detected before chamfer distance computation, so that the average quoted here is that computed across this number of edge points.)

5.2 Sequences with Added Synthetic Noise

To demonstrate the robustness of the system performance, salt and pepper noise was added to 10 normalised high-quality data sequences, each of length 20 images, from different walkers. A percentage p of pixels was randomly chosen; half of these were set to 1 and the remainder were set to 0, irrespective of their original values. Figure 5.2 shows an example frame with different levels of noise added. We then fitted models to these noisy data sequences.



(a) 0% (b) 10% (c) 20% (d) 30% (e) 40% (f) 50% (g) 60%

FIGURE 5.2: Examples of normalised silhouettes with added salt and pepper noise.

Results for 50% noise are shown in Figure 5.1(b) for the same (noise-free) images as in Figure 5.1(a), allowing easy comparison of the two cases and giving further insight into the interpretation of the chamfer distance values. As expected, there is an overall increase in error that is easily seen to result largely from poorer fitting of the dynamic parameters.

The means and corresponding error bars of the chamfer distance between the best-fit model and the original (noise-free) images are shown in Figure 5.3 as a function of percentage of added noise. First, we show the best-fit model obtained by optimising both static and dynamic parameters as described in Section 4.6. The results are shown by a full line. As can be seen, the system is almost completely unperturbed up to 30% noise and deteriorates thereafter. Second, to see how much the static parameters can affect the overall fitting performance, we fit the model to the images by optimising the dynamic parameters only. The results are shown in a dotted line. It is obvious that the optimisation of the static parameters helps improve the performance significantly. Last, to determine the relative contribution of the HMM decoding and the parameter optimisation, we show the results using only the six mean model walkers (dashed line) as in Figure 4.3. It can be seen that the HMM is almost completed unperturbed up to 50% noise and fails catastrophically thereafter. As expected, the gain from the additional parameter optimisation is maximal under low noise conditions and degrades gracefully up to the point of failure. Although not shown on the figure, 100% added noise gives an average error of just below 2.5 pixels. This apparently low value can be understood from the fact that normalization was done before adding the synthetic noise, so that the average static model is automatically placed at approximately the right place. In other words, results for 60% noise are not really distinguishable from those for 100% noise. The fact that an error of about 2.5 pixels corresponds to complete misfit can be checked with reference to the bottom right image of Figure 5.1(b) where the chamfer distance is 2.56 pixels and the fitted limbs are in essentially random positions.

5.3 Sequences with Artificial Occlusion

We also tested the system on the indoor data with artificial occlusion, for the same sequences as in the previous subsection. This is a good exemplar of difficult, structured 'noise' as opposed to the previously-used random salt and pepper noise. The method of

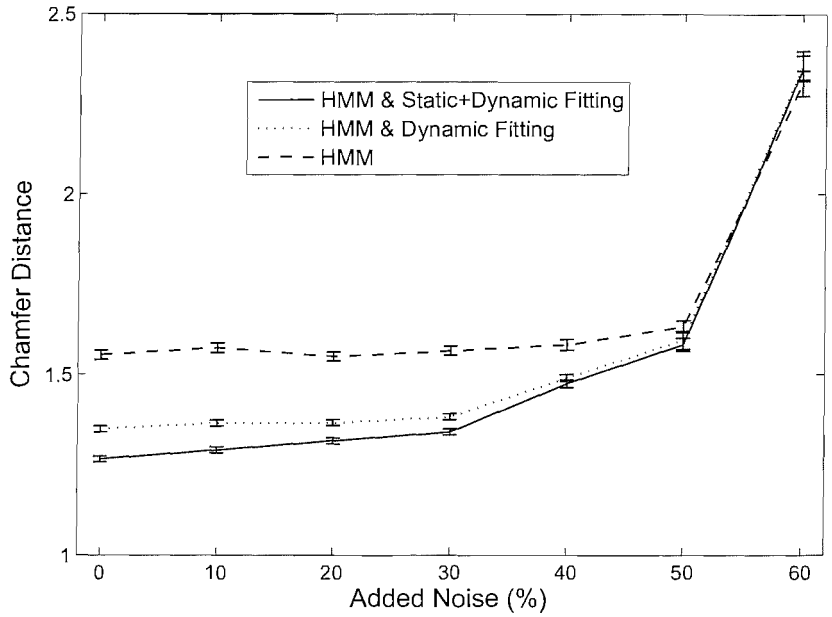


FIGURE 5.3: Means of the chamfer distances between the models extracted from the sequences to which have been added salt and pepper noise and the original clean data. Key: 'HMM' means we use only the six mean model walkers (exemplars) as in Figure 4.3; 'HMM & Dynamic Fitting' means we fit model walkers by optimising dynamic parameters only; and 'HMM & Static+Dynamic Fitting' means we fit using both static and dynamic parameters.

occlusion is illustrated in Figure 5.4(a). The walker is assumed to walk behind regularly-spaced vertical bars. The mean width of the silhouettes in a sequence was calculated and the mid-lines of neighboring bars arranged at intervals of this distance. The width of the bars is expressed as a proportion of this mean width. Figure 5.4(b) shows an example silhouette occluded by bars with different widths.

Chamfer distances were computed between the extracted models and the clean original images as before. Results are shown in Fig. 5.5 in terms of mean chamfer distances and estimated errors of these means versus the occlusion measure. The full line shows the fitting results by optimising both the static and dynamic parameters; the dotted line gives the results by optimising the dynamic parameters only; and the dashed line illustrates the contribution of the HMM decoding. Again, we can see that it is necessary to optimise both the static and dynamic parameters to maximise the system performance.

5.4 Simulated Sequences

Our system extracts human gait from a given sequence in a linear process: 1) the sequence is first normalised so that the walker is centralised in the images; 2) the

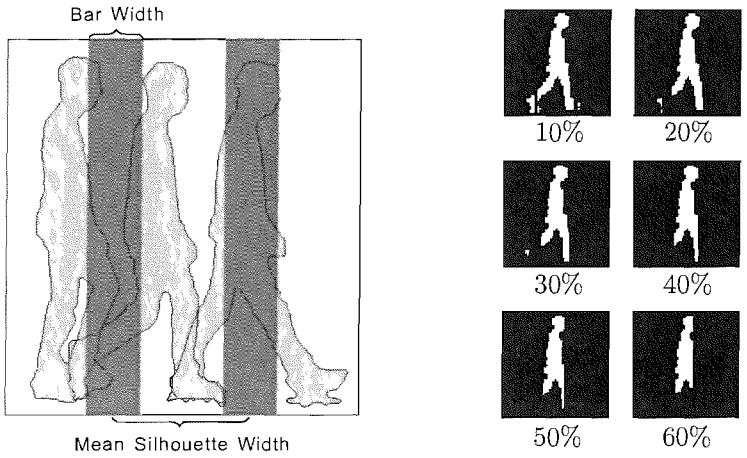


FIGURE 5.4: Artificially-occluded data: (a) illustrates how vertical bars are added to images; (b) shows a sample silhouette occluded by bars with different widths.

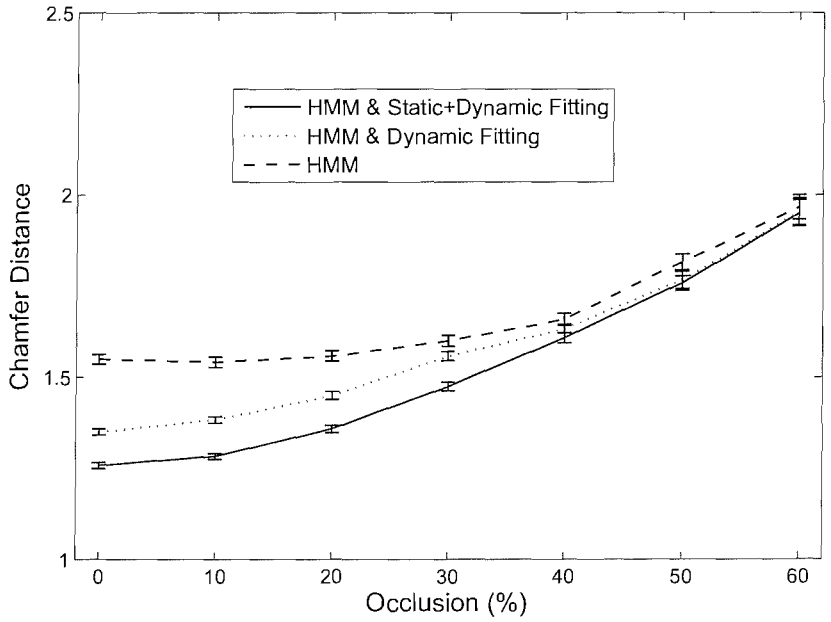


FIGURE 5.5: Means of the chamfer distances between the models extracted from the sequences which have been occluded and the original clean data. Key: 'HMM' means we use only the six mean model walkers (exemplars) as in Figure 4.3; 'HMM & Dynamic Fitting' means we fit model walkers by optimising dynamic parameters only; and 'HMM & Static+Dynamic Fitting' means we fit using both static and dynamic parameters.

normalised images are then labelle by the HMM in order to know the phase of each frame in the gait cycle; 3) after that, we fit the walker model to the images within a Bayesian framework and extract human gait through the optimised model parameters. To explore the effects of possible poor normalisation on the HMM labelling and model fitting, we tested our system on some simulated data.

The simulated data were generated from the 10 indoor image sequences as used in the previous two tests. We perturbed the positions of the silhouettes and added different levels of noise (salt and pepper noise and occlusions). The perturbations are added in the following way:

- Decide the maximum perturbation, e.g., n pixels.
- For each image, choose randomly an integer between $-n$ and n as the horizontal perturbation x_h. Obtain the vertical perturbation x_v in the same way.
- Change the position of the silhouette according to (x_h, x_v).

The artificial noise is added in the same way described in the previous two sections.

We first tested the HMM on the simulated data without any noise added. For each added perturbation (from 1 to 5 pixels), we generated 100 simulated sequences from each of the 10 indoor sequences. The section numbers labelled by the HMM for these indoor sequences were used as the ground truth. Figure 5.6 shows the results. The bars show the percentage of the frames belonging to section k_1 (numbers on the left) mislabelled by k_2 (numbers on the right). As the perturbation increases, we can see the increment of the number and height of the bars. In most of the cases, the images were mislabelled by the section number neighbouring its true section number.

To quantify the results, we define the error of an image belonging to section k_1 but mislabelled by k_2 as:

$$\text{error} = \min(|\mathrm{k_1} - \mathrm{k_2}|, \mathrm{K} - |\mathrm{k_1} - \mathrm{k_2}| + 1) \tag{5.1}$$

where K is the total number of sections in the gait cycle and $1 \leq k_1, k_2 \leq K$. Here we consider section 1 and section K to be adjacent. After determining a way to calculate the mislabelled error, we repeated the above experiment but added different levels of salt and pepper noise and occlusions. Again, the section numbers of the original indoor sequences were used as the ground truth to calculate errors. Figure 5.7 shows the mean errors (per image sequence) for the simulated data with salt and pepper noise added. A group of bars shows the errors of the HMM labelling the simulated data with a particular perturbation added. Each of them is illustrated in a unique grey colour and gives the result for the data with a certain percent of salt and pepper noise added. We also tested

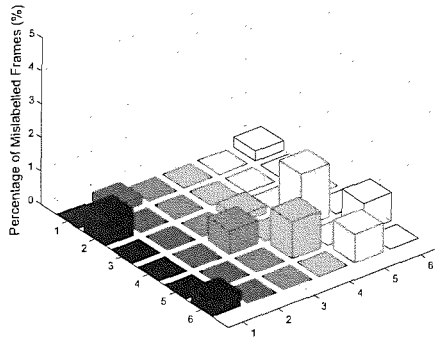
the HMM using the data with 50% noise and the errors where larger than 20, which meant total mislabelling. Therefore, we showed the results for the data with a level of noise up to 40% in the figure. As can be seen, the errors remain reasonably small with small perturbations added. They become larger when more normalisation noise is introduced in the simulation. Figure 5.8 shows the errors of the HMM labelling the simulated sequences with occlusions. We have slightly larger errors from the occlusion data than those from the data with random salt and pepper noise. It is coherent with the experimental results shown in the previous two sections. The HMM broke down when 50% occlusions were added to the perturbed image sequences.

Having tested the HMM on the simulated data, we are now at a stage to simulate the effect of poor normalisation on the model-fitting component. Because of the computational limit, we did not generate as many simulated sequences from the clean sequences as we did to test the HMM component. As before, we added perturbations from 1 to 5 pixels to the images. Only one sequence was generated from each of the clean sequence by adding a certain perturbation and some level of noise (salt and pepper noise or occlusoin).

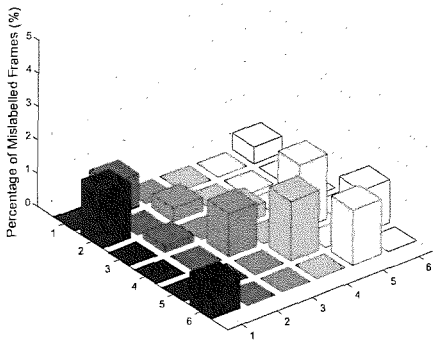
We quantify the results using chamfer distances in the same way as in the previous two sections. Figures 5.9 and 5.10 show the means of chamfer distances between the models extracted from the simulated sequences and original clean data. Six different lines were used to show the results for the image sequences without or with different amount of perturbations in the figures. The level of noise added in the image was up to 40% since the HMM component would fail at 50%. It can be seen that the system performance become worse when we increase the perturbation in the images.

5.5 Outdoor Sequences

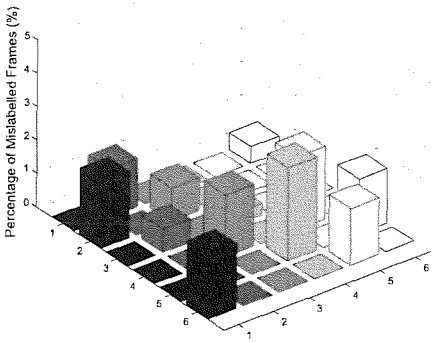
The ultimate test of our approach is how well it performs on image sequences with realistic amounts of noise, exemplified by the outdoor data. Figure 5.11(a) illustrates the complexity of the problem by showing the raw silhouettes used as inputs to the algorithm. These are clearly contaminated with extraneous detail such as a passing bus and another walker at some distance. Although the silhouettes could be 'repaired' (e.g., Grant et al. 2004), the techniques for so doing are *ad hoc* and we wished to avoid using them because our method is intended to cope with challenging data. freffig-noisy-result-model shows the models extracted from the data. To show the fidelity of the fit, Figure 5.11(c) illustrates the original images with outlines of these models superimposed. As can be seen, an accurate fit of the model to the walker is obtained. There are some systematic errors; for example, the forward leg position in frame 4 is slightly misaligned. However, given the degree of noise, this test illustrates the robustness of our algorithm.



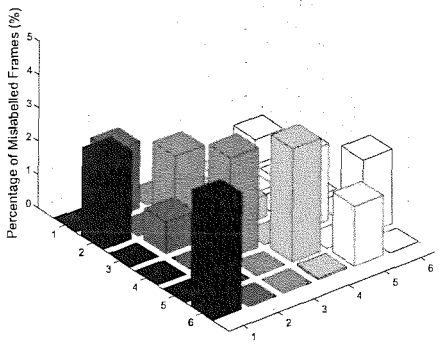
(a) 1 pixel



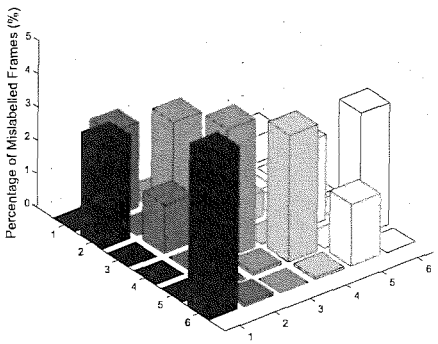
(b) 2 pixel



(c) 3 pixel



(d) 4 pixel



(e) 1 pixel

FIGURE 5.6: Results of the HMM labelling for the simulated sequences without any noise added. For each perturbation (from 1 to 5 pixels), we have 100 simulated sequences generated from each of the 10 indoor clean sequences. Each bar shows the percentage of the frames belonging to section k_1 (numbers on the left) mislabelled by k_2 (numbers on the right).

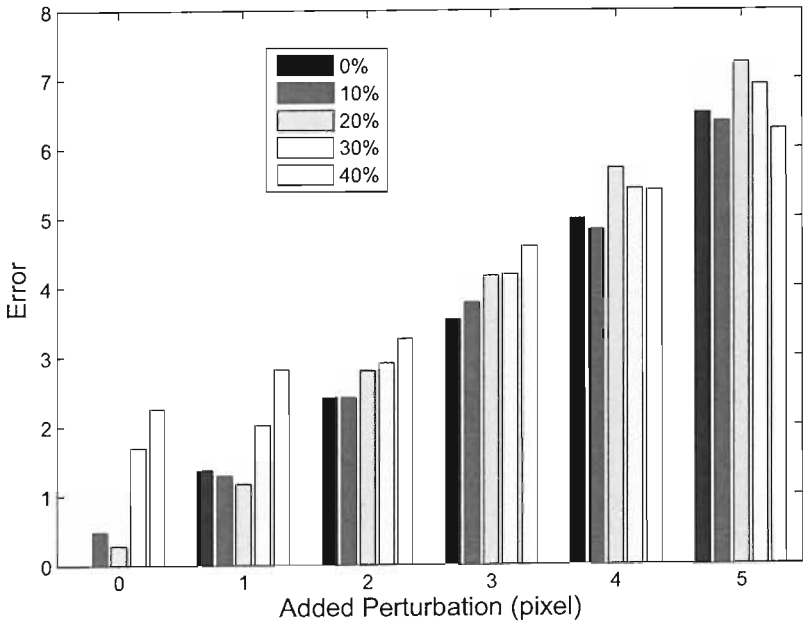


FIGURE 5.7: Means of the errors (per image sequence) of the HMM mislabelling on the simulated data with salt and pepper noise added. The grey colours represent different levels of salt and pepper noise.

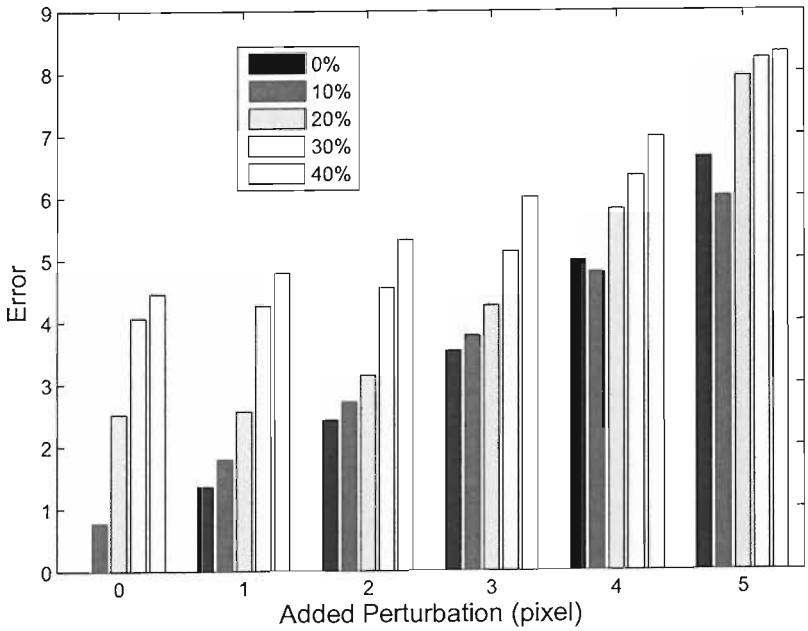


FIGURE 5.8: Means of the errors (per image sequence) of the HMM mislabelling on the simulated data with occlusions added. The grey colours represent different levels of occlusions

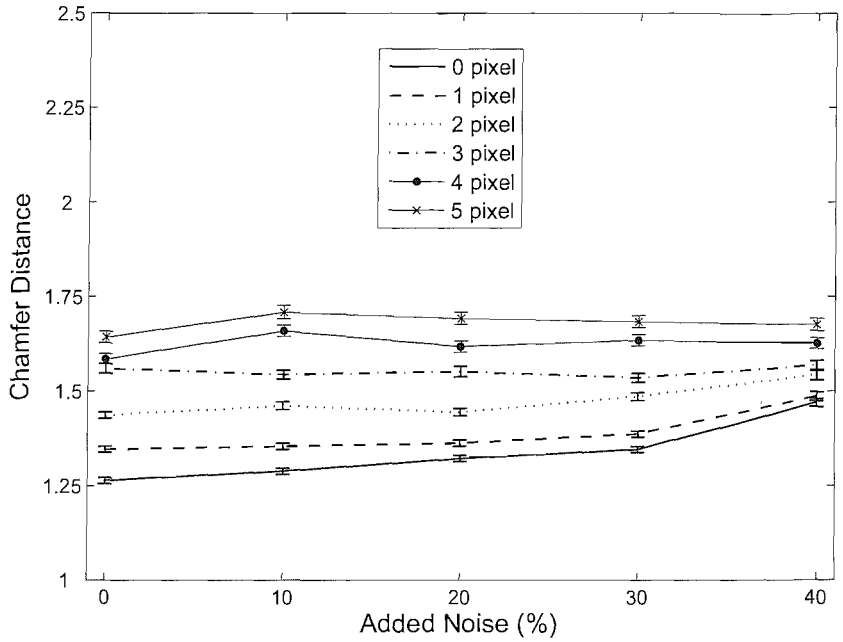


FIGURE 5.9: Means of the chamfer distance between the models extracted from the perturbed sequences to which have been added salt and pepper noise and original clean data. KEY: the amount of perturbation added in the images.

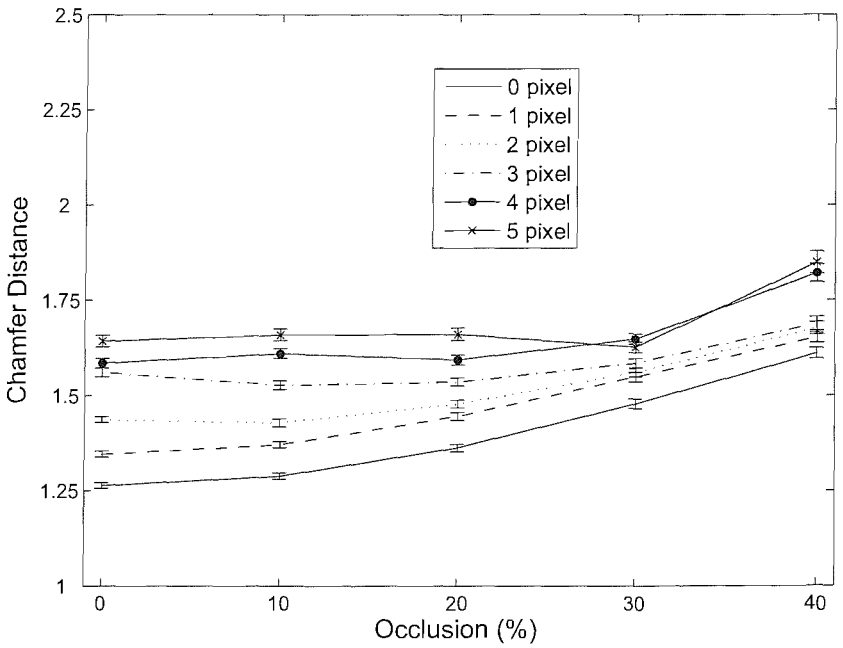


FIGURE 5.10: Means of the chamfer distance between the models extracted from the perturbed sequences to which have been occluded and original clean data. KEY: the amount of perturbation added in the images.

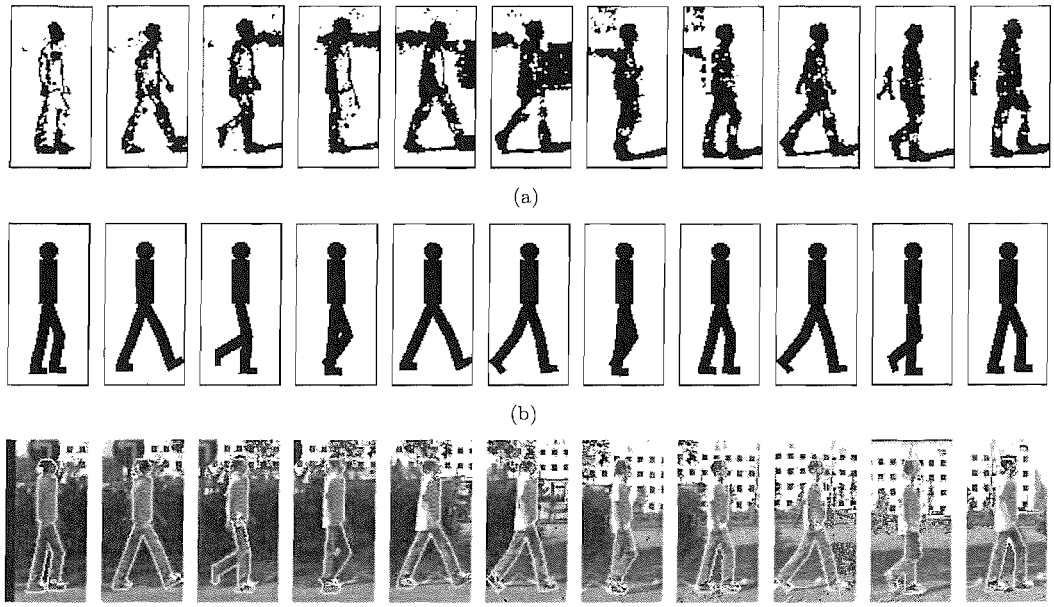


FIGURE 5.11: Typical model fitting results: (a) cropped sample silhouettes; (b) extracted models, and (c) extracted models superimposed on the raw images.

These results, although typical, are illustrative only. Ideally, we wish to quantify performance on as large a set of data as possible. The real problem in so doing is to know what counts as correct. To approximate a 'gold standard', we manually marked the positions of five fiducial points, namely the hip, front/back knee and front/back ankle, in the original data. Although the manual labelling can never be perfect, we feel this is a reasonable, practical compromise. The labelling was done for 10 sequences of 5 walkers from the database, each of which had 50 frames. The gait-extraction algorithm was then applied to these same sequences. We calculated the distances from the hip, front/back knee and front/back ankle positions obtained from the extracted models to the fiducial points in the corresponding images.

In Figure 5.12(a), we illustrate an example frame that has been marked manually and the corresponding points on the model fitted to the walker in that frame are shown in Figure 5.12(b). Table 5.1 shows the means and standard deviations (SD) of the distances computed for each of the five points plus the overall mean and SD. Note that the overall mean and SD for the example sequence shown in Figure 5.11 were 2.49 and 1.35 pixels respectively; this justifies our earlier description of this sequence as 'typical'.

We are unaware of any other system in the literature which is able to cope with this level of 'realistic' noise, especially when no attempt is made at repairing (cf. Grant et al. 2004) the sort of silhouettes seen in Figure 5.11(a). The reason that the results are as good as they are is because we have a consistent way of exploiting the constraints of the problem by treating them as prior knowledge within the Bayesian framework. We

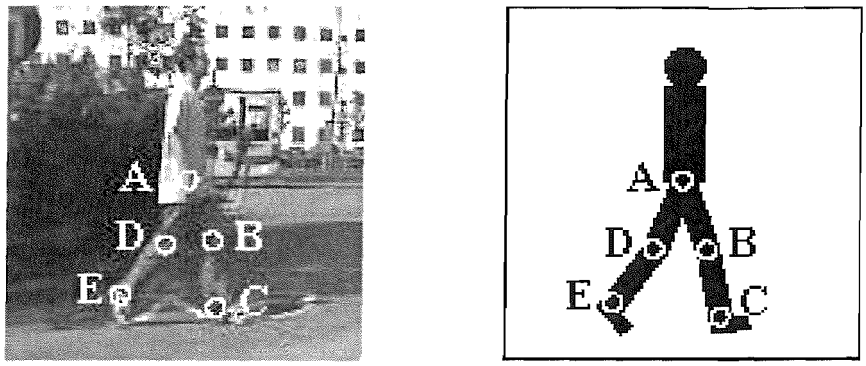


FIGURE 5.12: Joint positions on walker: (a) shows an example frame from one of the video sequence where the joints were marked manually; (b) illustrates the joint positions calculated from the model best fitting the walker in the corresponding frame.

(pixel)	Hip	Front Knee	Front Ankle
Mean	2.94	2.27	2.43
Horizontal Mean	1.12	0.89	1.60
Vertical Mean	2.51	1.92	1.48
SD	1.57	1.40	1.50
Horizontal SD	0.81	0.70	1.43
Vertical SD	1.69	1.47	1.15
(pixel)	Back Knee	Back Ankle	Overall
Mean	2.19	2.62	2.49
Horizontal Mean	1.19	1.29	1.22
Vertical Mean	1.64	2.04	1.92
SD	1.30	1.53	1.49
Horizontal SD	0.91	1.00	1.03
Vertical SD	1.25	1.55	1.48

TABLE 5.1: Means and standard deviations (SD) of the distances between the five joints marked manually on the outdoor data and those obtained by fitting the model.

accept that these constraints (i.e., uninterrupted walk orthogonal to camera direction) do simplify the problem considerably, but this kind of data is very typical of that used in current studies. Such data are available to us in a large database which was expensive and time-consuming to collect, and is starting to become widely-used in gait studies, so it is only sensible to use it at this stage of research.

5.6 Supplemental Data

We tested the framework on the supplemental data from the Southampton HiD database. Tests used 4 sequences of 20 frames each from one walker for each condition (added rucksack, long skirt or trench coat). Figures 5.13, 5.14 and 5.15 show typical results for

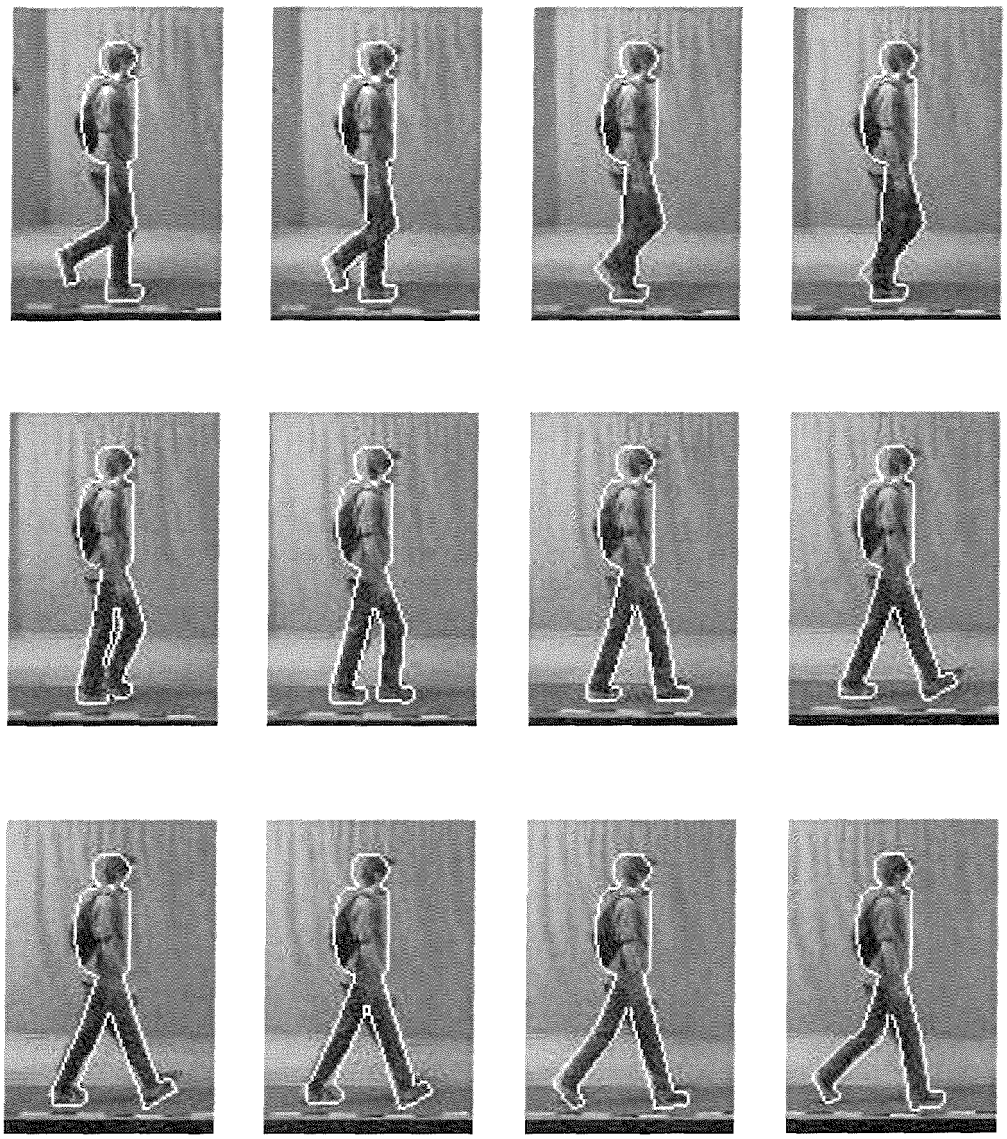


FIGURE 5.13: Gait extraction results for a walker carrying a rucksack.

each of the cases of rucksack, skirt and trenchcoat. It can be seen that the extended articulated models fit reasonably well to the walkers in the images.

We quantified the fitting as for the outdoor sequences, using average joint pixel-error (Table 5.2). As expected, for walkers wearing a long skirt or a trench coat, the errors at the knees are larger than at other points. In spite of this, the overall results (as judged visually) remain good. We believe the level of disruption of body shape which occurs in these sequences would defeat most current approaches to gait extraction.

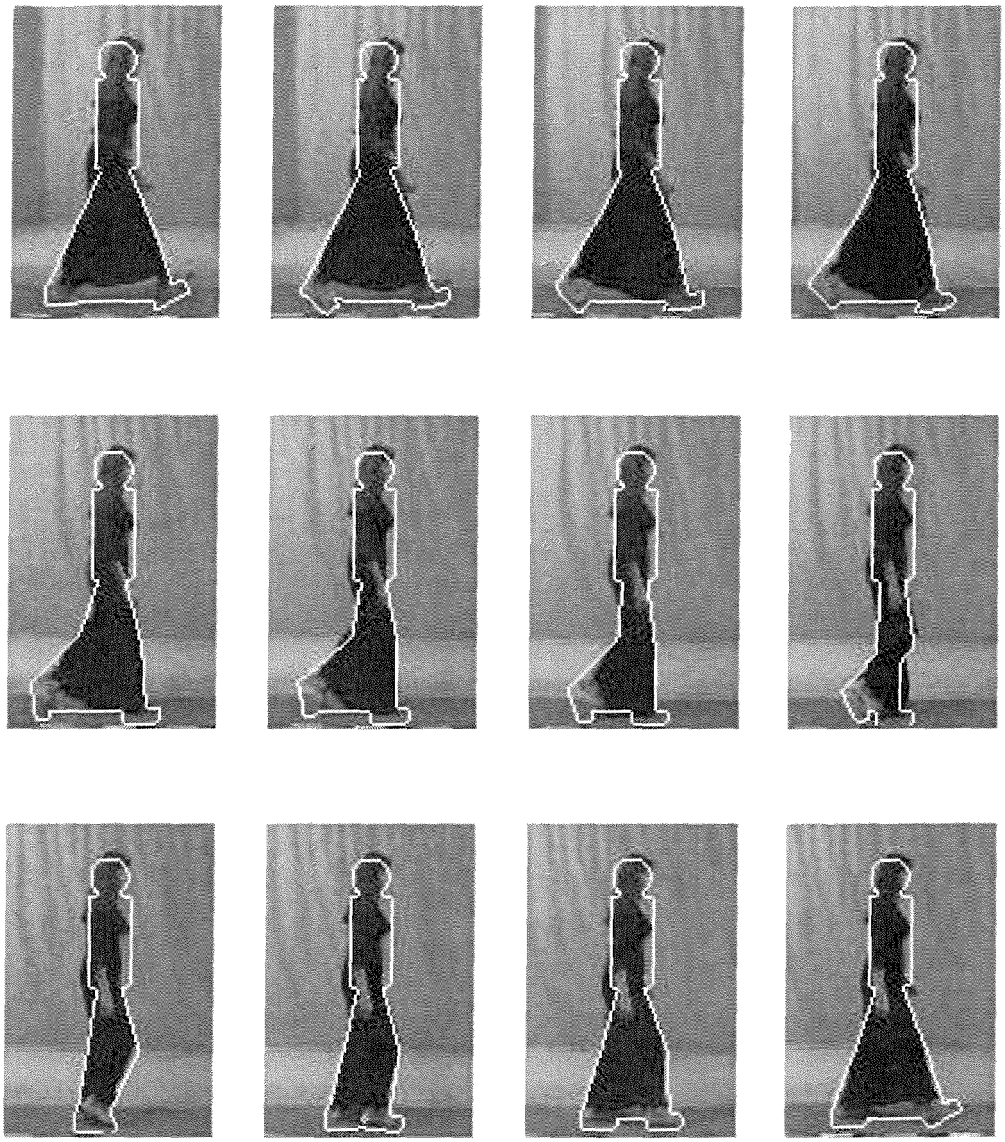


FIGURE 5.14: Gait extraction results for a walker wearing a long skirt.

5.7 Summary

We have demonstrated the robustness and extensibility of the Bayesian framework for extracting human gait through various experiments in this chapter. We first tested this system on the clean indoor data with synthetic noise added to simulate the kind of noise encountered in real world. Two kinds of noise were added: salt and pepper noise and occlusions. We quantified the fitting results using the chamfer distances between the models extracted from the sequences to which had been added noise and the original clean data. We also carried out simulations to explore the effect of poor normalisation on gait extraction. The simulated data were generated by perturbing the positions of silhouettes and adding artificial noise. We showed the errors of the HMM labelling and

FIGURE 5.15: Gait extraction results for a walker wearing a trenchcoat.

the chamfer distances between the models extracted from the simulated sequences and the original clean data. We then tested the system on the outdoor sequences with real-world noise. Good fitting results were achieved and presented visually. The results were quantified by the pixel errors measured between some manually marked joints of the walkers in the original sequences and the ones delivered by the best-fit models. The statistics of the errors demonstrated the accuracy and robustness of the fitting. Finally, we designed an experiment on the difficult supplemental data to test the extensibility of the system. We modified the basic articulated model to deal with the walkers with a rucksack, a long skirt, or a trench coat. Minimum changes were made in the framework and some highly encouraging results were shown in the end.

(pixel)		Hip	Front Knee	Front Ankle
Rucksack	Mean	2.24	3.53	3.79
	SD	1.57	1.80	1.39
Long Skirt	Mean	3.17	6.26	3.72
	SD	2.19	2.70	2.48
Trench Coat	Mean	2.01	6.19	3.63
	SD	1.04	1.63	2.24
(pixel)		Back Knee	Back Ankle	Overall
Rucksack	Mean	2.47	2.71	2.95
	SD	1.58	1.40	1.66
Long Skirt	Mean	6.33	4.68	4.83
	SD	2.64	2.41	2.80
Trench Coat	Mean	5.23	3.01	4.02
	SD	1.71	3.13	2.55

TABLE 5.2: Means and standard deviations (SD) of the distances between the five points marked manually on the sequences with rucksack, long skirt or trench coat and those obtained by fitting the extended model.

Chapter 6

Human Gait Recognition

We have described a Bayesian framework for extracting human gait from video sequences. In this chapter, we extend our system to be able to identify walkers using the extracted gait information. There are three major issues to be considered for any gait recognition system:

- What kind of gait data are used in the system and what experiments are designed on the data?
- What gait features are extracted for recognition?
- How does the system do recognition using the extracted features?

This recognition system has used a subset of the Southampton human HiD database (Shutler et al. 2002). The identification algorithm is tested on silhouettes extracted from video sequences filmed under both indoor and outdoor conditions. The silhouettes from indoor data are of good quality, while those from outdoor data contain real-world noise. We have designed three experiments to test the system, testing on

- Indoor data (silhouettes).
- Outdoor data (silhouettes).
- Outdoor data using the gait information extracted from indoor data as references.

The silhouettes in the first experiment are noise-free and each walker's data were filmed within the same session. The same experiment has been done in some previous work (Hayfron-Acquah et al. 2002; Foster et al. 2003) with high recognition rates reported. Veres et al. (2004) discussed what image information was important for a silhouette-based gait recognition algorithm on the indoor data. The results showed that high

recognition rates could be achieved on such high-quality data by only using body shapes (especially from the head and upper body). The second experiment poses a much more challenging problem for a gait recognition system, that is, how to handle real-world noise, although each walker's data were also filmed within one session. Since the silhouettes were extracted only using background subtraction, the quality of the silhouettes could be affected by various factors, such as illumination, moving objects at background, moving tree leaves, etc. Here, body shapes are no longer reliable and the use of dynamic information is increasingly demanded. The third experiment is the most difficult one. Besides the difficulty involved in the second experiment, we try to increase the difficulty by using the gait information extracted from data filmed in a different setting as references to identify walkers. No results have been reported for the second and third experiments by any other group.

6.1 Recognition Algorithm

We have managed to find a best-fitted model for each frame of a given walking sequence through the Bayesian framework. The parameters of these models are used to construct a feature vector. Since the static parameters are time-invariant and the dynamic parameters are time-variant, we can obtain the static parameters and some time series showing the changes of the joint angles over time from an image sequence. To characterise these time serious, we fit a Fourier series to each of them. The fundamental frequency, amplitudes and phases are viewed as the dynamic features and used to form the gait signature (a feature vector) together with the static parameters.

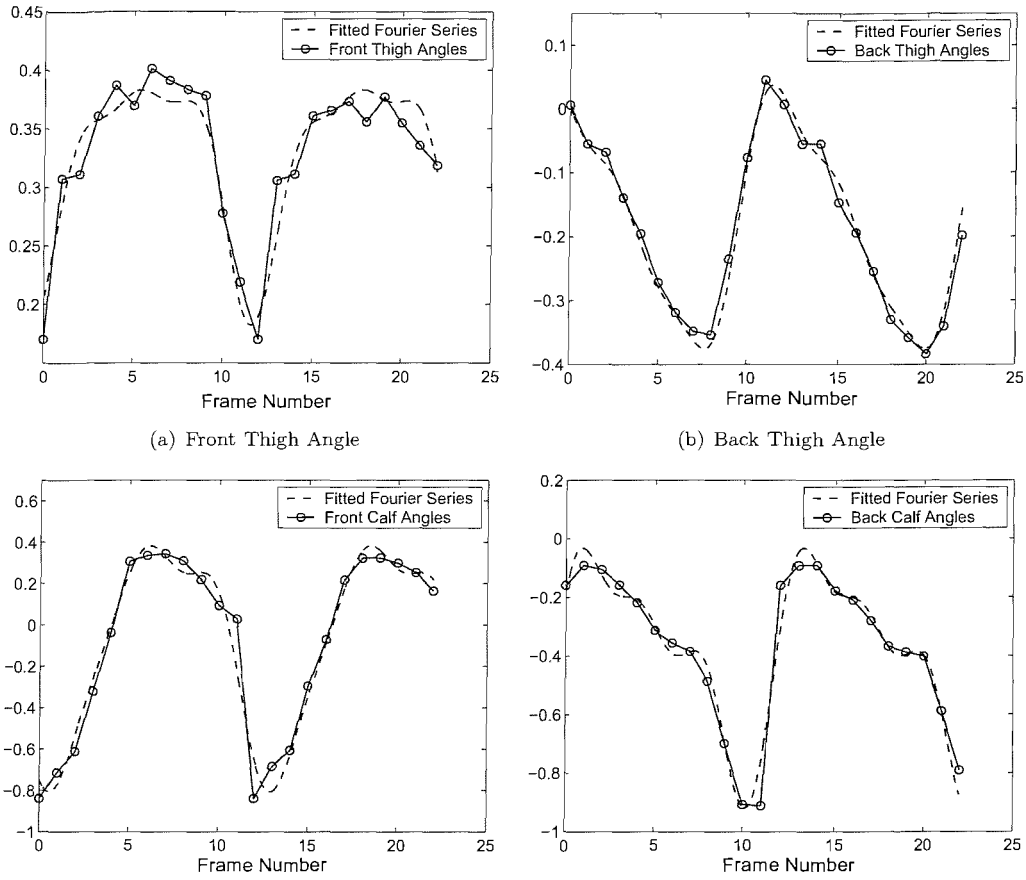
6.1.1 Fitting Fourier Series

We use Fourier series having a limited number of harmonics to approximate the extracted time series. The function $F(t; f_0, \boldsymbol{a}, \boldsymbol{b})$ returns the value of the Fourier series at time t, that is

$$F(t; f_0, \boldsymbol{a}, \boldsymbol{b}) = a_0 + \sum_{k=1}^{K} a_k \cos(2\pi k f_0 t) + \sum_{k=1}^{K} b_k \sin(2\pi k f_0 t). \tag{6.1}$$

where f_0 is the fundamental frequency, $\boldsymbol{a} = \{a_k\}_{k=1}^{K}$ and $\boldsymbol{b} = \{b_k\}_{k=1}^{K}$ are the Fourier series coefficients and K is the highest rank of the harmonics.

To estimate these parameters, we find the best fit between the joint-angle trajectories extracted from the model, $\phi_i(t)$, $1 \leq t \leq T$, and the four Fourier series symbolised by $F(t; f_0, \boldsymbol{a}_i, \boldsymbol{b}_i)$, $1 \leq i \leq 4$. Here the four trajectories describe the angular changes of the front thigh angle, back thigh angle, front calf angle and back calf angle. We have found



(a) Front Thigh Angle

(b) Back Thigh Angle

(c) Front Calf Angle

(d) Back Calf Angle

FIGURE 6.1: Results of fitting Fourier series to the joint-angle trajectories extracted from a sample sequence. The trajectories are displayed by the circles and solid lines while the crosses and dashed lines show the fitted FSs.

that it is sufficient to use $K = 3$. Since the trajectories share the same fundamental frequency, it is sensible to fit them simultaneously. The fitting is done by minimising:

$$\mathcal{E}(f_0, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_4, \boldsymbol{b}_1, \ldots, \boldsymbol{b}_4) = \sum_{i=1}^{4} \sum_{t=1}^{T} \Big(\phi_i(t) - F(t; f_0, \boldsymbol{a}_i, \boldsymbol{b}_i)\Big)^2 \tag{6.2}$$

It is a standard non-linear least square problem and can be well solved by the Levenberg-Marquardt algorithm (Moré 1977). Figure 6.1 shows the fitting results for the time series computed from a sample sequence. The joint-angle trajectories are labelled by circles and solid lines while the dashed lines show the fitted Fourier series.

6.2 Gait Signature

Having fitted the Fourier series to trajectories $\phi_i(t)$, we have the coefficients $\boldsymbol{a}_i = \{a_i^k\}_{k=0}^K$ and $\boldsymbol{b}_i = \{b_i^k\}_{k=0}^K$. We can compute amplitudes $\boldsymbol{A}_i = \{A_i^k\}_{k=0}^K$ and phases $\boldsymbol{\Psi}_i = \{\psi_i^k\}_{k=0}^K$ by:

$$\begin{aligned} A_i^0 &= a_i^0 \\ A_i^k &= \frac{1}{2} \mid a_i^k - b_i^k \mathrm{j} \mid \\ \psi_i^k &= \arg(a_i^k - b_i^k \mathrm{j}) \quad 1 \leq \mathrm{k} \leq \mathrm{K}. \end{aligned} \tag{6.3}$$

Since a gait sequence can start at any phrase in the gait cycle, the phases can not be used directly for recognition. To align them, we use ψ_2^1, the phase of the first-order harmonic of back thigh angle, as a reference and subtract it from other phases. To avoid discontinuity in the angular space, we compute the absolute values of the phases after subtraction. Let $\hat{\boldsymbol{\Psi}}_i = \{\hat{\psi}_i^k\}_{k=1}^K$ be the new aligned phases.

We construct a feature vector z from 36 features including the extracted fundamental frequency f_0, 16 amplitudes, 11 phases and static parameters $\boldsymbol{\tau} = \{\tau_n\}_{n=1}^8$:

$$\boldsymbol{z} = \left(f_0, \boldsymbol{\tau}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_4, \hat{\boldsymbol{\Psi}}_1, \hat{\psi}_2^2, \hat{\psi}_2^3, \hat{\boldsymbol{\Psi}}_3, \hat{\boldsymbol{\Psi}}_4\right). \tag{6.4}$$

Note that $\hat{\psi}_2^1$ is always equal to zero and therefore not added in the feature vector.

The feature vector $\boldsymbol{z}$ is not directly used to identify walkers because of two factors:

- Its elements are evaluated using different metrics (i.e., f is measured in the frequency domain while the static parameters are in pixels.).
- We do not know whether the features have equal discriminatory abilities or not.

Considering these two factors, we first normalise each element of the feature vector by subtracting the mean and then being divided by the standard deviation. To quantify the discriminative abilities of the features, we perform analysis of variance (ANOVA) and weight them by their F-statistics. See Appendix F for details of ANOVA and F-statistic. Using F-statistics to weight features has been proven to be effective in Lee (2003).

Mathematically, we can express the normalised and weighted new feature $\hat{z}_i$, $1 \leq i \leq 36$, as:

$$\hat{z}_i = w_i \frac{z_i - \mu_i}{\sigma_i} \tag{6.5}$$

where w_i is the F-statistic, μ_i the mean and σ_i the standard deviation of z_i.

6.3 Classification

We use the nearest-mean classifier to classify the normalised and weighted feature vectors. Assume that there N walkers to be recognised and each of them has M training sequences. After the model fitting and feature extraction, we have M feature vectors $\{\boldsymbol{z}_n^m\}_{m=1}^M$ for walker n. We then calculate the class mean $\boldsymbol{c}_n = \frac{1}{M}\sum_m \boldsymbol{z}_n^m$. Given an unknown sequence, we first compute the feature vector $\boldsymbol{z}'$. After that, we calculate the Euclidean distance between $\boldsymbol{z}'$ and each of the class means. The more likely the sequence is classified as walker n, the less distance there is between $\boldsymbol{z}'$ and the class mean $\boldsymbol{c}_n$.

Leave-one-out cross validation is used to test the performance of our gait-recognition system. Cross validation partitions the data we have into subsets such that a single subset is considered as the unknown testing data, while other subsets are the classified training data. Leave-one-out cross validation uses a single data point, which in our case is an image sequence, from the original data as the testing data, and the rest of the data as the training data. This process is repeated such that each data point is used once as the testing data.

6.4 Experiments and Results

We tested our gait recognition system first on the indoor data. Fifty subjects in the gait database were selected, each with 10 sequences. We computed the best-fitted models for the dataset and from the model parameters constructed a feature vector for each sequence. As mentioned in the previous section, F-statistics were calculated as weights to quantify the discriminatory capability of each gait feature. ANOVA was implemented and the output p-values and F-statistics are listed in Table 6.1. It can be seen that most features have relatively small p-values which indicate certain discriminative capabilities. The fundamental frequency f_0 has the biggest F-statistic. Among the static features, only τ_4 which stands for the leg width has a relatively large F-statistic. For the dynamic features, the amplitudes have larger F-statistics than the phases do. Leave-one-out cross validation was performed on the chosen data by a nearest-mean classifier with a Euclidean distance metric. The classifier calculates the Euclidean distances between a feature vector and class means, and chooses the class whose mean is nearest to the vector.

	f_0			
p-value	$< 10^{-20}$			
F-statistic	67.71			
	τ_1	τ_2	τ_3	τ_4
p-value	3.48×10^{-013}	$< 10^{-20}$	$< 10^{-20}$	$< 10^{-20}$
F-statistic	3.65	7.20	6.65	23.85
	τ_5	τ_6		
p-value	9.44×10^{-15}	2.66×10^{-15}		
F-statistic	3.93	4.03		
	A_1^0	A_1^1	A_1^2	A_1^3
p-value	$< 10^{-20}$	$< 10^{-20}$	$< 10^{-20}$	$< 10^{-20}$
F-statistic	36.60	17.06	5.11	4.80
	A_2^0	A_2^1	A_2^2	A_2^3
p-value	$< 10^{-20}$	$< 10^{-20}$	$< 10^{-20}$	9.11×10^{-6}
F-statistic	23.23	16.58	6.03	2.26
	A_3^0	A_3^1	A_3^2	A_3^3
p-value	$< 10^{-20}$	4.44×10^{-16}	$< 10^{-20}$	1.13×10^{-11}
F-statistic	13.81	4.17	11.85	3.38
	A_4^0	A_4^1	A_4^2	A_4^3
p-value	$< 10^{-20}$	$< 10^{-20}$	$< 10^{-20}$	1.53×10^{-10}
F-statistic	10.31	4.97	4.42	3.17
		ψ_1^1	ψ_1^2	ψ_1^3
p-value		$< 10^{-20}$	1.08×10^{-2}	0.89
F-statistic		9.41	1.57	0.58
			ψ_2^2	ψ_2^3
p-value			0.11	0.20
F-statistic			1.27	1.18
		ψ_3^1	ψ_3^2	ψ_3^3
p-value		$< 10^{-20}$	0.14	8.12×10^{-2}
F-statistic		1.05	1.24	1.32
		ψ_4^1	ψ_4^2	ψ_4^3
p-value		$< 10^{-20}$	0.13	4.17×10^{-2}
F-statistic		4.63	1.24	1.41

TABLE 6.1: F-statistics and p-values of the gait features extracted from noise-free indoor data.

To make a comparison, we implemented a baseline algorithm (Sarkar et al. 2005) on the silhouettes as well. It has been tested on various challenging datasets and proven to be robust and efficient on silhouette data. The algorithm evaluates the similarity between two sequences by computing temporal correlation of silhouettes. One deficiency of this algorithm is that it requires relatively accurate normalisation, that is, finding bounding boxes around walkers, which involved manual work in their work. We have provided an effective normalisation process in the framework and therefore can run the baseline algorithm directly on the normalised silhouettes. Details of the baseline algorithm can be found in Appendix E.

Following the way of presenting face recognition results in Phillips et al. (2000), we use cumulative match characteristics (CMC) to report the recognition results. Figure 6.2 shows the two CMC curves for the baseline and for our algorithm on the noise-free data. Each point in the figure corresponds to a value of k on the abscissa (labelled 'rank') and a probability p. This is the probability that the correct walker is included within the k top-ranking candidates. The performance of the baseline algorithm is superior in this case. It achieves 79%, 97% and 100% at ranks 1, 5 and 10, whereas the corresponding recognition rates for our algorithm are 61%, 92% and 99%. We interpret these results in the light that a simple identification algorithm can do rather well on the relatively 'easy' indoor data; there is no advantage to using anything more complicated. Further, we believe our algorithm is adversely affected by poor estimation of the dynamic parameters (having a strong dependence on f_0). By contrast, the baseline algorithm performs well on this high-quality database because it can exploit discriminative information about the shape of the head and upper torso.

We chose the same walkers as in the first experiment for outdoor data. Each of them had 10 sequences in the testing dataset. Following the same process as for indoor data, we performed ANOVA on each of the extracted gait features and results are listed in Table 6.2. Compared with Table 6.1, we can see that F-statistics of most features become smaller, which is expected since the human gait extracted from noisy outdoor data should be less accurate than from indoor data. Among the static features, the leg width is again the one having largest F-statistic.

Figure 6.3 shows the recognition results for outdoor data. The performance fell down in comparison with the performance on indoor data. For the baseline algorithm, there is 47% recognition rate drop at rank 1, while at rank 5, the decrement reduces to 23% and only 4% at rank 10. Our algorithm has a much smaller loss in performance (15%) in terms of recognition rates at the first rank, but encounters the largest fall (17%) at rank 5. As we mentioned before, the real test of our approach is how well it performs on the difficult (and much more realistic) outdoor data. Here, we can see that at the low ranks our new algorithm does significantly better than the baseline algorithm, which is known to perform well across a range of gait recognition scenarios.

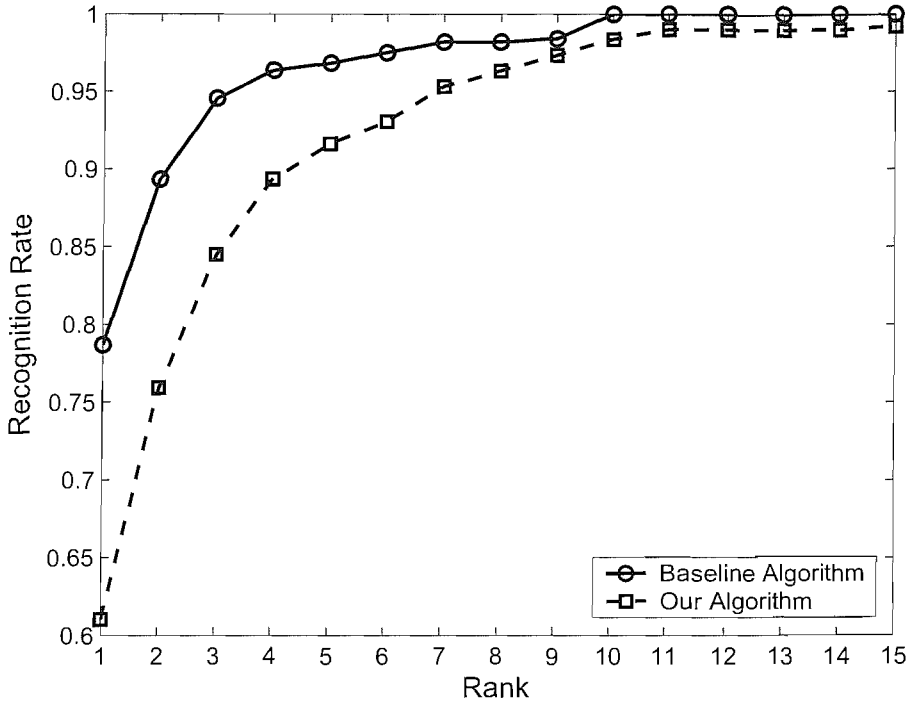


FIGURE 6.2: Cumulative match characteristics for the identification of walkers from indoor data for the baseline algorithm and the new Bayesian approach.

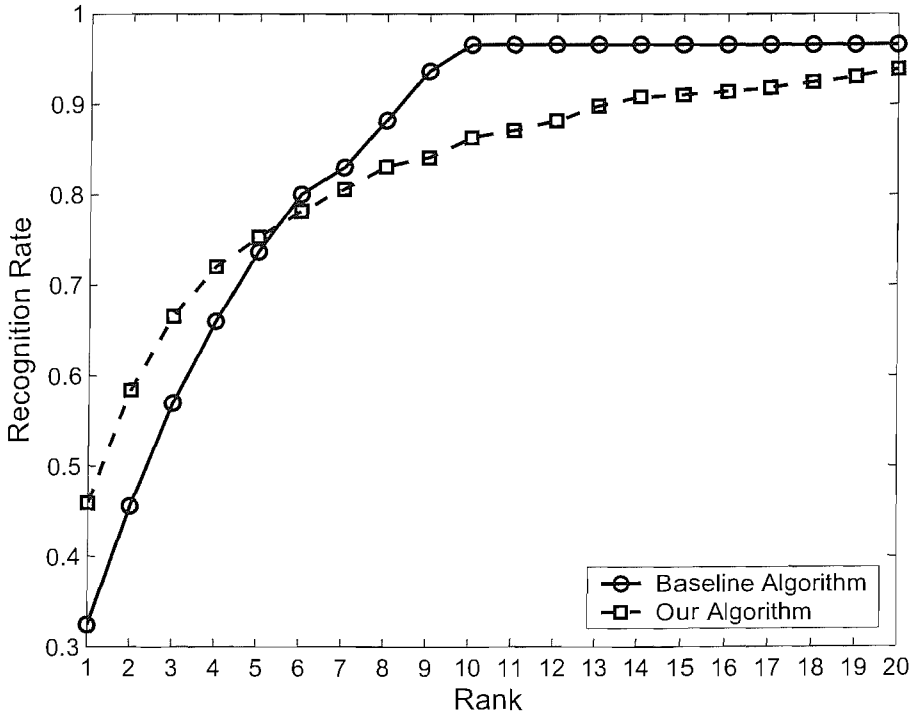


FIGURE 6.3: Cumulative match characteristics for the identification of walkers from outdoor data for the baseline algorithm and the new Bayesian approach.

	f_0			
p-value	$< 10^{-20}$			
F-statistic	20.46			
	τ_1	τ_2	τ_3	τ_4
p-value	1.41×10^{-2}	3.33×10^{-16}	$< 10^{-20}$	$< 10^{-20}$
F-statistic	1.54	4.21	4.96	42.08
	τ_5	τ_6		
p-value	6.89×10^{-8}	9.63×10^{-11}		
F-statistic	2.67	3.21		
	A_1^0	A_1^1	A_1^2	A_1^3
p-value	$< 10^{-20}$	$< 10^{-20}$	4.34×10^{-11}	7.36×10^{-10}
F-statistic	10.44	5.25	3.27	3.04
	A_2^0	A_2^1	A_2^2	A_2^3
p-value	$< 10^{-20}$	$< 10^{-20}$	3.67×10^{-7}	4.69×10^{-12}
F-statistic	13.71	5.76	2.53	3.44
	A_3^0	A_3^1	A_3^2	A_3^3
p-value	$< 10^{-20}$	2.79×10^{-11}	$< 10^{-20}$	2.04×10^{-3}
F-statistic	6.75	3.31	7.01	1.75
	A_4^0	A_4^1	A_4^2	A_4^3
p-value	1.67×10^{-15}	2.53×10^{-8}	1.66×10^{-12}	7.36×10^{-3}
F-statistic	4.07	2.76	3.53	1.62
		ψ_1^1	ψ_1^2	ψ_1^3
p-value		$< 10^{-20}$	0.15	0.69
F-statistic		7.01	1.23	0.89
			ψ_2^2	ψ_2^3
p-value			4.46×10^{-2}	0.19
F-statistic			1.40	1.19
		ψ_3^1	ψ_3^2	ψ_3^3
p-value		1.24×10^{-3}	1.65×10^{-3}	8.73×10^{-2}
F-statistic		1.81	1.78	1.31
		ψ_4^1	ψ_4^2	ψ_4^3
p-value		6.77×10^{-13}	2.30×10^{-3}	0.12
F-statistic		3.60	1.74	1.27

TABLE 6.2: F-statistics and p-values of the gait features extracted from noisy outdoor data.

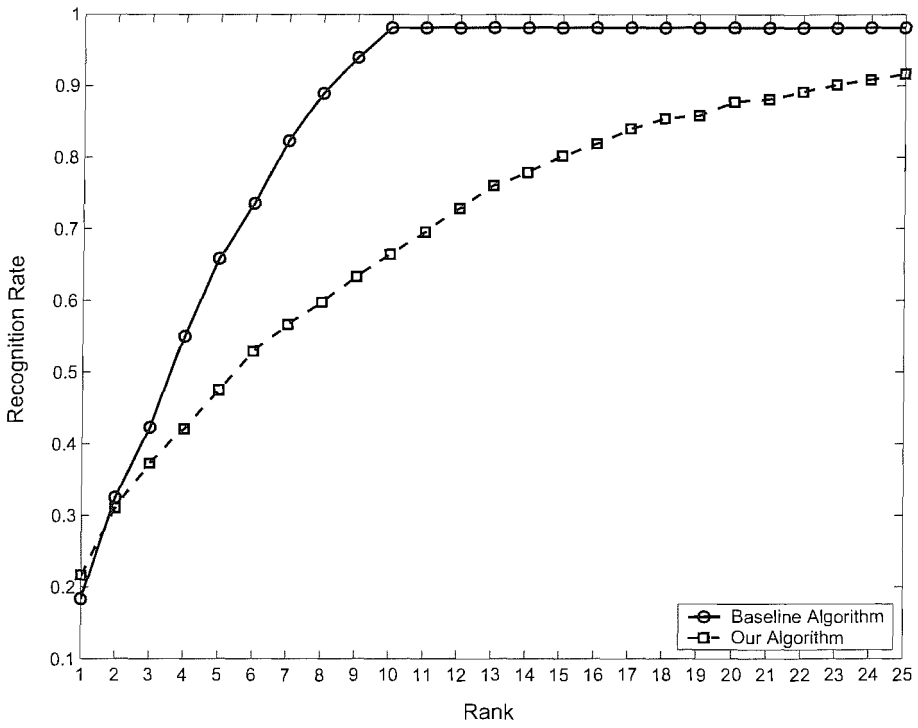


FIGURE 6.4: Cumulative match characteristics for the experiment aiming to identify walkers from outdoor data using indoor data as references.

The last experiment is to use the gait information extracted from the indoor data as references to identify walkers in outdoor data. In other words, we learned the class means from the feature vectors extracted from the known and noise-free indoor sequences and used these means to classify the feature vectors calculated from the unknown and noisy outdoor sequence. Note that we used the F-statistics of the feature vectors from outdoor data as the weights to generate gait signatures. Both indoor and outdoor datasets in the previous two experiments are used here. As mentioned at the beginning of this chapter, this experiment poses the most challenging problem since indoor and outdoor data are filmed in a different setting on the same day and the extent of noise differs significantly in two kinds of data. Figure 6.4 shows the results. Both methods had a recognition rate around 20% at the first rank. The CMC curve of the baseline algorithm then approaches the high-recognition-rate area much faster than our algorithm.

The baseline algorithm outperformed this system on the clean indoor data. It employs the shape information of the high-quality silhouettes simply by calculating the correlations between silhouettes. In contrast, the simple articulated model used in our system is only a crude representation of a walker and there, cannot extract enough silhouette information from the clean silhouettes to achieve high recognition rates. For the outdoor data, the silhouettes are much more noisy. It is more reliable to use the dynamic information (joint angles). Our recognition system had a better performance on the outdoor noisy sequences since we used a Bayesian framework for extracting the

dynamic information and such a framework could tolerate real-world noise by exploiting our strong prior knowledge. In both of the experiments, the leg width was reported to have the largest F-statistic. It is coincident with the results shown in Wagg and Nixon (2004a). Further research work is required to interpret this interesting result.

6.5 Analysis of the Contribution of Different Features

It has been demonstrated by the results shown in the previous section that the gait features we extract have discriminatory capabilities for gait recognition. However, we do not know whether the gait features contributed to the recognition rates equally or not. In this section, we extend the previous experiments by only using certain gait features to test their capabilities of identifying walkers.

We are interested in three kinds of features: the fundamental frequency f, the static features which are just the static parameters of the extracted models, and the dynamic features which are the amplitudes and phases obtained by fitting Fourier series to joint-angle trajectories. Each kind was used to recognise walkers on both indoor and outdoor data alone. Results are shown in Figures 6.5 and 6.6. We can see that the dynamic features provided most of the biometric information. The static features did not contribute much to the overall performance. This was expected since the 2D articulated model we build for walkers is very simple. Some parts of the model are not appropriate from the anatomical point of view. However, it indicates the potential improvement for our method as the body-shape information could result in a high recognition rate (Veres et al. 2004).

6.6 Summary

Human gait extracted by the Bayesian framework has been used to identify walkers in this chapter. Three experiments were designed to test the discriminatory capabilities of the extracted gait information, that is, testing on indoor data, testing on outdoor data, and testing on outdoor data using the gait information extracted from indoor data as references.

To build the gait signature, a feature vector, for each sequence, we fit Fourier series to the time series formed by the extracted joint angles. The computed fundamental frequency, amplitudes and phases together with the static parameters are used as the gait features. ANOVA is then performed on each feature to test its discriminatory ability. We normalised the features to have a zero mean and a unit deviation and weight them by their F-statistics.

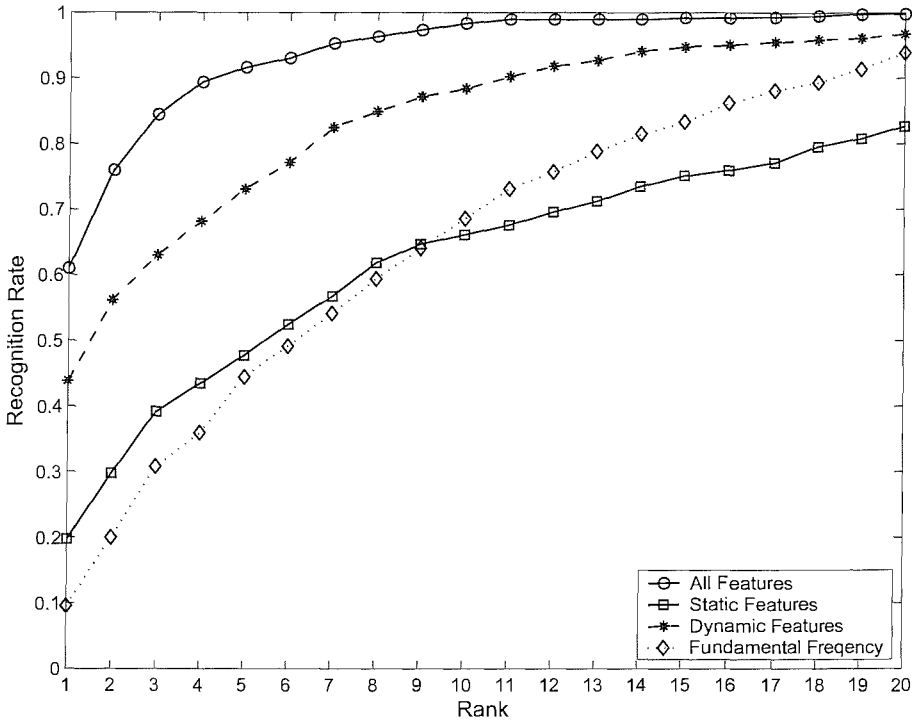


FIGURE 6.5: Comparisons of the recognition performance on indoor data using fully or partially the extracted gait features.

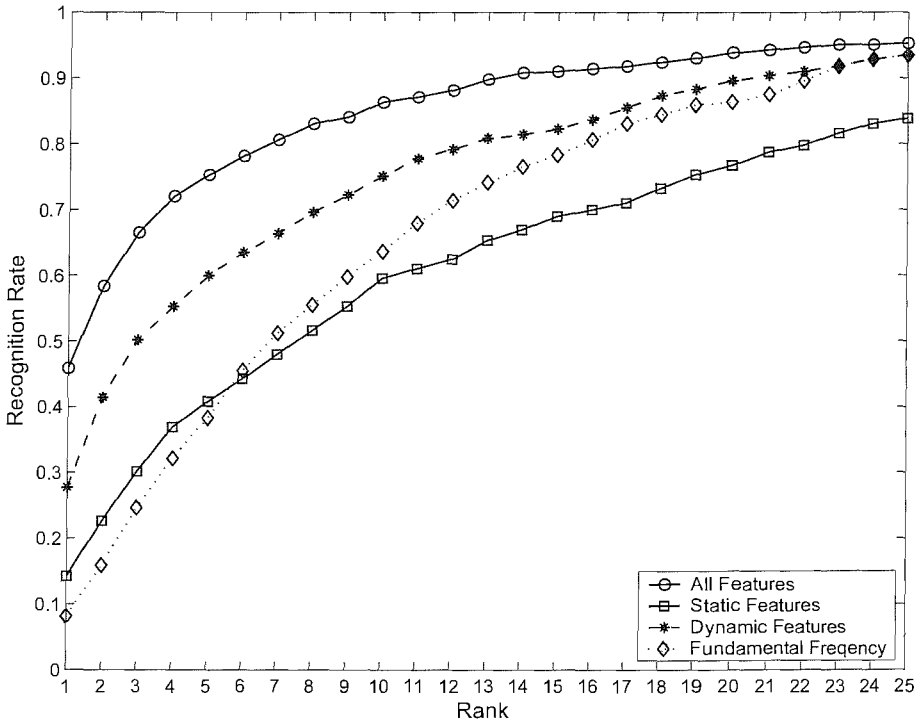


FIGURE 6.6: Comparisons of the recognition performance on outdoor data using fully or partially the extracted gait features.

We compared the performance of our gait recognition system with the baseline system (Sarkar et al. 2005). The baseline algorithm has a better performance on indoor data, but the recognition rates drop rapidly on outdoor data that contain real-world noise. In contrast, our method is more robust to the noise in the images. The discriminatory capabilities of certain kinds of features were also tested. It is found that the dynamic features (amplitudes and phases) provide most of the biometric information. The low recognition rates for the static features indicate the possibility of improving the performance of our system by introducing a more sophisticated model for walking people.

Chapter 7

Conclusion

Capturing the motion of walking people from image sequences is increasingly being demanded by applications in computer vision, computer animation, biometrics, etc. Many systems have been built for this problem, extracting the articulated motion or tracking the contours of the walkers in the images. However, most of them lack the capability to handle the inherent uncertainties in this problem (e.g., the motion often occurs in a complex condition, corrupted by real-world noise, or can be severely occluded by the clothing of the walker.) This thesis focuses on developing a general framework to cope with these uncertainties to extract the motion accurately from image sequences. This concluding chapter summarises the work presented in this thesis, analyses the limitation of this method, and suggests some future directions.

7.1 Summary of Work

We have adopted a Bayesian framework for extracting human gait. The framework is designed to be capable of handling the uncertainties including noise and occlusion. We have built our system using the Southampton HiD gait database. It consists of indoor noise-free, outdoor noisy and supplemental data. We used the indoor image sequences as the training data and tested the framework on the other two kinds of data. The outdoor images test the framework's ability to cope with real-world noise, while the supplemental data challenge the system to cope with the significant changes of the body shapes caused by carrying rucksacks and severe occlusion caused by wearing skirts or trench coats. Details of the database were given in Chapter 3. The chapter also involved the normalisation for the input images.

Chapter 4 described the construction of the framework. To compensate for the complexity of the gait-extraction problem, we exploited our prior knowledge of human walk within a Bayesian formalism. A simple articulated model was built to model the

articulation of the human body. Moreover, the periodicity of the walking motion was well captured by a hidden Markov model. We used a new technique, the PDF projection theorem to learn the high-dimensional observation probabilities from a one-dimensional image-distance metric space, which avoided the curse of dimensionality. The articulated model was fitted to the images by maximising the posteriors (MAP). For this MAP problem, we constructed a strong prior based on the statistics of the parameters of the articulated model. The prior imposes a soft constraint when searching for the optimal set of parameters, but gives a large penalty when the search process goes out of the feasible region in the parameter space. The statistics of the parameters were bootstrapped from a small amount of indoor data and refined by the updating component using more indoor sequences as training data. In the end, we demonstrated the flexibility and extensibility of the framework by modifying the articulated model to cope with walkers with different body configurations.

In Chapter 5, we designed various experiments to demonstrate the robustness and extensibility of the framework. We first tested the system on synthetic noisy images. These images were generated from the indoor data with two kinds of artificial noise added: salt and pepper noise and occlusions. The results were quantified by the image distance (chamfer distance) and showed the robustness of the framework against the artificial noise. To explore the effect of possible poor normalisation on the performance of rest of the system, we tested our system on some simulated sequences. These sequences were generated by perturbing the positions of silhouettes and adding artificial noise. We then tested the system on the outdoor sequences with real-world noise. Good fitting was achieved by the framework and both visual and quantitative results were given in this chapter. The final experiment examined the extensibility of this framework to handle the significant changes of body shapes and severe occlusions of the limbs. The articulated model was modified slightly (adding one or two static parameters) to cope with a rucksack, a long skirt or a trench coat. Without learning new priors, the framework achieved a reasonable fit between the modified model and the walkers.

We have argued several times in this thesis that the Bayesian framework is designed for accurate extraction of human gait from noisy image sequences. Although we have presented the fitting results in Chapter 5, it would be helpful if we studied a typical application using the extracted gait information and compared the results with other methods for this application. In Chapter 6, we attempted to use the extracted models to identify walking individuals. For each walker, the gait features consist of the static parameters and the Fourier descriptors (amplitudes and phases) of the trajectories of the dynamic parameters. These descriptors were obtained by fitting Fourier series to the trajectories. To make a comparison, we implemented a baseline algorithm (Sarkar et al. 2005) which has been shown to be robust against noise. Based on the test results of the two systems on both indoor and outdoor data, we found that our system had higher

recognition rates on outdoor data and less loss of performance from indoor to outdoor data, indicating the robustness of the framework against noise.

7.2 Limitations of Work

The limitations of this work are discussed in this section. First of all, the system is built on two assumptions that walkers are viewed from the side in the image sequences and only one walking individual is analysed by the system at a time. Secondly, the system is automated based on the probability distributions (transition and observation probabilities of the HMM) and the statistics of the parameters of the articulated model learned from the training data. Such prior knowledge makes the framework effective and robust. However, it is difficult for the system to deal with the motion that is not represented by the training data (e.g., subjects in image sequences walk very fast or slowly). To cope with, for instance, the fast walking speed, it would be necessary to relearn the system parameters from new training data for the fast walk. To reduce the ambiguity when fitting the articulated model to a walker, we distinguish only front/back legs instead of the traditional left/right legs. Finally, the system can not do the motion extraction at real time. Such a characteristic is determined by the way we optimise the model parameters, which requires the entire image sequence to be available (see Section 4.6 for details).

7.3 Future Work

We have described a Bayesian framework for capturing the motion of walking people. The framework is robust and powerful since it provides a natural way to incorporate our prior knowledge of the motion. Each component of the framework is simple and relatively independent from the others, which makes it easy to extend this work to tackle more difficult problems. In this section, we discuss some future work.

7.3.1 Extracting Human Gait in 3D

The current framework is view-dependent and can only extract 2D motion from the images. We believe that the system could be extended to handle different views within the same Bayesian framework. A 3D articulated model could be introduced and the same process be performed on the training data to learn the statistics of its parameters. Once we have obtained the means, we can project the 3D model onto the image plane easily from a given view angle to generate the exemplars for the HMM. We could then learn the observation probability distributions using the PDF projection theorem. One thing to be concerned is the computational cost of the optimisation. Since there are

more parameters introduced by the new model, we might need to use a more efficient optimisation process.

7.3.2 Detecting Changes of Walking Speed

We can extend our system to detect different walking speed automatically. At the moment, for a subject walking fast or slowly, we need to rebuild the HMM and relearn the parameters of the HMM. Lan and Huttenlocher (2004) presented a system that could capture the dynamics of a subject walking along a circle by a hidden Markov model. The HMM contains multiple state cycle each of which models the dynamics of the walker viewed from a particular angle. There are transitions connecting these cycles to cope with the changes of the view in the images. A similar idea could be used to detect the changes of the walking speed. For a particular speed (e.g., very fast, fast, slow, or very slow), we can build a state cycle to model the dynamics of the walk as we have done for the normal speed in the framework. We then connect these cycles to form an HMM with a more complex structure. Given an image sequence, we try to find the state sequence with the largest likelihood. The output of the HMM would tell us not only the position of the image in the gait cycle but the speed of the walker.

7.3.3 Model Selection

In Section 5.6, we showed that we could change the model to capture rucksacks, long skirts and trench coats. This raises the issue of how to select the correct model for a walker. As we are using a Bayesian framework, we can in principle compute the evidence to allow us to perform model selection. One way to evaluate this evidence could be that we build an HMM for each of the models representing various body appearances. For a particular walking speed, the HMMs could share the same structure, transition probabilities and initial probabilities. They differ from each other as to the observation probability distributions. Different exemplars are to be generated from the models to learn the distributions using the PDF projection theorem (see Chapter 4 for details of the learning process). Given an image sequence, we calculate its likelihood given each of the HMMs and choose the model corresponding to the HMM with the largest likelihood.

7.3.4 Other Applications Using the Bayesian Framework

We have demonstrated the robustness of the Bayesian framework against noise when extracting human gait from image sequences. We believe that such a framework can be well applied to other computer vision applications dealing with periodic motion. Once a parametric model is constructed, we can follow the learning process as described in Chapter 4 and make the system fully automatic. For example, we can implement the

framework for extracting the motion of hearts. It involves building a parametric shape model for a heart, constructing an HMM, evaluating the probability distributions for the HMM, learning the statistics of the parameters of the heart model to deliver strong priors in the Bayesian framework, and optimising the parameters.

Appendix A

Silhouette Examples

We present more examples of the silhouettes used to test the framework in this appendix. Although some of the silhouettes have been shown in Chapter 3, we think that more examples will give a clearer picture of the variations of the image data in the HiD gait database (Shutler, Grant, Nixon, and Carter 2002). In the following, we will give 4 sample sequences for images filmed indoor and 5 of those filmed outdoor. It will be seen that there are large variations of the silhouette quality in outdoor data. The 5 sample sequences to be shown are representative of those with relative good quality to the very poor ones.

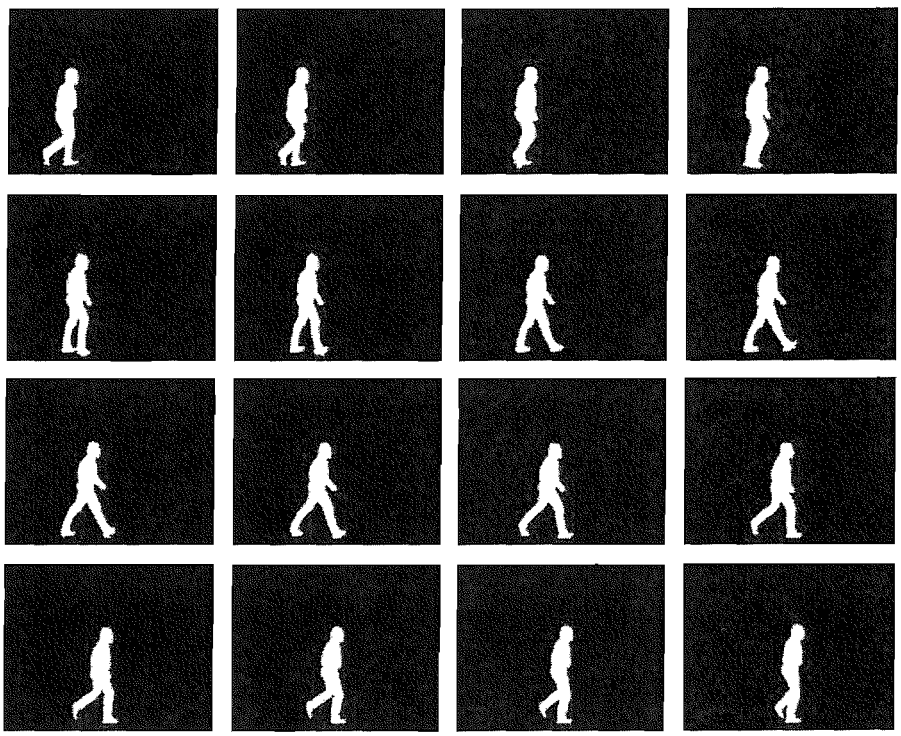


FIGURE A.1: Sample sequence 1 for clean indoor data.

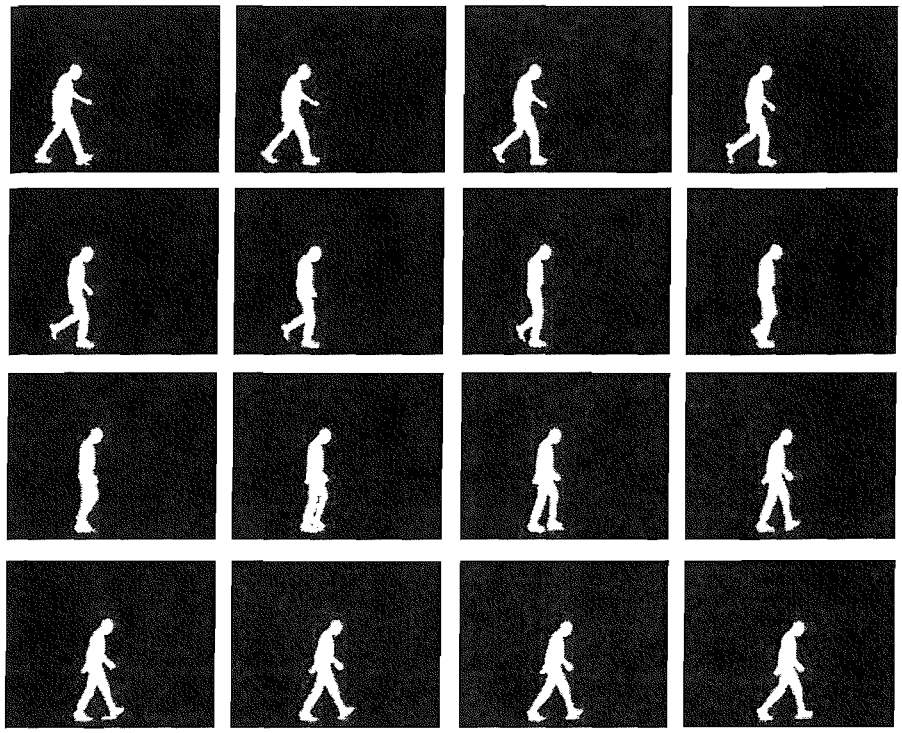


FIGURE A.2: Sample sequence 2 for clean indoor data.

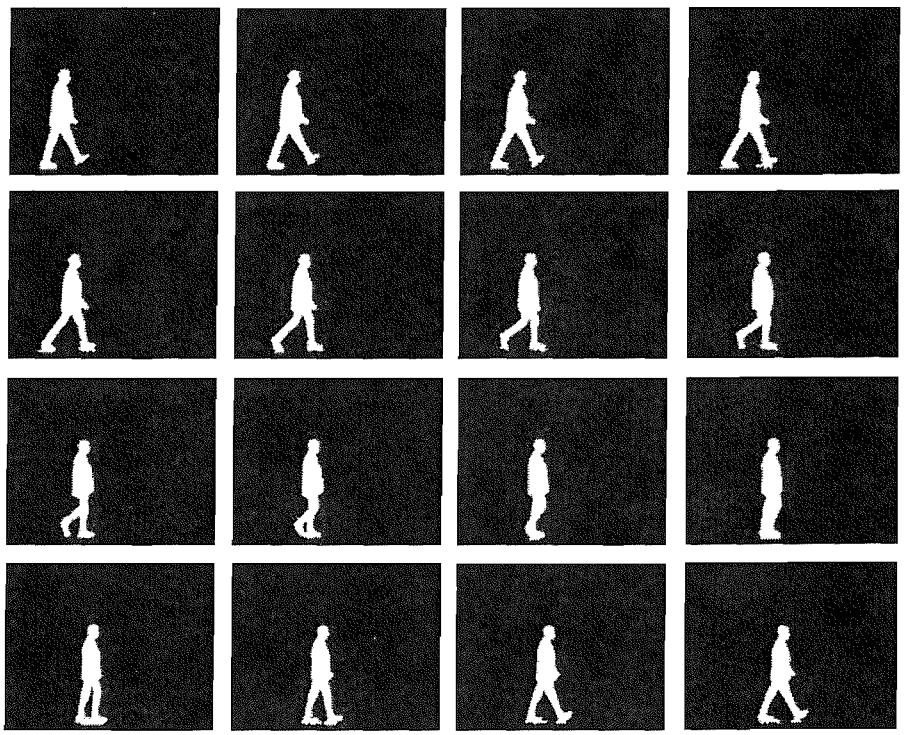


FIGURE A.3: Sample sequence 3 for clean indoor data.

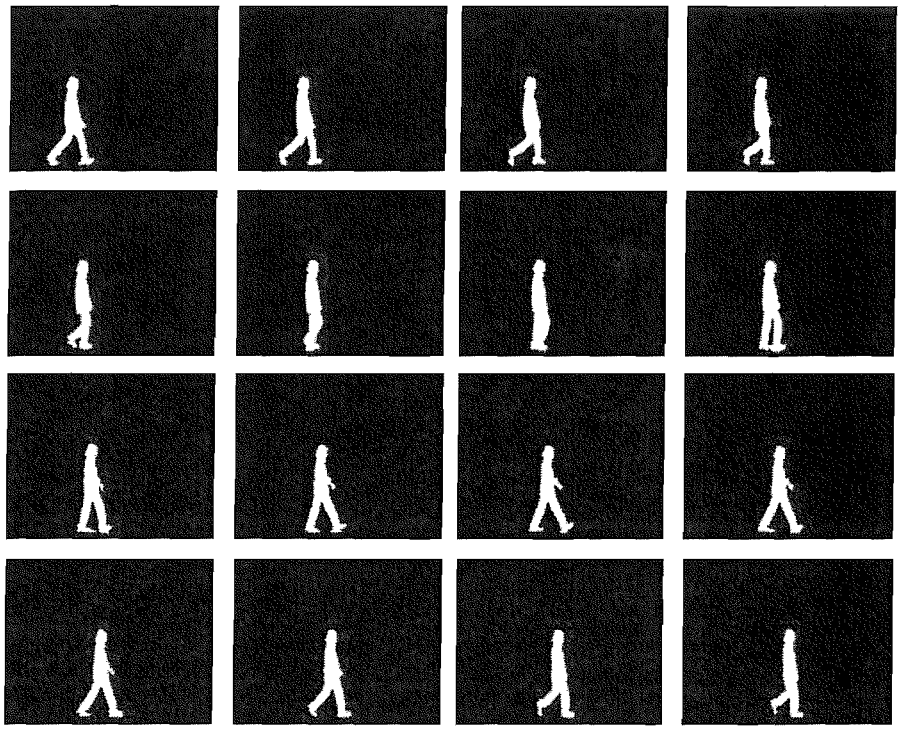


FIGURE A.4: Sample sequence 4 for clean indoor data.

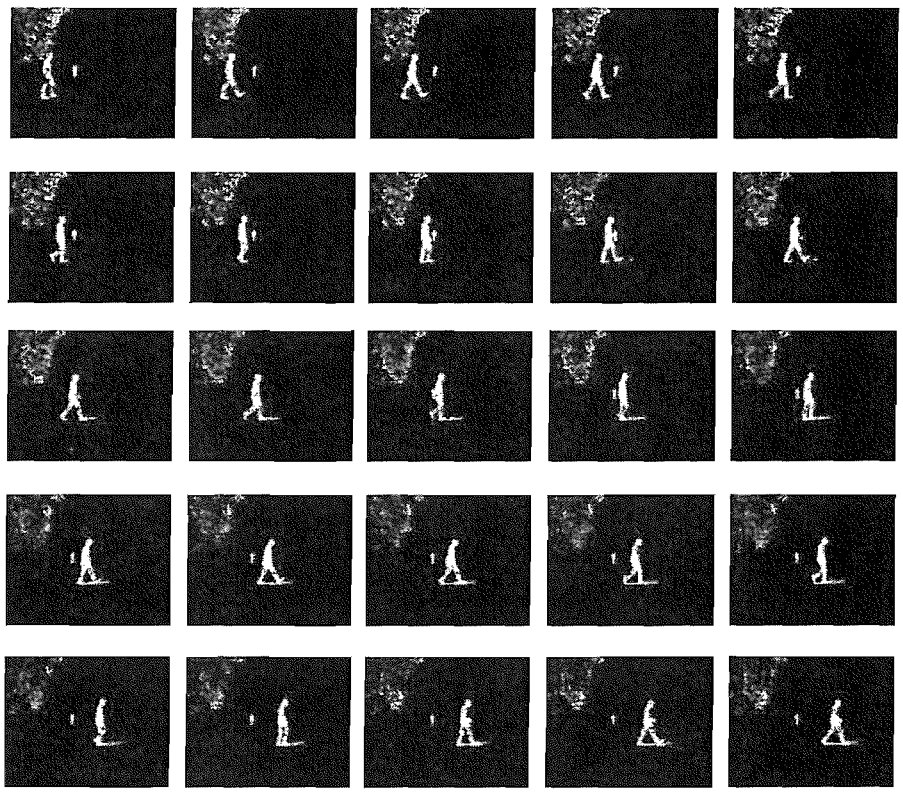


FIGURE A.5: Sample sequence 1 for noisy outdoor data.

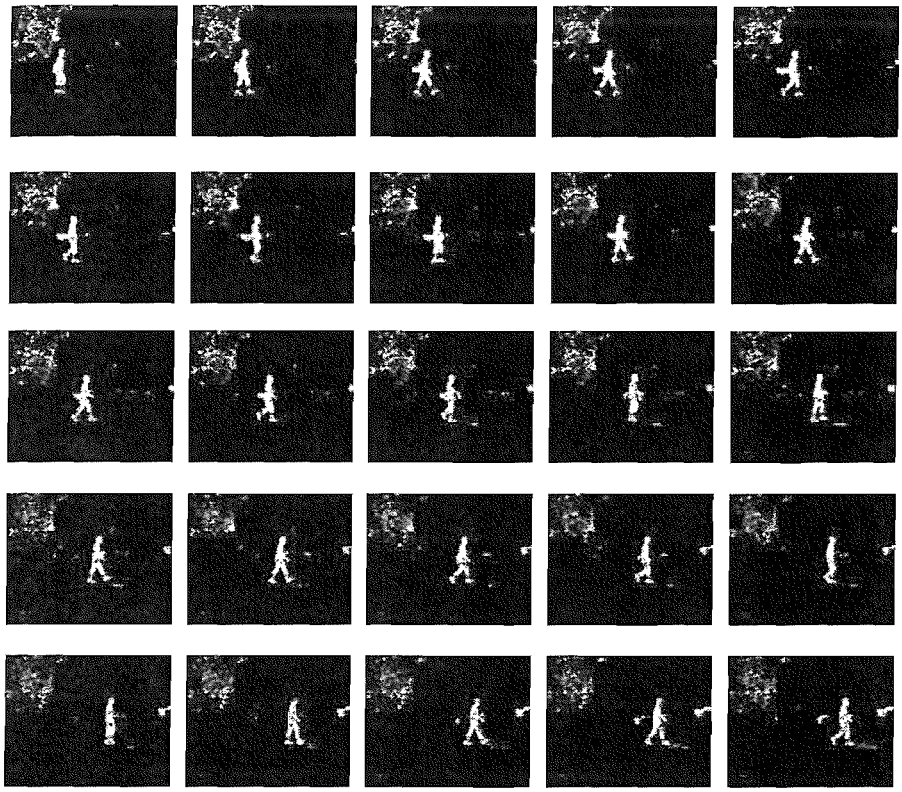


FIGURE A.6: Sample sequence 1 for noisy outdoor data.

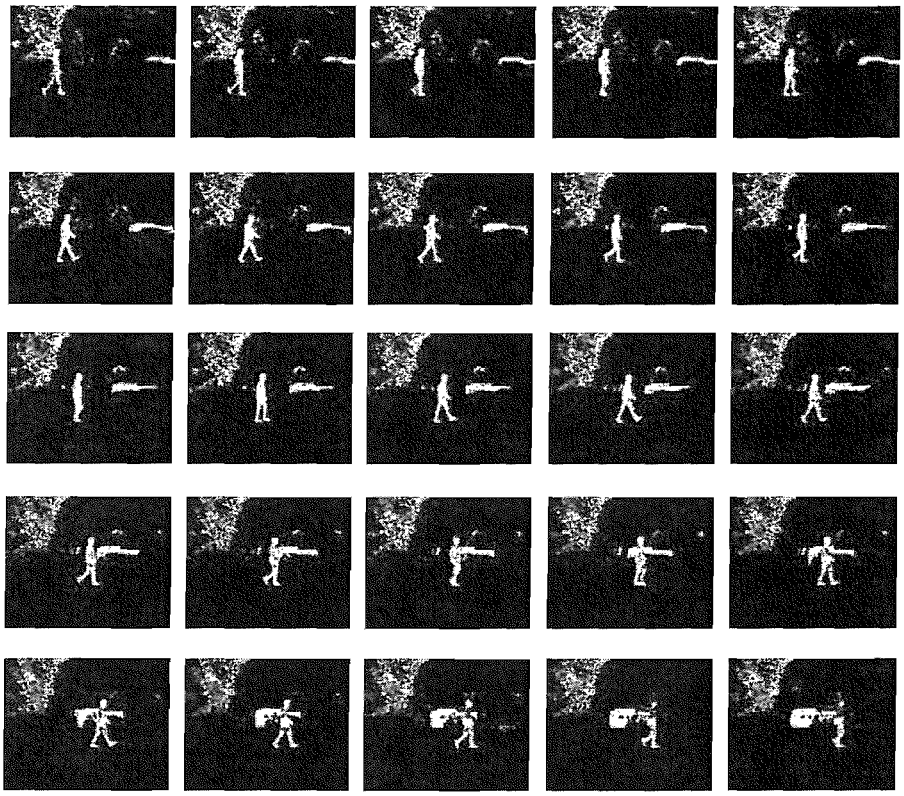


FIGURE A.7: Sample sequence 3 for noisy outdoor data.

FIGURE A.8: Sample sequence 4 for noisy outdoor data.

FIGURE A.9: Sample sequence 5 for noisy outdoor data.

Appendix B

Generalised Hough Transform

The Hough transform (HT) is a technique which can be used to detect a particular shape from the edges of an image. The classical Hough transform is used to isolate analytic curves (e.g., lines, circles, ellipses, etc.). The generalised Hough transform (GHT) is an extension of the classical Hough transform which can deal with non-analytic curves.

B.1 Classical Hough Transform

The basic idea behind the HT is to transform the edge points in the Cartesian image space to the Hough parameter space which is spanned by the parameters controlling the curves. Each edge votes for the curves passing it, that is to increase the value of the points corresponding to the curves in the parameter space. Those points having large values indicate the curves we want to detect from the images.

We give two examples of how the HT works for some analytic curves below. In the first example, we show how to detect lines in images. Figure B.1 illustrates the way we parameterise a line. We use the length of the normal from the origin to the line, ρ and the orientation of the normal with respect to the x-axis, θ to describe the line:

$$x\cos\theta + y\sin\theta = \rho. \tag{B.1}$$

Given an image, we first detect its edges $\{(x_i, y_i)\}_{i=1}^{N}$. Each edge (x_i, y_i) votes for the pair (ρ, θ) which makes Equation B.1 hold and is therefore, transformed to a curve in the Hough parameters space spanned by ρ and θ. The curves transformed from the edges on the same line intersect at the same point (ρ^*, θ^*) which defines the line in the image space.

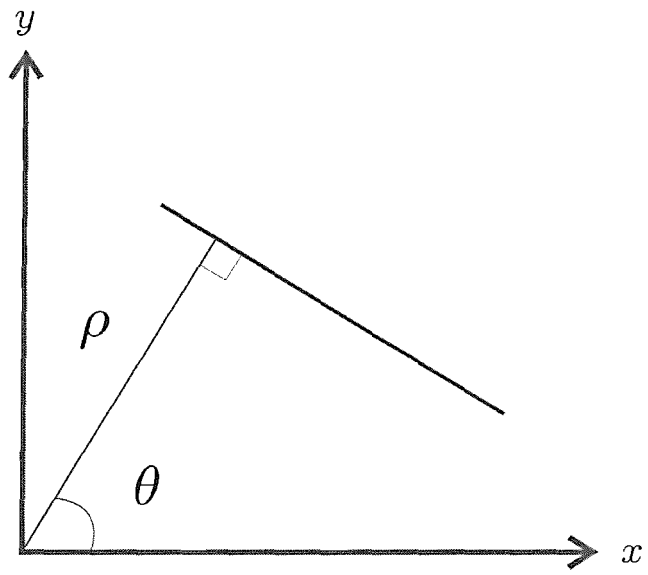


FIGURE B.1: Hough transform for a line.

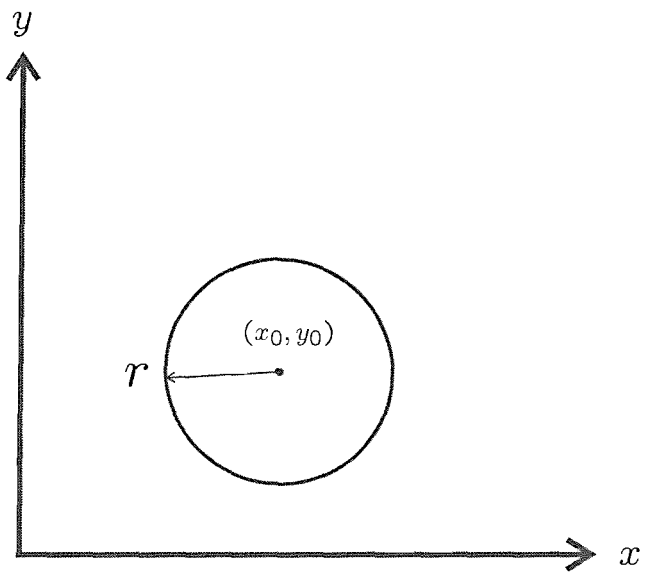


FIGURE B.2: Hough transform for a circle.

Figure B.2 shows the parametric description of a circle. We use three parameters, namely, the radius, the x coordinate x_0 and y coordinate y_0 of the center to describe a circle:

$$(x - x_0)^2 + (y - y_0)^2 = r^2. \tag{B.2}$$

In this case, the Hough parameter space is three dimensional.

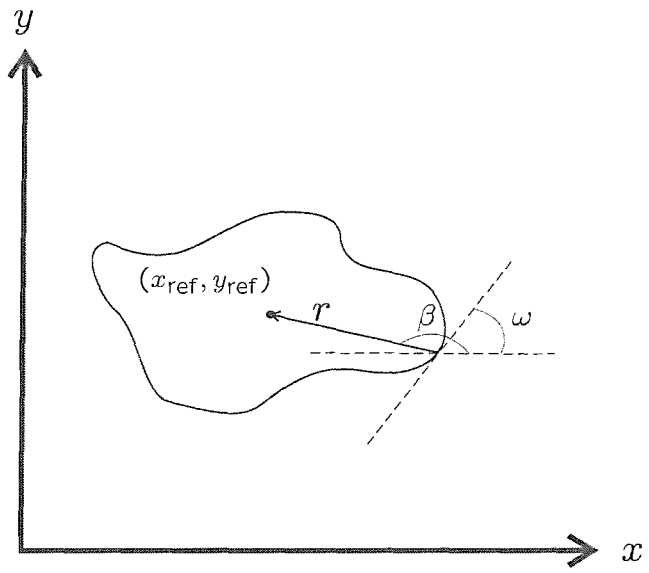


FIGURE B.3: Generalised Hough transform.

In reality, the number of curves going through an edge is infinite. However, for practical purposes, we discretise the Hough parameter space to make the number finite. The parameter space is therefore no longer continuous, but rather is represented by a rectangular structure of cells. This array of cells is called the accumulator array whose elements are accumulator cells (Sonka, Hlavac, and Boyle 1998). For any pixel, the voting is carried out by increasing the values of those accumulator cells corresponding to the lines going through it. The curves in the image may be detected by the cells with high value. The computational cost of the classical HT is proportional to the product of the number of accumulator cells and the number of edges detected in the image.

B.2 Generalised Hough transform

The classical Hough transform can only detect analytic curves in images. For non-analytic curves, the generalised Hough transform provides a solution. Figure B.3 shows the way an analytic curve is parameterised. Every edge (x, y) can be located by a reference point $(x_{\text{ref}}, y_{\text{ref}})$, the distance r from (x, y) to $(x_{\text{ref}}, y_{\text{ref}})$ and the angle β between the x-axis and the line going through both points:

$$\begin{aligned} x &= x_{\text{ref}} - r\cos\beta \\ y &= y_{\text{ref}} - r\sin\beta \end{aligned} \tag{B.3}$$

A reference table (referred as the R-table by Ballard 1981) is constructed to store the parametric edge information. In the R-table, each edge is remembered by a set $(x_{\text{ref}}, y_{\text{ref}}, r, \beta)$ and the set is indexed by the edge orientation ω (see Figure B.3). Consequently, there are two steps in the GHT. To use the R-table, we first do the edge detection on the given image to have the positions and orientations of the edge points. For each edge, we use its orientation as an index to look for the points with the same orientation on the curve. We then use the parameters in the R-table to locate the reference point and vote for it. The accumulator here is defined by the size of the image. Each accumulator cell is related to a pixel. Note that there could be zero or multiple entries indexed by an orientation. Therefore, an edge could vote for nothing or many positions. The high-value accumulator cells tell us directly the possible positions for the reference and therefore detect the curve in the image.

Appendix C

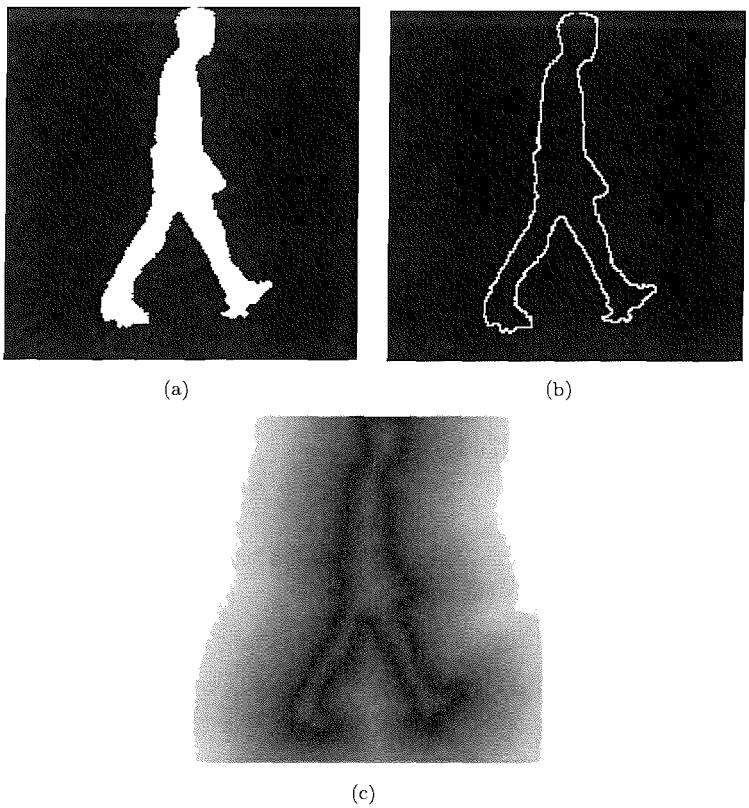
Chamfer Matching

Chamfer matching provides a robust and efficient measure of the similarity between two images: a test image and a reference image. After edge detection, we can have two sets of edge point, namely, the test edge point set $U = \{\boldsymbol{u}_n\}_{n=1}^{N}$ and the reference edge point set $V = \{\boldsymbol{v}_m\}_{m=1}^{M}$. Note that $\boldsymbol{u}_n$ and $\boldsymbol{v}_m$ are two-element vectors containing the coordinates of the edge points. For every u_n, we then estimate the minimal distance from u_n to the points in V. The chamfer distance is the average of these distances:

$$\rho(U, V) = \frac{1}{N} \sum_{\boldsymbol{u}_n \in U} \min_{\boldsymbol{v}_m \in V} \|\boldsymbol{u}_n - \boldsymbol{v}_m\| \tag{C.1}$$

To compute the chamfer distance efficiently, a distance transformation (DT) image is generated for the reference image. In such a DT image, each pixel value is assigned the nearest distance to the edge points in V. When computing the chamfer distance, the edge points of the test image are overlapped onto the DT image. The chamfer distance is simply the average of the corresponding pixel values. Figure C.1 illustrates how we generate the DT image from a given silhouette. In Figure C.1(c), the gray colors reflect the pixel values. It can be seen that the dark color corresponds to a small pixel value and vice versa. The computational efficiency of using DT images arises because if we want to estimate the chamfer distances of different test images to the same reference image, we only need to calculate the DT image for the reference image once.

To avoid using the true Euclidean distance that is computationally inefficient and unnecessary in an image space, some approximations have been used when creating DT images. For example, Borgefors (1988) introduced the utilization of a (3×3) mask with a (3,4)/3 distance measure to approximate the Euclidean distance. In our work, we use this approximation. We denote the pixel value of a DT image at the ith row and the jth column as $v_{i,j}$. The algorithm for creating the DT image from an input reference image is shown in Figure C.2.



(a) (b)

(c)

FIGURE C.1: Example of the chamfer transformation: (a) original silhouette, (b) edge image, and (c) DT image.

Input: A reference image having R rows and C columns.

Initialise: Edge detection for the input reference image. Create the initial DT image having R rows and C columns. All pixels corresponding to the reference edge points are set to zero and others infinity.

Do (forward)
  for $i = 2, 3, \ldots, R-1$ **do**
    for $j = 2, 3, \ldots, C-1$ **do**
      $v_{i,j} = \min(v_{i-1,j-1} + 4,\ v_{i-1,j} + 3,\ v_{i-1,j+1} + 4,\ v_{i,j-1} + 3,\ v_{i,j})$
    end
  end
Until *No pixel-value changes for the DT image*

Do (backward)
  for $i = R-1, R-2, \ldots, 2$ **do**
    for $j = C-1, C-2, \ldots, 2$ **do**
      $v_{i,j} = \min(v_{i,j},\ v_{i,j+1} + 3,\ v_{i+1,j-1} + 4,\ v_{i+1,j} + 3,\ v_{i+1,j+1} + 4)$
    end
  end
Until *No pixel-value changes for the DT image*

Output: The DT Image.

Figure C.2: Algorithm for creating the DT image from a reference image.

Having had the DT image, given a test image having N edge points, the chamfer distance is computed as:

$$\rho = \frac{1}{3}\sqrt{\frac{1}{N}\sum_{n=1}^{N} v_n^2} \tag{C.2}$$

where v_n is the pixel value corresponding to the nth edge point in the DT image.

Appendix D

The PDF Projection Theorem

The probability density function (PDF) projection theorem (Baggenstoss 2003) provides a general framework for projecting PDFs in the high-dimensional raw data space from PDFs in some low-dimensional feature space. Let H_0 be some fixed reference hypothesis with known PDF $p_x(\boldsymbol{x}|H_0)$, $\mathcal{X}$ be the region including all $\boldsymbol{x}$, where $p_x(\boldsymbol{x}|H_0) > 0$, $\boldsymbol{z} = T(\boldsymbol{x})$ be a many-to-one transformation, $\mathcal{Z}$ be the image of $\mathcal{X}$ under the transformation $T(\boldsymbol{x})$, and the PDF of $\boldsymbol{z}$ when $\boldsymbol{x}$ is drawn from $p_x(\boldsymbol{x}|H_0)$ exist and be denoted by $p_z(\boldsymbol{z}|H_0)$. It follows that $p_z(\boldsymbol{z}|H_0) > 0$ for all $\boldsymbol{z} \in \mathcal{Z}$. Let $p_z(\boldsymbol{z})$ be any PDF with the same region of support $\mathcal{Z}$. The theorem states the PDF $p_x(\boldsymbol{x})$, which is given by:

$$p_x(\boldsymbol{x}) = \frac{p_x(\boldsymbol{x}|H_0)}{p_z(\boldsymbol{z}|H_0)} p_z(\boldsymbol{z}) \tag{D.1}$$

is a PDF on $\mathcal{X}$, that is

$$\int_{\boldsymbol{x}\in\mathcal{X}} p_x(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1. \tag{D.2}$$

The prove of this theorem was given in Baggenstoss (2001) and will be presented as follows. From Equation D.1, we have

$$\begin{aligned} \int_{\boldsymbol{x}\in\mathcal{X}} p_x(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= \int_{\boldsymbol{x}\in\mathcal{X}} \frac{p_x(\boldsymbol{x}|H_0)}{p_z(T(\boldsymbol{x})|H_0)} p_z(T(\boldsymbol{x}))\mathrm{dx} \\ &= E_{\boldsymbol{x}|H_0}\left(\frac{p_z(T(\boldsymbol{x}))}{p_z(T(\boldsymbol{x})|H_0)}\right) \end{aligned} \tag{D.3}$$

where E calculates expectations. Using the change of variables theorem (Kaplan 1984), we have

$$E_{\boldsymbol{x}|H_0}\left(\frac{p_z(T(\boldsymbol{x}))}{p_z(T(\boldsymbol{x})|H_0)}\right) = E_{\boldsymbol{z}|H_0}\left(\frac{p_z(\boldsymbol{z})}{p_z(\boldsymbol{z}|H_0)}\right). \tag{D.4}$$

Therefore

$$\begin{aligned}\int_{\boldsymbol{x}\in\mathcal{X}} p_x(\boldsymbol{x})\mathrm{d}\boldsymbol{x} &= E_{\boldsymbol{z}|H_0}\left(\frac{p_z(\boldsymbol{z})}{p_z(\boldsymbol{z}|H_0)}\right)\\ &= \int_{\boldsymbol{z}\in\mathcal{Z}} \frac{p_z(\boldsymbol{z})}{p_z(\boldsymbol{z}|H_0)} p_z(\boldsymbol{z}|H_0)\mathrm{d}\boldsymbol{z}\\ &= \int_{\boldsymbol{z}\in\mathcal{Z}} p_z(\boldsymbol{z})\mathrm{d}\boldsymbol{z}\\ &= 1 \end{aligned} \tag{D.5}$$

Appendix E

Baseline Algorithm

The baseline algorithm for gait recognition was designed and tested on some challenging data in Sarkar et al. (2005). It is proven to be efficient and robust according to the results reported in the paper. The whole algorithm can be divided into three parts: silhouette extraction, gait period detection and similarity estimation. Note that since silhouettes had already been extracted in our work, we did not use the first part when performing the algorithm.

Gait Period Detection

Gait period N_{gait} is estimated by the following steps:

- Count the number of foreground pixels of lower body in each frame over time, $N_f(t)$.
- Find out all minima of $N_f(t)$ at the time when two legs overlap.
- Gait period N_{Gait} is computed as the median of the distances between two consecutive minima.

Similarity Estimation

To measure the similarity between a probe (unknown) sequence $S_P = \{S_P(1), \ldots, S_P(M)\}$ and a gallery (known) sequence $S_G = \{S_G(1), \ldots, S_G(M)\}$, the gait period of the probe, N_{Gait} is first estimated. The probe sequence is then partitioned into disjoint subsequences of N_{Gait} contiguous frames. The kth subsequence is denoted by $S_{Pk} = \{S_P(k), \ldots, S_P(k + N_{\text{Gait}})\}$.

The similarity between two frames, F_1 and F_2, is defined as:

$$FrameSim(F_1, F_2) = \frac{Num(F_1 \cap F_2)}{Num(F_1 \cup F_2)} \tag{E.1}$$

where $Num(\cdot)$ returns the number of foreground pixels in a silhouette. Using the above definition, the correlation between S_{Pk} and S_G can be computed as:

$$Corr(S_{Pk}, S_G)(l) = \sum_{j=1}^{N_{\text{Gait}}} FrameSim(S_P(k+j), S_G(l+j)). \tag{E.2}$$

The similarity between S_P and S_G is defined as:

$$Sim(S_P, S_G) = Median_k\Big(\max_l Corr(S_{Pk}, S_G)(l)\Big). \tag{E.3}$$

Appendix F

ANOVA

Analysis of variance (ANOVA) (Terrell 1999) is used to test the significant difference between means of multiple groups of data (the number of groups is usually larger than two). The *null hypothesis* of the ANOVA test is that all the groups share the same mean. If we have m groups of data, we denote the data in group i as $\{x_{ij}\}_{j=1}^{n_i}$ where $j = 1, 2, \ldots, m$ and n_i is the size of group i. The mean of group i is $\bar{x}_i = \frac{1}{n_i}\sum_j x_{ij}$ and the total mean is $\bar{x} = \frac{1}{n}\sum_i \sum_j x_{ij}$ where $n = \sum_i n_i$. We then define the *sum of squares for treatment* (SST) and the *sum of squares for error* (SSE):

$$\begin{aligned} SST &= \sum_{i=1}^{m} n_i(\bar{x} - \bar{x}_i)^2 \\ SSE &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 \end{aligned} \tag{F.1}$$

The *mean of squares for treatment* (MST) and *mean of squares for error* (MSE) are computed as:

$$\begin{aligned} MST &= \frac{SST}{m-1} \\ MSE &= \frac{SSE}{n-m} \end{aligned} \tag{F.2}$$

where $m - 1$ and $n - m$ are the degree of freedom of SST and SSE respectively. We then calculate the *F-statistic*, $F_{m-1,n-m}$, which is the test statistic here:

$$F_{m-1,n-m} = \frac{MST}{MSE}. \tag{F.3}$$

The significance of the test can be measured by the p-value, which is the probability of the observed F-statistic given the null hypothesis is true. If the null hypothesis is true in the test, the F-statistic is somewhere near 1. On the other hand, if there is significant difference between the means, the F-statistic becomes much larger than 1.

References

Aggarwal, J. K. and Q. Cai (1999). Human motion analysis: A review. *Computer Vision and Image Understanding 73*(3), 428–440.

Baggenstoss, P. M. (1999). Class-specific feature sets in classfication. *IEEE Transactions on Signal Processing 47*(12), 3428–3432.

Baggenstoss, P. M. (2001). A modified Baum-Welch algorithm for hidden Markov models with multiple observation spaces. *IEEE Transactions on Speech and Audio Processing 9*(4), 411–416.

Baggenstoss, P. M. (2003). The PDF projection theorem and the class-specific method. *IEEE Transactions on Signal Processing 51*(3), 672–685.

Ballard, D. H. (1981). Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition 13*(2), 111–122.

Barr, A. (1984). Global and local deformation of solid primitives. *Computer Graphics 18*(3), 21–30.

Baumberg, A. M. and D. C. Hogg (1994). An efficient method for contour tracking using active shape models. In *Proceedings of IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, Austin, TX, pp. 194–199.

BenAbdelkader, C., R. Cutler, and L. Davis (2002). Motion-based recognition of people in eigengait space. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, Washington, DC, pp. 267–272.

BenAbdelkader, C. and L. S. Davis (2002). Detection of people carrying objects: A motion-based recognition approach. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, Washington, DC, pp. 378–383.

Bhanu, B. and J. Han (2002). Bayesian-based performance prediction for gait recognition. In *Proceedings of IEEE Workshop on Motion and Video Computing (MOTION'02)*, Orlando, FL, pp. 145–150.

Bissacco, A., A. Chiuso, Y. Ma, and S. Soatto (2001). Recognition of human gaits. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Volume 2, Kauai Marriott, HI, pp. 52–57.

Blake, A., R. Curwen, and A. Zisserman (1993). A framework for spatio-temporal control in the tracking of visual contours. *International Journal of Computer Vision 11*(2), 127–145.

Bobick, A. F. and J. W. Davis (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence 23*(3), 257–267.

Borgefors, G. (1988). Hierarchical chamfer matching: a parametric edge matching algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence 10*(6), 849–865.

Boyd, J. E. (2001). Video phase-locked loops in gait recognition. In *Proceedings of IEEE International Conference on Computer Vision*, Vancouver, BC, pp. 696–703.

Bregler, C. and J. Malik (1998). Tracking people with twists and exponential maps. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, pp. 8–15.

Cëdras, C. and M. Shah (1995). Motion-based recognition: a survey. *Image and Vision Computing 13*(2), 129–155.

Chalidabhongse, T., V. Kruger, and R. Chellappa (2001). The UMD database for human identification at a distance. Technical report, University of Maryland.

Cham, T. J. and J. M. Rehg (1999). A multiple hypothesis approach to figure tracking. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Ft. Collins, CO, pp. 239–245.

Cunado, D., M. S. Nixon, and J. N. Carter (2003). Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding 90*(1), 1–41.

Cutler, R. and L. S. Davis (2000). Robust periodic motion and motion symmetry detection. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Volume 2, Hilton Head, SC, pp. 615–622.

Cutting, J. E., C. Barclay, and L. T. Kozlowski (1978). Temporal and spatial factors in gait perception that influence gender recognition. *Perception and Psychophysics 23*(2), 145–152.

Cutting, J. E. and L. T. Kozlowski (1977). Rocognizing friends by their walk: Gait perception without familiarity cues. *Bulletin Psychonomic Soc. 9*(5), 353–356.

Foster, J. P., M. S. Nixon, and A. Prügel-Bennett (2003). Automatic gait recognition using area-based metrics. *Pattern Recognition Letters 24*, 2489–2497.

Gavrila, D. (2000). Pedestrian detection for a moving vehicle. In *Proceedings of European Conference on Computer Vision*, Dublin, Ireland, pp. 37–49.

Gavrila, D. M. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU 73*(1), 82–98.

Gavrila, D. M. and L. S. Davis (1996). 3-D model-based tracking of humans in action: a multi-view approach. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, pp. 73–80.

Grant, M. G., J. D. Shutler, M. S. Nixon, and J. N. Carter (2004). Analysis of a human extraction system for deploying gait biometrics. In *IEEE Southwest Symposium on Image Analysis and Interpretation*, Lake Tahoe, NY, pp. 46–50.

Gross, R. and J. Shi (2001). The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University.

Hayfron-Acquah, J., M. S. Nixon, and J. N. Carter (2003). Automatic gait recognition by symmetry analysis. *Pattern Recognition Letters 24*, 2175–2183.

Hayfron-Acquah, J. B., M. S. Nixon, and J. N. Carter (2002). Human identification by spatio-temporal symmetry. In *Proceedings of IEEE International Conference on Pattern Recognition*, Quebec, Canada, pp. 632–635.

Hogg, D. (1983). Model based vision: a program to see a walking person. *Image and Vision Computing 1*(1), 5–20.

Isard, M. and A. Blake (1998). CONDENSATION – conditional density propagation for visual tracking. *International Journal of Computer Vision 29*(1), 5–28.

Johansson, G. (1975). Visual motion perception. *Scientific American 232*(6), 76–88.

Johnson, A. Y. and A. F. Bobick (2001). A multi-view method for gait recognition using static body parameters. *Lecture Notes in Computer Science (LNCS) 2091*, 301–311.

Ju, S. X., M. J. Black, and Y. Yacoob (1996). Cardbord people: a parameterized model of articulated motion. In *Proceedings of International Conference on Automatic Face and Gesture Recognition*, Killington, VT, pp. 38–44.

Julier, S. J. and J. K. Uhlmann (1997). A new extension of the Kalman filter to non-linear systems. In *Proceedings of International Symposium on Aerospace/Defence Sensing, Simulation and Controls.*

Kakadiaris, L. and D. Metaxas (2000). Model-based estimation of 3D human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(12), 1453–1459.

Kale, A., A. Sundaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Krüger, and R. Chellappa (2004). Identification of humans using gait. *IEEE Transactions on Image Processing 13*(9), 1163–1173.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering*, 35–45.

Kaplan, W. (1984). *Advanced Calculus* (3rd ed.). Reading, MA: Addison–Wesley.

Lan, X. and D. P. Huttenlocher (2004). A unified spatio-temporal articulated model for tracking. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Volume 1, Washington, DC, pp. 722–729.

Lappas, P., J. N. Carter, and R. I. Damper (2002). Robust evidence-based object tracking. *Pattern Recognition Letters 23*, 253–260.

Lee, L. (2003). Gait analysis for classification. Technical Report 2003-014, Artificial Intelligence Laboratory, MIT, Cambridge, MA.

Lee, L., G. Dalley, and K. Tieu (2003). Learning pedestrian models for silhouette refinement. In *Proceedings of IEEE International Conference on Computer Vision*, Nice, France, pp. 663–670.

Lee, L. and W. E. L. Grimson (2002). Gait appearance for recognition. In *ECCV Workshop on Biometric Authentication*, Copenhagen, Denmark, pp. 143–154.

Little, J. and J. Boyd (1998). Recognizing people by their gait: The shape of motion. *Videre (online journal) 1*(2).

Meyer, D., J. Pösl, and H. Niemann (1998). Gait classification with HMMs for trajectories of body parts extracted by mixture densities. In *Proceedings of British Machine Vision Conference (BMVC'98)*, Southampton, UK, pp. 459–468.

Minka, T. (2004). Exemplar-based likelihoods using the PDF projection theorem. Technical report, Microsoft Research Ltd., Cambridge, UK.

Moeslund, T. B. and E. Granum (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding 81*(3), 231–268.

Moré, J. J. (1977). The Levenberg-Marquardt algorithm: Implementation and theory. *Lecture Notes in Mathematics 630, Springer Verlag*, 105–116.

Murray, M. P. (1967). Gait as a total pattern of movement. *American Journal of Physical Medicine 46*(1), 290–332.

Ning, H., L. Wang, W. Hu, and T. Tan (2002). Articulated model based people tracking using motion models. In *Proceedings of IEEE International Conference on Multimodal Interfaces*, Pittsburgh, PA, pp. 383–388.

Niyogi, S. A. and E. H. Adelson (1994). Analyzing and recognizing walking figures in XYT. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Seattle, WA, pp. 469–474.

Ohya, J. and F. Kishino (1994). Human posture estimation from multiple images using genetic algorithm. In *Proceedings of IEEE International Conference on Pattern Recognition*, pp. 750–753.

O'Rourke, J. and N. I. Badler (1980). Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 2*(6), 522–536.

Phillips, P. J., H. Moon, S. A. Rizvi, and P. J. Rauss (2000). The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence 22*(10), 1090–1104.

Powell, M. J. D. (1964). An efficient method for finding the minimum of a function in several variables without calculating derivatives. *The Computer Journal 7*, 155–162.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*(2), 257–286.

Rehg, J. M. and T. Kanade (1994). Visual tracking of high DOF articulated structure: an application to human hand tracking. In *Proceedings of European Conference on Computer Vision*, Stockholm, Sweden, pp. 35–46.

Rehg, J. M. and T. Kanade (1995). Model-based tracking of self-occlusion articulated objects. In *Proceedings of IEEE International Conference on Computer Vision*, Boston, MA, pp. 612–617.

Rohr, K. (1994). Towards model-based recognition of human movement in image sequences. *CVGIP, Image Understanding 59*(1), 94–115.

Sarkar, S., P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. Bowyer (2005). The humanID gait challenge problem: data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 27*(2), 162–177.

Shutler, J., M. Grant, M. S. Nixon, and J. N. Carter (2002). On a large sequence-based human gait database. In *Proceedings of 4th International Conference on Recent Advances in Soft Computing*, Nottingham, UK, pp. 66–72.

Shutler, J. D. and M. S. Nixon (2001). Zernike velocity moments for description and recognition of moving shapes. In *Proceedings of British Machine Vision Conference*, Manchester, UK, pp. 705–714.

Sonka, M., V. Hlavac, and R. Boyle (1998). *Image Processing, Analysis, and Machine Vision* (2nd ed.). Thomson-Engineering.

Stenger, B., P. Mendonca, and R. Cipolla (2001). Model-based hand tracking using an unscented Kalman filter. In *British Machine Vision Conference*, Manchester, UK, pp. 63–72.

Sundaresan, A., A. RoyChowdhury, and R. Chellappa (2003). A hidden Markov model based framework for recognition of humans from gait sequences. In *Proceedings of IEEE International Conference on Image Processing*, Volume 2, Barcelona, Spain, pp. 85–88.

Tanawongsuwan, R. and A. Bobick (2001). Gait recognition from time-normalized joint-angle trajectories in the walking plane. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Volume 2, Kauai Marriott, HI, pp. 726–731.

Terrell, G. R. (1999). *Mathematical Statistics: A Unified Introduction.* New York: Springer.

Thayananthan, A., R. Navaratnam, P. H. S. Torr, and R. Cipolla (2004). Likelihood models for template matching using the PDF projection theorem. In *Proceedings of British Machine Vision Conference (BMVC2004)*, Kingston, UK. pagination unknown.

Toyama, K. and A. Blake (2002). Probabilistic tracking with exemplars in a metric space. *International Journal of Computer Vision 48*(1), 9–19.

Veres, G., L. Gordon, J. N. Carterand, and M. S. Nixon (2004). What image information is important in silhouette-based gait recognition? In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Volume 2, Washington, DC, pp. 776–782.

Wagg, D. K. and M. S. Nixon (2004a). Automated markerless extraction of walking people using deformable contour models. *Computer Animation and Virtual Worlds 15*(3–4), 399–406.

Wagg, D. K. and M. S. Nixon (2004b). On automated model-based extraction and analysis of gait. In *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, Seoul, South Korea, pp. 11–16.

Wang, L., W. Hu, and T. Tan (2003). Recent developments in human motion analysis. *Pattern Recognition 36*(3), 585–601.

Wang, L., H. Ning, T. Tan, and W. Hu (2003). Fusion of static and dynamic body biometric for gait recognition. In *Proceedings of IEEE International Conference on Computer Vision*, Volume 2, Nice, France, pp. 1449–1454.

Yam, C. Y., M. S. Nixon, and J. N. Carter (2004). Automated person recognition by walking and running via model-based approaches. *Pattern Recognition 37*(5), 1057–1072.

Yamamoto, M. and K. Koshikawa (1991). Human motion analysis based on a robot arm model. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Maui, pp. 664–665.

Yoo, J. H., M. S. Nixon, and C. J. Harris (2002). Model-driven statistical analysis of human gait motion. In *Proceedings of IEEE International Conference on Image Processing*, Rochester, NY, pp. 285–288.

Zhang, J., R. Collins, and Y. Liu (2004). Representation and matching of articulated shapes. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Volume 2, Washington, DC, pp. 342–349.