University of Southampton
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

Local and Global Models for Articulated Motion Analysis

by

David Kenneth Wagg

A thesis submitted for the degree of
Doctor of Philosophy

June 2006

University of Southampton
Faculty of Engineering, Science and Mathematics
School of Electronics and Computer Science

Doctor of Philosophy

Local and Global Models for Articulated Motion Analysis

by David Kenneth Wagg

Abstract

Vision is likely the most important of the senses employed by humans in understanding their environment, but computer systems are still sorely lacking in this respect. The number of potential applications for visually capable computer systems is huge; this thesis focuses on the field of motion capture, in particular dealing with the problems encountered when analysing the motion of articulated or jointed targets, such as people. Joint articulation greatly increases the complexity of a target object, and increases the incidence of self-occlusion (one body part obscuring another). These problems are compounded in typical outdoor scenes by the clutter and noise generated by other objects.

This thesis presents a model-based approach to automated extraction of walking people from video data, under indoor and outdoor capture conditions. Local and global modelling strategies are employed in an iterative process, similar to the Generalised Expectation-Maximisation algorithm. Prior knowledge of human shape, gait motion and self-occlusion is used to guide this extraction process. The extracted shape and motion information is applied to construct a gait signature, sufficient for recognition purposes.

Results are presented demonstrating the success of this approach on the Southampton Gait Database, comprising 4820 sequences from 115 subjects. A recognition rate of 98.6% is achieved on clean indoor data, comparing favourably with other published approaches. This recognition rate is reduced to 87.1% under the more difficult outdoor capture conditions. Additional analyses are presented examining the discriminative potential of model features. It is shown that the majority of discriminative potential is contained within body shape features and gait frequency, although motion dynamics also make a significant contribution.

Contents

Page Number

Declaration of Authorship

I, **David K Wagg**, declare that this thesis and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University;
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- Where I have consulted the published work of others, this is always clearly attributed;
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- Some parts of this work have been published in the following articles:

D K Wagg and M S Nixon: "Model-Based Gait Enrolment in Real-World Imagery." *Proc. Multimodal User Authentication*, 189-195, 2003.

D K Wagg and M S Nixon: "On Automated Model-Based Gait Extraction and Analysis." *Proc. Automatic Face and Gesture Recognition*, 11-16, 2004.

D K Wagg and M S Nixon: "Automated Markerless Extraction of Walking People Using Deformable Contour Models." *Computer Animation and Virtual Worlds*, **15**(3-4):399-406, 2004.

Signed:_____

Date:_____

Acknowledgements

First and foremost, my thanks go to my supervisor Prof. Mark Nixon. I have enjoyed our weekly meetings, and his guidance and advice throughout the term of my Ph.D. has been invaluable.

I also gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council, who have provided the funding that made this research possible.

Finally, I would like to thank my parents for their support during my time at the University of Southampton, and for their trust when, halfway through my A-levels, I decided that I was better suited to science than literature! Needless to say, I would not be in this position today without their support.

List of Symbols

$C_{base}(s, t)$ – Base contour

$C_{def}(s, t)$ – Deformed contour

COM – Centre of Mass of subject

H – Height of subject

$I_{grey}(t)$ – Sequence of greyscale image data

$I_{edge}(t)$ – Sequence of pre-processed image data (edges extracted and background removed)

M – Model complexity

MHS – Mean human shape model

$MS(s)$ – Mean contour shape model

n – Model sample index

ω – Gait frequency

ϕ - Gait phase

PM – Periodicity mask

s – Contour point index

t – Time index (frame number)

T – Number of frames in gait sequence

θ – Set of posterior (local) models for joint rotation $[\theta_{px}, \theta_{py}, \theta_h, \theta_k, \theta_a]$

$\theta_a(t)$ – Rotation of the ankle joint

$\theta_h(t)$ – Rotation of the hip joint

$\theta_k(t)$ – Rotation of the knee joint

$\theta_{px}(t)$ – Rotation of the pelvis about the x-axis

$\theta_{py}(t)$ – Rotation of the pelvis about the y-axis

Θ – Set of prior (global) models for joint rotation $[\Theta_{px}, \Theta_{py}, \Theta_h, \Theta_k, \Theta_a]$

$\Theta_a(n)$ – Rotation of the ankle joint

$\Theta_h(n)$ – Rotation of the hip joint

$\Theta_k(n)$ – Rotation of the knee joint

$\Theta_{px}(n)$ – Rotation of the pelvis about the x-axis

$\Theta_{py}(n)$ – Rotation of the pelvis about the y-axis

$X(t)$ – Position of COM along the x-axis

$Y(t)$ – Position of COM along the y-axis

Chapter 1. Introduction

1.1. Motivations for Automated Motion Analysis

There is an increasing need in modern society for computers to have the ability to understand and interact naturally with people. Computer systems now store and control an enormous amount of information covering all aspects of our lives, from entertainment and education to administration and security. However, computers generally have a limited understanding of the information they store. This information is increasingly inaccessible, due to the simple fact that there is so much of it that we cannot find the information we need. The limiting factor in this problem is the sensory capability of computer systems, that is, their ability to interpret visual, audio or tactile information.

The area addressed by this thesis is computer vision, and in particular the ability to automatically extract human motion from video data, often referred to as *motion capture*. Many diverse fields such as film and TV production, human-computer interfaces and smart surveillance would benefit immensely from an improved motion capture capability [Hu 04, Moeslund 01, Wang 03a].

(a) Film and TV production (b) Computer game control (c) Clinical gait analysis (d) Smart surveillance

Figure 1: Sample applications for automated motion analysis

Film and TV productions increasingly employ computer-generated graphics (CG) to create scenes that would have been impossible or too expensive to produce in the past. However,

whilst realistic graphics can now be produced as a matter of routine, animating CG characters so that they appear to move naturally is much more difficult. The solution is often to capture the motion of a human actor and apply it to the virtual character [Dontcheva 03]; such a system was recently employed in the 'Lord of the Rings' films to animate the CG character 'Gollum'. In this case, an actor was clothed in a coloured suit covered with white reflective markers (Figure 1a), so that his motion could be tracked relatively easily by the position of these markers and transferred to the CG character. These marker-based systems are currently cumbersome, expensive and require a great deal of expertise to set up and monitor. A markerless motion capture system would not only expedite the process, but would allow a much wider range of activities to be captured.

The field of human-computer interfaces may also benefit from markerless motion capture systems. A recent commercial development in this field is the Sony EyeToy [Sony 05], shown in Figure 1b. An addition to the games console Playstation 2, the EyeToy consists of a small camera placed on top of the TV and accompanying software. This system is capable of detecting coarse limb motion of the player under restricted capture conditions, translating this motion into commands to control computer games [Larssen 04].

In a clinical context, motion information is useful to physicians treating diseases or injuries affecting a patient's ability to walk [Ayyappa 97, Inman 81, Perry 92, Winter 91]. In such situations, the problem of obtaining accurate and reliable motion information is solved by the use of reflective markers, similar to the solution found in film and TV production (Figure 1c). Marker-based tracking systems circumvent the difficulties encountered in tracking highly variable targets by locating clearly defined markers, instead of the subjects themselves. Although expensive, this technology is mature and proven to operate successfully under controlled conditions. However, the cooperation of the subject is required, and it takes a great deal of time and expertise to attach the markers, making this class of system useless in authentication or surveillance applications.

Surveillance, defined as monitoring an area for unauthorised or illegal activity, is a routine and tedious task (Figure 1d). A human operator may become tired or bored, may take a break, and there is always the possibility of misconduct or criminal behaviour. It is also expensive to hire sufficient people to monitor a large area. 'Smart surveillance' aims to replace many of the human operators with a computer system capable of markerless motion capture, which would not suffer from these disadvantages. However, the variability in appearance and range of motion possible in typical human activities make this problem

very difficult to solve. Consequently, it is advantageous to focus initial research on more restricted application scenarios.

1.2. Gait as a Biometric

This thesis focuses on the capture and analysis of human walking motion, known as *gait*. An important application for gait analysis is in *biometric* identification and authentication [Nixon 99]. A biometric is defined as a set of biomechanical measurements that can uniquely identify an individual. Most established biometrics, such as face, fingerprint and iris recognition, are dependent on spatial patterns for recognition. By way of contrast, a gait 'signature' is spatio-temporal, consisting of spatial patterns and their variation over time. A person's gait signature is determined by their body shape, musculo-skeletal structure and joint dynamics during walking motion. This gives gait some unique characteristics amongst biometrics. Gait analysis is usable from a distance, and does not require the subject to be aware of or cooperate with its use. This makes it particularly valuable for surveillance, building entry control, and other applications in which non-interactive operation is required [Nixon 03, Wang 03a]. Unlike many appearance-based biometrics, gait information can be extracted from a wide range of camera viewpoints, and from low-resolution imagery. It is also very difficult to convincingly imitate another person's gait, making this form of identity verification robust to impostor attempts. However, gait is not yet widely accepted in the field of biometrics and many important research questions remain unanswered, such as the variability of a person's gait signature over time, the effect of aging and other covariate factors (see Section 7.2).

Early clues to the potential of gait for recognition purposes came from the literature of clinical gait analysis. In [Winter 91] the author suggests that gait may be usable as a cue to identity: "A given person will perform his or her walking pattern in a fairly repeatable and characteristic way, sufficiently unique that it is possible to recognize a person at a distance by their gait." Psychological studies have also examined this possibility [Johannson 73, Stevenage 99], demonstrating that it is possible to recognise people by their walking motion, when all other cues to identity have been removed. In biometric applications, gait bears most similarities to facial features. Table 1 summarises the main characteristics of each biometric.

Table 1: Comparing face and gait biometrics

	Face	Gait
Advantages	Medium distance operation	Long distance operation
	Only one image required	Possible with low resolution imagery
		High viewpoint invariance
		Robust to occlusion
		Robust to impostor attempts
Disadvantages	High resolution imagery usually required	Video imagery required
	Near-frontal viewpoint usually required	Additional sensing modalities required for non-lateral viewpoints
	Can be obscured by glasses, hairstyle, headgear, facial hair	Can be obscured by clothing, carried objects
	Features can be distorted by facial expression, aging	Features can be distorted by footwear, load carried, injury, mood, aging

Of course, face recognition is the more mature technology, and has the additional advantage of familiarity. We are all accustomed to identifying people by their facial features and can do so with a high degree of accuracy. Gait does not have this advantage. However, many of the strengths of face recognition are complementary to gait, so a biometric system incorporating both face and gait analysis is likely to be much more effective than a system using either alone [Shakhnarovich 02, Kale 04a, Zhou 05].

1.3. Literature Review

1.3.1. Markerless Motion Capture

The field of motion capture is aimed at the ability to automatically extract human motion from video data [Aggarwal 99, Gavrila 99, Moeslund 01]. Current commercial motion capture systems operate at the cost of attaching markers to the subject. These markers are then tracked using optical or electro-magnetic sensing apparatus, avoiding the issues involved in tracking the people themselves. While successful in some applications, marker-based systems suffer from a number of disadvantages that limit their deployment. The most obvious is their often prohibitive cost, but additionally, the markers employed can restrict the range of motion of the subject, and require expertise and a great deal of time to attach

correctly. The cooperation of the subject is also required, making such systems useless for surveillance applications, and reducing their utility in many other scenarios. These limitations motivate the development of markerless motion capture systems, capable of extracting shape and motion information unaided from video data. Current approaches typically fall into one of two categories: those that attempt to recover general, full-body motion, and those that focus on specific, limited activities.

Recovering full 3-dimensional body motion is naturally more difficult, due to the increased range of possible body configurations and the greater incidence of self-occlusion. Many recent approaches to this problem have employed multiple cameras to resolve pose ambiguities. These approaches already have some practical applications in film and TV production, but the use of multiple cameras is expensive both in monetary terms and computational complexity. In [Bregler 04] human and animal kinematics (joint positions and rotations) are extracted using a region-based motion estimation framework, employing data from one or more cameras, although manual initialisation in the first frame is required. Multiple cameras are used in [Cheung 05a, Cheung 05b] to derive shape information from silhouettes, using the acquired human model to track subject motion. Unconstrained human motion is captured in [Davison 01], employing data from multiple cameras in a probabilistic particle filtering approach. A similar approach is employed by [Drummond 02], coupling a statistical tracking framework to trinocular imagery for human tracking. In [Plänkers 03], a detailed human body model is created using metaballs (soft ellipsoidal primitives). Tracking of human motion using this model is demonstrated with stereo and trinocular imagery, although some manual initialisation is required. A probabilistic model of human motion is constructed using information from multiple cameras in [Sidenbladh 02], and subsequently used in monocular tracking of human walking and arm gestures.

Although the capability to extract unconstrained human motion is a desirable goal for markerless motion capture technology, it can be beneficial to solve more constrained problems first, with a view to applying the techniques learned to unconstrained motion. Most research to date following this approach has focused on the analysis of monocular imagery (involving only one camera) of people walking and running. These activities account for the majority of human motion in surveillance imagery, an important application for this technology. The use of monocular imagery reduces both the cost and the complexity of implementation of tracking systems. These constraints also aid performance evaluation, partly because the storage requirements for test data are lower for monocular imagery, and partly because the range of motion of the human subjects is more

narrowly defined. Most of the approaches described above for extraction of unconstrained human body motion are only evaluated on ten or fewer test sequences; many of the approaches described below are tested on hundreds of sequences.

The W^4 surveillance system [Haritaoglu 00] is capable of tracking multiple pedestrians, distinguishing people from other objects such as cars through shape and periodic motion cues. The model extracted is simple, tracking only the head, torso, hands and feet, as the focus of this research is determining the activities and relationships of people in crowded scenes. The system is also capable of detecting objects carried. This approach is tested on around 200 sequences, operating at a rate of approximately 25Hz for low-resolution imagery. The approach described in [Ning 04a] employs the Condensation framework [Isard 98] to track walking people, using a learned dynamical model and a manually defined shape model to assist the recovery of full-body kinematics. Successful extraction is demonstrated with 6 subjects walking in an indoor environment and 20 outdoor subjects, although the outdoor environment is relatively clean. Learned dynamical models are also employed in [Urtasun 04a], using PCA to extract a motion model for walking and running activities. This approach is evaluated on a small number of indoor video sequences. [Yoo 03] derives a skeletal model of a walking person from their silhouette using a mean anatomical segmentation. Although this approach is restricted to clean indoor data, it is evaluated on a large number of subjects (100, with 3 sequences per subject). In [Zhang 04] a 2D polygonal mesh is fitted to walking subjects, using a learned human shape model to guide a Bayesian template matching approach. This approach is demonstrated to operate effectively on the outdoor portion of the Southampton Gait Database (see Section 1.4), although the computational requirements of the approach are somewhat high, operating at a rate of around one frame per minute.

Regardless of approach, almost without exception recent approaches have utilised some form of anatomical shape model to aid the motion capture process. For constrained motion, it is often possible to apply models of motion as well (for example, sports video analysis or gait analysis). However, models present their own problems. The more accurate a model is required to be, the greater the number of parameters that are required to define the model. Computational demands will increase exponentially with model complexity when an exhaustive search strategy is employed, limiting such approaches to very simple models. In order to resolve this conflict and achieve fast, robust and accurate motion capture, it is necessary to employ highly sophisticated search strategies (see Chapter 2).

1.3.2. Gait Recognition

For the main application areas under consideration (entry control and surveillance), a markerless tracking system faces a number of difficulties. In outdoor conditions lighting variation is inevitable, as are variations in the orientation of the subject and occlusion of the subject by other objects. The inherent self-occlusion of the subject's limbs during gait must also be considered. Current approaches to extraction of gait information for recognition purposes are generally divided into two categories: *appearance-based* and *model-based* [Nixon 03, Wang 03a].

Appearance-based approaches attempt to extract a gait description without resolving body pose, avoiding the most complex part of this problem. Typically these approaches use the subject's silhouette and derived features as the basis for recognition. Methods for statistical analysis such as Principle Components Analysis (PCA) and Linear Discriminant Analysis (LDA) [McLachlan 92] are employed to automatically extract information useful for recognising people. The advantage here is chiefly simplicity, compared to model-based approaches. However, these approaches are essentially data-driven, and recognition performance is limited by the quality of the input data. The main difficulty is that in recognising people, humans are capable of employing a vast array of previous experiences and learned principles to find information that is useful for identification. Without using this kind of prior knowledge, it is very difficult to separate useful information from irrelevant data.

Model-based approaches incorporate knowledge of the shape and dynamics of human gait into the extraction process, constraining the expected shape and motion of the subject to a known set of possible behaviours. In principle, using a model ensures that only image data corresponding to allowable human shape and motion is extracted, reducing the effect of noise (irrelevant data). In practice, this can be difficult to achieve. Designing appropriate models of human shape and motion is difficult and time-consuming, and often error-prone; the construction of a suitable model is an additional potential failure point in the motion extraction process. Model-based approaches must also contend with computationally intensive, multi-dimensional searches to match image data to possible model configurations. Consequently, model-based gait recognition is generally more difficult to implement than appearance-based approaches.

Appearance-based approaches are quite common in the current literature. Silhouette self-similarity over time is measured in [BenAbdelkader 04], using PCA to extract a usable

gait description. Results are presented for a number of datasets, demonstrating a Correct Classification Rate (CCR) of 70% for 108 sequences from 44 subjects walking fronto-parallel to the camera view-plane, under relatively clean outdoor conditions. In [Foster 03] simple area-based masks are used to generate gait features, achieving a CCR of 75.4% on a subset of the Southampton Gait Database (see Section 1.4), comprising 912 sequences from 114 subjects captured under indoor conditions. In [Han 04] features extracted from the average silhouette using PCA are employed as the basis for recognition. This approach is evaluated on the Human ID Gait Challenge dataset [Phillips 02, Sarkar 05], which includes 1870 sequences from 122 walking subjects captured in outdoor conditions. A number of tests are included, examining the effects of covariates such as camera viewpoint, shoe type, loading and surface type. The approach of [Han 04] achieves a maximum CCR of 94% on Test B, in which only shoe type is varied. Recognition performance drops to only 9% on Test K, which was captured 6 months later than the training sequences, with resulting changes in shoe type and clothing.

An alternative mode of silhouette-based gait recognition is introduced in [Hayfron-Acquah 03], based on features extracted using the spatio-temporal symmetry operator. A CCR of 97% is achieved on a subset of the Southampton Gait Database (see Section 1.4), comprising 112 sequences of 28 subjects captured under indoor conditions. In [Kale 04b] features derived from the silhouette are coupled to Hidden Markov Models to enable gait recognition. This approach achieves a CCR of 89% on Test B of the Gait Challenge dataset, dropping to 17% for Test K. The approach of [Lee 04] uses a bilinear model to separate silhouette-based gait information into two classes, extracting identifying characteristics and discarding other gait variables. A CCR of 86% is achieved on Test B of the Gait Challenge dataset, though the more complex tests are not examined. In [Mowbray 04] Fourier descriptors are employed to model the periodic deformation of a walking person's silhouette boundary. Recognition results are presented for a subset of the Southampton Gait Database (see Section 1.4) comprising 1062 sequences from 115 subjects walking indoors, demonstrating a CCR of 86%.

In [Shakhnarovich 02] the probabilistic combination of face and gait analysis is investigated, employing a multiple camera motion capture system. Experiments on 206 sequences from 26 subjects demonstrate a CCR of 68% for gait alone, 73% for face alone and 85% when the biometrics are fused. The approach of [Shutler 06] relies on Zernike velocity moments extracted from silhouette data for recognition, achieving a CCR of 95.5% on a subset of the Southampton Gait Database (see Section 1.4) comprising 200

sequences from 50 subjects walking indoors. In [Tolliver 03] a clustering algorithm is employed to generate average silhouettes for a selected number of key gait poses. Recognition performance based on these features is evaluated on the Gait Challenge dataset, attaining a CCR of 81% on Test B. A statistical framework for silhouette analysis is described in [Vega 03], demonstrating a CCR of 90% on Test A of the Gait Challenge dataset (in which the camera viewpoint is varied by 30 degrees between the training and test sets). In [Wang 03b] PCA is applied to spatio-temporal features derived from the silhouette of a walking person. A CCR of 83% is attained on the NLPR database, which comprises 240 sequences from 20 subjects captured under relatively clean outdoor conditions. However, a CCR of only 70% is attained on Test A of the Gait Challenge dataset.

Model-based approaches to gait recognition are far more rare. In [Cunado 03] the motion of the thigh is modelled using a Fourier series, extracting a global gait signature using a Genetic Algorithm-based Velocity Hough Transform (GaVHT). A CCR of 100% was achieved on a small indoor dataset of 40 sequences from 10 subjects. In [Meyer 98] a statistical tracking framework is used to extract gait features. These features are used to train Hidden Markov Models to distinguish between walking, running, limping and hopping gaits. Results presented for 96 sequences from 12 subjects demonstrate an average CCR of 62% in distinguishing the different gait types. In [Ning 04b] a hierarchical approach to the recovery of motion parameters is presented, combining edge and region-based tracking, though the shape model is manually defined. Experimental results are presented for indoor and outdoor data, demonstrating a CCR of 88% on the NLPR database (240 sequences from 20 subjects). In [Yam 04] a temporal template tracking algorithm is employed to generate a global gait signature for walking and running subjects. Results are presented for 100 walking and 100 running sequences from 20 subjects on a treadmill. A CCR of 85% is attained for walking subjects, increasing to 91% for running subjects. It is suggested that this gait signature could be normalised for mode of gait, making it invariant to walking and running styles.

In addition to the main areas identified above, some recent approaches have combined appearance-based and model-based gait analysis. In [Bazin 05], features extracted using an appearance-based approach [Veres 04] are fused with features extracted using a model-based approach [Wagg 04b]. Performance is evaluated on a subset of the indoor portion of the Southampton Gait Database (see Section 1.4), comprising 1079 sequences from 115 subjects. An equal error rate (EER) of 15.5% is reported for the appearance-based features,

and 7.3% for the model-based approach. Fusion of the two feature vectors yields an improved EER of 5.9%. In [Veres 05] a fusion process is used to overcome the difficulties presented by time-dependent covariates, using the same two feature extraction approaches as employed in [Bazin 05]. Performance is evaluated on the Southampton Gait Database (see Section 1.4), training on the indoor portion of the LDB (2163 sequences from 115 subjects) and testing on the SDB (3177 sequences from 10 subjects). A CCR of 23.5% is reported for the appearance-based approach, and 14.2% for the model-based approach. Fusion of the two feature vectors yields an improved CCR of 27.7%. In [Wang 04] a similar process is described, using the appearance-based approach of [Wang 03] and the model-based method of [Ning 04a]. Evaluation is performed on a subset of the NLPR database, comprising 80 sequences from 20 subjects. An equal error rate (EER) of 8.4% is reported for the model-based features, and 10.0% for the appearance-based approach. Fusion of the two feature vectors yields an improved EER of 3.5%.

Some recent research has examined the matter of which image features are relevant to recognition by gait, and how covariate factors can affect the recognition process. In [Liu 04] the silhouettes of walking people are manually segmented according to body part, allowing examination of the discriminatory potential of different parts of the body during gait. For Test B of the Gait Challenge dataset, a CCR of 22% is achieved using only the arms, 37% using only the upper body and 49% using only the legs compared to a CCR of 49% using all of the features. The impact of gait speed is examined in [Tanawongsuwan 03], showing that stride length and cadence increase approximately linearly with increasing gait speed. A normalisation process is demonstrated allowing recognition across four different walking speeds. In [Veeraraghavan 04] the roles of shape and kinematic features are investigated, concluding that kinematics boost the recognition performance of shape features, but are insufficient on their own. Features derived from the average silhouette and from the average differential silhouette are analysed in [Veres 04]. It is shown that upper body shape features contribute most to recognition capability, as the articulation of the legs destroys any shape information in the averaged lower body.

1.4. Gait Database and Assumptions

All testing of this thesis is carried out on the Southampton Gait Database [Shutler 02], consisting of 115 subjects (24 female and 91 male). The gait database is split into three sections; the Large Database (LDB), which includes sequences captured under indoor and outdoor conditions, the Small Database (SDB) which investigates covariate factors (such as walking speed, footwear, clothing and loading) for a sample of 10 subjects from the LDB under indoor capture conditions, and the Temporal Database (TDB), which includes the same sample of 10 subjects, filmed at a later date. Each video sequence is stored in Digital Video (DV) format, encoded in colour PAL format at a resolution of 720 by 576 pixels (though each sequence is converted to a greyscale colour format for the purposes of this thesis). The sequences were recorded at a rate of 25 frames per second in progressive scan mode. The average number of frames for the indoor dataset is approximately 60 per sequence, or about two complete gait cycles. For the outdoor dataset, the subject is further away from the camera and thus is visible for longer, averaging approximately 100 frames per sequence. The first, middle and final frames of two example sequences from the LDB are provided in Figures 2 and 3.

Figure 2: LDB indoor dataset, example sequence '008a013s00R'

Figure 3: LDB outdoor dataset, example sequence '008e013s00R'

The LDB is employed in testing this thesis, limited to only those sequences where the subject walks parallel to the camera's plane of view (Appendix C includes some additional results for the SDB). There are approximately 20 such sequences available for each subject in each condition, totalling 2163 indoor sequences and 2657 outdoor sequences. The indoor filming environment was constructed so as to minimise sources of variation and noise (Figure 2). The subject is filmed against a static (chromakey fabric) background. The lighting is constant and approximately uniform in order to minimise the appearance of shadows. The subjects are constrained to a narrow walking path, and no other moving objects are included in the scene. The outdoor environment is far more variable (Figure 3). Lighting conditions change according to cloud cover and time of day, and shadows are clearly apparent. In particular, the tree at the left-hand side of the scene casts a large shifting shadow over the walking path. Pedestrians and vehicles pass by in the background of the scene, and other trees and foliage generate additional motion sources.

Based on the properties of this database, it is possible to make certain simplifying assumptions about the motion capture task. In terms of subject motion, it is assumed that:

1. **Gait is uninterrupted (approximately constant velocity)**
2. **The ground plane is approximately level (mean *y*-displacement is small)**
3. **Motion in the *z*-plane is negligible (constant scale)**
4. **There are no other people following the same path as the subject**

Some of these assumptions are violated in a small number of the outdoor sequences. It is estimated around 2% of the sequences in the outdoor database significantly deviate from these assumptions, which should not impact any conclusions to be drawn from analysis of this database. These assumptions are quite limiting in terms of the motion that can be analysed, but it should be noted that the approach detailed within this thesis is model-based, and can be generalised to other viewpoints and less restrictive path assumptions. These assumptions allow this thesis to focus on the basic problem of capturing shape and motion from noisy video data, leaving aside for the moment additional problems of variable viewpoints, paths and multiple subjects.

1.5. Thesis Outline and Contributions

This thesis presents a model-based approach to automatic extraction and analysis of gait patterns. It is shown that these patterns carry information sufficient to enable recognition of walking people, within the constraints of the test database. The focus of this thesis is the development of computer vision algorithms that may operate on noisy, cluttered imagery.

Chapter 2 lays the foundation of this thesis, describing the strategies by which image data may be translated into meaningful descriptions of a walking person's body shape and motion. Two main strategies for model-based analysis are identified, one taking a global view of the presented data and working down to local estimates, and the other starting at a local level, building up a complete description piece by piece. These strategies are employed in combination within this thesis as appropriate, to make best use of their distinctive advantages. A model of human shape and motion is introduced, combining an articulated geometric model with deformable contours for accurate and efficient shape representation. An additional model component describes self-occlusion of the legs, predicting which parts of the legs will be visible given the approximate shape and pose of the subject.

Chapter 3 describes the process of building an initial model of the subject. The aim is not to resolve everything in one effort, but rather to generate a simple description of the subject with the greatest possible reliability. Low model complexity reduces the computational requirements of this approach, enabling the enrolment of large numbers of subjects into a gait recognition system. This research was presented as a poster at the 2003 Workshop on Multimodal User Authentication [Wagg 03].

Chapter 4 is concerned with how this initial gait model may be extended and adapted to better describe the subject, increasing the recognition potential of the extracted model. The first approach operates purely on a global level, adjusting a generic model of gait until it matches the motion of the subject. This research was presented orally at the 6th International Conference on Automatic Face and Gesture Recognition [Wagg 04a], and repeated with some extensions for a domestic audience at the 2004 BMVA Symposium on Biometrics. Concluding that a purely global approach cannot achieve sufficient accuracy within reasonable computational requirements, a second approach employs a snake-based local adaptation process, exploiting prior knowledge of gait patterns to predict self-occlusion of the legs. This research was presented orally at the 2004 International

Conference on Computer Animation and Social Agents, and published in the Journal of Computer Animation and Virtual Worlds [Wagg 04b]. However, this approach cannot recover from large errors in initialisation. It is concluded that neither approach is sufficient alone, and for greatest performance the two strategies should be combined.

Chapter 5 describes a hybrid adaptation algorithm, combining local and global modelling strategies in an iterative process. Observing that local strategies perform well if correctly initialised, a locally maximal model configuration is determined at the beginning of each iteration of the adaptation algorithm. A global expectation process computes the global model best fitting those local configurations deemed reliable. This global model is used to generate a new model initialisation, thereby improving at each stage the gait description that may be attained. This research is to be published in a forthcoming journal article.

Chapter 6 is concerned with evaluation of these techniques, focusing on the final and most successful approach of Chapter 5. A number of gait sequences are summarised, showing the extracted model resulting from each approach. The computational demands of each approach are briefly detailed, followed by an analysis of the discriminatory potential of the extracted gait features. The performance of each approach in recognising subjects within the gait database is included, demonstrating the superiority of a hybrid approach. Additional error metrics are employed to indirectly measure the reliability of model extraction.

Chapter 7 details the conclusions reached in the process of this research. Some directions for future research are suggested, aimed at reducing the restrictive capture conditions assumed for this thesis, and examining some of the covariate factors that may impede recognition in real-world application scenarios.

The results presented within this thesis have been employed by other researchers in further analyses. The fusion of model-based gait features with appearance-based features is examined in [Bazin 05], demonstrating the improvement in recognition capability achieved by combination of different gait modalities. This research made use of the gait features extracted using methods described in Chapters 3 and 4.2. In [Veres 05] the same features are employed in a fusion framework, examining the effect of time-dependent covariate factors on recognition performance. These papers confirm the utility of the approach described within this thesis for the application scenario of recognition by gait.

Chapter 2. Modelling Strategies

2.1. Introduction

This chapter provides an overview of the modelling strategies applied later in this thesis. In a model-based computer vision approach there are generally two models involved, the *prior* and the *posterior*. The prior model represents what is known about the class of target objects in general, before examining any available image data. For example, in the case of people the prior model may include mean anatomical proportions or gait patterns. The prior model can guide extraction of the posterior model by defining its expected properties, allowing rejection of image data that does not fit this model. The posterior model is a description of the target object, constructed from its image and guided by the prior model. This chapter is primarily concerned with modelling theory and definition of the prior model; extraction of the posterior model is covered in Chapters 3, 4 and 5.

Strategies for modelling shape and motion are categorised according to whether they operate at a *local* or *global* scale. In general, these terms describe the scope of the model, or the extent of the target object that is represented by a single model. In the context of image sequence analysis, these terms are generally employed to refer to temporal scope. A local model describes the appearance and position of an object within a single frame (position in time within a video sequence). A global model includes temporal as well as spatial information, describing the shape and motion of the target within a whole sequence. An object moving through T frames could be modelled by a series of T independent local models, or by a single global model. It is also possible to compromise by defining models over two or more frames [Lappas 02], combining some of the properties of each strategy. However it is difficult to define an acceptable compromise, as the computational advantage of a local approach is lost rapidly as temporal dependencies are increased.

There are two main factors to consider in model construction: *complexity* and *constraints*. A basic measure of complexity is M, the number of possible states or configurations the model can represent:

$$M = \prod_{i=0}^{F-1} V_i \tag{1}$$

Where F is the number of model features and V_i is the number of possible values feature i can take. Each model feature is a parameter describing a single aspect of the target object. By way of example, a simple model of a car could be defined as a single 2D rectangle. In this case, $F = 5$, as the state of the model is completely specified by (x, y) position, length, width and rotation. V is determined by how these parameters are quantified. Often parameters will be given a finite range of discrete values, for example, rotation could be constrained to vary from 1° to 360° in 1° increments, yielding $V_i = 360$ for this feature. Note that V may be difficult to measure if an adaptive or continuous feature representation is employed.

The model must be sufficiently complex to adequately represent all of the important features of the target object, but complexity has a considerable impact on computational requirements, particularly for global models. Constraints control which configurations are allowable, and are employed to eliminate implausible or uncommon model configurations. This reduces computational requirements and increases the reliability of posterior model extraction.

To illustrate these factors, we may consider an object moving through a sequence of T frames. For an object described by a model with complexity M, the total *model space* comprises M^T states. Model space is defined as the set of possible solutions or model configurations for a given length of sequence. M is usually very large; often for any practical application, the model space is far too large to be searched exhaustively. Consequently, a *search space* is defined, a subset of the model space that is most likely to contain a good solution. It should be noted that a search space is only required for posterior model extraction, as the configuration of the prior model is, by definition, already known.

The two main modelling strategies take very different approaches to defining a search space. In a local strategy, each frame is considered to be independent and unconnected to other points in time, resulting in a search space of only $M \times T$ states. In a global model, all points are connected, and without constraints the search space comprises M^T states. This demonstrates the main advantage of local models: for a given limit on computational requirements, local models can afford much greater complexity ($M \times T << M^T$ for typical values of M and T where $M >> 1$, $T >> 1$), so that a more accurate object description can be attained. The complexity of a global model must be much lower to retain reasonable computational requirements; the advantage of a global approach is that due to temporal dependencies, constraints are easier to apply to the search space and missing data is easier

to reconstruct. This means that a global approach is more reliable when processing noisy or incomplete image data. These properties are summarised in Table 2.

Table 2: Properties of modelling strategies

	Local	**Global**
Domain	Spatial	Spatio-temporal
Model space	M^T	M^T
Search space	$M \times T$	M^T
Primary Advantage	Low computational cost	Low sensitivity to noise

Some examples of model constraints are given in Table 3. Spatial constraints can be applied to both strategies, typically including limitations on position, size and relative proportions, and the range of motion of joints. These constraints have the effect of reducing M and thus the size of the search space. Temporal constraints can further reduce the size of the search space, and play an important role in ensuring robust posterior model extraction. An important temporal constraint is continuity of motion, which ensures that the model moves in a fashion consistent with physical laws. Other temporal constraints can include periodicity of motion, whereby some motion is repeated in a predictable manner, and symmetry of motion (similarities in motion of different object features, for example, bilateral symmetry in leg motion). Model shape can also be constrained temporally, restricting changes in appearance over time.

Table 3: Common model constraints

Spatial	**Temporal**
Position constraints	Continuity of motion
Size constraints	Periodicity of motion
Anatomical proportions	Symmetry of motion
Range of joint motion	Shape rigidity

These constraints are relatively easy to apply to a global model, as variation over time is explicitly defined. A local model is by definition independent from all other points in time, which limits the application of temporal constraints. Means of applying some degree of temporal constraints to local approaches are discussed in Section 3.1, exploiting the temporal dependencies implicit in the ordering of local models.

This chapter discusses only models of the upper body and the legs; no attempt is made to model the arms. The arms are not essential to gait analysis and are less predictable than

the legs; for example, carrying a load may eliminate arm motion entirely. In sagittal views (where the subject's side is presented to the camera), the arm furthest from the camera is also almost completely occluded by the torso, reducing the amount of useful data that can be collected. For these reasons the arms are not included in the models introduced within this chapter. For more general application scenarios the arms may be important, and could be modelled in the same manner as the legs.

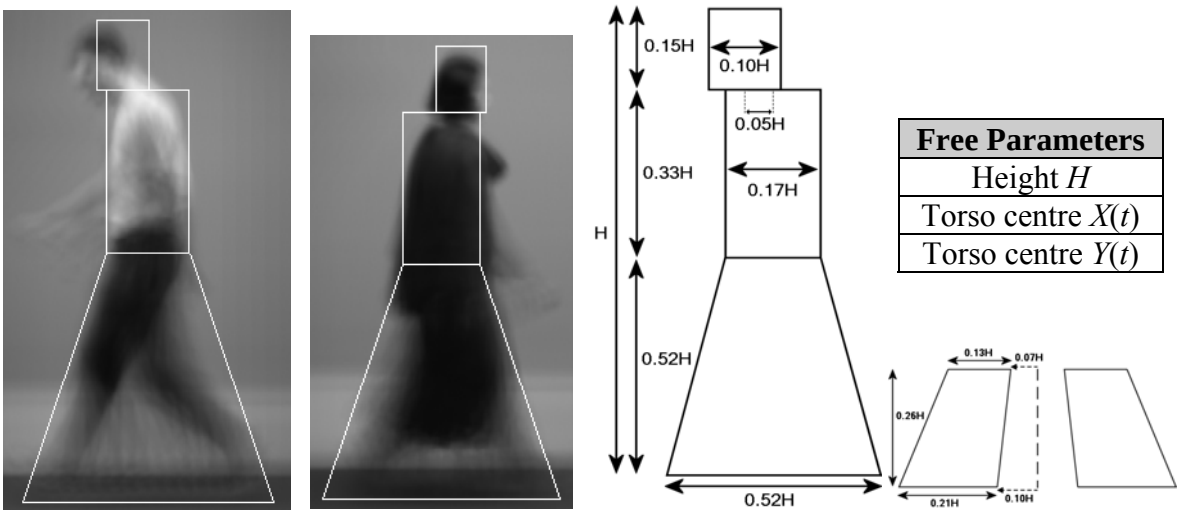
The approach described within this thesis employs both local and global modelling strategies to extract shape and motion features. In general terms, simple global models are employed to provide reliable initialisations for more complex local models. In this way, the local models inherit some of the reliability and constraints of a global approach, while retaining their own speed advantage.

2.2. Shape Models

2.2.1. Geometric Basis

A geometric basis is particularly well suited to global approaches. A shape model based on simple geometric structures has few parameters, countering the high cost of resolving temporal correlations. In order to further reduce computational requirements, the model extraction process is divided into two stages. The first stage is concerned with locating the subject within the video sequence and determining their approximate size. This requires only a very simple shape model, approximating the average shape of a walking human (Figure 4a-c). The second stage involves estimating limb motion and detailed anatomical features, employing a more sophisticated model (Figure 5). Figures 4a and 4b were generated by temporally averaging an image of the subject centred within a bounding box over the gait sequence. The advantage of the mean human shape model (Figure 4c) is that it is very simple, having only three free parameters: height and (x, y) coordinates. All other model parameters are fixed in proportion to the subject's height, based on mean anatomical data [Winter 90]. The head offset additionally requires knowledge of the direction of the subject's facing, which is easily determined from their velocity (assuming the subject is walking forwards). An alternative to a geometric basis is described in [Baumberg 94], in which an Active Shape Model [Cootes 92] is used to represent the shape of walking

people. This model is constructed using PCA to extract the mean shape and principle modes of variation in a training set of pedestrian images. However, this approach requires a large set of aligned silhouette images, and it introduces an unknown degree of dependency on the training set.



(a) Sequence '008a013s00L' (b) Sequence '012a031s00R' (c) Mean human shape model *MHS* (d) Periodicity mask *PM*

Figure 4: Simple geometric model of average human shape during normal gait

The limitation of this model is that little individual variation is taken into account, leading in some cases to significant matching errors. However, approximate height and position may be reliably established using this model (Section 3.2.1). The periodicity mask (Figure 4d) approximates the outer region of the leg swing during normal gait, so that sum edge strength within the mask varies periodically during the subject's gait. This mask is used within Section 3.2.2 to estimate gait frequency and phase.

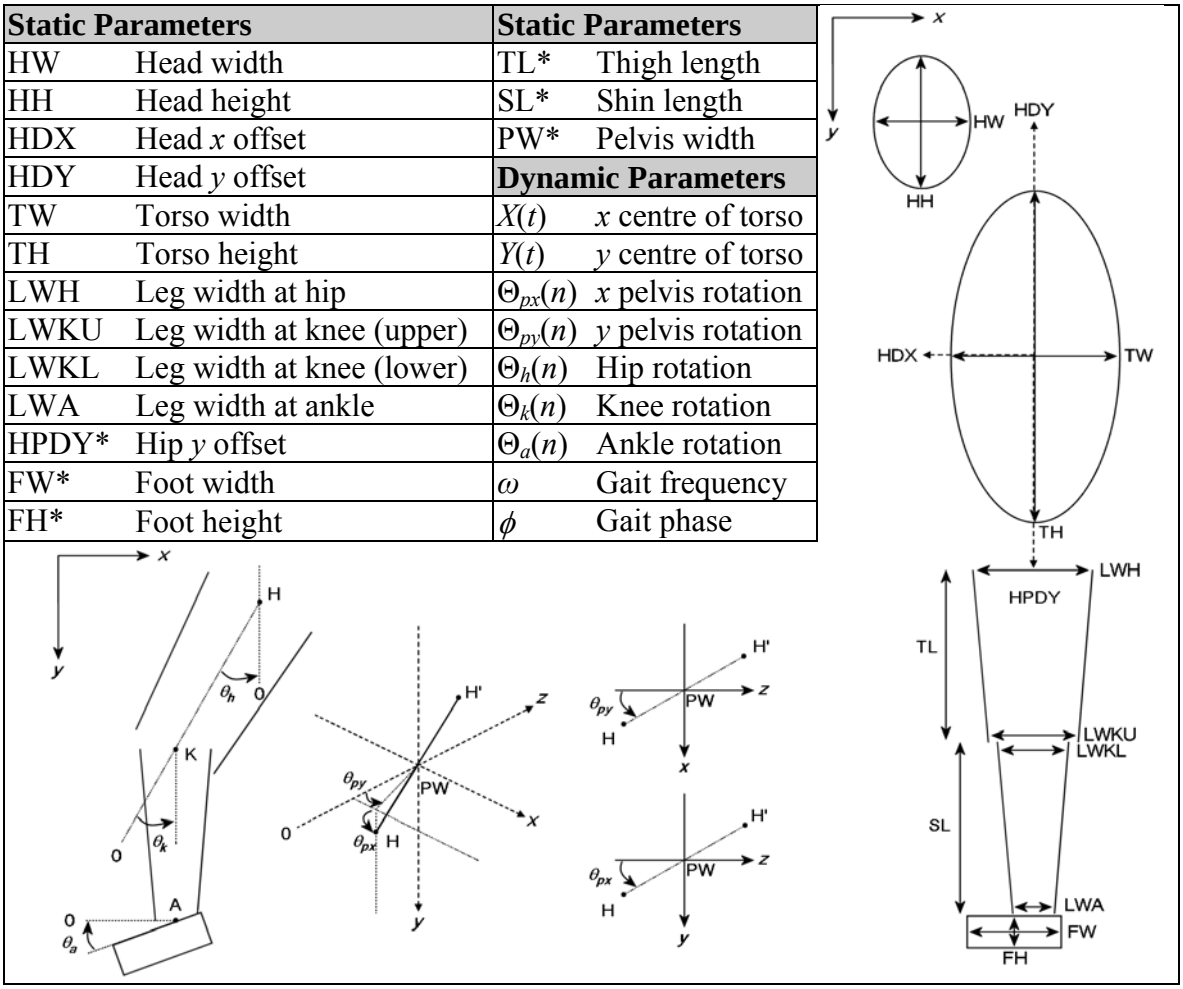
The model employed within the second stage is more detailed, which permits a better match with the subject's body (Figure 5 illustrates some example model configurations). The head and torso are modelled using ellipses; each leg is modelled by two tapered pairs of lines, and the foot by a rectangle. These model parameters are detailed in Table 4. Note that $\Theta(n)$ describes joint rotations over an average gait cycle. The joint rotations in the sequence time domain $\theta(t)$ are obtained by scaling and replicating $\Theta(n)$ using the gait frequency and phase (see Section 2.3.2).

The upper and lower leg segments are modelled separately, so there can be discontinuities in leg shape at the knee joint where the thigh and shin segments do not meet cleanly. This is not always undesirable, as it is an efficient method of representing the leg

shape of a subject wearing shorts or a skirt. Leg, foot and pelvis segment lengths (marked with an asterisk in Table 4) are fixed in proportion to the subject's height according to mean anatomical proportions [Winter 90], as these parameters are difficult to estimate accurately.

Table 4: Geometric human body model parameters

Static Parameters		**Static Parameters**	
HW	Head width	TL*	Thigh length
HH	Head height	SL*	Shin length
HDX	Head *x* offset	PW*	Pelvis width
HDY	Head *y* offset	**Dynamic Parameters**	
TW	Torso width	$X(t)$	*x* centre of torso
TH	Torso height	$Y(t)$	*y* centre of torso
LWH	Leg width at hip	$\Theta_{px}(n)$	*x* pelvis rotation
LWKU	Leg width at knee (upper)	$\Theta_{py}(n)$	*y* pelvis rotation
LWKL	Leg width at knee (lower)	$\Theta_h(n)$	Hip rotation
LWA	Leg width at ankle	$\Theta_k(n)$	Knee rotation
HPDY*	Hip *y* offset	$\Theta_a(n)$	Ankle rotation
FW*	Foot width	ω	Gait frequency
FH*	Foot height	ϕ	Gait phase



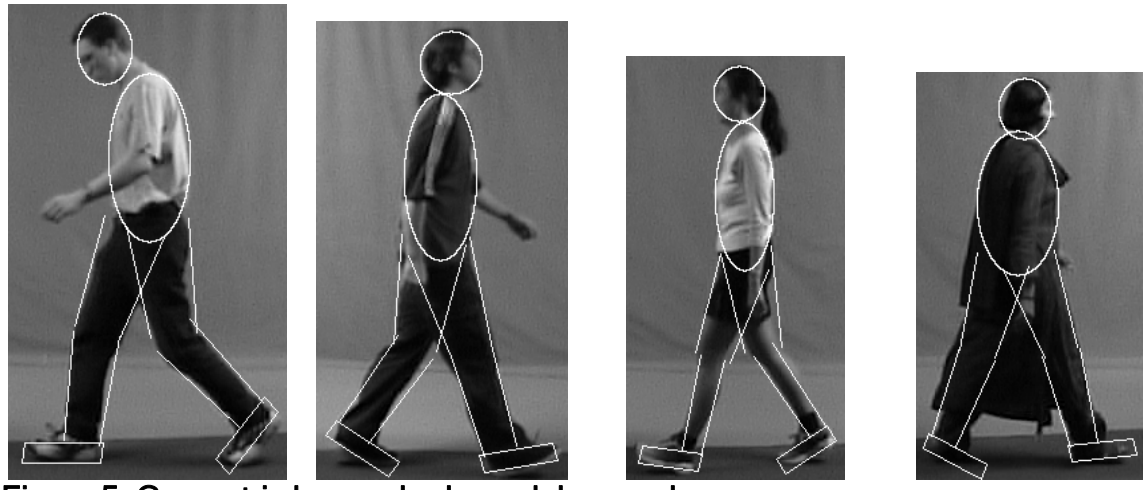


Figure 5: Geometric human body model examples

This model has 10 free parameters approximating the mean shape of the subject, and 7 free dynamic parameters that describe body pose and position. This model is still relatively simple and thus incurs low computational requirements, but has limited adaptation potential. To accurately represent individual variation and clothing effects, a more flexible basis is required.

2.2.2. Deformable Contour Basis

To account for more complex body shapes and changes in shape over time, a deformable contour layer is added to the geometric basis. A deformable contour C may be described by a sequence of points in 2D space indexed by distance along the contour s and time t:

$$C(s,t) = [x(s,t), y(s,t)] \tag{2}$$

Where $0 \leq s < 1$ and $0 \leq t < T$, and T is the number of frames in the gait sequence. The geometric model described in Section 2.2.1 is converted to three open-ended contours, modelling the base shape of the upper body and legs:

$$C_{base} = contour(X(t), Y(t), \theta(t), MS(s)) \tag{3}$$

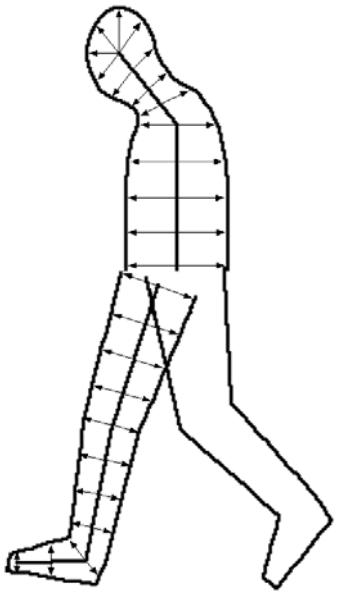


Figure 6: Base contour model construction

Where C_{base} is the set of base upper body and leg contours produced given position vectors X and Y (see Section 2.3.1), joint angles $\theta = [\theta_{px}, \theta_{py}, \theta_h, \theta_k, \theta_a]$ and mean contour shape model MS. This model is derived from the static parameters of the geometric model, defining the position of each contour point relative to the originating skeletal point. Figure 6 illustrates the construction of the base contour model from the skeleton and mean shape model (the rear leg is omitted for clarity). The *contour* operation is essentially a matter of merging and connecting geometric components to obtain a continuous contour for each body segment. A sampling rate of one contour point per four pixels is employed throughout this thesis, which yields acceptable results while keeping computational requirements low. Some examples of base contours are given in Figure 7, derived from the geometric models depicted in Figure 5.

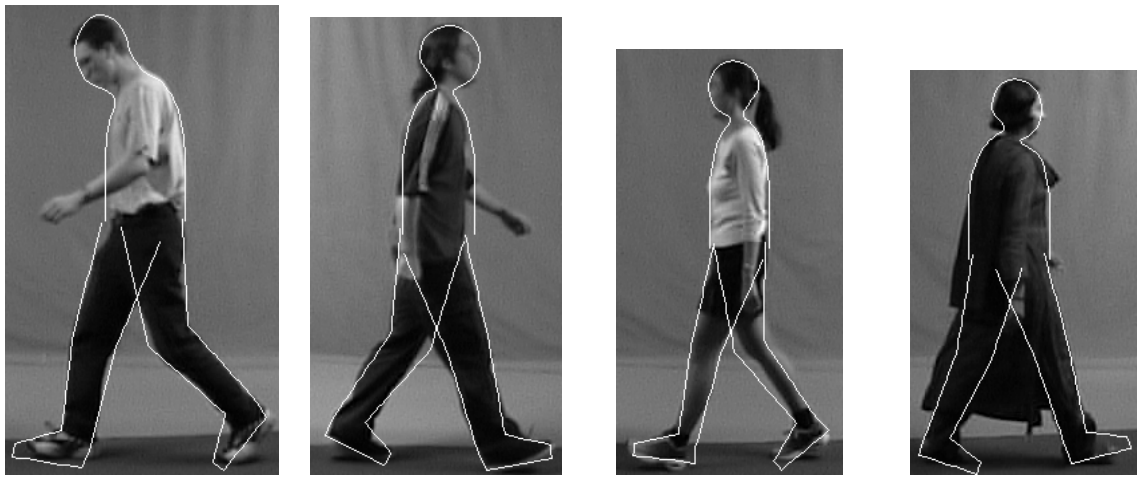


Figure 7: Base contour model examples

A local deformation is computed for each frame to fit the base contour model to the image data. Deformation is allowed only along the local normal to the base contour, which greatly simplifies the contour optimisation process. This restriction reduces the search space for each contour point, and guarantees approximately constant contour point spacing. It is possible to make this simplification because the base contours are globally initialised (Chapter 3), ensuring that the snake is started in a reasonable configuration. Any large errors are corrected on a global basis (Section 5.3), leaving only smaller errors for the deformable layer to correct. Figure 8 depicts the deformed contour model computed from the base contour model (Figure 7), using the method of Section 5.2.3.

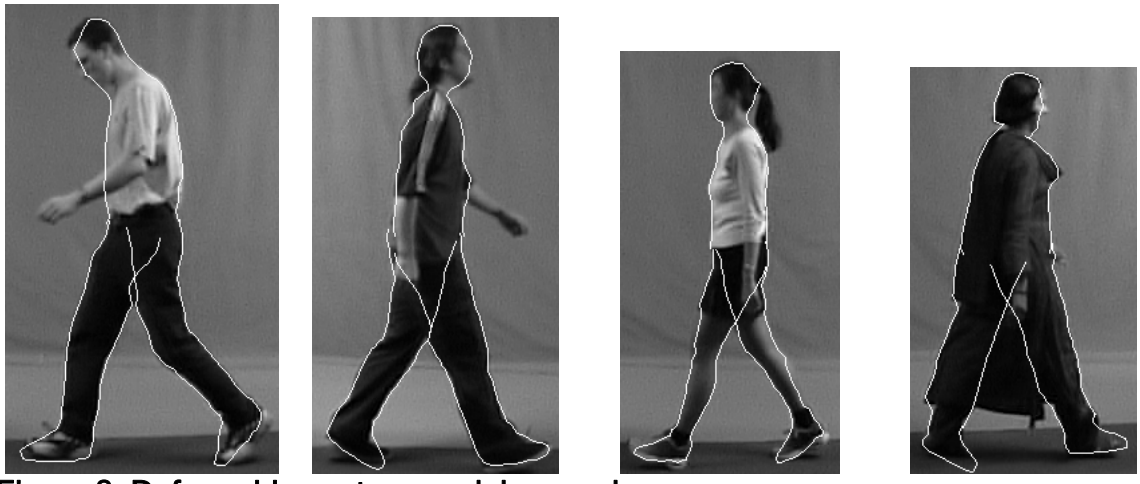


Figure 8: Deformable contour model examples

Local deformation is defined for each frame as the displacement from the base contour in the direction of the local normal vector:

$$C_{def}(s,t) = C_{base}(s,t) + d(s,t) \cdot normal(C_{base}(s,t)) \quad (4)$$

Where C_{def} is the deformed contour (Figure 9) and d is the deformation magnitude. The local normal to the base contour C_{base} is computed by taking the cross product of the local difference vectors:

$$normal(C_{base}(s,t)) = (C_{base}(s-1,t) - C_{base}(s,t)) \times (C_{base}(s,t) - C_{base}(s+1,t)) \quad (5)$$

Some of the local normal computations can be avoided for the leg contours by using a single normal for each segment (between the hip and knee, or knee and ankle). This approximation can be made because the leg contours are approximately straight between joints, constrained by the rigidity of the leg bones.

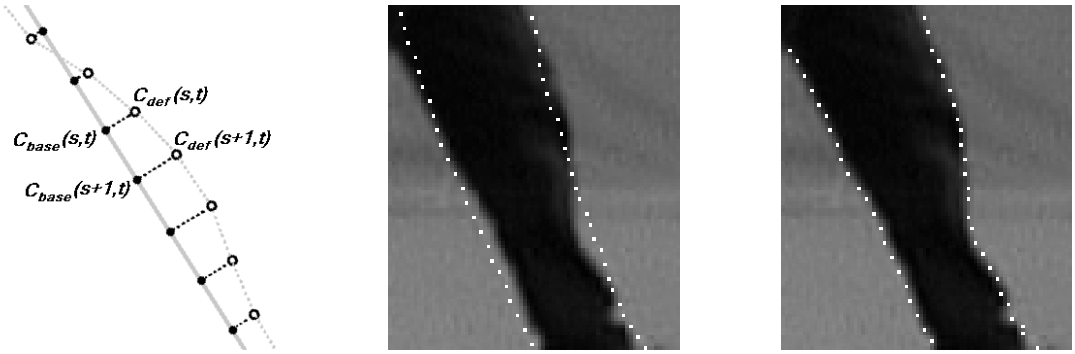


Figure 9: Deformation from base contour model in direction of local normal

The deformation magnitude is determined using the energy minimisation framework described for the snake algorithm [Kass 87, Williams 92]. Additional model and temporal constraints are incorporated within this framework to make best use of the prior knowledge available (Sections 4.2.2 and 5.2.3).

2.3. Motion Models

2.3.1. Centre of Mass Motion

In a hierarchical ordering of motion components, displacement of the body *centre of mass* (COM) is of greatest importance; all other components are defined relative to the COM:

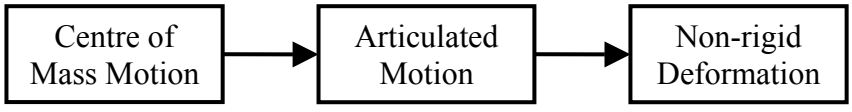


Figure 10: Motion hierarchy for human gait

Non-rigid deformations are changes in shape due to clothing or soft tissue effects (Section 2.2.2). Articulated motion covers all shape deformations due to joint rotations (Section 2.3.2). COM motion includes translation of the whole body along the *x*, *y* and *z* axes, and also any recognisable global characteristics of motion, for example, vertical oscillation. In reality COM motion is a consequence of limb motion, but from a computer vision point of view, limb position is difficult to resolve accurately, whereas COM is relatively easy to find. Consequently, it is more effective to extract COM motion first, and subsequently resolve limb motion.

COM motion is extracted in three stages, employing more complex constraints only after basic motion has been successfully extracted. As the first stage (Section 3.2.1) has little prior information to build on, simple global models are employed to quickly determine the approximate position and scale of the subject. This stage is more concerned with reliability of extraction than with accuracy. Each subject is walking along approximately level ground, so motion in the *y*-plane is at this stage assumed to be negligible ($Y_1(t) = y_0$, where y_0 is a constant) and the following constraint equation is applied to motion in the *x*-plane:

$$X_1(t) = vt + x_0 \tag{6}$$

Where *v* is the subject's velocity, *t* is a time index and x_0 is the starting *x*-coordinate. These assumptions are insufficient for later stages, which require more accurate localisation of the subject's COM to resolve limb position and shape deformations.

The second stage (Section 3.2.3) determines COM motion more accurately, using the first stage as a starting point. The motion constraints now allow for acceleration in the *x*-plane:

$$X_2(t) = at^2 + vt + x_0 \tag{7}$$

Where *a* is the subject's acceleration. Motion in the *y*-plane is assumed to be oscillatory about a fixed point. A person's COM will rise and fall in a periodic fashion during gait due to limb articulation, with a frequency double that of the gait cycle [Gard 01]:

$$Y_2(t) = A_y \sin 2\left(\omega t + \phi + \phi_y\right) + y_0 \tag{8}$$

Where A_y is the amplitude of vertical oscillation, ω is the gait frequency, ϕ is the gait phase, ϕ_y is the *y*-phase difference and y_0 is the centre of motion in the *y*-plane. ϕ_y is approximately $\pi/8$, representing the phase difference between the cycle of vertical displacement and the cycle of leg motion. This quantity was empirically determined from a small subset of the indoor dataset.

The third stage (Chapter 5) is an iterative process, successively estimating local deviation from a global prediction and then re-estimating a new global model from the collected local model configurations. Motion in the *x*-plane is constrained in the same manner as in the previous stage ($X_3 = X_2$). In the outdoor dataset, the ground plane is not quite level and there may be some deviation from the walking path, as it is not as clearly marked in the outdoor dataset as in the indoor dataset. Camera distortion may also cause curvature in the ground plane. To account for these variations, the final constraint equation adds the *y*-motion gradient, v_y:

$$Y_3(t) = A_y \sin 2\left(\omega t + \phi + \phi_y\right) + v_y t + y_0 \tag{9}$$

Motion in the *z*-plane (movement towards or away from the camera) is assumed to be negligible at all stages.

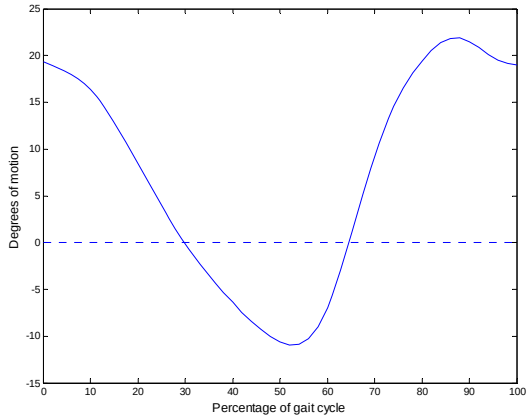
2.3.2. Articulated Motion

The articulation (jointed motion) of the limbs during gait greatly increases the difficulty of tracking the human body. An enormously complex model is required to fully account for the variability in body pose and appearance. To reduce the computational burden of this model, it is important to take advantage of any prior knowledge available.

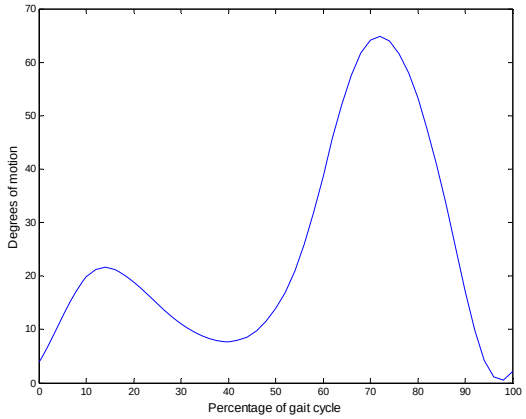
Marker-based clinical studies [Whittle 99, Winter 91, Gard 01] have measured mean gait patterns for normal people. This data forms the basis of prototypical models for leg articulation and consequent vertical displacement of the upper body. Figure 11 displays these models for a single gait cycle, from right heel-strike to right heel-strike (see Section 2.2.1, Table 4 for conventions in measuring leg joint rotation). Beginning with a prototype model has two main advantages; firstly that approximate limb configuration can be established very quickly, with minimal knowledge of other body parameters. Secondly, it is far easier to ascertain an individual's variation from the norm than it is to determine their gait pattern from first principles.

This raw data is sampled at 15 points over a single gait cycle to generate a set of models suitable for use in posterior model extraction (Chapters 3-5): $[\Theta_{px}(n), \Theta_{py}(n), \Theta_h(n), \Theta_k(n), \Theta_a(n)]$, $0 \leq n < 15$. This sampling rate is motivated by a clinical study [Angeloni 94] indicating that the majority of gait information is contained below 5Hz, implying a minimum sampling rate of 10Hz for accurate reconstruction. For a typical gait cycle

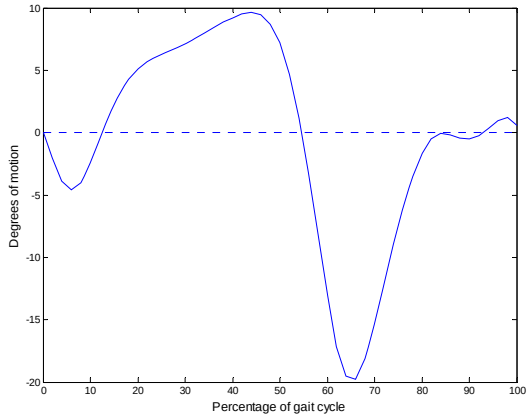
duration of 1 second, 15 samples per cycle is equivalent to a sampling frequency of 15Hz, which allows some degree of latitude for longer gait cycles (people who walk more slowly than average).



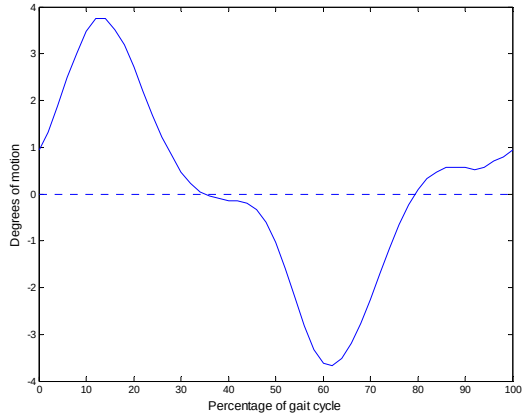
(a) Hip rotation Θ_h



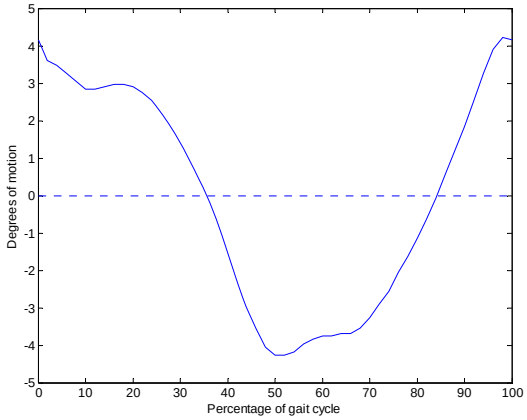
(b) Knee rotation Θ_k



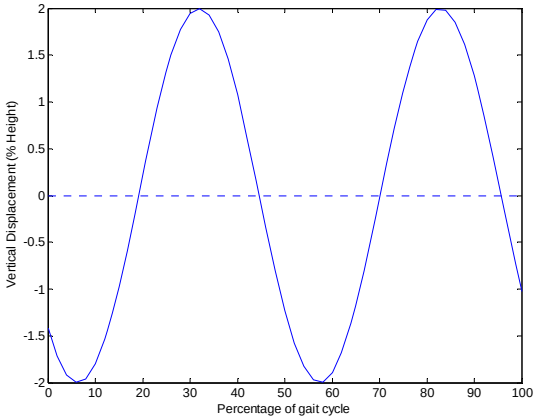
(c) Ankle rotation Θ_a



(d) Pelvic rotation about x axis Θ_{px}



(e) Pelvic rotation about y axis Θ_{py}



(f) Vertical displacement

Figure 11: Mean joint rotation patterns for a single gait cycle

The minimum amount of information required to employ these models is the subject's gait frequency ω and phase ϕ, which measure cadence and starting leg pose. Section 3.2.2 details an efficient means of estimating periodicity, using simple area-based edge strength measurements. Using these parameters, each prototype is scaled in frequency to fit the subject using Hermite spline interpolation [Kochanek 84], shifted according to phase and replicated periodically to fill the gait sequence. This process generates a set of predicted joint rotations for each leg at each frame in the gait sequence: $[\theta_{px}(t), \theta_{py}(t), \theta_h(t), \theta_k(t), \theta_a(t)]$, where $0 \leq t < T$, T is the number of frames in the gait sequence and $\theta_{right}(t) = \theta_{left}(t + P/2)$, where P is the gait period ($P = 2\pi/\omega$).

The above application of the prototypes assumes that the left and right legs move identically at a phase difference of π radians, and that each gait cycle repeats exactly the same pattern. Although gait motion will vary under normal conditions, for uninterrupted sections of gait, the tracking error associated with these assumptions is small. Gait asymmetry may prove to be a useful discriminant for recognition purposes, but is not considered within the scope of this thesis.

This approach has the additional advantage of increasing the degree of averaging in shape and motion extraction. If part of the leg is obscured for any reason, we are not limited to purely spatial means of reconstructing the missing data. Image data can be substituted from the same part of the gait cycle at other points in time, or from the other leg. This capacity greatly increases the robustness of extraction. Of course, these models contain little individual gait information, but they are a good basis for adaptation techniques (Chapters 4 and 5), which aim to adapt these generic models to fit the data observed for each subject.

2.4. Occlusion Models

Self-occlusion of the legs during motion is a major tracking problem. Local tracking algorithms will typically fail when image data is missing, because they do not incorporate sufficient prior knowledge or surrounding data to fill in the missing values. Global tracking algorithms will fare better when missing data is encountered, because data is gathered across the entire gait sequence, but computational requirements are greatly increased due to this extra data. As a consequence, it is desirable to use local algorithms where reliable

data is available, falling back on global algorithms only when there is too much missing data. This is possible because self-occlusions are predictable to some extent; leg motion is periodic and conforms to a known pattern, so it is known approximately which parts of the leg will be occluded at each point in the gait cycle. Occlusions from other objects can also contribute to the occlusion model if they are included in the tracking framework. Although the object state may not be known precisely, any information describing the position and size of the object is potentially useful in predicting possible occlusions.

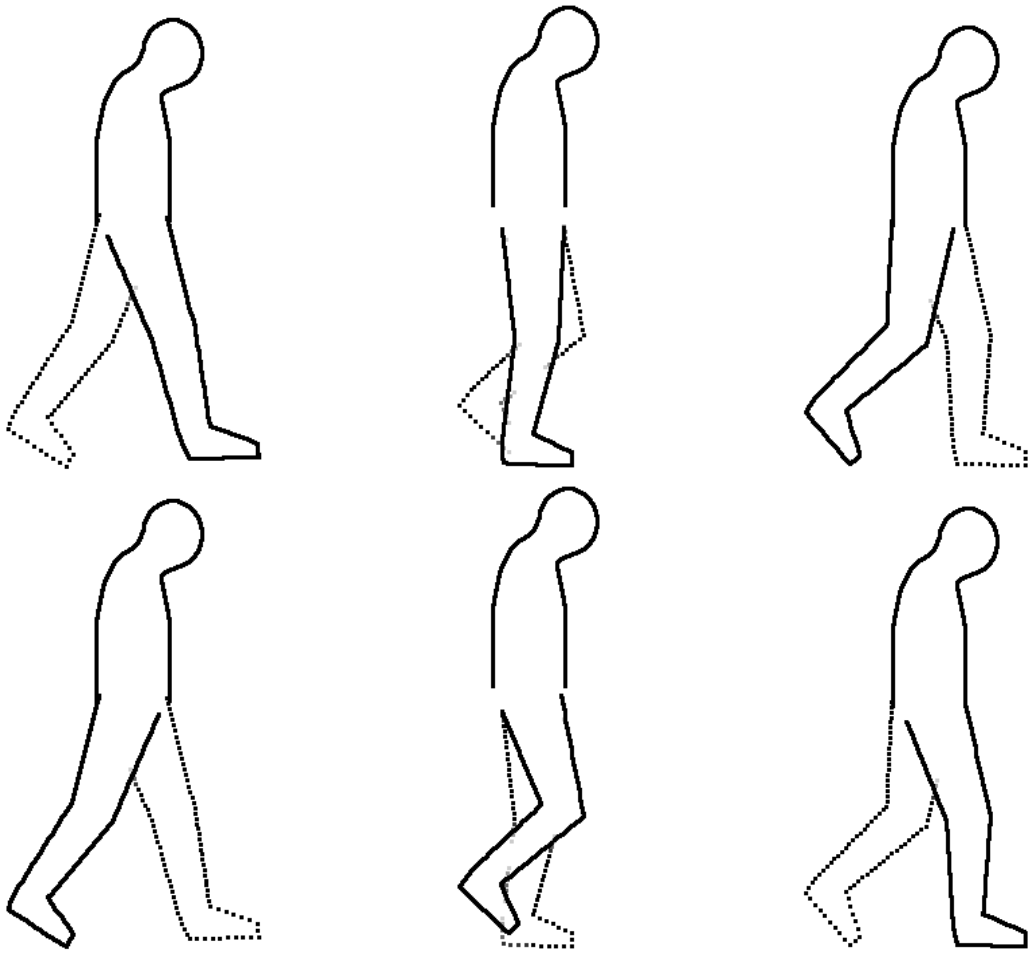


Figure 12: Occlusion model predicting visibility of parts of the leg during gait – left leg (top row) and right leg (bottom row)

An occlusion model o is computed by overlaying the contours representing each leg, assigning to each point on the model an occlusion value of '1' if it is completely visible and '0' if it is completely occluded (obscured by the other leg):

$$0 \le o\left(C_{base}(s,t)\right) \le 1 \tag{10}$$

The legs are assumed to be mutually occluding, as where the legs intersect, edges will not be visible for either leg. This assumption arises from the likelihood that both legs will be the same colour, and thus there will be no significant edges where the legs coincide. This computation generates a binary occlusion model, to which spatial smoothing is applied to generate a continuous model. Figure 12 illustrates the occlusion model for some example leg configurations, where the dotted line represents visible parts of the occluded leg.

The occlusion model is employed when computing shape deformations (Sections 4.2.2 and 5.2.3) to ensure that image data in occluded regions is not used to generate the posterior model. The contour shape is determined partly by the strength and location of image edges, and partly by globally driven spatial and temporal constraints. The occlusion model balances these two determinants, ensuring that in occluded areas, where image edges are likely to be absent or unreliable, the contour shape is dominated by global constraints. These constraints can be relaxed where the leg is expected to be visible, so that image data plays a greater part in determining leg contour shape.

The arms are not yet tracked, and consequently are not included in these occlusion models. It is expected that some errors in contour shape will be evident at hip level, where the hands pass in front of the legs and the torso. See Sections 6.2.2-6.2.3 for examples of models extracted with the aid of an occlusion model.

2.5. Conclusions on Modelling Strategies

The model-based computer vision paradigm seeks to separate a scene into those parts that can be understood and explained according to prior knowledge, and those parts that are extraneous. In the case of gait extraction, this prior knowledge is of normal human body shape, gait motion and the expected (scene-dependent) path of the subject. The modelling strategy is the means by which this knowledge is applied: in this chapter, two different classes of strategy are identified.

A global modelling strategy generates a single model that can describe the whole sequence, using all of the image data to determine its configuration. This approach is highly robust, because any local inadequacies in data due to noise or occlusion can be overcome by interpolation between surrounding areas, both in the spatial and temporal domains. There is no initialisation problem, because the model is initialised at all points in

time simultaneously, using all the available data. The main difficulty when using a global modelling strategy is the computational requirements. The model space is very large, because the size of the space increases in geometric proportion to the number of model parameters and the number of frames in the sequence. Additionally, each model evaluation must consider the whole sequence, making each model state more expensive to evaluate. Consequently global approaches perform best when a great deal of prior knowledge is available and the problem can be highly constrained, thus reducing the number of valid model states.

The opposing strategy is to model the sequence at a local level, describing a single frame at a time. The primary motivation for this approach is to reduce computational requirements. Given real-world limitations on computer power, lower computational requirements mean that a more complex model can be used, thus more accurately representing the subject. However, the lack of temporal information renders this strategy vulnerable to missing data, as it is difficult to reconstruct occluded features. It is worth noting that both the initialisation and missing data problems are significantly reduced if a global strategy is used to construct an initial model that a local strategy can build upon.

This chapter has considered global and local modelling strategies as two opposing extremes; it is of course possible to define a strategy somewhere in between, exploiting the temporal domain to a greater or lesser extent. Most approaches in the recent literature operate on a local level, considering a bare minimum of temporal information to solve the problems of initialisation and missing data. The justification for such approaches is usually that the computational burden of including more temporal information is too high. This does not do the global paradigm justice, as these requirements can be managed by reducing model complexity. Although one may baulk at the amount of data and dependencies considered in a global approach, it is worth considering that this information is always present, it is simply ignored in a local approach.

There is one final point to consider in the area of modelling strategies. In a model-based approach, it is necessary to gather prior knowledge of body shape and motion, and to decide what is allowable and probable in each of these cases. This is an additional source of error that should be considered. If any of the prior models are inaccurate, it may become difficult or even impossible to determine the correct model configuration, and if any of the assumptions made in constructing the model are violated, the resulting gait description is unlikely to be acceptable.

Chapter 3. Posterior Model Initialisation

3.1. Introduction

This chapter is concerned with initialisation of the posterior model. The posterior model is a description of the target object, constructed from the observed image data and guided by the prior model.

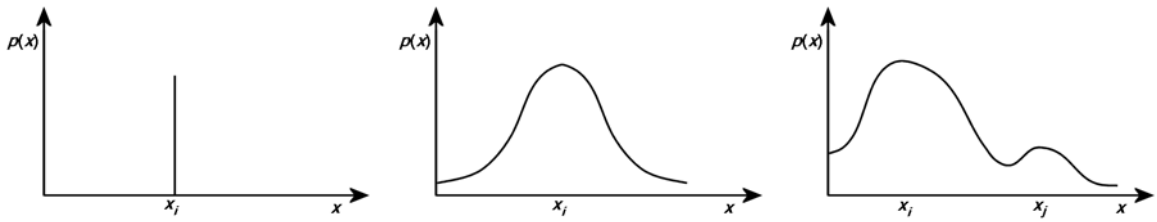
Recent reviews in the area of posterior human model extraction include [Moeslund 01, Wang 03a]. A typical approach to posterior human model extraction involves *tracking*, extracting a local model for each frame in the sequence and deriving constraints at each frame from the previous model configuration. This limited form of temporal dependency allows continuity of motion to be enforced over a short period of time, while retaining the favourable computational requirements of a local approach. However, there are some difficulties in implementing this solution.

The first problem is *initialisation*; tracking must start somewhere, and wherever the start point is, there will be no previous frame of reference from which to derive constraints. Unreliable image data in the first frame can also prevent a good initialisation. The second problem is *missing data*, causing a loss of tracking. This problem occurs when other objects obscure the target and the model state is not determined correctly. Such errors can be difficult to recover from, because each model is limited by the previous configuration through temporal continuity constraints. If the error is large enough, tracking can fail entirely as the correct model configuration is eliminated from the search space.

The global evidence gathering paradigm was formulated as a solution to these problems, exploiting the whole image sequence to realise a robust tracking algorithm that is effective in very noisy imagery [Nash 97]. All available image data is considered simultaneously, so that localised noise cannot upset the extraction process. The main difficulty in the global approach is that it is impossible to implement in a naïve fashion; the search space is too large to be evaluated exhaustively. Generally there are two approaches to solving this problem: the first is to simplify the models until the search space is tractable, and the second is to employ optimised search algorithms that evaluate only a subset of the search space, such as a genetic algorithm [Nash 98] or gradient descent.

The latter approach allows a global approach to be used with minimal alteration, even for relatively complex models. However, this does come at the cost of some reliability; no non-exhaustive search algorithm can guarantee an optimal solution, and the larger the search space becomes the less likely it is that a good solution will be found. The simplification approach is generally more robust, as the whole search space can be evaluated. The drawback in this case is that simplifying the model reduces its information content, limiting potential applications.

An alternative approach is to use probabilistic models to overcome the missing data problem, such as in the Condensation algorithm [Isard 98] or other particle filtering algorithms [MacCormick 00]. A deterministic model assumes that the state of the target object can be determined exactly (Figure 13a), whereas a probabilistic model assumes some degree of uncertainty in the estimated state of the target object (Figure 13b).



(a) Single state (b) Single hypothesis (c) Multiple hypothesis

Figure 13: Deterministic and probabilistic models of object features

Loss of tracking is possible in standard tracking algorithms when missing data is encountered because only one hypothesis (model state) is considered for each frame, and the current model state is assumed to be within a small distance of the previous model state. This means that when the model state is determined incorrectly, subsequent frames will inherit a poorly defined search space and tracking may be lost entirely. In a probabilistic approach, the search space is derived by sampling the model space according to the computed probability of a match, concentrating the search in more likely areas. If the model can accommodate multiple hypotheses (Figure 13c), a much wider variety of model states can be considered for the next frame, allowing for recovery of tracking. The drawback to allowing multiple hypotheses is that computational requirements can be significantly higher. This is tackled in particle filtering approaches by the use of statistical sampling to concentrate sampling in more likely regions of the model space, thereby reducing the number of model states evaluated.

Probabilistic models are not utilised within this thesis due to the additional complexity in their implementation. The issue of probabilistic versus deterministic models does not affect the central argument of this thesis regarding the utility of local and global models; the approaches described assume a deterministic gait model, but could equally be implemented using a probabilistic model.

This thesis proposes an approach to posterior model extraction combining local and global models. To this end, extraction is split into two stages; an initialisation stage and an adaptation stage. This chapter describes the first stage (Figure 14), using global models to overcome the initialisation problem encountered when using local models. The adaptation stage is covered in Chapters 4 and 5.

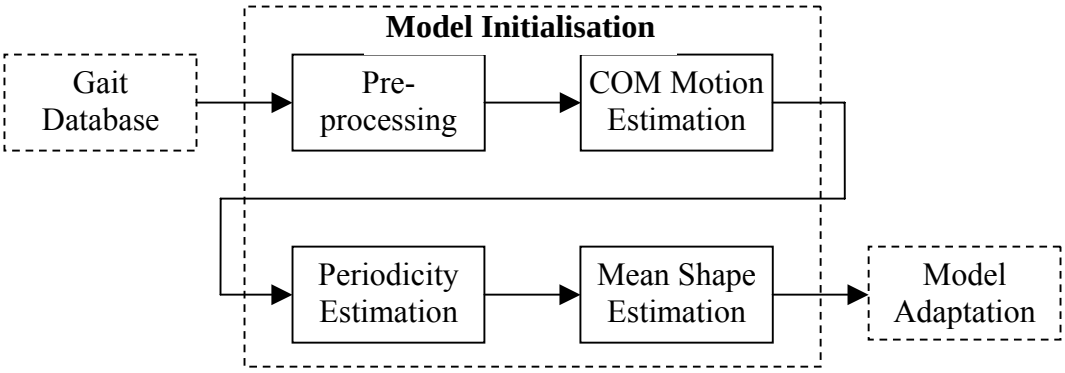


Figure 14: Overview of Posterior Model Initialisation

The aim for this chapter is to abstract a low-dimensional description of the subject's shape and motion from their image data. This model is an initial approximation that later stages will build upon, so although it does not have to be too precise, it must be reliable. If the initial model fails to capture important subject features, then any attempt at refining the model will also fail. Of course, an accurate initialisation is desirable, but reliability is the primary goal for this stage.

3.2. Global Evidence Gathering

3.2.1. Centre of Mass Motion

The most basic task in motion capture and surveillance applications is detecting the presence of a person, and tracking their position throughout the video sequence. In the Southampton Gait Database (Section 1.4), each subject walks at an approximately constant speed along a fixed horizontal track. As a consequence it is possible to apply a simple and robust means of global tracking through the temporal accumulation of edges [Wagg 03]. Edges are detected using the Sobel operator, and background subtraction is applied using Yang and Levine's method [Yang 92]. This pre-processing step reduces the sequence to only moving edges (Figures 15 and 16). Appendix A includes some further examples of the edge data generated using this method.

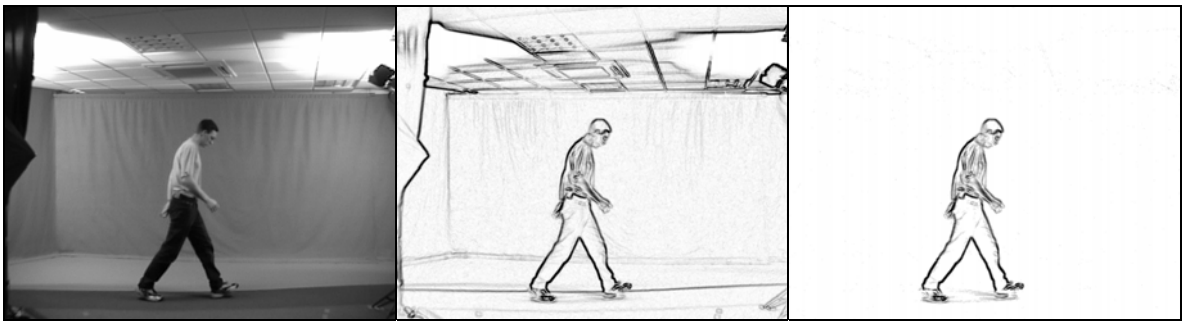


Figure 15: Pre-processing to reduce quantity of data (indoor dataset)

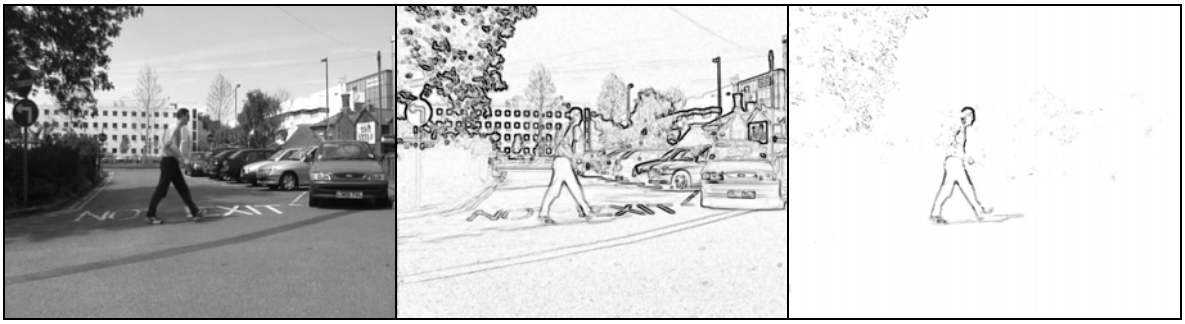


Figure 16: Pre-processing to reduce quantity of data (outdoor dataset)

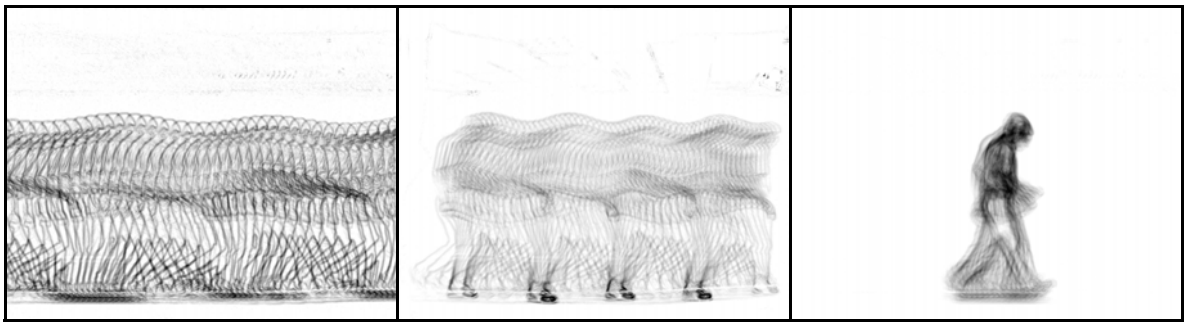
This method of pre-processing is very simple, and performance on the outdoor dataset is not that impressive. Some important edges are lost, and there is still a great deal of irrelevant data present, due primarily to the shadows cast by the tree at the left-hand side of the scene. Many more advanced methods of background removal exist [Al-Mazeed 03, Elgammal 02, Stauffer 99], which can be applied where performance is a critical requirement. However, this thesis is concerned with the development of robust computer vision algorithms rather than systems directly applicable to a commercial environment, in which maximising performance would be an important consideration. For the purpose of

testing extraction algorithms, sub-optimal image data quality is not an important issue, as it is performance under adverse conditions that is of interest.

To find objects moving in the horizontal plane, the edge sequence is temporally accumulated according to the expected motion of the subject's centre of mass (see Section 2.3.1, Equation 6):

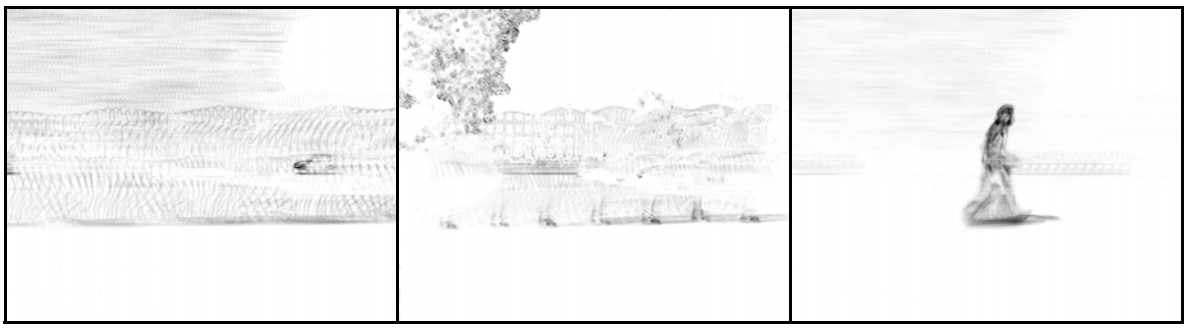
$$A_v(i,j) = \sum_{t=0}^{T-1} I_{edge}\left(i + v\left(\frac{T}{2} - t\right), j, t\right) \tag{11}$$

Where A_v is the accumulation for velocity v (pixels per frame), I_{edge} is the pre-processed edge image sequence, i and j are coordinate indices, t is a time index and T is the number of frames in the sequence. The offset $T/2$ ensures the subject's edges accumulate in the middle of the frame rather than the side, so that data is not lost. This function generates an accumulation that represents the average shape of the object(s) moving at the given velocity. Some example accumulations are given in Figures 17 and 18 (these images have been normalised to improve visibility).



(a) v = -10.0 (b) v = 0 (c) v = 10.0

Figure 17: Temporal accumulation of indoor sequence '008a013s00R'



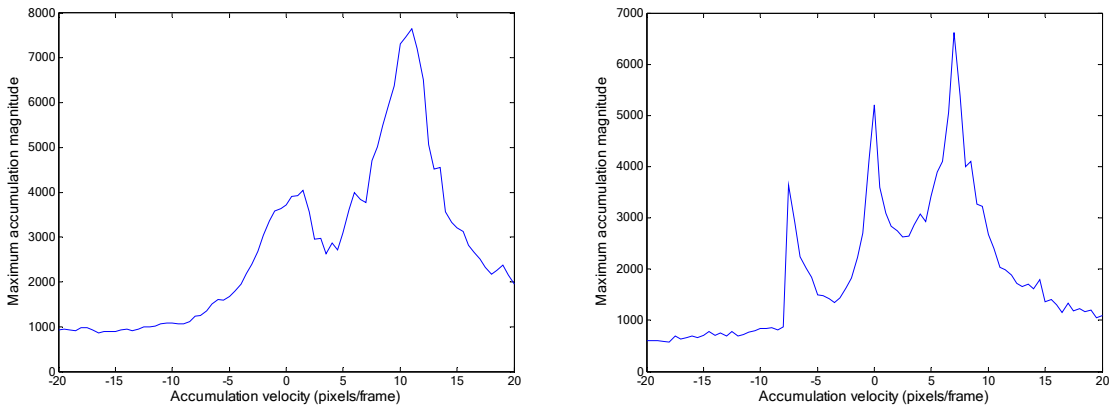
(a) v = -7.0 (b) v = 0 (c) v = 7.0

Figure 18: Temporal accumulation of outdoor sequence '008e013s00R'

The accumulations generated for the indoor example clearly shows the subject at a velocity of 10 pixels per frame, while there are no distinct objects at v = -10. At a velocity of zero,

only the lower part of the supporting leg is visible, as this is the only moving object that remains in the same position for extended periods of time. In the outdoor sequence, more objects are visible at a velocity of zero, as trees and other foliage have no net velocity, but will move and sway in the wind. The subject is clearly visible at $v = 7$, but there is also a car visible moving in the opposite direction ($v = -7$).

If the accumulation velocity matches that of an object within the scene, its edges will accumulate to a localised region, with the result that $max(A_v)$ will be high. Conversely when no objects are moving at the accumulation velocity, object edges will be spread out across the accumulation, resulting in a low value for $max(A_v)$. A number of peaks is usually evident in the variation of $max(A_v)$ with v (Figure 19), indicating the presence of multiple objects or multiple velocities in the scene. Positive velocities correspond to objects moving from left to right, and negative velocities correspond to objects moving from right to left.



(a) Indoor (b) Outdoor

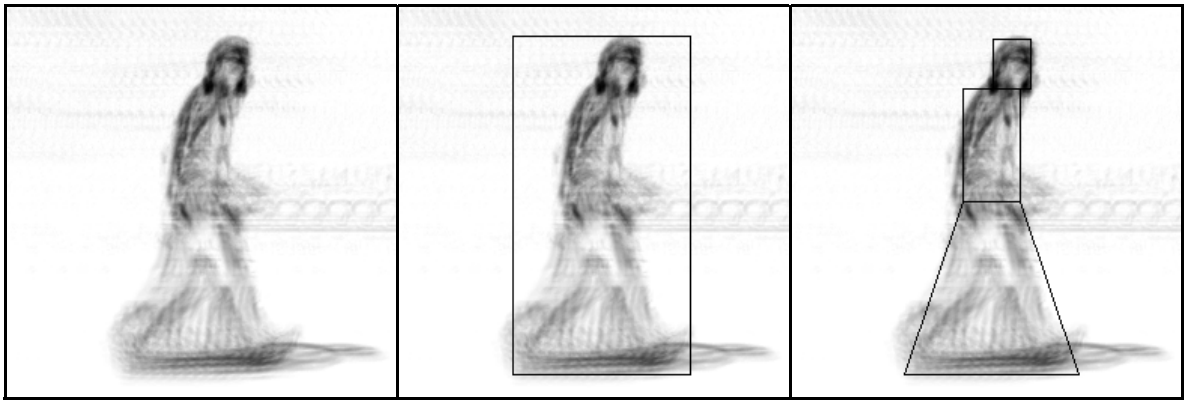
Figure 19: Variation of max(A_v) with accumulation velocity

Since under normal conditions a walking person must have non-zero velocity, the peaks at $v = 0$ can be discounted; these peaks represent stationary motion sources, such as trees and foliage. The lower part of the supporting leg is also apparent in these accumulations, as it is stationary for extended periods of time. This means that for the indoor dataset, the velocity of the subject can be adequately determined from the highest peak in this plot, as the subject is known to be the only moving object in the scene. This is insufficient for the outdoor dataset, as there can be other significant peaks in this plot due to the presence of other objects (for example, the peak at $v = -7$ in Figure 19b is caused by a car in the background, visible in Figure 18a). In order to distinguish people from other moving

objects, it is necessary to score each accumulation according to both peak magnitude and correlation with average human shape. The best fitting subject velocity v^* is given by:

$$v^* = \arg\max_v \left(\max(A_v) \times corr(A_v, MHS_v)\right) \quad (12)$$

Where MHS_v is the mean human shape model (Section 2.2.1) fitted to accumulation A_v. This process is illustrated in Figure 20. A bounding box is fitted to the accumulation, constraining its aspect ratio to approximate human proportions (2:1). The bounding box is initialised at the position of $max(A_v)$, and iteratively expands along each edge while there is sufficiently high edge strength present along the edge, and the aspect ratio remains close to human proportions. The bounding box is then shifted so as to maximise the sum edge strength within the box, to reduce the effect of shadows cast on the ground. The height and centre coordinates of this bounding box set the position and scale of the mean human shape model. It can be difficult to fit a bounding box to noisy data, particularly if there are other objects moving at the same velocity in the same region as the subject. This is an important source of error within the posterior model initialisation process, but could be reduced if the tracking framework considered other moving objects as well as the subject.



(a) Accumulated region (b) Bounding box (c) Mean human shape model

Figure 20: Fitting a mean human shape model to accumulated edge data

Temporal accumulation effectively performs the same global evidence gathering process as the Velocity Hough Transform [Nash 97]; however, decoupling shape parameters from motion parameters improves the efficiency of the process. Noting that Equation 11 simply shifts and accumulates each frame, computational efficiency can be improved by first run-length encoding the edge data. This representation is shift-invariant and, as runs of zero magnitude edge strength can simply be discarded, the order of the algorithm is reduced to

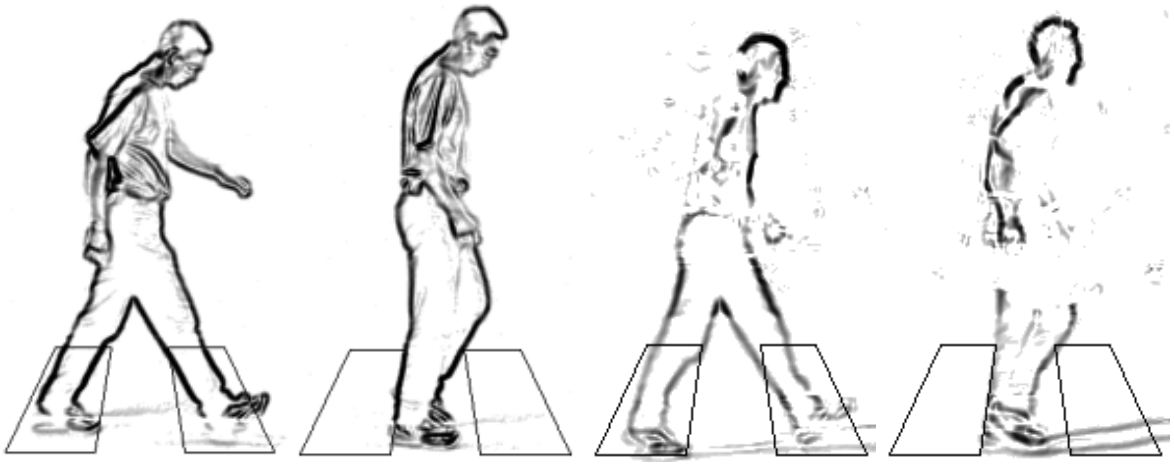
O($V \cdot E \cdot T$); where V is the number of possible velocities, E is the mean number of edge points in a frame and T is the number of frames in the gait sequence.

The computational requirements of this algorithm are low, because motion is constrained to a constant velocity in the horizontal plane and mean human shape is assumed. Robust operation is derived from the global nature of the accumulation, as motion and scale are estimated from an average measure of the whole sequence. Local deficiencies in image data are implicitly compensated for by more reliable data in other frames. This algorithm relies on a number of simplifying assumptions that make it unsuitable for more generic application scenarios, which may require the capability to deal with unconstrained motion and crowded scenes. There are many other approaches that specifically target these capabilities [Haritaoglu 00, Zhao 04], but this is an unnecessary complication to the central argument of this thesis.

3.2.2. Periodicity

The motion of a person's limbs during normal gait creates a complex periodic pattern, formed from many different components. Gait frequency and phase are particularly useful components, because together they describe a large part of this motion, and can be easily extracted without resolving limb dynamics. In simple terms, the gait frequency describes the speed at which the subject moves their legs, and phase describes their starting pose.

Methods for estimating motion periodicity are well established. Cutler and Davis present a general method for periodicity detection [Cutler 00], based on measurements of silhouette self-similarity over time and using autocorrelation-based analysis to extract periodicity. The main drawback of this generic approach is that the computational demands are quite high due to the cost of computing silhouette self-similarity, particularly for long sequences of video data. More recent approaches have typically measured simpler silhouette features such as width and height [BenAbdelkader 02, Collins 02], resulting in much lower computational requirements. This thesis employs edge features rather than silhouette features, being better suited to extraction of detailed limb dynamics.



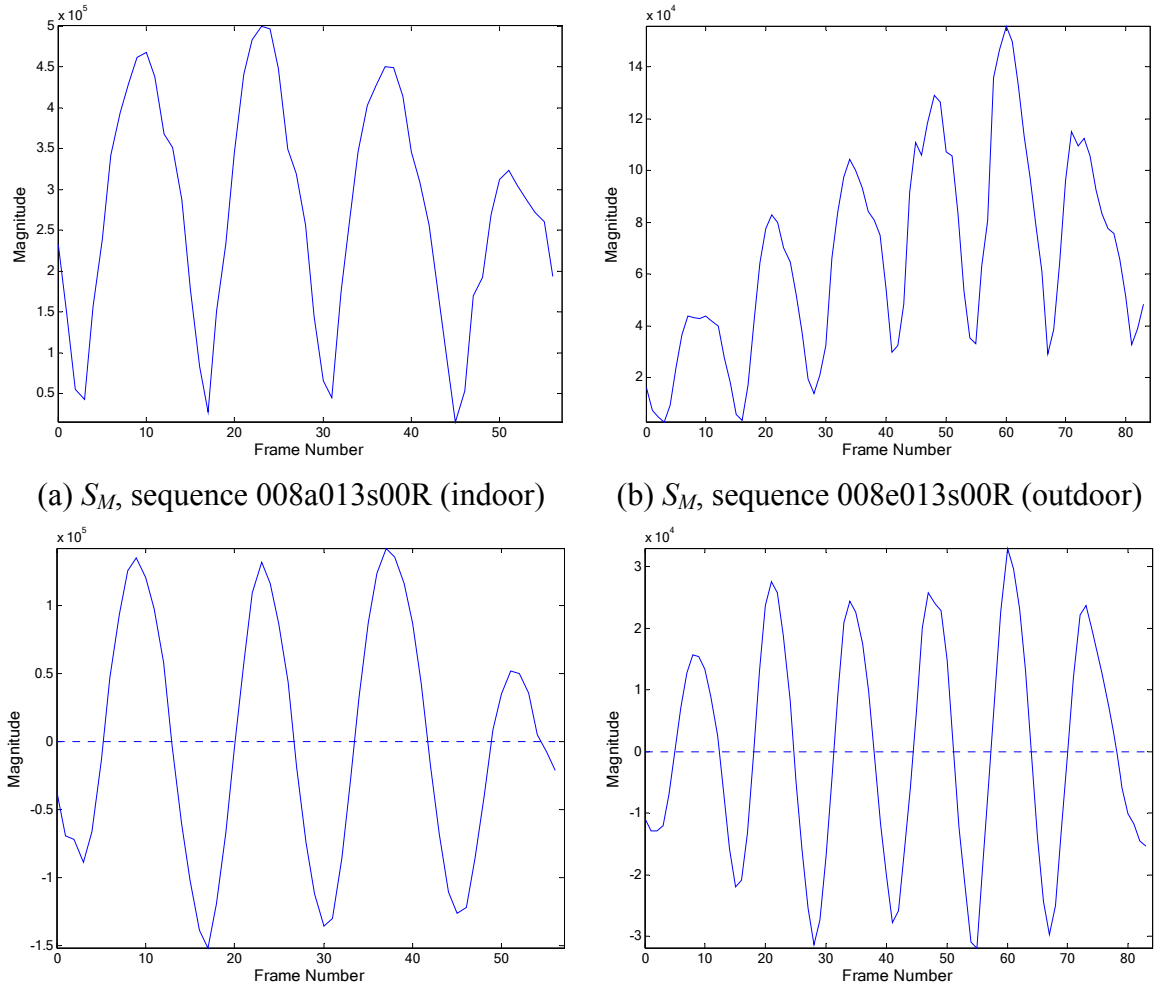
(a) Indoor data (b) Outdoor data

Figure 21: Periodicity extraction using a simple area mask

Assuming the subject is of mean anatomical proportions, a binary periodicity mask (*PM*) is constructed to overlay the outer regions of a person's stride (this mask is defined in Section 2.2.1, Figure 4d). A periodic sequence of measurements S_M is generated by computing the sum edge strength within the periodicity mask *PM*:

$$S_M(t) = \sum_{i,j} I_{edge}(t) * PM \tag{13}$$

This total varies according to leg pose, being greatest at heel-strike frames when the legs are separated and lowest when the legs are crossing (depicted in Figure 21a-b). The sequence of measurements obtained for an example indoor and outdoor sequence are depicted in Figure 22a-b. Both sequences clearly show a periodic pattern of variation in sum edge strength as the legs sweep through the area of the mask, which may be modelled as a sinusoid plus noise sources. The main sources of noise (non-periodic variation) are shadows, occlusions and edges belonging to other objects moving in the vicinity of the subject. This may be observed in Figure 22b, in which the reduction in magnitude over the first half of the sequence is caused primarily by the shadow cast by a large tree, located at the left-hand side of the scene (see Section 1.4, Figure 3). The distance of the subject from the camera may also be a factor, as the person will be closer to the camera at the midpoint of the sequence and this may increase the strength of the subject edges. Any errors made in estimating the motion of the subject may additionally increase the noise level of this sequence.



(a) S_M, sequence 008a013s00R (indoor)

(b) S_M, sequence 008e013s00R (outdoor)

(c) S_F, sequence 008a013s00R (indoor)

(d) S_F, sequence 008e013s00R (outdoor)

Figure 22: Periodicity measurements for example gait sequences

The effects of noise on a sinusoidal sequence may be grouped into three categories: variation in local mean (the position of the sinusoid centre), variation in sinusoid amplitude, and small-scale fluctuations. These noise effects can reduce the accuracy of period and phase extraction, and explicitly modelling and removing noise sources before performing periodicity analysis can improve performance. Small-scale noise is easily removed using a running-average antialiasing filter. For short gait sequences, the two remaining noise sources can be modelled by low-order polynomials that represent the effect of scene features on the measured edge strength of the subject's legs. For longer gait sequences, simple polynomials may be insufficient to model the possible noise effects, and it may be necessary to analyse the sequence piece-wise. It may also be possible to obtain an independent estimate of the local noise level by measuring variation over the sequence in edge strength in the area surrounding the subject, avoiding the area affected by the

person's gait motion. However, it may be difficult to completely avoid the influence of the subject and any shadows they cast from this noise estimate, and the estimate computed for the area surrounding the subject may not accurately reflect the noise level of the area occupied by the subject. For this reason, the noise level of the sequence is estimated directly using low-order polynomials, and the sequence is filtered to remove its influence:

$$S_F = \frac{S_M - p_1(S_M)}{p_2(\,|S_M - p_1(S_M)|\,)} \tag{14}$$

Where S_F is the filtered sequence, S_M is the measured sequence and $p_1(x)$ and $p_2(x)$ denote the best polynomial fit to sequence x, computed by least-squares linear regression [Fox 97]. The order of the polynomials p_1 and p_2 may be varied according to the degree of noise filtering required. In practice however, little benefit is gained from the second term p_2 (see Appendix B), and the filtering operation can be simplified to:

$$S_F = S_M - p(S_M) \tag{15}$$

Figure 22c-d shows examples of the resulting filtered data (order of $p = 3$). Although some non-periodic variation is still evident, the filtered sequence is more regular and closer to a pure sinusoid than the measured sequence, increasing the likelihood of successful period extraction. There are two common approaches to extracting periodicity information from sequences of noisy data, Fourier analysis and autocorrelation-based analysis. The Fourier transform decimates a sequence into a sum of sinusoids, so extraction of periodicity information from a Fourier representation is quite straightforward. The main drawback of Fourier analysis is in the complexity of implementation, and the need to resample the input to achieve sufficiently high accuracy in periodicity determination. Autocorrelation-based analysis is far simpler, relying on comparisons of the sequence with delayed (shifted in time) versions of itself to derive periodicity information:

$$autocorr(k) = \frac{\sum_{i=0}^{T-1}\left(S_F(t) - \overline{S}_F\right)\left(S_F(t+k) - \overline{S}_F\right)}{\sum_{i=0}^{T-1}\left(S_F(t) - \overline{S}_F\right)^2} \tag{16}$$

Where $autocorr(k)$ is the autocorrelation of the sequence S_F at delay k and $\overline{S}_F$ is the sequence mean. A high autocorrelation coefficient is likely to indicate periodicity in the sequence of the corresponding time delay, so the dominant period of the sequence can be determined by finding the delay that maximises autocorrelation.

A simple alternative method for detecting the period of an approximately sinusoidal sequence is direct matching of prototype sinusoids. Periodicity is determined by finding

the frequency and phase that minimises the difference between the sequence and a prototype sinusoid. This minimisation can be performed very quickly for the range of frequency and phase expected for a walking person:

$$X_s^* = \min_{\omega_i,\phi_j}\left(\sum_{t=0}^{T-1}\left(S_F(t) - A_s \sin 2\left(\omega_i t + \phi_j + 0.58\right)\right)^2\right) \tag{17}$$

Where X_s^* is the minimal error, A_s is the sinusoid amplitude (a fixed proportion of the mean sequence magnitude) and ω_i and ϕ_j are the proposed gait frequency and phase. Note the small phase offset that is required to align the cycle of measured edge strength with the gait cycle, as peaks in S_F do not coincide exactly with heel-strikes (peaks in the gait cycle). This phase offset was determined experimentally using a small subset of the indoor database, suggesting an optimal value of around 0.58 radians. It is also important to note that the frequency of the edge strength cycle is twice that of the gait frequency (the periodicity mask cannot distinguish between the left and right legs).

For the indoor dataset, there is approximate ground truth data available for periodicity, as the frames in which a heel-strike occurs are labelled. The accuracy of the extracted gait frequency and phase can be estimated by comparing the predicted heel-strike frames with the labelled frames:

$$MD = \frac{1}{NS}\sum_i\left(\frac{1}{H(i)}\sum_j\left|HS_{auto}(i,j) - HS_{label}(i,j)\right|\right) \tag{18}$$

Where MD is the mean difference between the automatically extracted heel-strike labels HS_{auto} and the ground-truth labels HS_{label}, NS is the total number of sequences (NS = 2163 for the indoor dataset), $H(i)$ is the number of heel-strikes in sequence i and j is the heel-strike index. MD is shown in Figure 23 for a range of filter polynomial orders. The heel-strike labels were generated for the indoor dataset semi-automatically from chromakey-extracted silhouettes and checked manually [Shutler 02]. Each heel-strike was marked with the nearest integer frame number, so a minimum error rate of around ±0.5 frames can be expected. The automatically extracted period differs from the ground-truth labels by a similar level, indicating comparable performance. The indoor dataset was designed to yield clean data, and here there is little benefit to performing any kind of filtering.

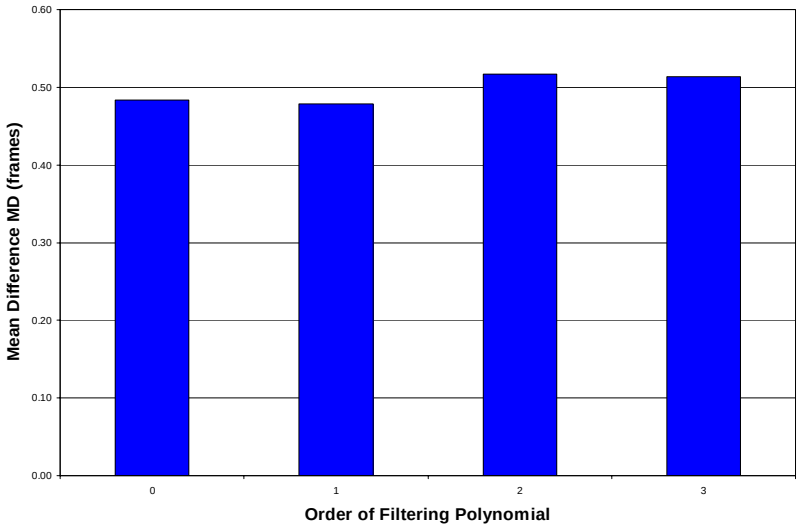
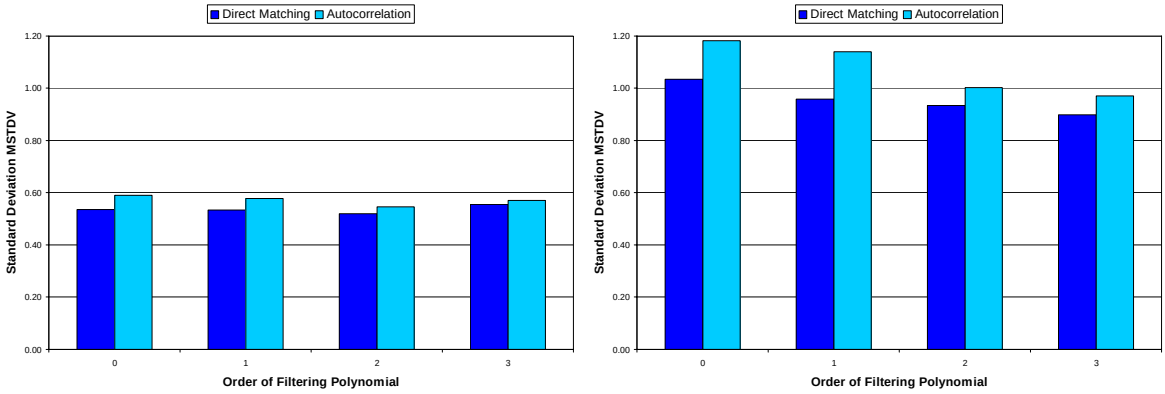


Figure 23: Comparison of automatic extraction of heel-strike frames with manually labelled frames

The outdoor dataset provides the real challenge for periodicity extraction, and there is no ground truth data available for this dataset. Instead, statistical testing is employed to indirectly estimate periodicity extraction performance. Consistency of extraction is a useful measure indicative of reliable performance. We would expect to measure approximately the same gait period for each sequence of the same subject, so the standard deviation of period within each subject should be low. This quantity is measured for all subjects in the database using the following equation:

$$MSTDV = \frac{1}{TS}\sum_{i} stdv(i) \tag{19}$$

Where $MSTDV$ is the mean within-subject standard deviation in extracted gait period, TS is the total number of subjects (TS = 115), and $stdv(i)$ is the standard deviation in period for subject i. Figure 24 shows this measurement for the indoor and outdoor datasets, comparing the performance of direct matching with autocorrelation-based periodicity extraction. As we would expect, periodicity extraction is far more consistent for the indoor dataset. These results also demonstrate that direct matching has a slight advantage over autocorrelation. The most likely reason for this is that it is already known that the shape of the sequence is approximately sinusoidal, and direct matching takes advantage of this fact. Autocorrelation does not assume any particular form of periodicity a priori, and so cannot achieve the same level of accuracy in this situation.



(a) Indoor dataset (b) Outdoor dataset

Figure 24: Average per-subject standard deviation in gait period

The results for the outdoor dataset show a general improvement in consistency as the order of the filtering polynomial is increased. For the indoor dataset, periodicity is extracted with greatest consistency at an order of 2. As the primary focus is performance under adverse conditions, an order of 3 is employed for periodicity extraction throughout the rest of this thesis.

3.2.3. Mean Shape

Mean shape can be reliably determined from noisy video data by temporally accumulating edges, according to the target object's estimated position at each frame. This has the effect of collating all the subject's edge data present in the sequence, so that data missing in some frames due to occlusion can be filled in using data present in other frames. Any error in estimated position has the effect of blurring the accumulation, so it is desirable to use a more sophisticated motion model than was employed in the initial accumulation (Section 3.2.1). Equations 7 and 8 (Section 2.3.1) are used to model the subject's COM motion:

$$A_v(i,j) = \sum_{t=0}^{T-1} I_{edge}\left(i + v\left(\frac{T}{2} - t\right) + a\left(\frac{T^2}{4} - t^2\right), j - A_y \sin 2\left(\omega t + \phi + \phi_y\right), t\right) \quad (20)$$

Where A_v is the accumulation for velocity v and acceleration a (pixels per frame), I_{edge} is the edge strength image at time index t, i and j are coordinate indices and T is the number of frames in the sequence. A_y is the amplitude of vertical oscillation, ω is the gait frequency, ϕ is the gait frequency and ϕ_y is a fixed phase offset (see Section 2.3.1). The gait frequency and phase are determined using the method outlined in the previous section,

and A_y is initially set to a fixed proportion of the subject's height. Figure 25 shows the improvement in accumulation quality gained by this process, evidenced by a reduction in blurring of the edges of the subject.

(a) Initial accumulation (b) Improved accumulation (c) Geometric shape model

Figure 25: Mean shape from temporal accumulation

Mean shape can be estimated robustly from this accumulation, using geometric components to represent segments of the body. The head and torso are modelled with ellipses, each leg by two pairs of lines and the foot by a rectangle (Section 2.2.1). Mean anatomical proportions [Winter 90] define the initial state of the model. The head and torso models are refined by adjusting the ellipse parameters so as to maximise model correlation with the accumulation:

$$M_e^* = \max_{x,y,W,H}\left(corr\left(ellipse(x,y,W,H), A_v\right)\right) \tag{21}$$

Where M_e^* is the maximal correlation for an ellipse defined by centre (x, y), width W and height H fitted to accumulation A_v. Although there are 4 free parameters to be determined for each ellipse, the initialisation provided by the previous stage is sufficient to allow small search bounds, keeping computational demands low.

Leg shape is lost during the accumulation process due to articulation of the leg joints. It is possible to generate a similar global accumulation of leg shape, but this can only be done with accurate knowledge of leg dynamics. It is more efficient to retrieve mean leg shape from a number of local estimates than to attempt resolution of leg dynamics without accurate knowledge of leg shape. An improved estimate is necessary because mean anatomical proportions are not appropriate for certain types of clothing (baggy trousers, shorts or skirts for example).

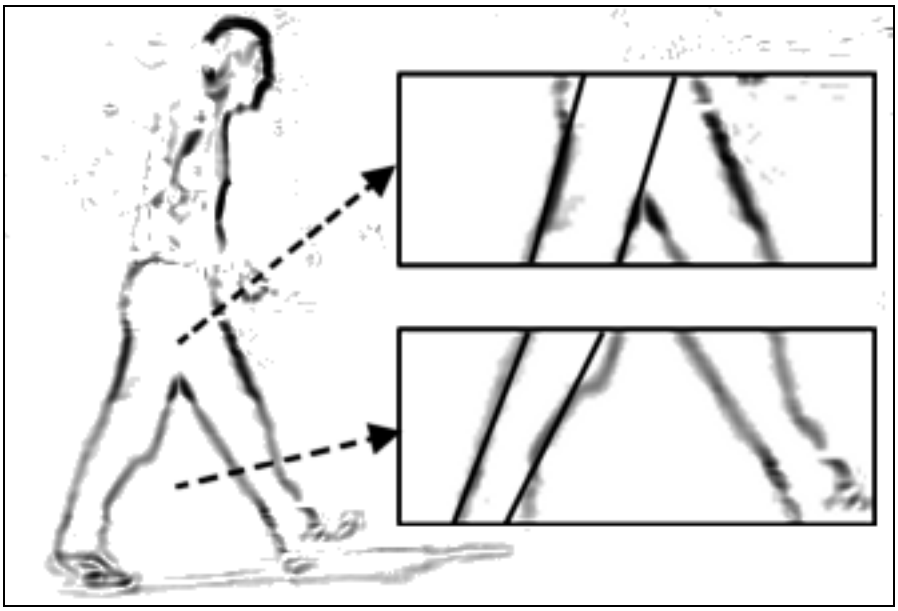


Figure 26: Locating pairs of lines using the Hough transform

Mean leg shape is obtained by computing a line Hough transform [Illingworth 88] for each frame within the upper and lower leg regions (above and below knee level, see Figure 26). Each peak in the Hough space indicates the presence of a linear structure at the corresponding position and angle, and each leg will manifest a pair of peaks. Using the expected rotation and width of the leg to constrain the Hough space, the pair of peaks with highest combined magnitude yields a local estimate of leg width. A final estimate for leg shape is determined by computing a weighted mean of the leg width estimates computed in each frame:

$$W_m = \sum_{t=0}^{T-1} p_t W_t \Big/ \sum_{t=0}^{T-1} p_t \tag{22}$$

Where W_m is the mean width of the leg (at the hip, knee or ankle) over T frames, and W_t and p_t are the estimated leg width and the peak Hough space magnitude respectively at time t. The weighting allows more reliable estimates (those with high peak magnitude) to dominate the final estimate, reducing the effect of poor local estimates. These four estimates of leg width (at the hip, above and below the knee and at the ankle) provide a simple initial estimate of leg shape. The two estimates of leg width at the knee allow for discontinuity in leg shape, due to clothing such as shorts or skirts.

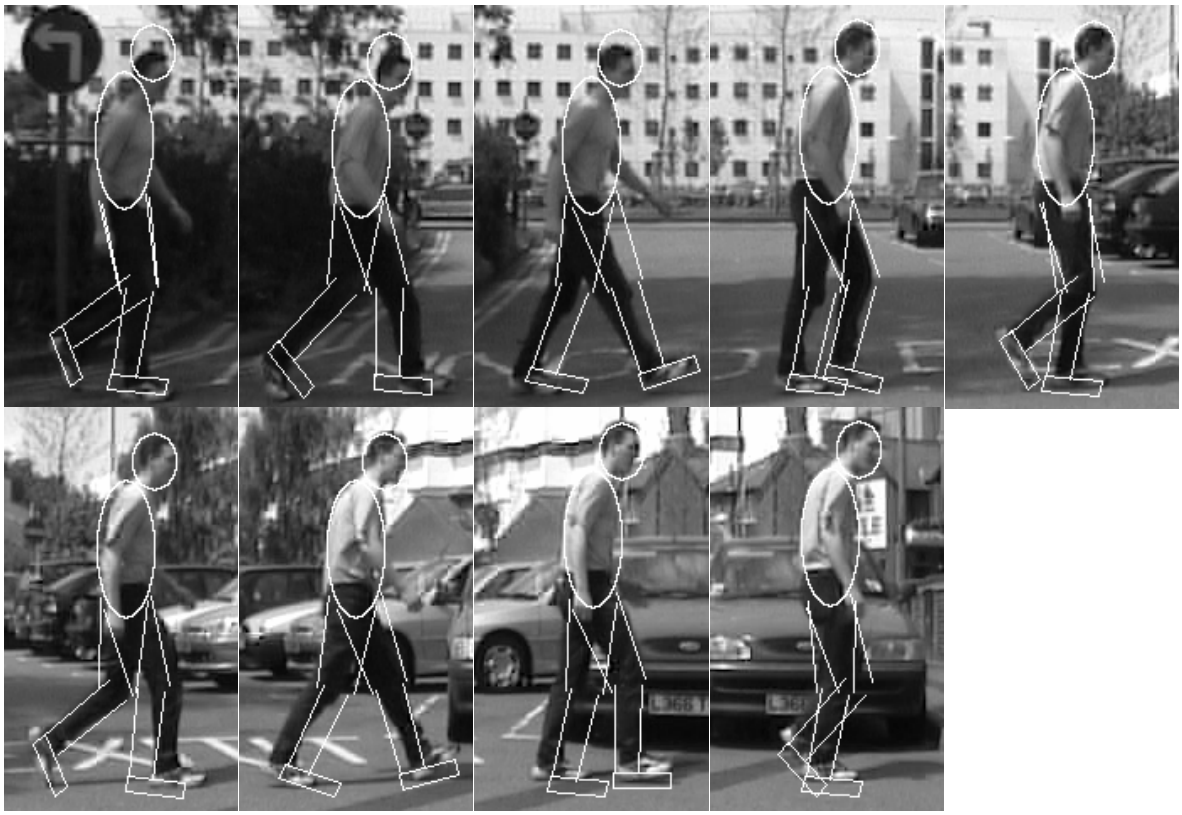


Figure 27: Posterior model initialisation, sequence '008e013s00R'

Figure 27 summarises the initial posterior model extracted for an example outdoor gait sequence, sampled every 10 frames. There are some obvious discrepancies in leg position, bearing in mind there are only two free parameters in the initial articulation model (gait frequency and phase). As the slope of the ground plane is not yet taken into account, there is some degree of error in vertical position, and height may be slightly overestimated. However, the errors in model configuration are small compared to the possible range of the model, and provide a good basis for adaptation to improve the model parameters.

3.3. Conclusions on Posterior Model Initialisation

The aim of this stage is essentially to extract the minimum amount of information from a gait sequence required to apply the models detailed in Chapter 2. This prior knowledge can then guide the remainder of the extraction process. To meet these aims, robust methods are required; if this stage fails it is unlikely that a good gait description will be obtained. This is only an initialisation of the posterior model, so high accuracy is not essential, and global

methods are ideally suited to this sort of problem. Since a large amount of data and prior knowledge is applied to the extraction process, global algorithms are highly robust. Equally, the quantity of data processed restricts the accuracy that can be obtained, as model complexity must be limited to reduce computational requirements.

The method presented in Section 3.2.1 is a development of the Velocity Hough Transform [Nash 97], separating the temporal accumulation stage from shape extraction. This approach retains the algorithm's well-known property of robust operation, but greatly reduces its computational requirements.

Section 3.2.2 presents a method of periodicity detection for walking people, exploiting prior knowledge of where periodic motion is expected and the form it is likely to take. The method of normalisation presented is suitable for short gait sequences, where non-periodic variation can be modelled with a simple polynomial sum. This is quite sufficient for gait recognition applications, where only short sequences (a few seconds) of gait are required, and longer sequences could be analysed piece-wise if necessary.

The third section of this chapter deals with approximation of the mean body shape of the subject. The upper body (excluding the arms) is largely immobile during gait and so upper body shape can be robustly estimated from a global temporal accumulation. The lower body is in constant motion and so an estimate of mean leg shape is generated from many local estimates.

This initialisation provides a base model that can be adapted to better fit the image data, through the incorporation of additional gait and body shape parameters describing the individual characteristics of the subject. Approaches to model adaptation are discussed in Chapters 4 and 5.

Chapter 4. Model Adaptation Strategies

4.1. Introduction

This chapter provides an overview of the model adaptation problem and presents two basic approaches, distinguished by the choice of a local or global modelling strategy. The new hybrid approach proposed by this thesis is described in Chapter 5.

The complexity of computer vision tasks is such that it is often impractical or even impossible to extract all the required information in one step. This difficulty arises when a high degree of complexity is required in the posterior model to adequately represent the target object, resulting in a very large search space. An efficient strategy for evaluating a large search space is to break the process down into multiple stages, often referred to as *hierarchical* or *pyramidal* processing:

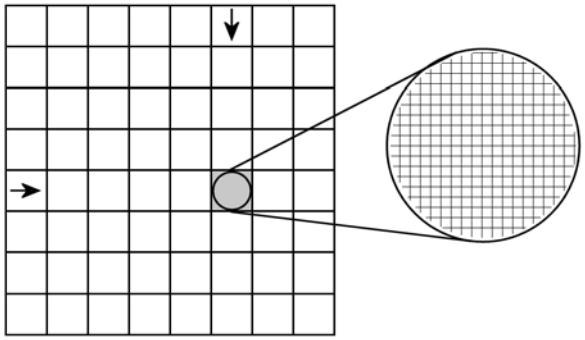


Figure 28: Hierarchical search process

For the purposes of this thesis, the process of posterior model extraction is split into two stages: *initialisation* and *adaptation*. The initialisation stage (Chapter 3) covers the whole extent of the search space, using a widely spaced sampling grid (i.e. a simple model) to quickly locate the general area of the solution. In the adaptation stage, this area is sampled more thoroughly to achieve the required level of accuracy.

The two basic approaches to model adaptation differ according to how the posterior model is formulated. As the initialisation stage employs a global model, the most straightforward approach is to retain this formulation, performing model adaptation in the global domain (Section 4.2.1). The second approach is to break this global model down into many local models and adapt each individually, taking advantage of the consequent

reduction in computational demands to employ a more complex shape representation (Section 4.2.2). Observing that neither approach is without fault, a hybrid model adaptation strategy is described in Chapter 5 that attempts to combine the best features of each strategy, obtaining a more accurate final gait description.

4.2. Model Adaptation

4.2.1. Global Model Adaptation

The initialisation process employed within Chapter 3 relies on a global modelling strategy to generate a reliable approximation of the subject's shape and motion. The simplest approach to adaptation is to modify the global model itself, adjusting model parameters to better fit the observed image data. This early approach was published in [Wagg 04a].

Performing adaptation on a global basis has the advantage of exploiting the spatio-temporal correlations explicit in global models, gathering structural evidence from the whole sequence to determine local model states. This means that the correct configuration can be ascertained even when the subject is completely occluded, provided that there is sufficient information in neighbouring frames. The problem associated with resolving temporal correlations is that it greatly increases the computational requirements of the adaptation process, consequently restricting the degree of accuracy that can be attained.

The mean body shape model computed in Section 3.2.3 is sufficient to represent major shape characteristics at this stage, but little information is known about the subject's leg motion. The initial model assumes average gait (Section 2.3.2), modified only by the subject's gait frequency and phase. The model adaptation process operates on the mean gait model, adjusting rotation parameters until the model correctly predicts the pose of the legs at each frame in the gait sequence.

Adaptation is performed via a simple iterative gradient descent procedure. Of the parameters contained within the gait model, only the hip, knee and ankle rotation models $[\Theta_h, \Theta_k, \Theta_a]$ are considered for adaptation to restrict computational demands. As the amplitude of ankle rotation is relatively small the hip and knee rotation models are adapted first, assuming mean ankle motion. Once optimal rotation patterns have been determined for the hip and knee, the ankle model is subsequently adapted using the same process. The

following pseudo-code describes the gradient descent process for hip and knee joint rotation model adaptation:

For i = 0 to maximum number of iterations - 1
. $\Theta_h^i = \Theta_h^{i-1}$
. $\Theta_k^i = \Theta_k^{i-1}$
. For n = 0 to number of samples - 1
. . *maxS* = 0
. . For δ_h = -1 to 1
. . . For δ_k = -1 to 1
. . . . Compute $S(\delta_h, \delta_k)$
. . . . If $S(\delta_h, \delta_k)$ is greater than *maxS*
. . . . . $max\delta_h = \delta_h$
. . . . . $max\delta_k = \delta_k$
. . . . . $maxS = S(\delta_h, \delta_k)$
. . $\Theta_h^i = \Theta_h^i + max\delta_h \cdot G_n$
. . $\Theta_k^i = \Theta_k^i + max\delta_k \cdot G_n$
. If Θ_h^i is equal to Θ_h^{i-1} and Θ_k^i is equal to Θ_k^{i-1}
. . Stop

The number of samples per gait cycle is 15, chosen to limit computational requirements while retaining useful gait information (see Section 2.3.2). The model update function employed in the above algorithm is a 1D Gaussian function:

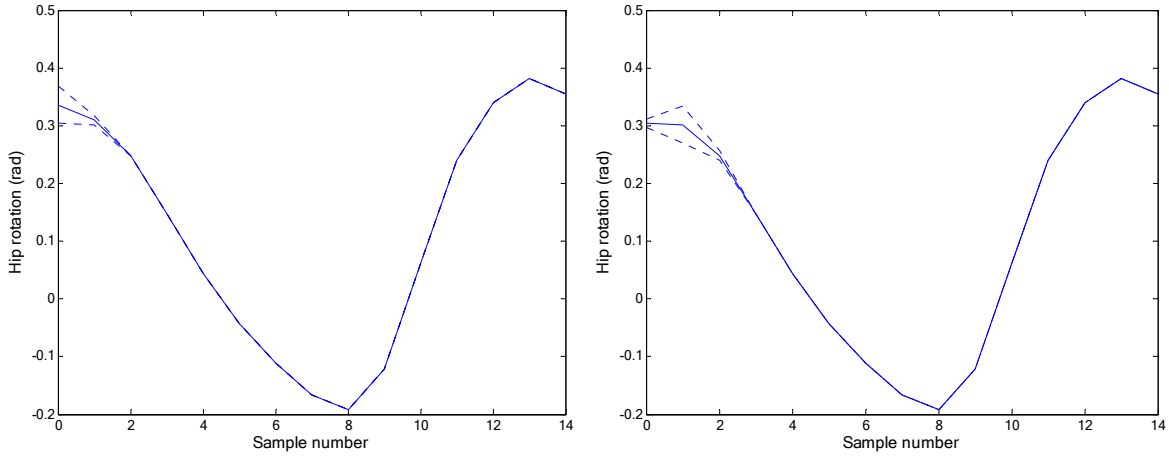
$$G_n(m) = A_G e^{-\frac{(m-n)^2}{2\sigma^2}} \tag{23}$$

Where m is the Gaussian sample index, n is the index of the current model sample, A_G is the amplitude of the Gaussian (controlling the rate of descent) and σ is the variance (controlling temporal continuity). If the variance is small, each model sample is updated almost independently, whereas a large variance means that each update affects a large number of neighbouring model samples. This ensures that temporal continuity can be enforced, as any change in sample magnitude will proportionally affect neighbouring samples. For the results presented in Chapter 6, $A_G = \pi/128$ (approximately 1.4 degrees) and $\sigma = 0.6$.

The score S is determined by computing the correlation of the leg model with the edge image sequence:

$$S(\delta_h, \delta_k) = \sum_{t=0}^{T-1} corr\left(GSM(\Theta_h^i, \Theta_k^i, \Theta_a, t), I_{edge}(t)\right) \tag{24}$$

Where $S(\delta_h, \delta_k)$ denotes the score for the model adjustments (δ_h, δ_k) and *GSM* is the geometric shape model at time t given by the adjusted hip and knee rotation models (Θ_h^i, Θ_k^i) at the i^{th} iteration and the mean ankle rotation model Θ_a. Figure 29 demonstrates the model update process for the hip rotation model.



(a) $n = 0$ (a) $n = 1$

Figure 29: Hip joint rotation model adaptation

The ankle rotation model adaptation follows the same process as the hip and knee, except that only a single quantity requires optimisation:

$$S(\delta_a) = \sum_{t=0}^{T-1} corr\left(GSM(\Theta_h, \Theta_k, \Theta_a^i, t), I_{edge}(t)\right) \tag{25}$$

The gradient descent process continues iteration until the joint rotation models reach a constant state, or until a maximum number of iterations are reached (though in practice this process typically converges within 10 iterations).

The main limitation of this approach is that no local deviation is allowed from the global model, which leads to an approximation error when the subject varies their gait pattern, or when their body shape changes (due to clothing for example). As the global model must be relatively simple to restrict computational requirements, this limitation can lead to significant inaccuracies in the gait description. Section 6.2 provides some examples of the gait models extracted using this method of model adaptation.

4.2.2. Local Model Adaptation

An alternative approach to adaptation is to break the initial global model down into many local models and adapt each individually. Although this means that temporal correlations cannot be fully exploited, the local models should begin in a good initial configuration due to the global initialisation. The main advantage of this approach is that fewer model states need to be evaluated; less image data is processed for each model state evaluation, and so a much more complex shape model can be employed. This later approach was published in [Wagg 04b].

The initial global model is used as the starting point for local deformable contour models (see Section 2.2.2). To adapt the contour shape to fit the image data, a relatively simple gradient descent formulation based on the greedy snake [Williams 92] is employed, whereby contour adaptation is expressed as a process of energy minimisation. The snake energy incorporates internal constraints on local curvature and contour point spacing, and external constraints that are used to attract the snake to image features:

$$E^*_{snake}(C) = \min \int_0^1 \left(E_{int}(C) + E_{ext}(C)\right)\, ds \tag{26}$$

Where E^*_{snake} is the minimal snake energy, C is the snake contour (Section 2.2.2), E_{int} is the internal snake energy, and E_{ext} is the external energy. The internal energy for all contours is described by:

$$E_{int}(C) = \alpha E_{c1} + \beta E_{c2} + \gamma E_t \tag{27}$$

Where E_{c1} corresponds to normalised first-order continuity of the snake, E_{c2} corresponds to normalised second-order continuity of the snake and E_t corresponds to normalised first-order temporal continuity. E_{c1} has the effect of enforcing even contour point spacing, E_{c2} restricts contour curvature and E_t enforces continuity of motion. The weighting coefficients α, β and γ control the balance of these three energy contributions.

The upper body and leg contours differ greatly in shape and expected level of occlusion. In order to account for this difference, slightly different external energy terms are employed to optimise snake adaptation. For the upper body contours the external energy is described by:

$$E_{ext,\ body}(C) = \lambda I_{attr}(C) \tag{28}$$

Where $I_{attr}(C)$ is the image attraction term for the contour C and λ is a weighting term. The image attraction term is defined as $I_{attr} = 255 - I_{edge}$, so that the snake is attracted towards

image edges. For the leg contours an occlusion weighting term o is added, and an additional constraint E_{side} forcing the front and back leg contours to remain within an expected distance:

$$E_{ext,leg}(C) = o(C)\lambda I_{attr}(C) + \rho E_{side}(C) \tag{29}$$

Where $o(C)$ is the occlusion model prediction (Section 2.4) for the contour C and ρ controls the weighting of E_{side}, which is equal to the difference between the expected and measured distance between the front and back contours of the leg:

$$E_{side}(C) = \left| W_m(C) - \left| C_{front} - C_{back} \right| \right| \tag{30}$$

Where W_m defines the expected leg width at each contour point (computed from the geometric model parameters), and C_{front} and C_{back} are the contours defining the front and back of the leg. All snake control parameters were determined empirically; optimal values will vary depending on the specific application.

The occlusion model (Section 2.4) is computed a priori for this approach, assuming mean gait motion and leg shape. This model defines the expected level of occlusion at each contour point for each leg position during a gait cycle. The purpose of the occlusion model is to reduce the contribution of image features at points in the gait cycle where the legs occlude each other. At these points we would not expect to see reliable edge information, and so the snake is constrained to rely more on initialisation, and on internal and temporal constraints (the temporal constraint effectively allows some degree of interpolation over occluded frames). The snake contours are driven to a minimal energy state by an iterative process of gradient descent. Note that due to the inclusion of a temporal constraint in a local search process, the order of iteration in performing the minimisation is important. A single iteration of gradient descent is performed for each frame in the gait sequence, before repeating the process for subsequent iterations.

This local adaptation process can quickly generate an accurate model of the person's shape and leg pose. However, if the global initialisation is too far from the correct shape and pose this process will fail, as the adaptation process only operates within a local scope. This approach proved unreliable for extraction of foot shape and ankle dynamics, and consequently the global adaptation approach (Section 4.2.1) is employed for the feet. Section 6.2 provides some examples of the gait models extracted using this method of model adaptation.

4.3. Conclusions on Model Adaptation Strategies

The model adaptation process is tasked with extending and refining the initial model that approximates the subject. The primary objective is to obtain a gait description that can be used to recognise the individual. A secondary objective is to achieve sufficient accuracy in capturing the shape and motion of the subject that this representation is useful for other purposes, such as automated rotoscoping or motion capture for film and TV production.

The first approach to model adaptation operates on a global level. At each iteration of the adaptation process, appropriate adjustments to the global gait model are determined by reference to edge data across the whole sequence. The quantity of data referenced at each model evaluation necessarily limits the number of adjustments that can be made, meaning that the model must be of relatively low complexity to limit computational requirements. This increases the difficulty of the recognition process (there is less to distinguish different subjects), and this approach is effectively useless for applications demanding highly accurate extraction.

The second approach considers a much smaller quantity of data to make each model adjustment, which means that a far more complex shape model can be employed. The problems associated with poor data quality are partially surmounted through the global initialisation, and temporal correlation with neighbouring frames. However, in instances where the initial global model is highly inaccurate this approach will fail, leading to poor performance in a small number of cases.

Neither approach is sufficient alone, the first because it places too great a limit on model complexity, and the second because of its sensitivity to poor initialisation. As the strengths of each strategy are complementary, improved performance can be gained by combining the two approaches into a single process (Chapter 5).

Chapter 5. Hybrid Model Adaptation

5.1. Introduction

This chapter describes the new approach proposed by this thesis to the model adaptation problem. Both local and global modelling strategies have distinctive advantages. Global evidence gathering approaches are able to exploit temporal correlations in video imagery to enable robust model extraction. The drawback is that such approaches are slow, due to the increased dimensionality of the search space, and must compensate by using relatively simple models. By contrast, local evidence gathering approaches are fast, due to the greatly reduced amount of data considered, and consequently are able to extract highly complex shapes. However, it is difficult to apply global constraints on structure and temporal consistency to local search algorithms. This leads to extraction errors in high levels of noise, or when the initial model configuration is inadequate.

This thesis proposes an integrated approach of global and local evidence gathering. Global shape and motion models are iteratively updated according to locally maximal model configurations. These global models are then used to re-initialise local model configurations, repeating the process until a convergent point is reached in the global model or a maximum number of iterations are reached. This approach is similar to the generalised EM algorithm [Dempster 77, Wu 83], differing in that it is formulated directly in terms of maximising model to image correlation.

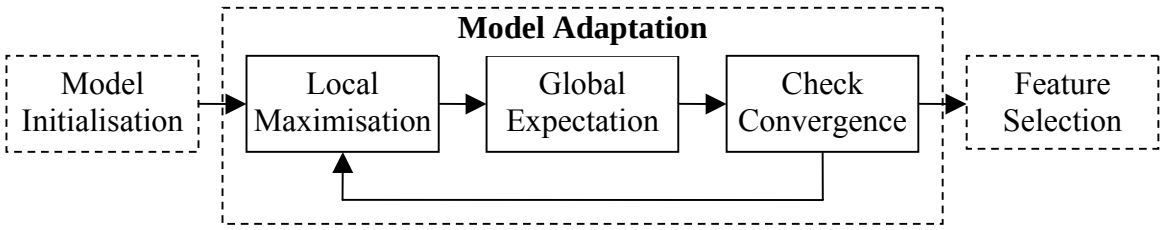


Figure 30: Overview of Hybrid Model Adaptation

The main difficulty associated with this approach is controlling the balance between local and global information. When local information is reliable, it is desirable to use a local adaptation technique, as it can extract a more accurate model than a global technique for

the same computational requirements. If local information is missing or corrupted, a local adaptation technique would yield a poor result, and so a global technique is preferable in this case. This suggests that the perceived reliability of image data in a particular area should be the controlling factor in balancing local and global adaptation. These issues are discussed further in the following sections.

5.2. Local Maximisation

The global initialisation computed in Chapter 3 predicts, to a limited degree of accuracy, the appropriate model configuration for each frame. This geometric model is first converted to a set of contour models (see Section 2.2.2), so that shape deformation may be computed. The initial model can then be greatly improved by computing a locally maximal model configuration at each frame, adjusting model parameters to better fit the observed image data. Figure 31 illustrates a single iteration of the local maximisation process for an example outdoor sequence.

(a) Global initialisation (b) Dynamic adjustment (c) Contour deformation

Figure 31: One iteration of local maximisation

The local maximisation process is not able to locate an optimal configuration on its own, as it is required to search only a small area, both to limit computational demands and to ensure that global constraints are not violated. This means that after a single iteration of local maximisation some large errors are visible (Figure 31c). The global expectation process (described in Section 5.3) is then employed to correct these errors, generating a new initialisation for the next iteration of local maximisation.

5.2.1. Search Methods

This section is concerned with how the model space should be divided up and evaluated in order to efficiently find a locally maximal configuration. It is impractical to attempt optimisation of all the model parameters simultaneously, as the size of the model space is geometrically proportional to the number of free parameters and thus very large. It is preferable to define a search space, a subset of the model space thought likely to contain a good solution (see Section 2.1). To illustrate this concept, consider a model of the hip, knee and ankle joints. If 10 possible states are allowed for each joint, the model space comprises 1000 states (10×10×10). If the search is ordered hierarchically, such that an optimal value is determined first for the hip, then the knee and last the ankle, a search space of only 30 states must be evaluated, avoiding evaluation of the full model space. Although this means that an optimal solution cannot be guaranteed (not all model states are evaluated), usually an acceptable solution can be quickly determined.

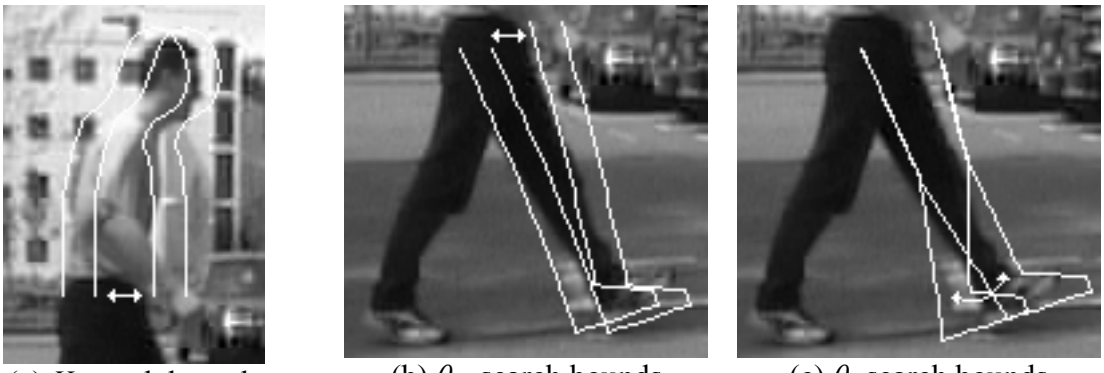
Following this strategy, a hierarchical decomposition of the model space is applied, ordering dynamic model parameters (those describing position and joint rotations) from the centre of mass downwards [X, Y, θ_{px}, θ_{py}, θ_h, θ_k, θ_a]. Static parameters (those describing shape) are not considered at this stage; shape adaptation is covered in Section 5.2.3. Each parameter is sampled uniformly within the region enclosed by the search bounds, evaluating each sampled configuration to determine a locally maximal parameter value:

$$X^* = \arg\max_i \left(corr\left(C^i_{test}, I\right)\right), \quad C^i_{test} = contour\left(X + i, Y, \theta, MS\right), \quad -S_i \le i \le S_i \qquad (31)$$

Where X^* is the optimal parameter value within the locality of the initial value X, C^i_{test} is the test contour defined by the modified dynamic parameters and the mean shape model, i is the search index and S_i is the search bound. The model correlation *corr* is a measure of how well the current model configuration matches the image data I (which includes edge data I_{edge} and greyscale data I_{grey}, see Section 5.2.2 for further details). The *contour* operation is described in Section 2.2.2. Locally maximal values can be computed for the remaining parameters in the same manner, replacing X and X^* in Equation 31.

Figure 32 illustrates some typical search bounds for an example gait sequence (Figure 31a illustrates the initial state of the model). Determination of appropriate search bounds can be problematic. It is desirable to keep search bounds as small as possible to reduce computational requirements, and to enforce global constraints. However, small search bounds will increase susceptibility to noise or poor initialisation, as fewer alternate

configurations will be evaluated. For this work, appropriate search bounds for each parameter were determined manually, but it may be possible to determine search bounds semi-automatically based on the estimated reliability of the local image data.



(a) X search bounds (b) θ_{py} search bounds (c) θ_k search bounds

Figure 32: Search bounds for example dynamic parameters

Although it is possible to apply the same process to all of the model parameters, performance can be improved by maximising the hip and knee joint parameters jointly:

$$\left(\theta_h^*, \theta_k^*\right) = \arg\max_{i,j}\left(corr\left(C_{test}^{i,j}, I\right)\right), \qquad -S_i \le i \le S_i, \qquad -S_j \le j \le S_j, \tag{32}$$

$$C_{test}^{i,j} = contour\left(X, Y, \theta_{px}, \theta_{py}, \theta_h + i, \theta_k + j, \theta_a, MS\right)$$

Where S_i and S_j are the search bounds for hip and knee joint rotation respectively. Although computational requirements are increased, a combined optimisation process is advantageous in order to overcome the unreliability of edge data at hip level. This unreliability stems from self-occlusion of the legs as they cross over; as both legs are likely to be the same colour and texture, there is often little to distinguish the left and right legs at hip level. Lower down, the legs are generally further separated and so this is less of a problem. Occlusion by the hands can also contribute to unreliable data at hip level. By optimising the hip and knee joints simultaneously, edge data from the upper and lower leg can be considered, decreasing the likelihood of a poor match. Once a locally maximal model configuration has been determined for time t, the process is repeated for time $t + 1$, until all the frames in the gait sequence have been locally maximised.

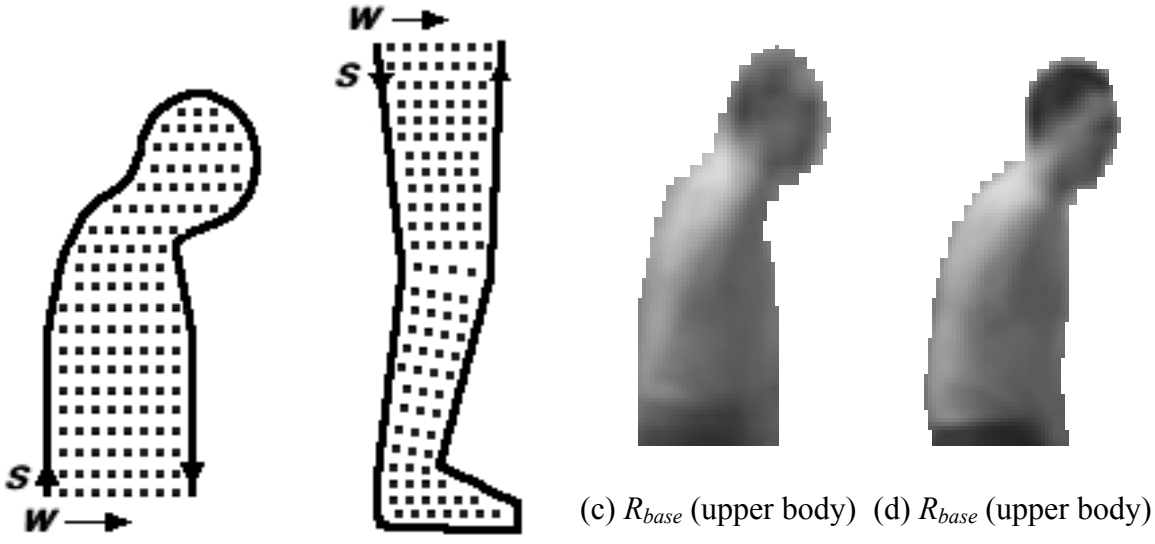
5.2.2. Model Evaluation

A good model configuration is defined as one that yields a high correlation between the model and the subject's image. Useful measures for computing model-image correlation include *edge correspondence* and *region correspondence*. Edge correspondence is a measure of how closely model edges coincide with image edges. Region correspondence is a measure of similarity between the image region enclosed by the model and the region corresponding to the image of the subject. As the subject image is unknown, it is estimated by computing a global mean of the model regions (see Equation 34). Edge correspondence is computed by convolving the edge image with the model contour:

$$corr_{edge}(i) = \sum_{s} I_{edge} * C^{i}_{test} \tag{33}$$

Where $corr_{edge}$ is the edge correspondence and C^{i}_{test} is the model contour to be evaluated.

A high edge correspondence indicates that the model is closely aligned with image edges; however, it does not guarantee that the model matches the correct edges. If the initial model configuration is poor, or the subject is occluded, the match may be coincidental. For this reason, region correspondence is also required. Region correspondence is not as precise as edge correspondence, as the measure is computed over a much larger number of pixels, but it is consequently more robust and can help to disambiguate subject edges from other edges. In order to measure region correspondence, a sampling grid G_{base} is defined from the contour model C by sampling uniformly between opposite contour points (Figure 33a-b).



(a) G_{base} (Upper body) (b) G_{base} (Leg) (c) R_{base} (upper body) at iteration 0 (d) R_{base} (upper body) at iteration 10

Figure 33: Base region model construction for sequence '008e013s00R'

The base region model is determined by computing the mean grey-level intensity of each point in the region over the whole sequence:

$$R_{base}(s,r) = \frac{1}{T}\sum_{t} I_{grey} * G_{base} \tag{34}$$

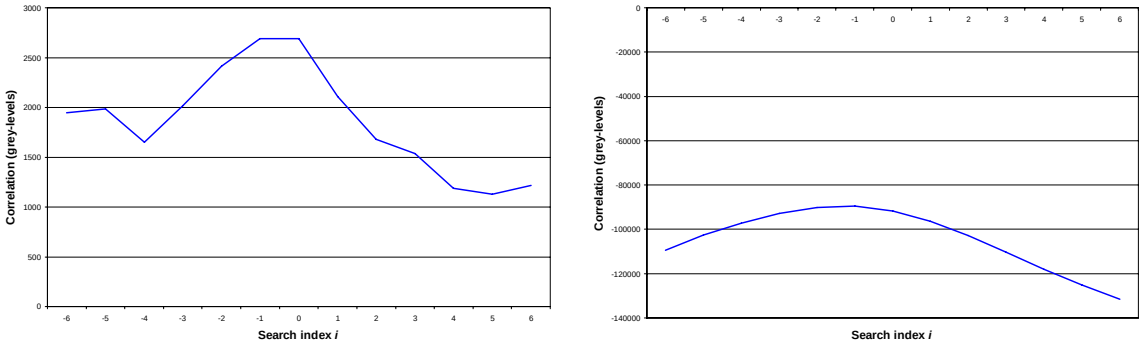
Where R_{base} is the base model region, s and r are contour and width indices respectively and I_{grey} is the greyscale image data. Figure 33c illustrates the initial region model obtained for the upper body in an example outdoor sequence. The initial estimate is quite indistinct, due to the blurring effect of errors in localisation of the subject's COM. The region model is updated at the beginning of each iteration of the adaptation algorithm, improving as the accuracy of model localisation improves. Figure 33d shows the region model attained after 10 iterations of hybrid model adaptation. The boundary between the subject's shirt and trousers, and his hairline, is clearly visible in this model. The mean shape of the model has also improved, due to increasing accuracy in local shape deformation (see Section 5.2.3). These improvements aid computation of region correspondence, increasing the accuracy in model localisation that may be achieved. Using this model, region correspondence is defined as:

$$corr_{reg}(i) = -\sum_{s,w}\left(\left(I_{grey} * G_{test}\right) - R_{base}\right)^2 \tag{35}$$

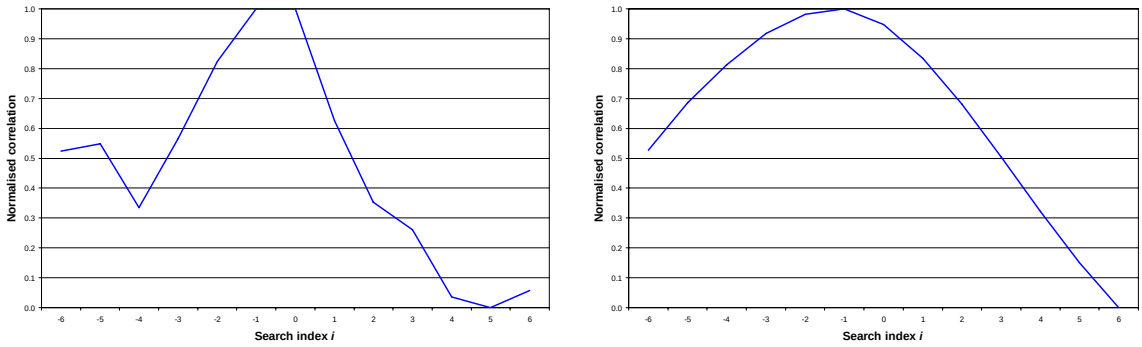
Where G_{test} is the sampling grid generated for the test model contour C_{test}. The raw outputs of the two correspondence measures differ considerably in magnitude and range, and must be normalised before they can be combined effectively. Normalised edge correspondence is given by:

$$ncorr_{edge}(i) = \frac{corr_{edge}(i) - \min_i\left(corr_{edge}\right)}{\max_i\left(corr_{edge}\right) - \min_i\left(corr_{edge}\right)} \tag{36}$$

Normalised region correspondence is computed in the same way, replacing $corr_{edge}$ with $corr_{reg}$. For the purposes of this thesis, each measure is assumed to have equal importance, and an equal weighting is applied to edge and region correspondence measures. It is likely that performance could be improved by altering the balance between these two measures, but it is difficult to specify an appropriate balance. In common with search bound determination, it may be beneficial to use the perceived reliability of the image data to control the balance between these correspondence measures. Figure 34 plots edge and region correspondence for the upper body when the COM position X is varied in frame 33 of sequence '008e013s00R' (see also Figure 32a, which displays the relevant model and image data).



(a) $corr_{edge}$ (b) $corr_{reg}$



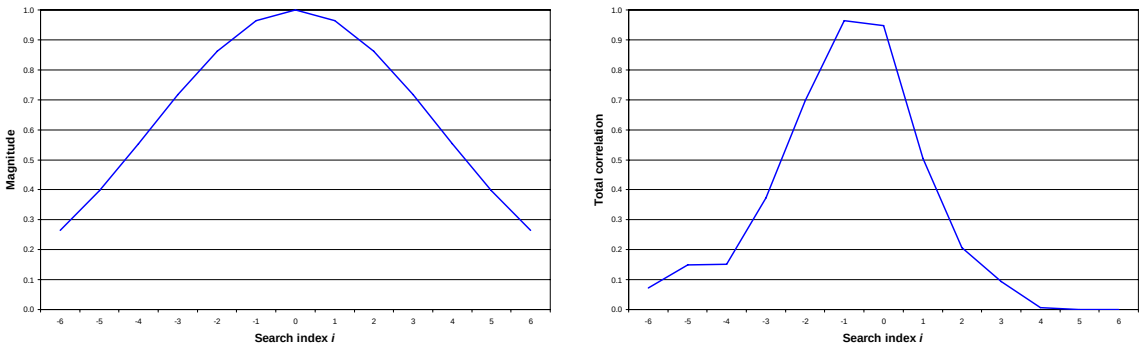
(c) $ncorr_{edge}$ (d) $ncorr_{reg}$

Figure 34: Normalisation of model correspondence measures

An overall correspondence is computed according to the following equation:

$$corr_{tot}(i) = ncorr_{edge}(i) \times ncorr_{reg}(i) \times e^{-\frac{i^2}{4\sigma^2}}, \quad \sigma = \frac{2S_i + 1}{5} \tag{37}$$

Where S_i is the search bound for the current parameter i and σ is the variance of the Gaussian weighting that is applied to bias the evaluation towards the initial model configuration, ensuring that the model will only be altered if there is a significant amount of evidence to support an alternate configuration. This reduces the sensitivity of the local maximisation process to noise, preventing unnecessary changes in model configuration.



(a) Central Gaussian weighting (b) Total correlation $corr_{tot}$

Figure 35: Generating an overall model correspondence measure

The variance of the Gaussian determines the strength of the bias. If the variance is set too high there will be no constraining effect, and if it is too low then no adaptation will be allowed. For the Southampton Gait Database, the equation for the variance σ was defined so that solutions at either extreme have a weighting of 0.25 compared to the central (initial) configuration (Equation 37 and Figure 35). For this example, only a small change in model configuration is necessary and so the Gaussian weighting has relatively little effect. The weighting plays a more important part in the determination of leg parameters. The greater degree of freedom in these parameters means that it is easier to select a poor model configuration based on small fluctuations in correspondence, and the Gaussian weighting reduces the likelihood of this occurrence.

5.2.3. Contour Deformation

Once the dynamic model parameters have been optimised, it is possible to compute the contour deformation best matching the subject's image. Recall that deformation is defined in the direction normal to the base contour (Section 2.2.2):

$$C_{def}(s,t) = C_{base}(s,t) + d(s,t) \cdot normal(C_{base}(s,t)) \tag{4}$$

A locally optimal shape deformation is computed by treating d as a set of three open-ended 1D snakes (the upper body and two legs), adding a global model constraint and an occlusion model constraint to the traditional formulation. Snake energy is defined by:

$$E^*_{snake}(d(s,t)) = \min \int_s (\lambda E_{\text{int}}(d(s,t)) + (1-\lambda) E_{ext}(d(s,t)))\ ds \tag{38}$$

Where E^*_{snake} is the minimal snake energy and λ controls the relative importance of internal energy E_{int} and external energy E_{ext}. Internal energy applies spatio-temporal constraints on snake shape, and is defined by:

$$E_{int}(d(s,t)) = \alpha E_{curv}(C_{def}(s,t)) + (1-\alpha)|d(s,t)| \tag{39}$$

Where E_{curv} corresponds to normalised second-order continuity of the contour C_{def} (restricting contour curvature). The model constraint $|d|$ has the effect of minimising deformation magnitude, thus enforcing global model constraints on shape and motion through the base contour. The weighting coefficient α controls the balance between the two energy contributions. External energy attracts the snake towards nearby edges, and is defined by:

$$E_{ext}(d(s,t)) = o(C_{base}(s,t))I_{attr}(C_{def}(s,t)) \quad (40)$$

Where the image attraction term is defined as $I_{attr} = 255 - I_{edge}$. The occlusion model o controls the relative contribution of image features and internal constraints (see Section 2.4), given the current model configuration and mean contour shape. For a fully visible point $o = 1$ and for a fully occluded point $o = 0$. This ensures that when parts of the leg are not visible due to self-occlusion, the snake is driven by internal forces rather than image forces. Using an occlusion model in this way is an effective way of applying global constraints to the typically local snake algorithm, as there is little computational burden added to the snake algorithm. Global constraints are applied in a separate algorithm that employs a greatly simplified search process to limit computational requirements.

As arm motion is not yet considered, the upper body is assumed to be fully visible. Some errors can be expected in deformation at waist level, as occlusion by the hands is not taken into account in the occlusion model. Figure 36 shows the model contours extracted after one iteration of local maximisation for an example outdoor sequence:

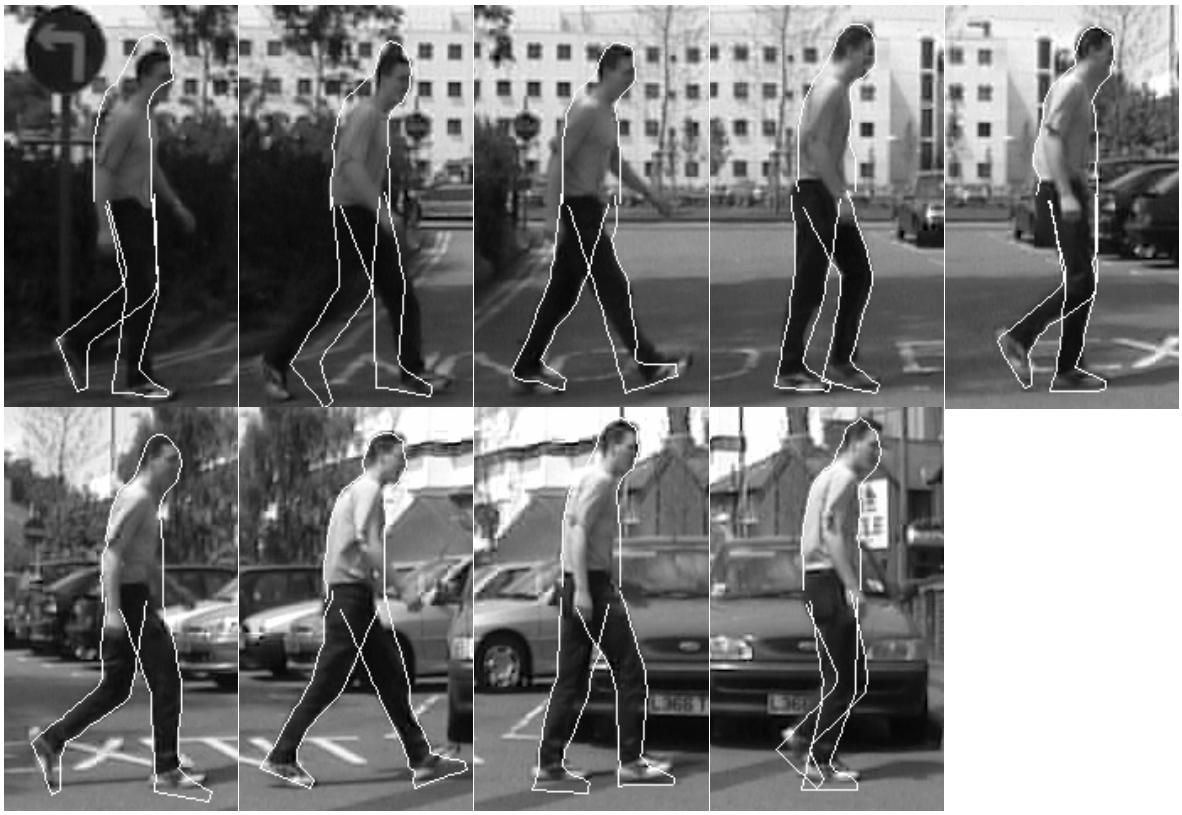


Figure 36: Local maximisation, sequence '008e013s00R'

Some errors in model configuration are clearly visible after one iteration of local maximisation. These errors occur when the initial model configuration is too far from the

true state of the subject (see Figure 27, Section 3.3), or where there is too much image data missing to support the local search process (for example when the subject is in deep shadow). The global expectation process is used to correct these errors, using global averaging and model fitting.

5.3. Global Expectation

The task for this section is to decide which local model configurations are reliable and which are not, and use the reliable data to generate a new, improved initialisation for the next iteration of the algorithm. Reliability is estimated by computing edge correspondence for each model configuration and normalising to the range [0, 1]:

$$w(t) = \frac{corr_{edge}(t) - \min_t\left(corr_{edge}\right)}{\max_t\left(corr_{edge}\right) - \min_t\left(corr_{edge}\right)}, \qquad corr_{edge}(t) = \sum_s I_{edge}(t) * C_{def}(t) \tag{41}$$

Where $w(t)$ is the reliability weight for the model configuration at time t. The weight vector w defines relative reliability, assigning the worst matching configuration a weight of zero. This ensures that reliable image data is used to best advantage and poor model configurations are rejected. A new initialisation for the global joint motion models Θ is generated by computing a weighted average of the local estimates θ:

$$\Theta'(n) = \frac{1}{\sum_k \mathrm{w}\left(n + kP\right)} \sum_k w\left(n + kP\right)\theta\left(n + kP\right), \qquad 0 \le n < 15, \quad n + kP < T \tag{42}$$

Where Θ' is the new joint motion model, n is a model index, P is the gait period and k is a positive integer. The left and right leg joint motion models are merged for increased reliability (see Section 2.3.2). Thus, if one leg is heavily occluded, image data from the other leg can still be used to predict its motion. Similarly, if the subject is visible for more than one gait cycle (~1 second), data from reliable gait cycles can be used to compensate for missing data in occluded or noisy cycles.

A new global model for x-motion is generated by fitting a third-order polynomial (Equation 6, Section 2.3.1) to the position vector X. The polynomial parameters are computed by weighted least-squares linear regression [Fox 97], using the estimated reliability w to control the influence of the model configuration computed for each frame. A new y-motion model is generated in a similar fashion, fitting a first-order polynomial to

the position vector Y to determine the y-motion gradient (Equation 8, Section 2.3.1). The amplitude of y-motion is then determined by:

$$A_y^* = \arg\min_i \left(\sum_{t=0}^{T-1} w(t) \left(Y(t) - A_y^i \sin 2\left(\omega t + \phi + \phi_y\right) - v_y t - y_0 \right)^2 \right) \tag{43}$$

Where A_y^* is the new amplitude of vertical oscillation, A_y^i is the test amplitude, ω and ϕ are the gait frequency and phase respectively, ϕ_y is the phase offset of vertical oscillation, v_y is the y-motion gradient and y_0 is the centre of oscillation. As shape deformation is defined as a vector displacement in the direction normal to the base contour, computation of mean shape is trivial:

$$MS'(s) = MS(s) + \frac{1}{\sum_t w} \sum_t w \cdot d(s,t) \cdot normal\left(C_{base}(s,t)\right) \tag{44}$$

Using these new global models, a more accurate local model configuration can be computed for each frame, and the local maximisation process can be repeated. Iteration of local maximisation and global expectation is continued until convergence is reached (there is no change in the global expectation) or until a maximum number of iterations is reached.

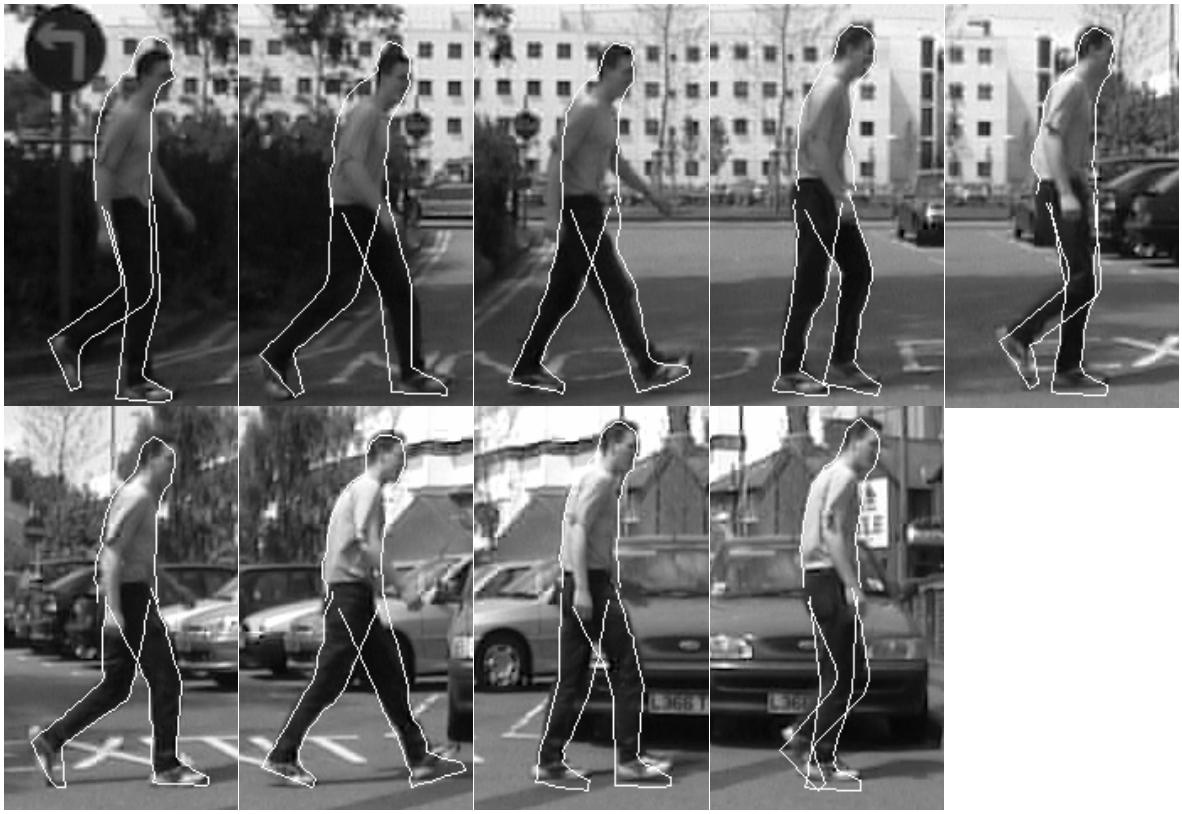


Figure 37: Global expectation, sequence '008e013s00R'

Figure 37 shows the result of the global expectation process after a single iteration for an example outdoor sequence. Compared with the locally maximised configuration (Figure 36), this figure shows that the global expectation process greatly reduces the degree of error in poor configurations, whilst accurate results are left mostly unchanged.

5.5. Conclusions on Hybrid Model Adaptation

The hybrid approach to model adaptation introduced in this chapter applies both purely global and purely local adaptation strategies, rather than defining a compromise between the two. Local maximisation is performed with no temporal constraints and where practical a single parameter at a time, to ensure that parameter optimisation is fast. Global constraints on shape and motion are enforced within this process by attraction to a single base model contour. This base model is computed using a global evidence gathering algorithm, and prevents the local maximisation process deviating from the global prediction unless there is sufficient image evidence to suggest it is necessary. This strategy is generally successful in coping with high levels of noise, because model constraints take over when evidence is poor. However, this strategy alone is vulnerable to poor initialisation, because the optimisation process is highly local and the correct model configuration can easily be excluded from the search space.

It is assumed that the local maximisation process will on average increase the similarity of the model configuration to the subject, even if some configurations are inadequate. The global expectation process generates a new global model from the local parameter estimates. This process uses relative edge correspondence to decide which configurations are reliable and which are not, allowing reliable estimates to dominate the global model prediction. The global model thus computed is then broken down into local models again and the process is repeated. This iterative process bears many similarities to the EM algorithm, and indeed it can be viewed as a specialised implementation of the Generalised EM algorithm. No attempt has yet been made to ensure the convergence criteria of the GEM algorithm [Wu 83] are met, although there is no reason why this should not be possible.

The primary difficulty in this approach to model adaptation is in balancing the local maximisation and global expectation processes. Inevitably in the expectation step, some of

the work done by local maximisation will be undone. Ideally, all the poor configurations would be eliminated and the remaining configurations would be used to generate a new global model, but this is not always the case. In the presence of noise and occlusions, a good model configuration will not always be supported by sufficient image evidence. Improved pre-processing would of course alleviate this problem, but this does not address the inherent difficulty of assessing the reliability of image data.

The new approach defined in this chapter offers an efficient model-based method of motion capture for walking people, capable of operating under cluttered outdoor conditions. This capability is shown to good effect in the chosen application scenario of recognition by gait, demonstrating a high recognition rate on a large database of walking people in Section 6.5.

Chapter 6. Performance Assessment

6.1. Introduction

This chapter is concerned with assessment of performance for the three different approaches to model adaptation described in Chapters 4 and 5. All of these approaches rely on the same initialisation process described in Chapter 3. The availability of ground truth data is highly limited. For the indoor portion of the LDB, chromakey-extracted silhouettes are available, and heel-strike frames are labelled. The latter corpus of data is used to validate the extraction of gait frequency and phase (Section 3.2). The silhouette data could potentially be used to validate the extracted model contours, but the arms are not yet included in the model, which would introduce an unknown error factor. For this reason, the silhouette data is not used to evaluate performance at this time. There is no ground truth data available for the outdoor dataset, which mandates the use of indirect measures to assess performance.

The first and most intuitive measure of performance is a visual assessment of a sample of sequences, presented in Section 6.2. This is also the least rigorous means of assessment, as it is impractical to assess all sequences in the database manually and it is a largely subjective measure. The example gait sequences included are sampled uniformly at a rate of one frame in ten, and cropped to the region containing the subject. The y-position of this region is constant, so that the slope of the subject's walking path may be observed.

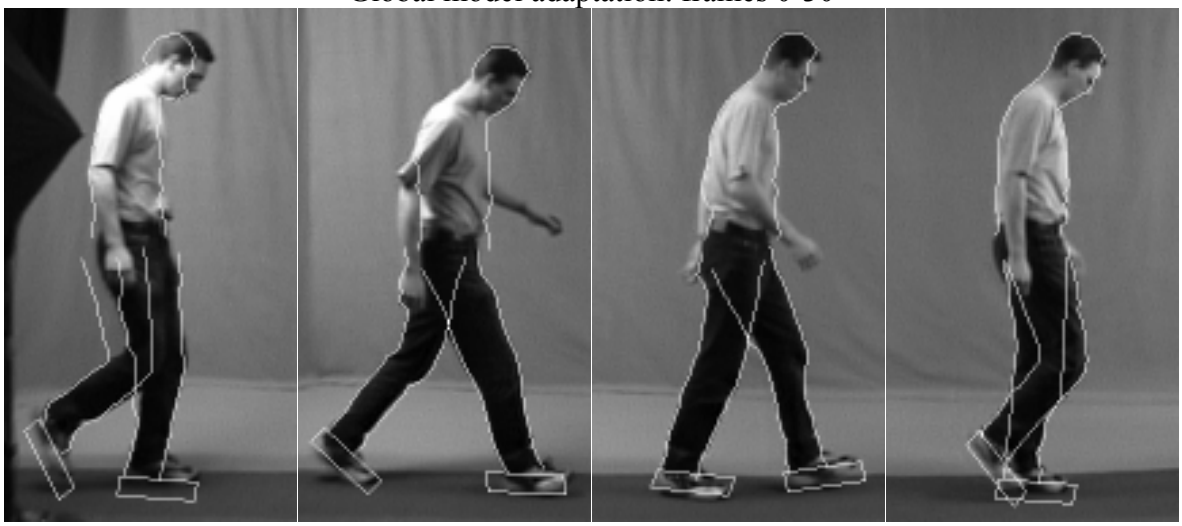
Section 6.3 summarises the computational requirements for each adaptation algorithm. Though not a central concern, it is important for many practical applications that a gait description can be extracted quickly. Section 6.4 presents a detailed analysis of the discriminatory capability of each model feature extracted using hybrid model adaptation (Appendix C lists a full analysis for each adaptation approach). This analysis allows each model feature to be weighting according to its importance in the recognition process.

Recognition capability is investigated in Section 6.5, comparing performance on the indoor and outdoor datasets of the LDB for different selections of features. Additional error metrics are included in Section 6.6, which measure the consistency of contour extraction on the indoor and outdoor datasets.

6.2. Example Gait Sequences

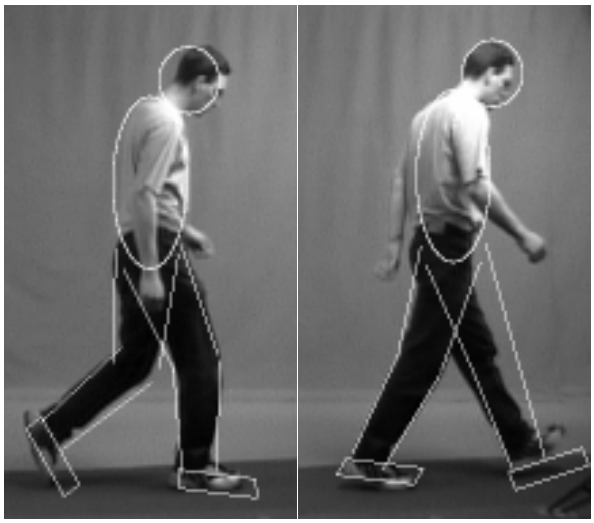
6.2.1. Indoor Dataset Comparisons

Global model adaptation: frames 0-30

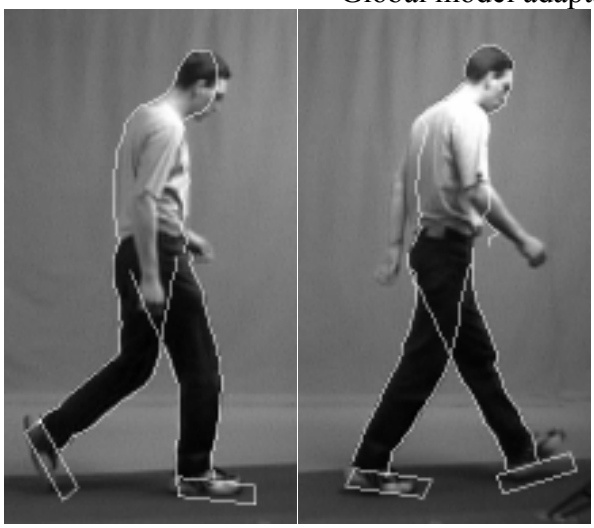


Local model adaptation: frames 0-30

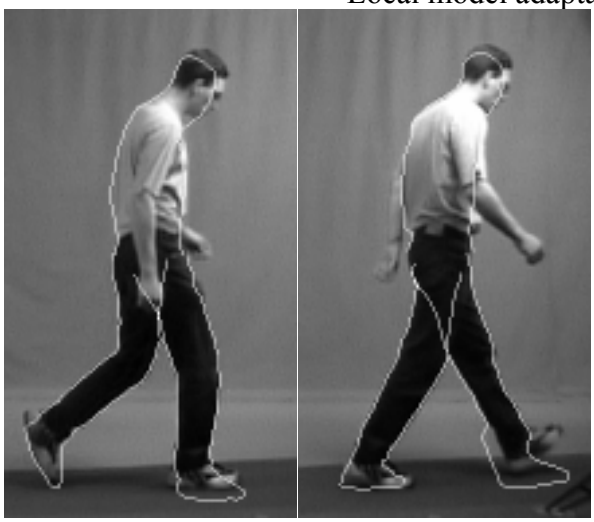
Hybrid model adaptation: frames 0-30



Global model adaptation: frames 40-50



Local model adaptation: frames 40-50



Hybrid model adaptation: frames 40-50

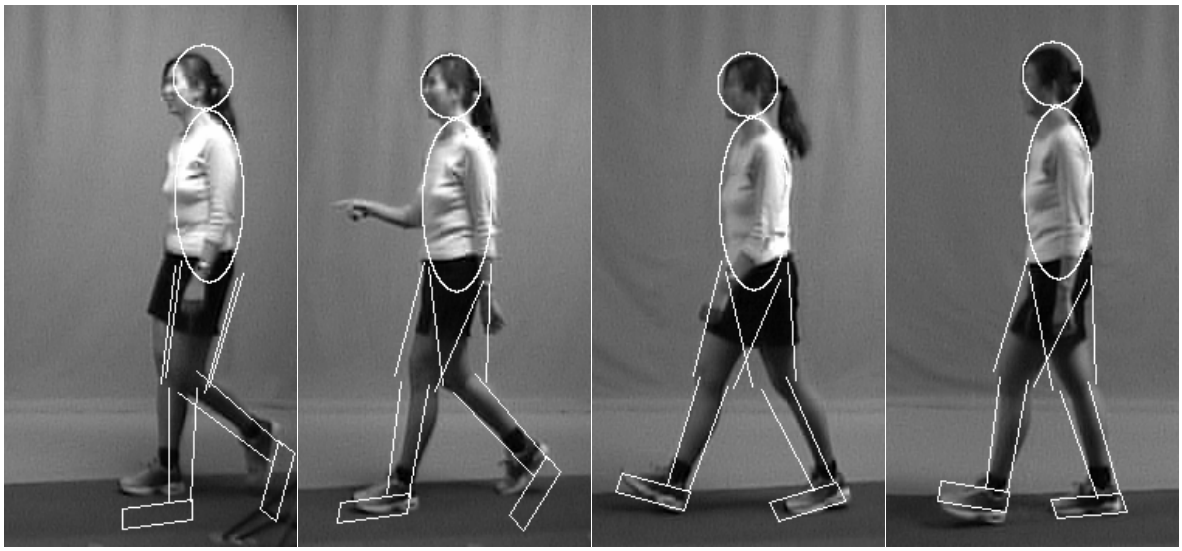
Figure 38: Comparison of extraction performance for sequence '008a013s00R'

Figure 38 demonstrates some of the main problems with a global approach to model adaptation. The most obvious source of error is the limited complexity of the shape model

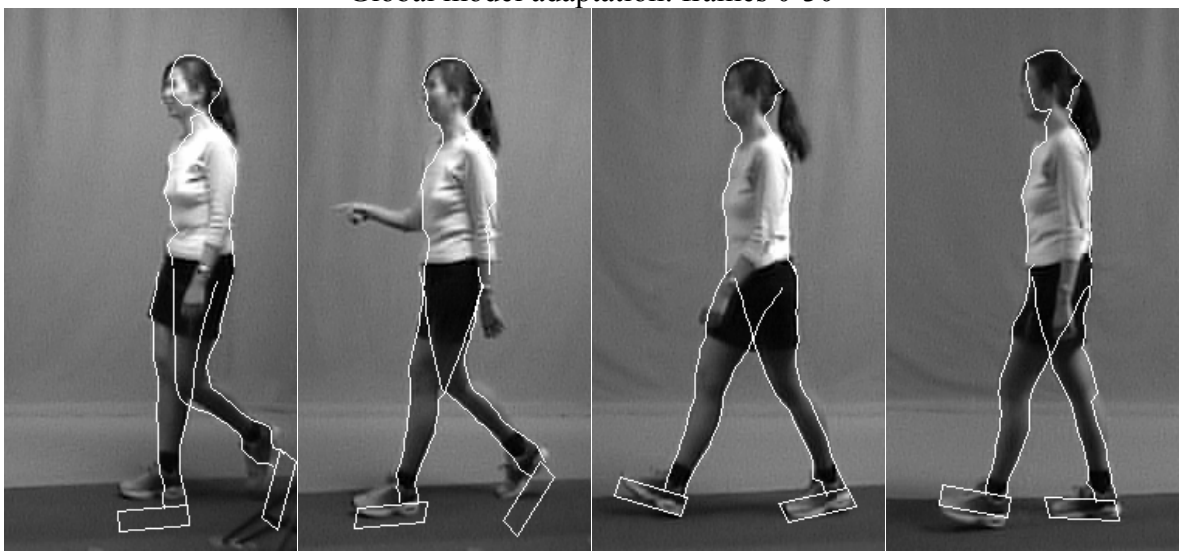
when using rigid geometric components. This model allows a reasonable approximation of body shape, but fine detail is lost. The local and hybrid approaches employ a deformable shape model achieving much better accuracy in modelling the shape of the subject, and local deviations from the global model are also naturally accommodated. However, some errors are apparent here, particularly in the head and feet regions. These areas incorporate relatively complex edge data, due to facial features or footwear, and when the initialisation of the contour is inadequate, the local maximisation process will adapt to fit the wrong edges. The internal contour forces employed to restrict unnecessary deformation also tend to encourage shrinkage of convex contours (see Equation 27, Section 4.2.2 and Equation 39, Section 5.2.3), leading to interior edges being matched in favour of the outer boundary. Some recent approaches to deformable contour fitting have employed alternative internal constraint forces to alleviate this problem [Perrin 01]; combining edge data with silhouette boundary data may also improve matters. An additional source of error for the upper body contours is variation in head inclination, which is assumed to be constant for all of the approaches. The hybrid approach does not demonstrate much of an advantage over the local approach in this example, although there is some improvement in shape consistency.

There are also errors in the extraction of leg position and pose, particularly in the first and final frames of Figure 38, for all of the extraction approaches. For the global adaptation approach, the pelvis is assumed to follow the mean rotation patterns established in Section 2.3.2, and there can be some significant individual deviation from this average pattern. The largest source of error however is the inability to account for local deviations from the global model. Each gait cycle is assumed to be identical, and this is generally not the case for real people. It is possible to design global models of joint rotation that do not rely on this assumption, but the resulting increase in model complexity is likely to make this approach computationally intractable. Curvature of the ground plane also contributes to errors in all three approaches, which again is most apparent in the first and final frames, where the subject is at the edges of the scene. Camera distortion can be corrected to remove such errors, but it may be more useful to allow for such distortions in the model, so that the approach can be used with uncalibrated cameras.

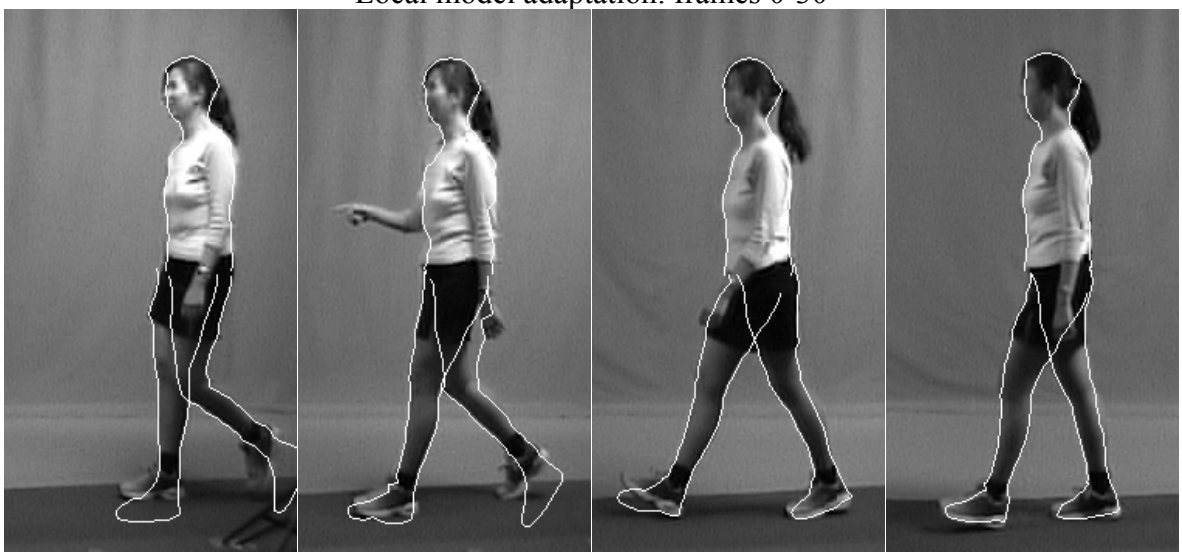
The ankle joint suffers from being at the bottom of the model hierarchy, so errors in pelvis, hip and knee rotation all accumulate down to the position of the ankle. This makes ankle rotation and foot shape very difficult to resolve. It may be possible to incorporate an additional bottom-up model adaptation process to improve ankle and foot parameter estimates, particularly as the supporting foot is static for extended periods of time.



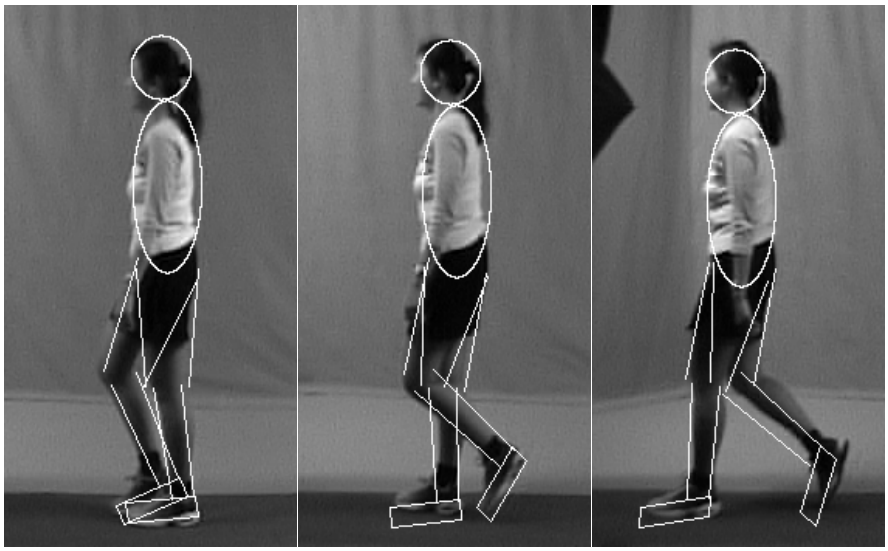
Global model adaptation: frames 0-30



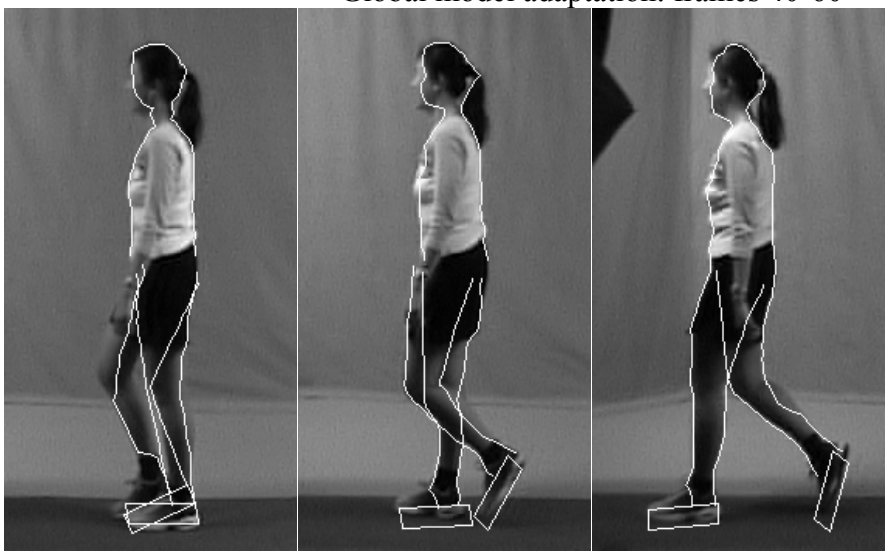
Local model adaptation: frames 0-30



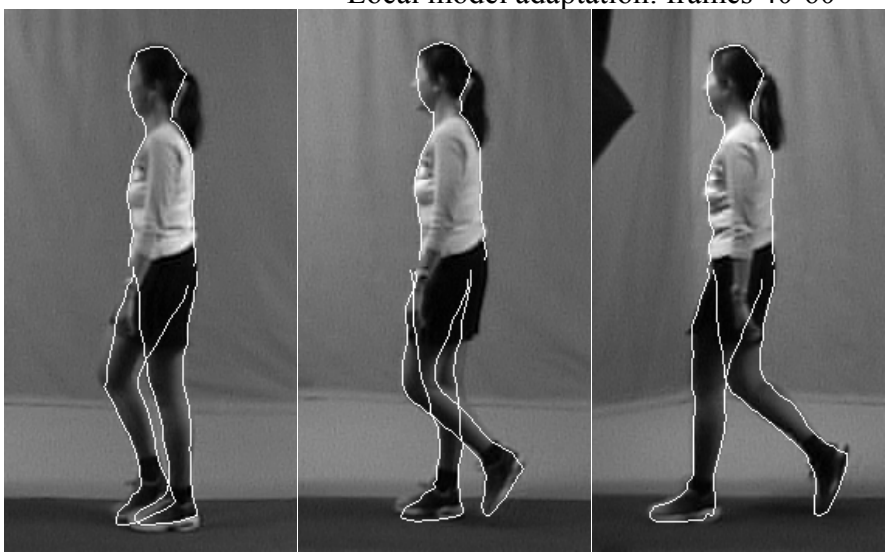
Hybrid model adaptation: frames 0-30



Global model adaptation: frames 40-60



Local model adaptation: frames 40-60



Hybrid model adaptation: frames 40-60

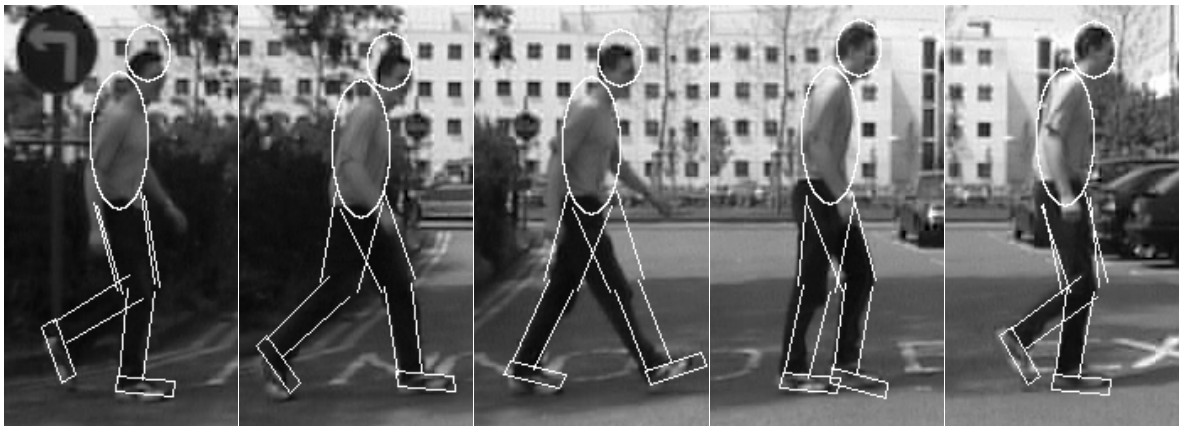
Figure 39: Comparison of extraction performance for sequence '012a033s00L'

Figure 39 illustrates the results obtained for a subject wearing shorts. For this subject, leg edges are poorly defined due to the soft boundaries of the legs, and the local adaptation approach shows numerous errors in contour shape as a result. The global adaptation approach fairs reasonably well, although the size of the legs is underestimated due to the limited complexity of the shape model. Hybrid model adaptation shows a clear advantage over the first two approaches for this sequence, as global constraints can be applied effectively to compensate for the poor edge data quality. Both leg and upper body contours match well to the subject and are extracted consistently. There is some error in height estimation for all three approaches, most noticeably in the first frame, caused partly by camera distortion and partly by the slope of the path taken by the subject. This error is smallest for the hybrid adaptation approach in which the *y* motion gradient is taken into account.

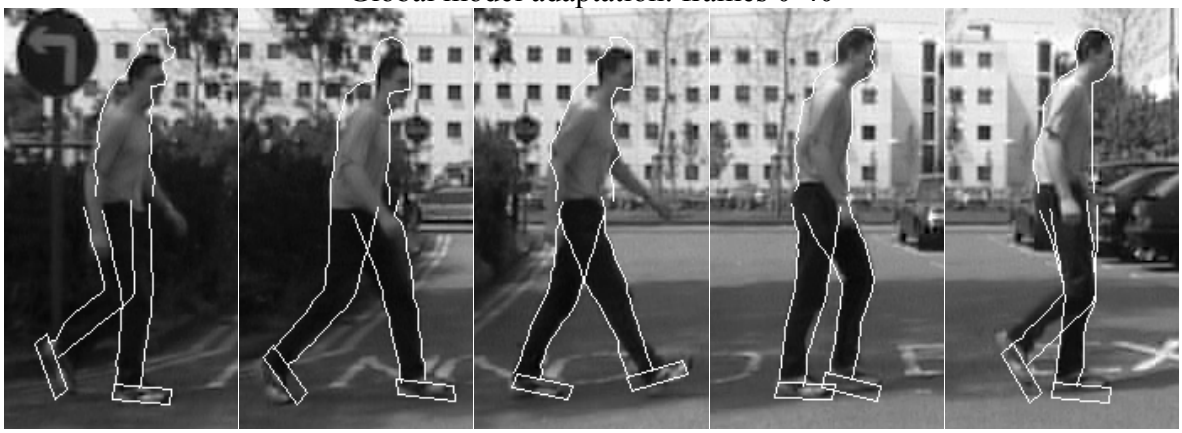
6.2.2. Outdoor Dataset Comparisons

The examples in this section show the same subjects as the examples in the previous section, filmed under outdoor conditions. The level of background clutter is much higher in these sequences, and shadows cast by the subject and by background objects are much more pronounced. These sequences generally exhibit the same problems as the indoor sequences, though errors due to the assumption of a flat ground plane are more serious here. The subject in Figure 40 moves on a slight slope, with the result that the global and local adaptation approaches display large mismatch errors in the first few frames, and to a lesser extent the final frames. The inclusion of the *y* motion gradient in the constraints applied in the hybrid adaptation approach demonstrates a clear advantage in this example.

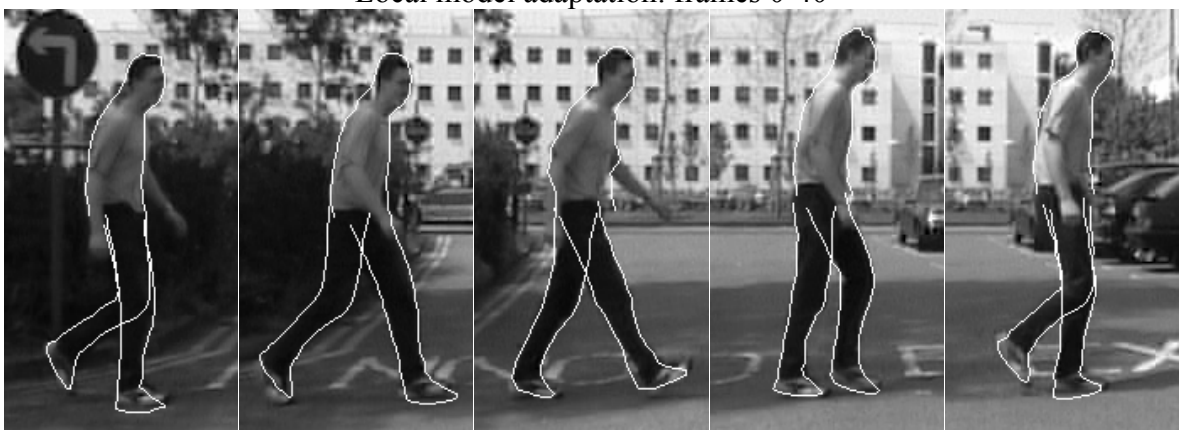
All approaches extract joint rotations with reasonable accuracy, which is encouraging considering the increased level of noise in the outdoor dataset. The reduction in performance is most pronounced in the case of the local approach, which does not have the capability to recover from large errors in initialisation. In contrast to the results obtained for the indoor sequence of the same subject (Figure 38), in this example the hybrid approach does significantly improve on the contour shape extracted by the local approach.



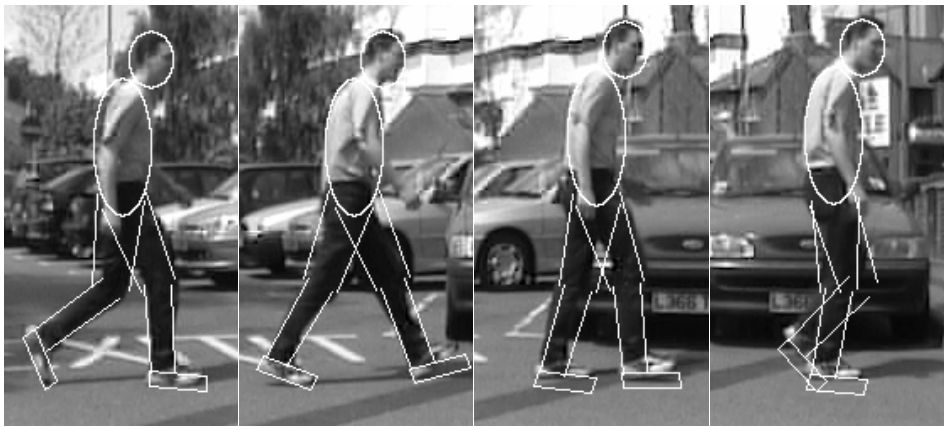
Global model adaptation: frames 0-40



Local model adaptation: frames 0-40



Hybrid model adaptation: frames 0-40



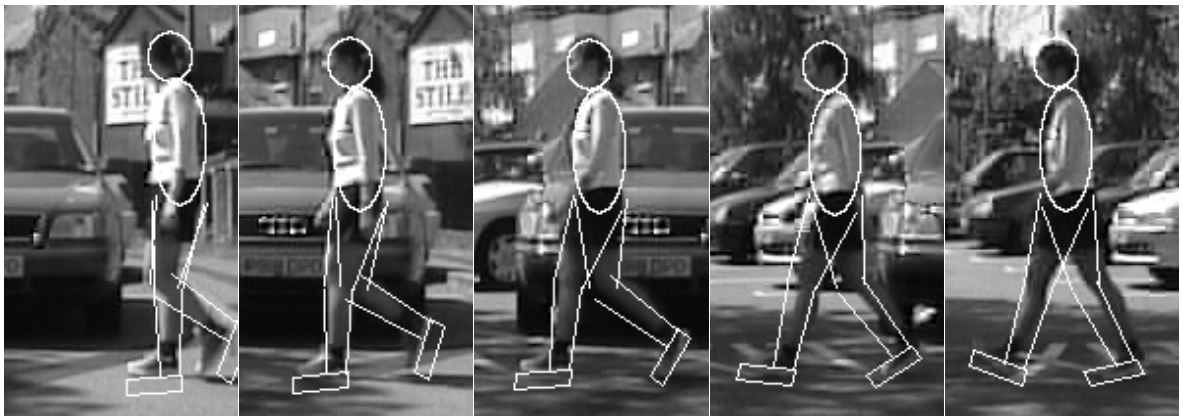
Global model adaptation: frames 50-80

Local model adaptation: frames 50-80

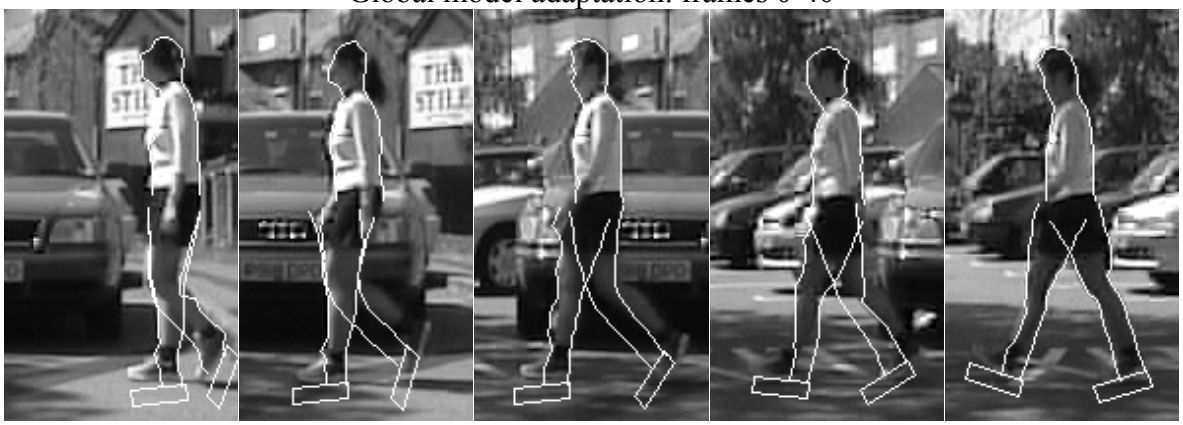
Hybrid model adaptation: frames 50-80

Figure 40: Comparison of extraction performance for sequence '008e013s00R'

Figure 41 further demonstrates the superiority of the hybrid adaptation approach. Although there are some minor errors in joint rotation estimation for the first frame, shape and motion is extracted reliably throughout the remainder of the sequence. By comparison, both the global and local adaptation approaches exhibit some errors in extracting the position and pose of the leg, and the local approach often fails to extract acceptable models of leg and head shape. Again, the slope of the subject's path is a problem for the first two approaches, which do not consider this source of variation.

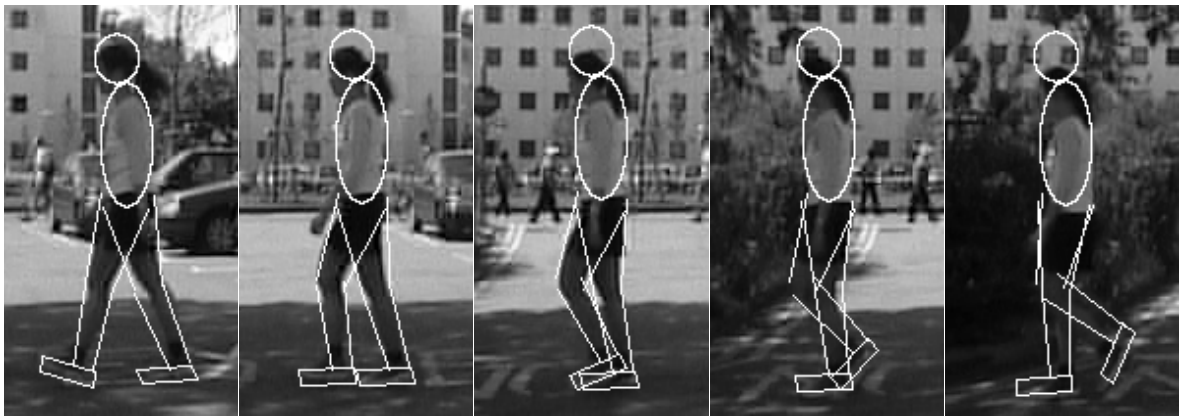


Global model adaptation: frames 0-40

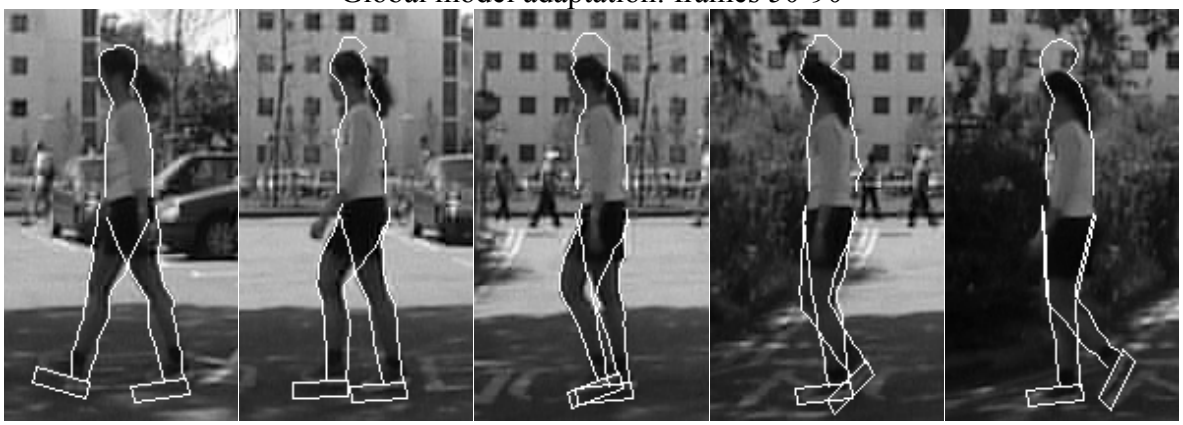


Local model adaptation: frames 0-40

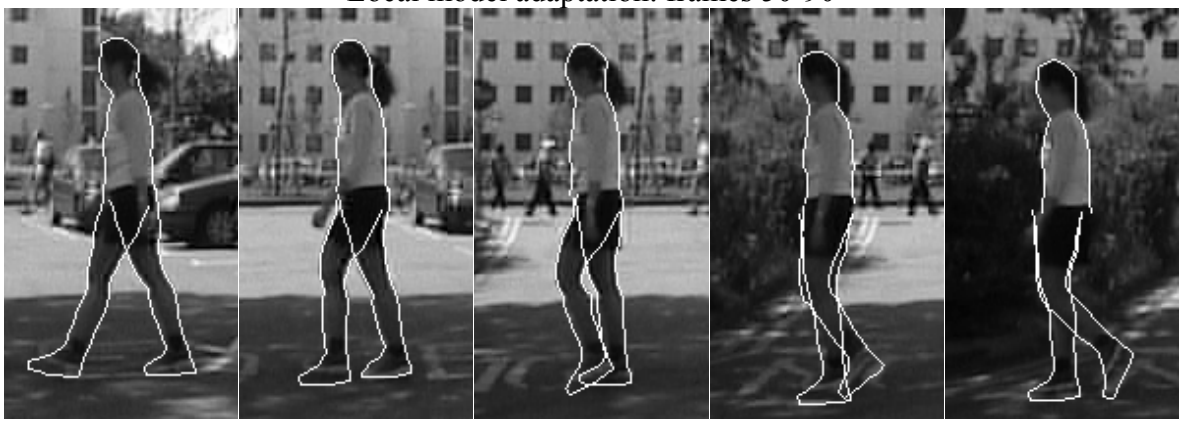
Hybrid model adaptation: frames 0-40



Global model adaptation: frames 50-90



Local model adaptation: frames 50-90



Hybrid model adaptation: frames 50-90

Figure 41: Comparison of extraction performance for sequence '012e033s00L'

6.2.3. Additional Examples for Hybrid Model Adaptation

This section includes a number of additional sequences from the outdoor dataset, demonstrating the performance of gait model extraction using the hybrid model adaptation approach.

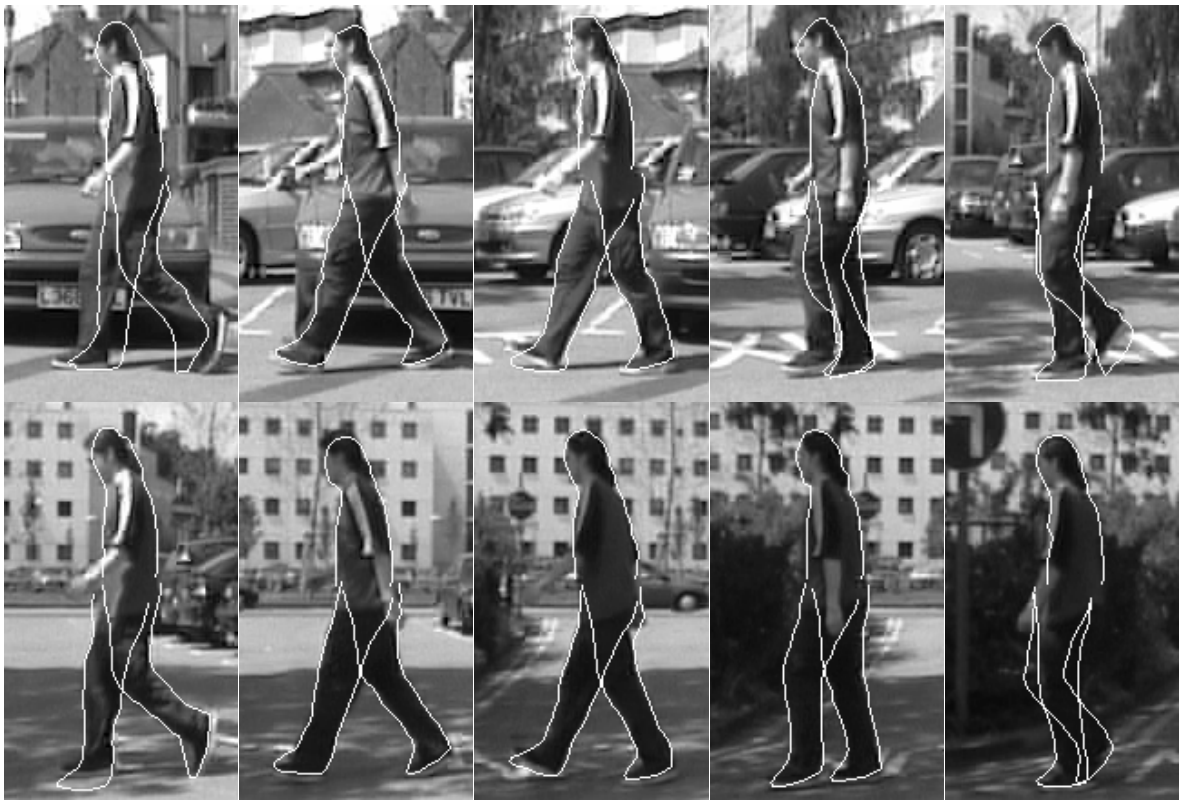
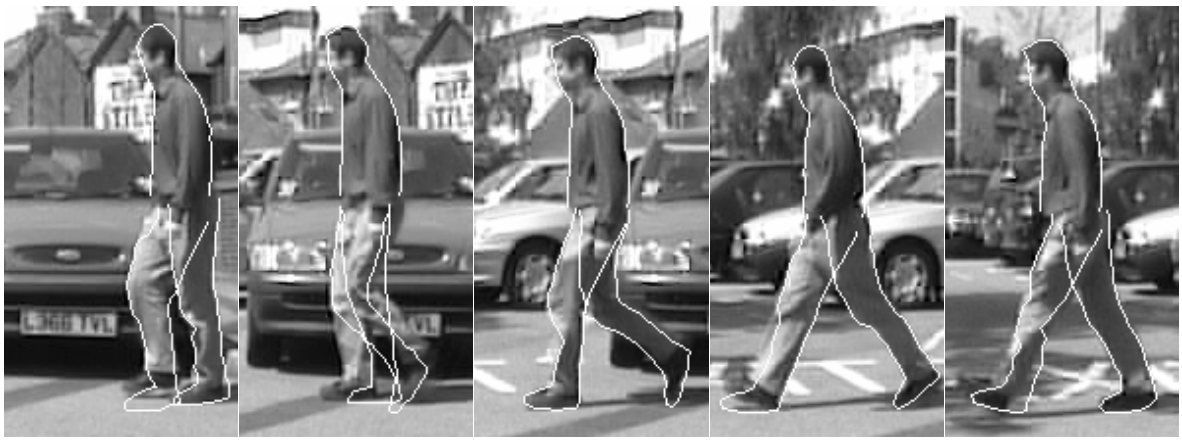


Figure 42: Hybrid model adaptation, sequence '008e014s00L'

Figure 42 shows some significant errors in extraction of the head and lower leg in the first part of this sequence. However, the gait signature derived for recognition (Section 6.5) is an average measure of body shape and gait motion, meaning that local errors will not significantly obstruct the recognition process, provided that they constitute only a small proportion of the sequence.



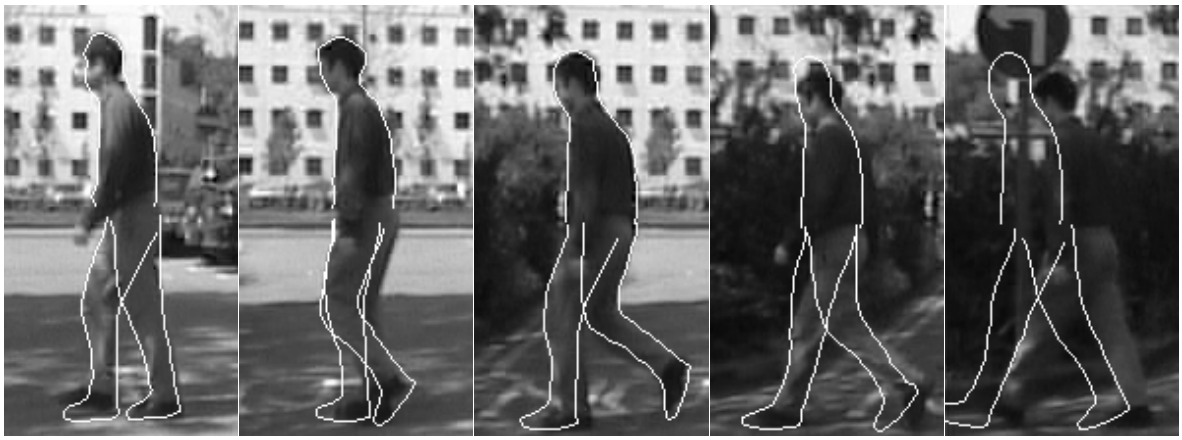


Figure 43: Hybrid model adaptation, sequence '008e015s00L'

Two additional sources of error may be observed in Figure 43. The subject's trousers are somewhat difficult to distinguish from the background, as no colour information is employed in the extraction process. The folds and shadows in the trousers also create additional edges that the leg contours may be attracted to. There is a tracking failure towards the end of the sequence, possibly because the subject slows down, or possibly due to the lack of edge data in this heavily shadowed region.

Figure 44: Hybrid model adaptation, sequence '008e016s00R'

Figure 44 depicts a generally good model extraction. The most significant error apparent in this example is the placement of the hip joints in the support phase of the gait cycle. There are also some errors in foot placement and pose in this sequence.

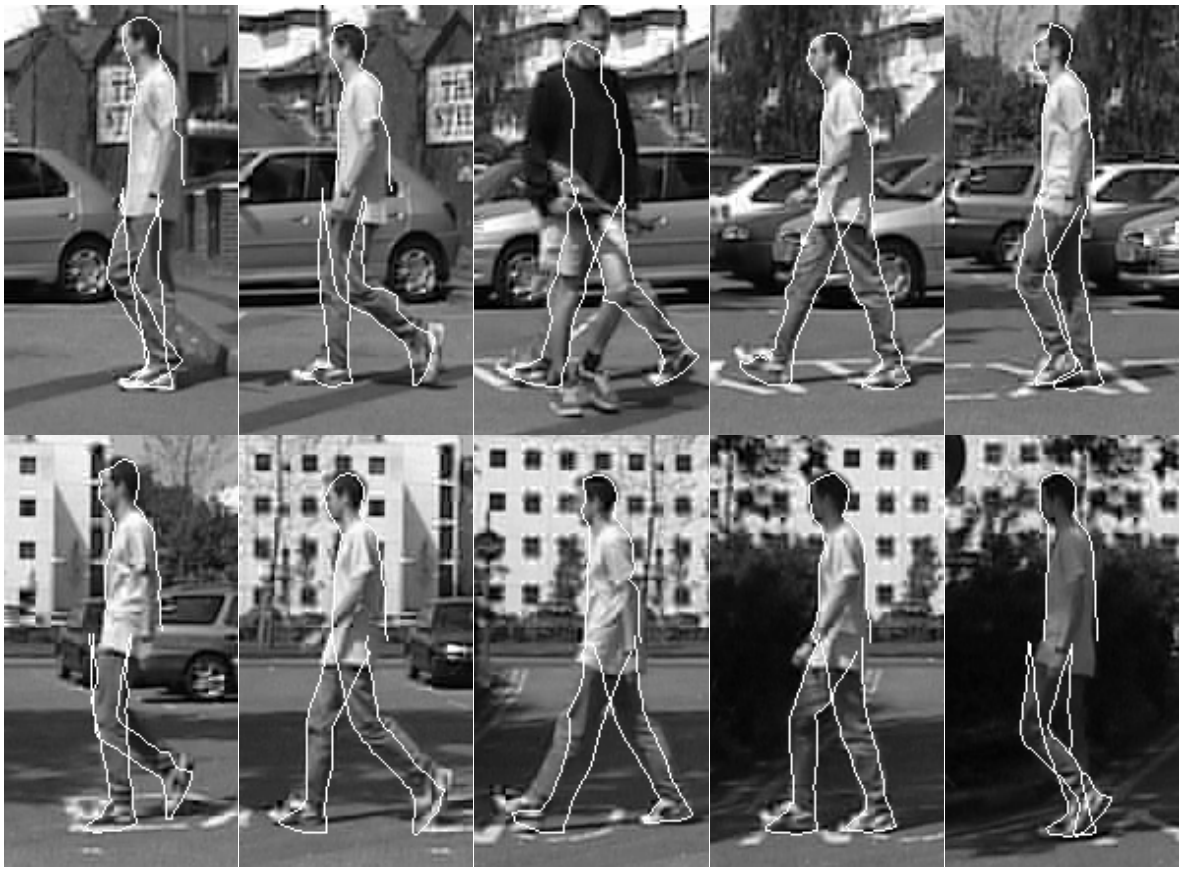


Figure 45: Hybrid model adaptation, sequence '009e017s00L'

Figure 45 demonstrates one of the advantages of including a global modelling strategy. In frame 20, the position and pose of the subject is extracted with surprising accuracy, even though the subject is almost completely occluded by another pedestrian. Errors in extraction of the lower leg are more numerous in this sequence, which indicates that the local maximisation process is unable to correct the degree of error present in the global initialisation, or that the local edge data available is unreliable.

Figure 46: Hybrid model adaptation, sequence '009e018s00R'

Extraction of the head contours is unreliable in Figure 46, and some difficulties are encountered in dealing with the bare legs of the subject. However, joint dynamics are generally extracted with a high degree of accuracy.

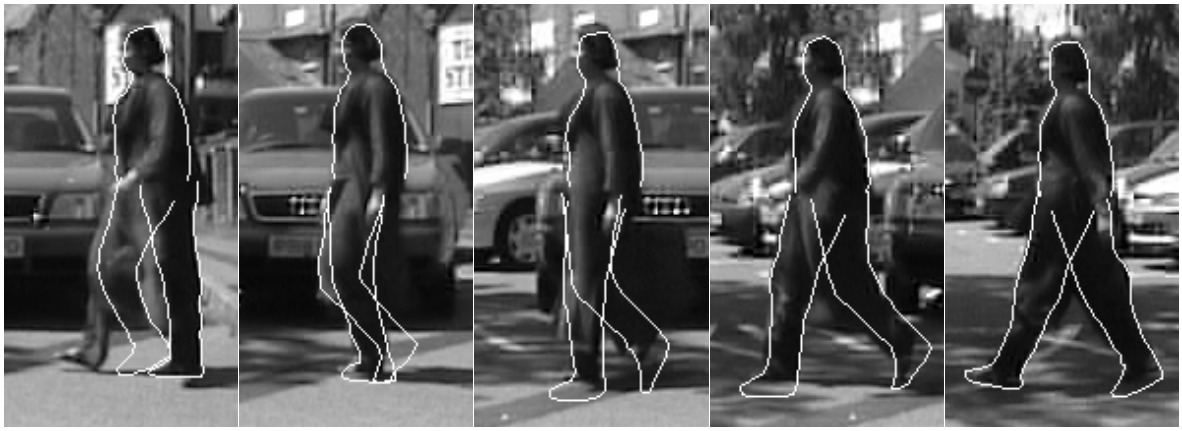


Figure 47: Hybrid model adaptation, sequence '012e031s00L'

The long, flowing robes of the subject in Figure 47 present a serious problem for gait extraction algorithms, as a large proportion of the leg region is obscured. This sequence is also heavily shadowed, increasing the difficulty of the extraction task. However, enough variation is visible to extract periodicity information in this sequence, which allows a reasonable approximation of the subject's gait. It is unclear in general what proportion of the legs must be visible to allow gait extraction, but even if the legs were completely obscured, the motion of the clothes may be sufficient to extract some gait information. Of course, these comments apply equally to subjects wearing a long skirt or trench coat.

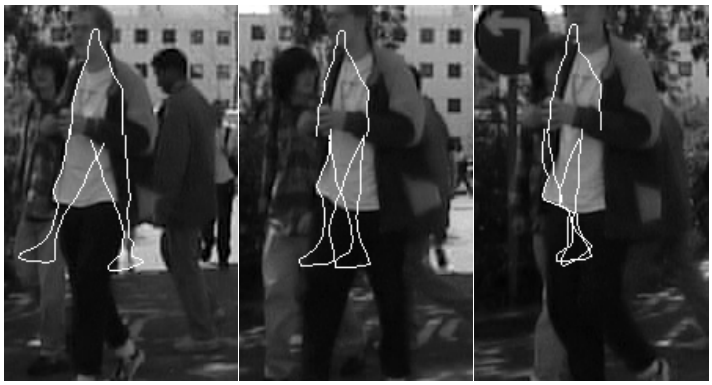


Figure 48: Hybrid model adaptation, sequence '012e032s00L'

Figure 48 demonstrates the resulting confusion when the subject is forced to slow down to avoid a collision with a group of people blocking the path. This violates the constant velocity assumption made by the COM tracking algorithm (Section 3.2.1), and prevents the computation of an accurate temporal accumulation. This in turn means that the model adaptation algorithm is poorly initialised, and an acceptable gait description cannot be extracted. This problem is solvable with a more sophisticated tracking algorithm, which would be required for more general application scenarios (see Section 7.2).

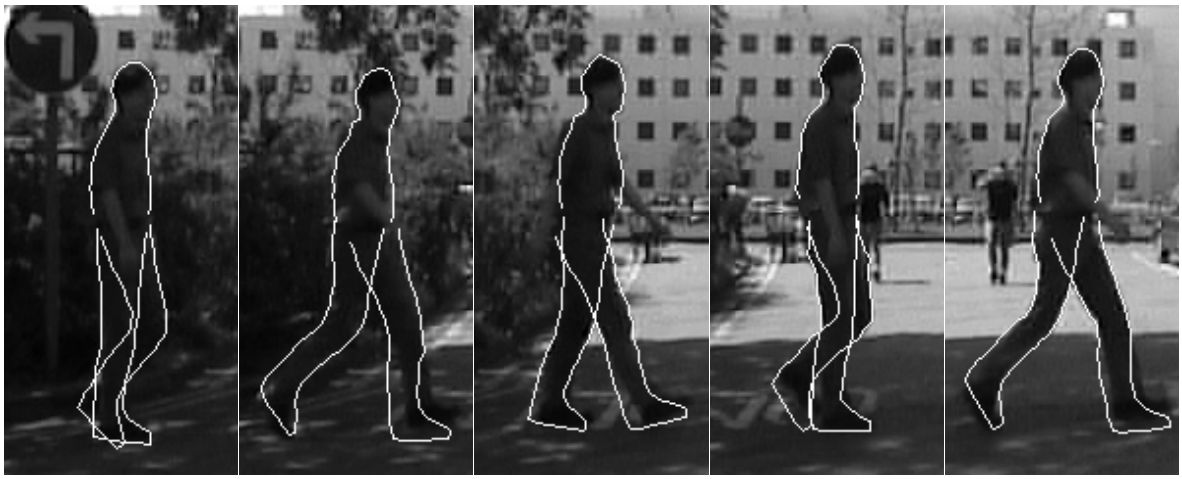


Figure 49: Hybrid model adaptation, sequence ‘012e034s00R’

Figure 49 shows a heavily shadowed sequence in which there is little contrast between the subject’s legs and the ground. This leads in some cases to errors in leg pose determination, but the upper body is extracted well.

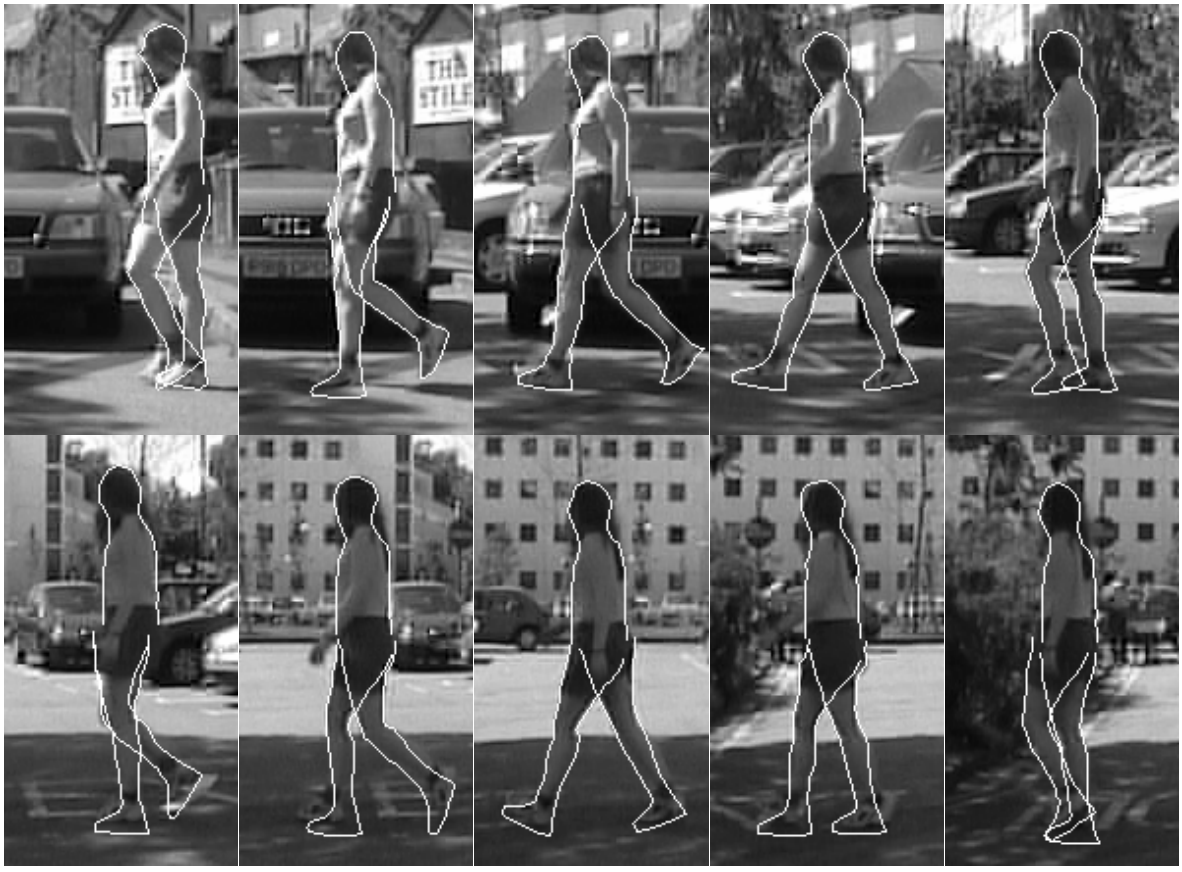


Figure 50: Hybrid model adaptation, sequence ‘012e035s00L’

The subject in Figure 50 is similarly in heavy shadow, although there are no major errors apparent in this extraction.

Figure 51: Hybrid model adaptation, sequence '012e036s00L'

Figure 51 shows a successful extraction despite the bus passing behind the subject, demonstrating that correlation with mean human shape is an effective means of distinguishing the subject from other moving objects (see Equation 12, Section 3.2.1). The only obvious error here is in the width of the upper body, which is overestimated to some extent due to the influence of edge data from the bus.

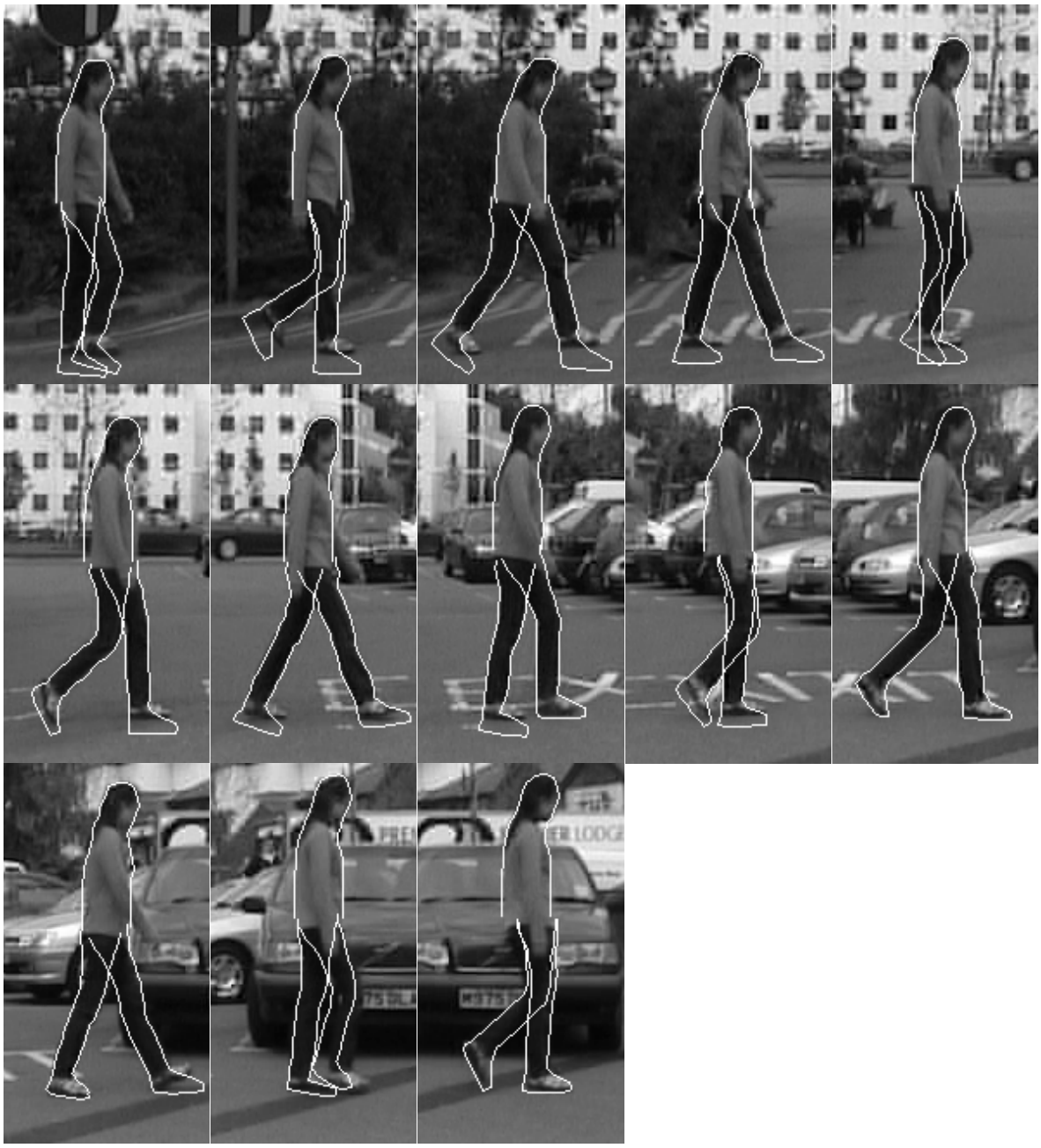


Figure 52: Hybrid model adaptation, sequence '013e037s00R'

The sky is clouded over in Figure 52, so there are no significant shadows in the foreground of this scene. Although most of the model is extracted well, there are some errors in the localisation of the feet. This suggests an additional problem related to the assumption of a flat ground plane. Although the hybrid model adaptation approach can correct the *y*-position of the COM, the height of the subject is assumed to be constant. When the ground plane is not flat height is likely to be over-estimated, as the temporal accumulation will be spread out over the *y*-plane. This problem could be solved by the inclusion of height as an additional free parameter in the adaptation process, but as all other shape parameters are

dependent on the subject's height, this may require too much additional computation. Again, relaxing the assumptions made on the motion of the subject and employing a more sophisticated COM tracking algorithm would alleviate this problem.

6.3. Computational Requirements

All testing for the algorithms described within this thesis was performed using a 2.4GHz Pentium 4 PC with 1GB RAM. Figure 53 shows the time taken in video pre-processing and gait model extraction for the indoor and outdoor datasets of the LDB (see Section 1.4).

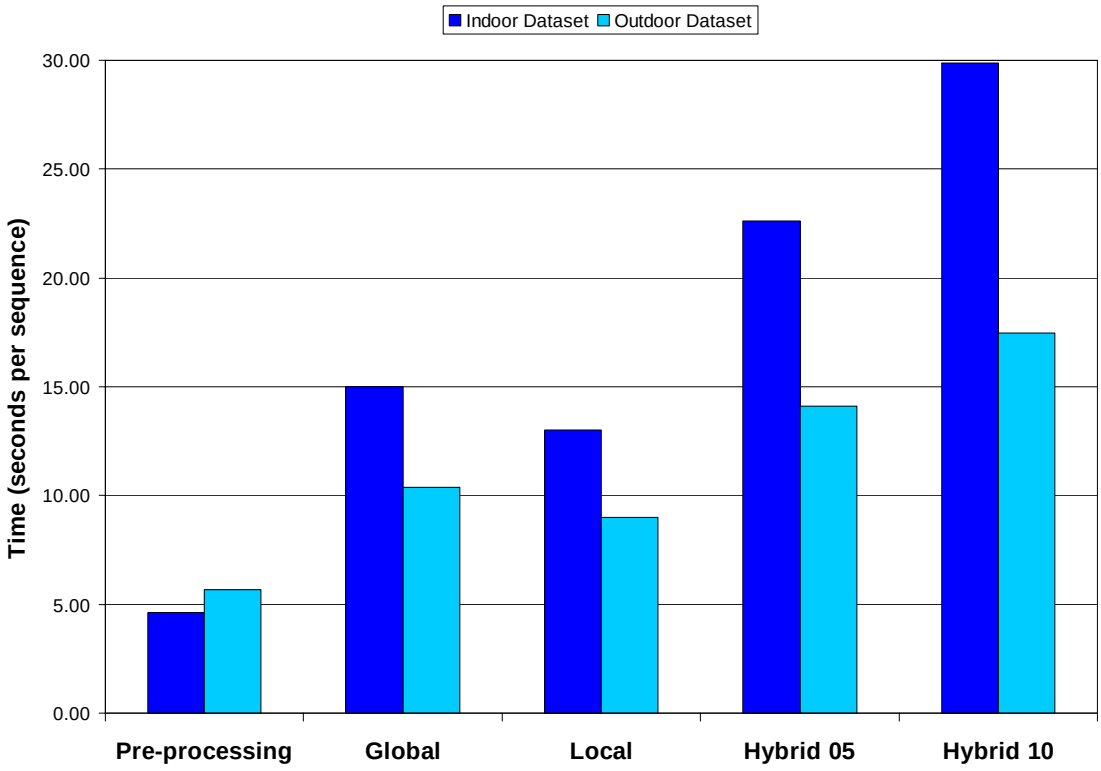


Figure 53: Average sequence processing time for LDB

The iterative process employed by the hybrid model adaptation approach is computationally expensive, but some compromises can be made by adjusting the (maximum) number of iterations (the computation times for 5 and 10 iterations are included above). The times given in Figure 53 are average processing times per sequence, where indoor sequences average 62.4 frames and outdoor sequences average 98.5 frames. Assuming 10 iterations of hybrid model adaptation, the total processing rate from raw

video footage to a complete gait description is 1.8Hz (frames per second) for the indoor dataset and 2.7Hz for the outdoor dataset. The difference is due to the increased apparent size of subject in the indoor dataset, which correspondingly increases the size of the search bounds in model parameter estimation.

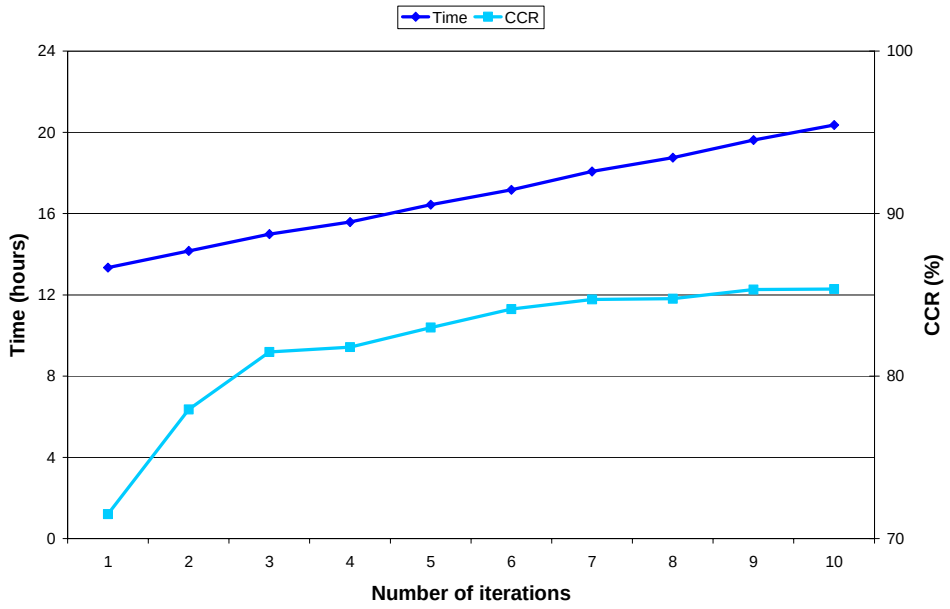


Figure 54: Hybrid model adaptation - effect of increasing number of iterations

Figure 54 illustrates the diminishing returns for increasing the number of iterations in the hybrid model adaptation approach, when processing the outdoor dataset. Computation time increases at an approximately linear rate, but the increase in recognition performance (Correct Classification Rate) slows considerably after 6 iterations. Refer to Section 6.5 for details of recognition performance evaluation. The maximum number of iterations was set to 10 for all the results presented for hybrid model adaptation (Sections 6.2.3 and 6.4-6.6).

6.4. Feature Selection and ANOVA

Chapters 3-5 describe how a model of the subject may be extracted from raw video data, describing the appearance and characteristic gait pattern of the subject. Not all of the parameters used in this model are useful for identification purposes, and so a *feature vector* is defined. The feature vector is a set of numbers describing various aspects of the subject's gait and appearance that are useful for recognition. Features are divided into two

categories: *static* and *dynamic*. Static features describe model constants, such as mean body shape, gait frequency and average walking speed. Dynamic features describe how the model changes over time, such as rotation of leg joints. Not all of these features are immediately useful, and some regularisation must be performed prior to recognition. Body shape features and walking speed are divided by the apparent height of the subject in order to make them scale-invariant, so that the distance of the subject from the camera is not an issue.

To distinguish one person from another, they must have different features and it must be possible to measure them in a reliable and repeatable fashion. These requirements may be stated formally for a given subject *i* and feature *x* as follows:

1. **Within-class variance of *x* is low**
   (the value of *x* measured for subject *i* is always similar)
2. **Between-class variance of *x* is high**
   (the value of *x* measured for other subjects is dissimilar)

Study of these quantities is known as *Analysis of Variance* (ANOVA). The ANOVA *f*-statistic, or Fisher criterion [McLachlan 92], is a useful estimate of discriminatory potential. The *f*-statistic is a ratio of between-class variance to within-class variance:

$$f = \frac{S_B^2}{S_W^2} \tag{45}$$

Where S_B^2 is the between-class variance and S_W^2 is the within-class variance. These quantities are defined by:

$$S_B^2 = \frac{\sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2}{k-1}, \qquad S_W^2 = \frac{\sum_{i=1}^{k} (n_i - 1) s_i^2}{NS - k} \tag{46}$$

Where *k* is the number of subjects in the dataset, n_i is the number of training sequences for subject *i*, *NS* is the total number of training sequences, $\bar{x}_i$ is the mean value of feature *x* for subject *i*, $\bar{x}$ is the mean feature value over all subjects and s_i^2 is the variance of feature *x* for subject *i*. It is important that the *f*-statistic is computed only from training set samples to avoid unfair bias in the normalisation process. A high *f*-statistic does not necessarily guarantee that the feature is useful for recognition, since variance is not always consistent with recognition capability, but it is highly suggestive and is easily computed.

The results of ANOVA for feature vectors extracted using hybrid model adaptation are summarised in Figures 56-58 and 59-60; the full listing is included in Appendix C. ANOVA results for global and local model adaptation are also included in Appendix C.

To aid in the understanding of joint rotation model features, Figure 55 provides an example of the approximate leg pose corresponding to each point in the sampled gait cycle. Samples 1-8 are taken from the float (or swing) phase of the gait cycle, and samples 9-15 are taken from the support phase.

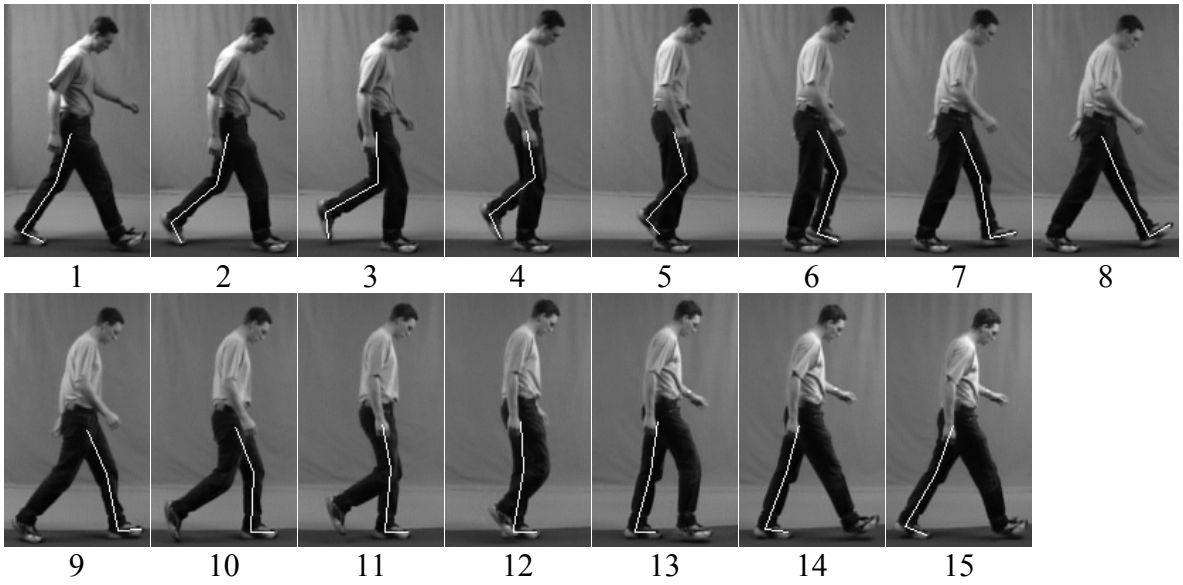


Figure 55: Examples of leg pose for a single gait cycle

Figures 56-58 show the rotation model features (Θ_h, Θ_k, Θ_a) for the hip, knee and ankle joints respectively. It is apparent from these graphs that some parts of the gait cycle are more useful than others for discriminating between subjects. This may be because these parts better characterise individual gait patterns, or that model contours can be more reliably extracted at these points.

For the indoor dataset, the discriminative potential of hip rotation features is greatest around samples 2, 8 and 14. These positions correspond approximately to heel-strike poses (see Figure 62), where there is minimal self-occlusion of the legs. This may indicate that ease of extraction is the dominant factor in discriminatory capability, or that the manner in which a subject's heel strikes the ground is strongly indicative of identity. This pattern is only partially replicated on the outdoor dataset, which suggests that occlusions by other objects or noise sources are a significant factor under these conditions.

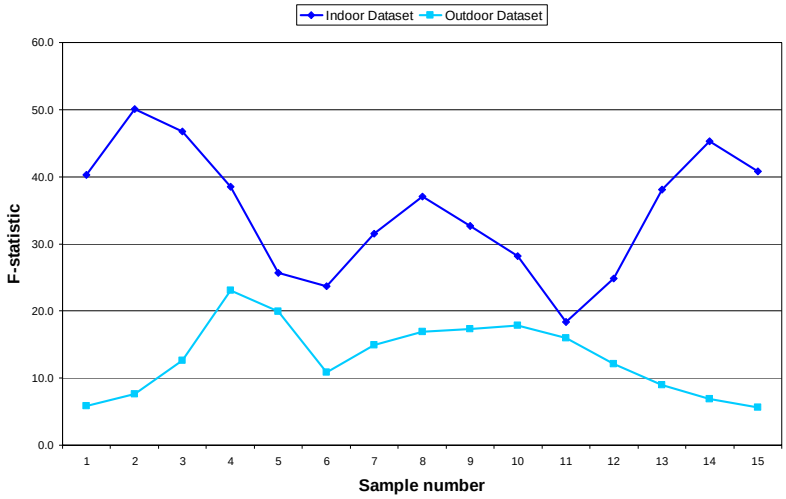


Figure 56: Hip rotation model features

Knee rotation features display a similar pattern of discriminative potential, peaking at samples 1, 10 and 15 for the indoor dataset. Again, this pattern is partially replicated in the outdoor dataset, reduced in magnitude under these more difficult extraction conditions.

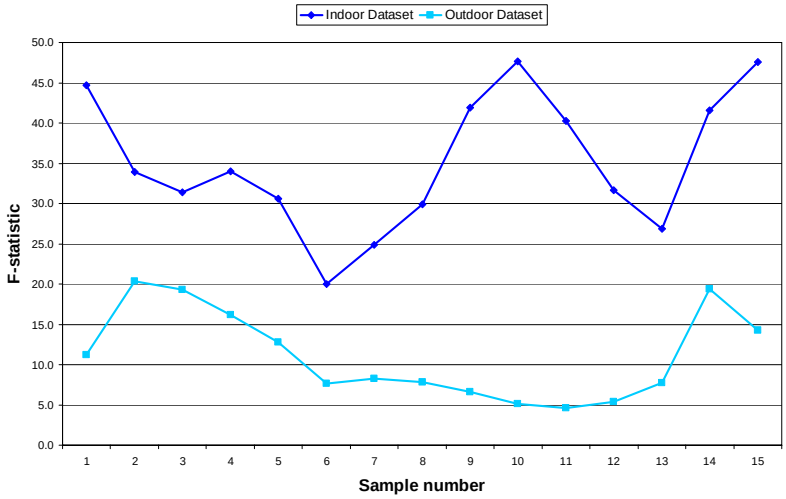


Figure 57: Knee rotation model features

Discriminatory potential is greatly reduced for ankle rotation features, reflecting the reduced accuracy in extraction of these features. This makes it difficult to draw any firm conclusions for ankle features, though a similar three-peak structure is discernable for the indoor dataset, at samples 3, 8 and 12. The minimal discriminatory capability of ankle rotation features on the outdoor dataset suggests that some work is necessary to improve performance. It is possible that there simply isn't much useful information in ankle

rotation, perhaps due to the restrictions of footwear, but increased extraction accuracy is required to confirm or deny this hypothesis.

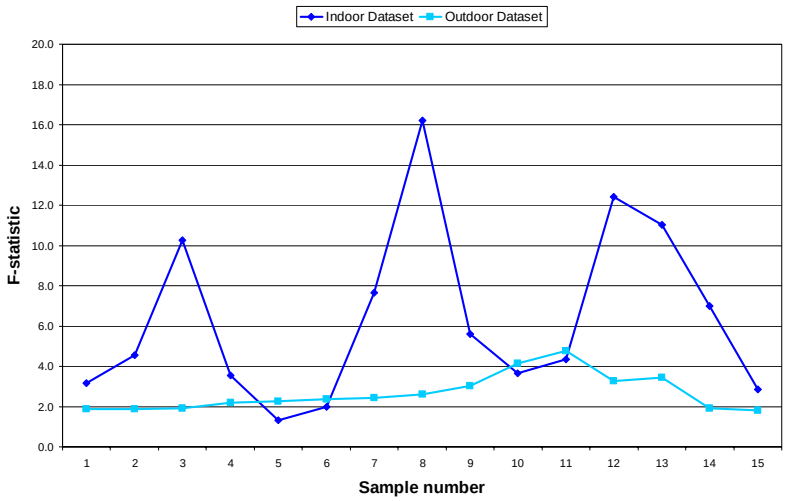


Figure 58: Ankle rotation model features

Figures 59 and 60 plot the ANOVA *f*-statistic of mean leg width and mean upper body width against leg length and upper body height respectively. The origin for both feature groups is the hip, extending down the length of the leg and up the length of the upper body respectively. An additional statistical analysis of these features is included in Section 6.6.

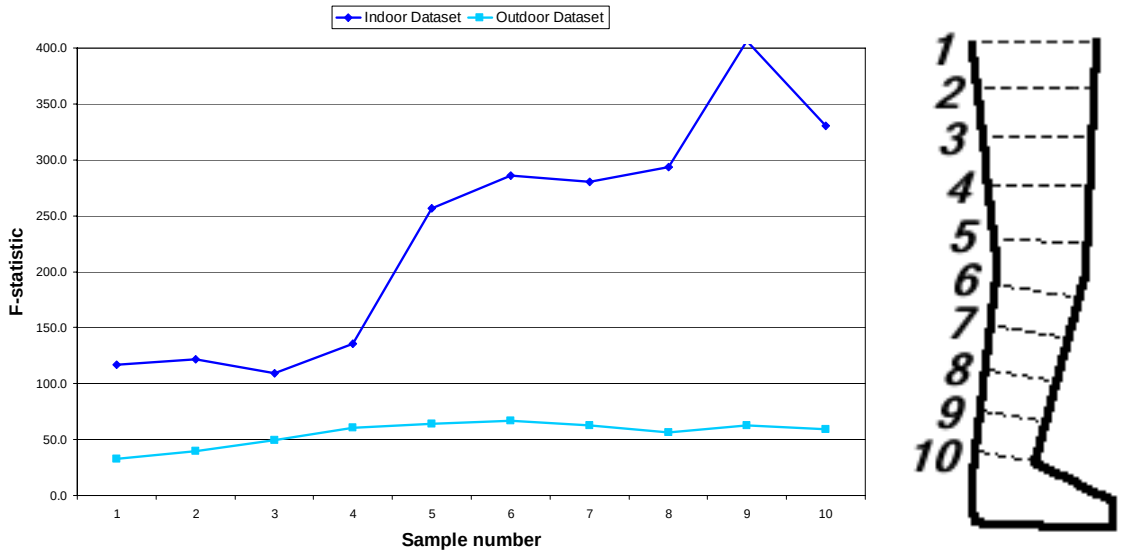


Figure 59: Leg width model features

Leg shape features demonstrate a clear reduction in discriminative potential around the hip and upper thigh level, and an increase towards ankle level. The most likely reasons for

these characteristics are self-occlusion of the legs, and occlusion of the legs by the hands. Self-occlusion is greater at hip level because the legs are closer together, and the hands will only occlude the legs at or slightly below hip level. The same trend is observable on both datasets, though discriminative potential is significantly lower for the outdoor dataset. In addition to the increased noise level in the outdoor dataset, the subject is also significantly further away from the camera, so that they appear around 30% smaller. This reduction in scale decreases the degree of precision that may be attained in model extraction, contributing to a reduction in discriminative potential.

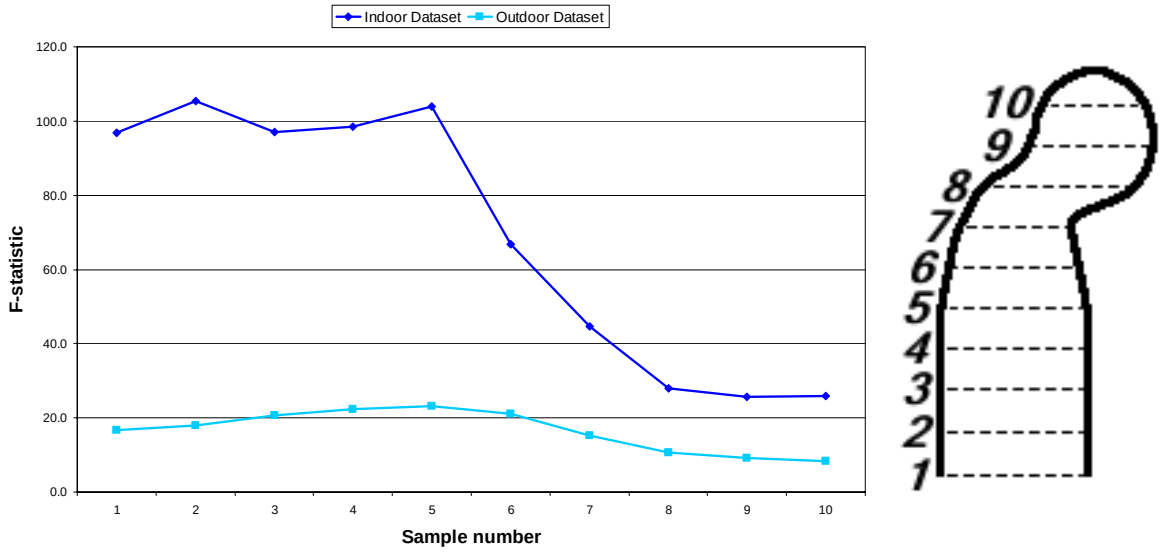


Figure 60: Upper body width model features

Interestingly, occlusion of the upper body by the arms does not appear to be an important factor here. Discriminative potential is greatest in the lower portion of the torso, dropping steeply above shoulder level in both datasets. This may indicate difficulty in extracting head shape features, and the magnitude of body width may also be a factor (the head is smaller than the torso, so the feature difference between subjects is proportionally smaller). An additional source of error is the inclination of the head, which will affect the shape of the body above shoulder level and may vary over different sequences of the same subject.

This analysis is used to bias the feature vector, so that features with greater discriminatory capability have a greater impact on the recognition process. The recognition algorithm (Section 6.5) employs a Euclidean distance metric to compute the similarity of different feature vectors, so the process is inherently biased by the size of each feature. The variation of high-magnitude features will dominate variation in smaller features. This bias

is removed by normalising each feature to the range [0, 1], and a weighting is applied by multiplying each feature by its *f*-statistic:

$$x' = f_x \frac{x - \min(x)}{\max(x) - \min(x)} \tag{47}$$

Where x is the current feature, x' is the normalised feature and f_x is the ANOVA *f*-statistic computed for feature x. This weighting makes the contribution of each feature directly proportional to its discriminatory capability. Although this process cannot be guaranteed to give precedence to the most useful features, in general it may be observed that features useful in recognition score highly in discriminatory capability, and this weighting can be expected to improve recognition capability.

6.5. Recognition Capability

For evaluation purposes, the database must be divided into a training set and test set. The classifier learns what distinguishes the subjects from each other using samples in the training set, and applies these principles to classify unknown samples from the test set. Different approaches to dataset division exist, but for this approach a simple method is sufficient. In 'leave-one-out' cross-validation [Ripley 96], one sample is picked for the test set, using the rest of the database for training. After classifying the sample, it is returned to the training set and the next sample is picked. This process is repeated for each sample in the database, yielding an estimate of the generalisation performance of the classifier.

The K-nearest neighbour method [Ripley 96] was chosen for the classifier, as it is simple, fast and performs reasonably well. In this method, the training set is ordered into a ranked list according to the distance of each sample from the test sample. A Euclidean distance metric is employed for computing the distance between feature vectors. The identity of the sample is inferred by examining the top K samples, choosing the most commonly occurring label (identity), or if there is a tie, the closest sample. A value of K = 1 was observed to yield optimal performance for all adaptation approaches; Figure 61 illustrates the decrease in recognition performance observed with increasing K for the hybrid model adaptation approach.

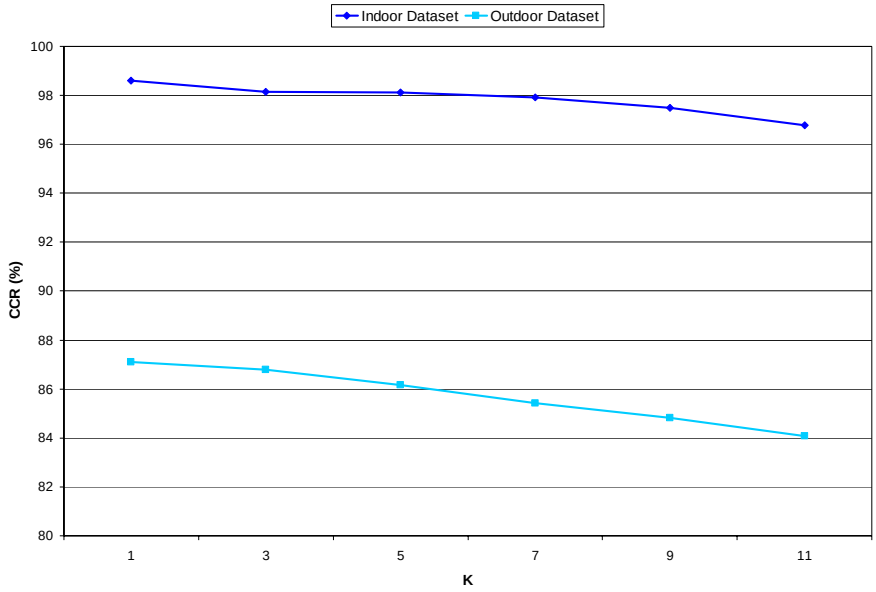


Figure 61: Variation of recognition rate with neighbourhood size K

Figures 62-64 plot the Cumulative Match Score (CMS) for the feature vectors extracted using global, local and hybrid approaches to model adaptation. The CMS curve indicates the probability that the correct identity for a given sample sequence is in the top *R* ranks. The CMS for $R = 1$ is equal to the Correct Classification Rate (CCR). The test database (Section 1.4) comprises 115 subjects, with 2163 gait sequences captured under controlled indoor conditions and 2657 gait sequences captured under outdoor conditions.

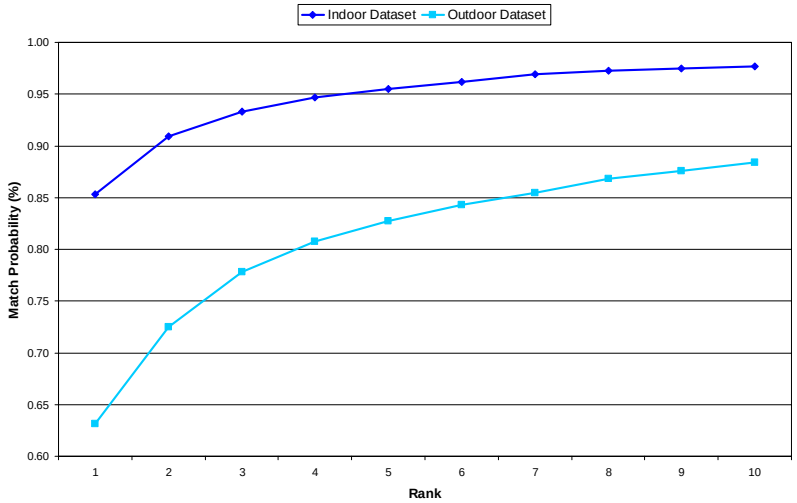


Figure 62: Cumulative match score for global model adaptation

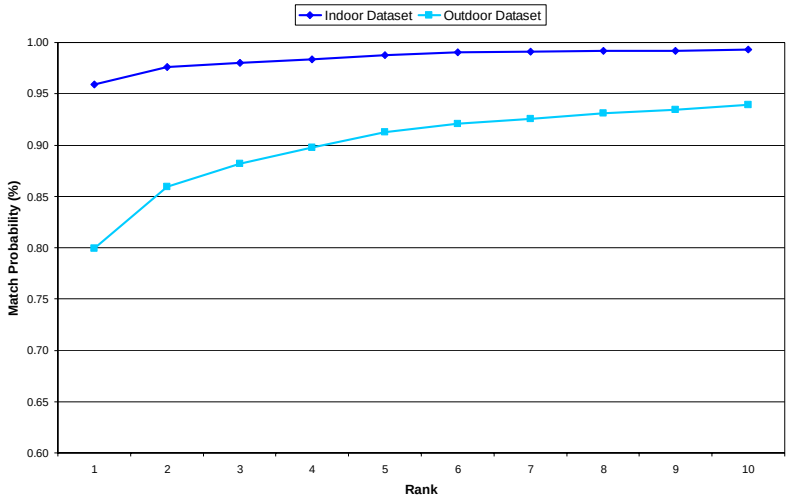


Figure 63: Cumulative match score for local model adaptation

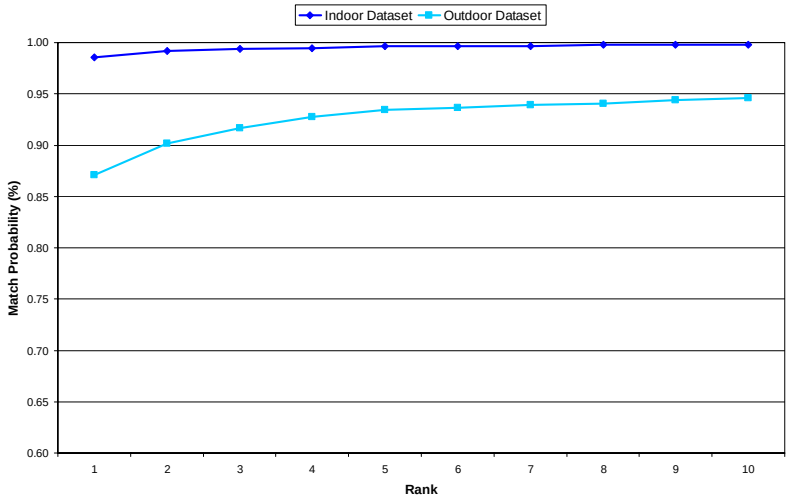


Figure 64: Cumulative match score for hybrid model adaptation

The performance of the global model adaptation approach (Section 4.2.1) equates to a CCR of 85.3% on the indoor dataset, reduced to 63.2% on the more difficult outdoor dataset. The local approach (Section 4.2.2) improves on this result, achieving a CCR of 95.9% on the indoor dataset and 79.9% on the outdoor dataset. Recognition performance is further improved using the hybrid approach to model adaptation (Chapter 5), resulting in a CCR of 98.6% on the indoor dataset and 87.1% on the outdoor dataset.

These results compare favourably with other published approaches tackling this database. In [Hayfron-Acquah 03] a CCR of 97.3% is achieved on a small subset of the indoor dataset, comprising 112 sequences from 28 subjects. In [Mowbray 04] a CCR of 86.2% is attained, from analysis of the half of the indoor dataset where the subject walks from right to left (1062 sequences from 115 subjects). In [Shutler 06] a CCR of 95.5% is

achieved on a subset of 200 sequences from 50 subjects from the indoor dataset. In [Veres 04] the whole indoor dataset is analysed, achieving a CCR of 100% for a feature vector of 4096 dimensions, and 95.2% when this feature vector is reduced to 100 features. By comparison, the hybrid model adaptation approach (Chapter 5) achieves a CCR of 98.6% with an 83-dimensional feature vector. At the time of writing, there are no other published approaches tackling gait recognition on the outdoor dataset for comparison.

The results presented in Figures 62-64 were generated using the whole feature vector, listed in Appendix C. This may be appropriate for some applications, such as comparing two sequences of video footage taken within a short period of time to ascertain if the subject is the same in both cases. However, it is not appropriate to use all of these features for some other applications. Some features, such as body shape, will change significantly from day to day according to clothing worn, and therefore may not be useful for applications such as entry control. Gait speed and frequency can be altered if the subject is in a hurry, or is tired. Different footwear, mood or injuries may alter some of the dynamic parameters, and all may change over extended periods of time. These covariate effects require further research to identify the extent of their influence and how these effects may be minimised (Section 7.2).

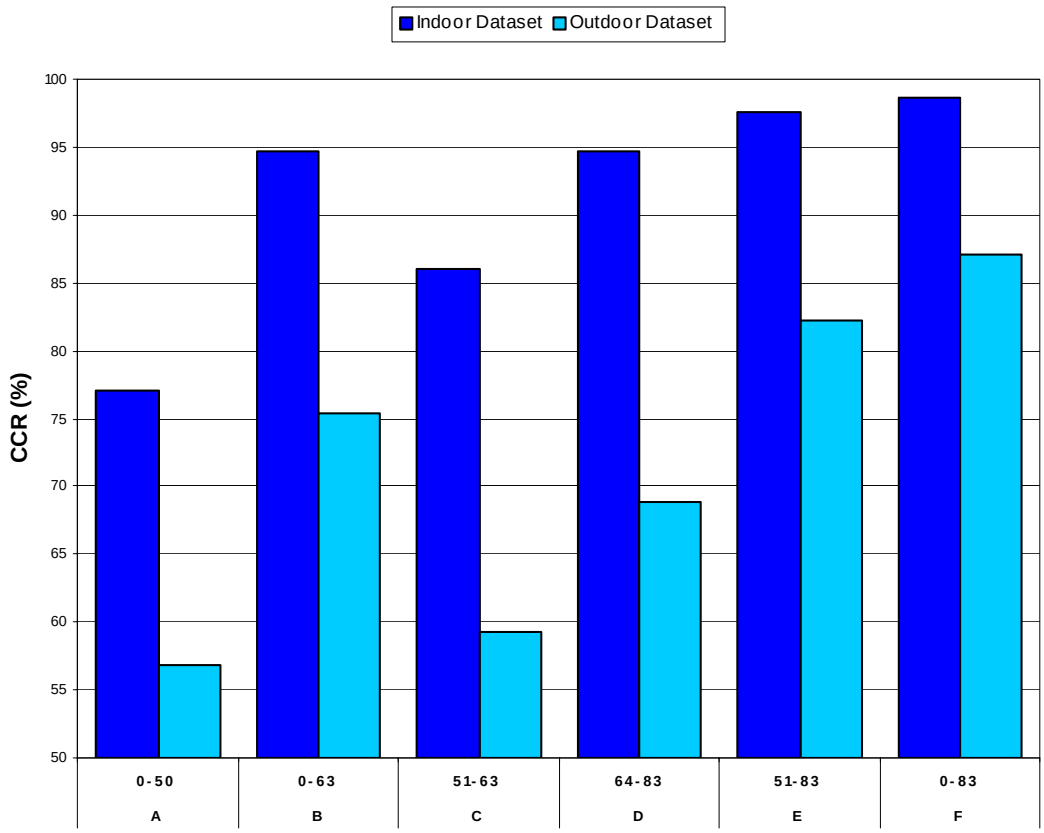


Figure 65: Recognition performance using subsets of the feature vector

Figure 65 summarises some additional results generated using the hybrid model adaptation approach, demonstrating recognition performance using subsets of the feature vector (Appendix C lists the full feature vector). Test A employs only dynamic parameters and derived measures (features 0-50), which are least likely to be affected by covariate factors. This test yields the worst result, faring particularly badly on the outdoor dataset. However, it may be possible to improve performance by performing some form of regularisation or data transformation on the joint rotation patterns extracted for each subject. Test B adds the static parameters computed in Chapter 3, including gait speed, frequency and geometric shape model parameters. Tests C and D isolate the respective performance of the global model parameters computed in Chapter 3 and the features derived from local contours in Chapter 5. These tests show that body shape alone is a potent feature for identification, although performance is greatly reduced on the outdoor dataset. Test E demonstrates recognition capability using both sets of static features, boosting performance particularly on the outdoor dataset. Test F is provided for reference purposes, illustrating the performance attained using the whole feature vector.

6.6. Additional Error Metrics

The performance of the hybrid model adaptation approach is evaluated further in this section using a simple statistical test. The standard deviation of leg or upper body width provides an approximate measure of the consistency of the extraction process. This is a relatively indirect method of measuring extraction performance, but it enables comparisons between the indoor and outdoor databases, and illustrates variation in extraction performance for different parts of the body. Standard deviation of contour width is computed as follows:

$$\sigma(n) = \frac{1}{NSub-1}\sum_{i}^{NSub-1}\sqrt{\frac{1}{NSeq(i)-1}\sum_{j}^{NSeq(i)-1}\left(x(n,i,j)-\mu(n,i)\right)^2} \tag{48}$$

Where $\sigma(n)$ is the standard deviation of mean contour width at sample index n, $NSub$ is the number of subjects, $NSeq(i)$ is the number of sequences featuring the subject i, $x(n, i, j)$ is the mean contour width at index n measured for sequence j of subject i and μ is equal to $mean_j(x(n, i))$. Low standard deviation is desirable, as this indicates that a similar body shape is extracted for each sequence of the same subject. High standard deviation in extracted shape would suggest inconsistencies in extraction for some sequences. Figure 66

plots the standard deviation of leg width at along the length of the leg, for the contours extracted using hybrid model adaptation. Figure 67 displays the same information for the upper body (the torso and head).

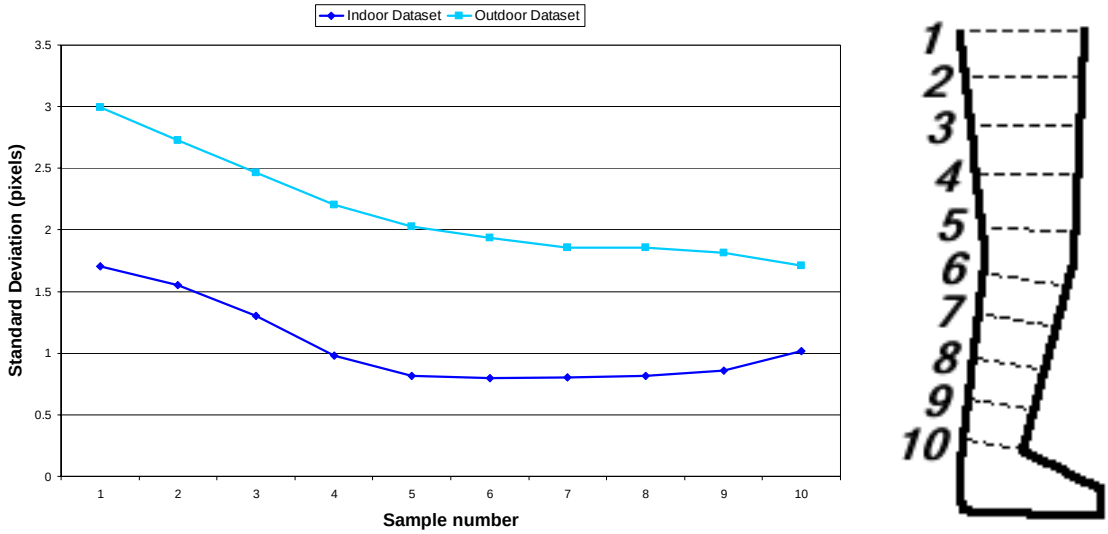


Figure 66: Consistency of leg shape extraction

The analysis of leg shape consistency reinforces the conclusions drawn from a visual examination of extraction performance, showing that estimation of leg shape is generally less reliable around hip level, where the arms occlude the legs and the thighs occlude each other. The reduced reliability of model estimation on the outdoor database is also clear from this analysis. For reference, note that the mean subject height extracted from the indoor dataset is approximately 330 pixels, and in the outdoor dataset where the subject is further from the camera, the mean extracted height is around 210 pixels.

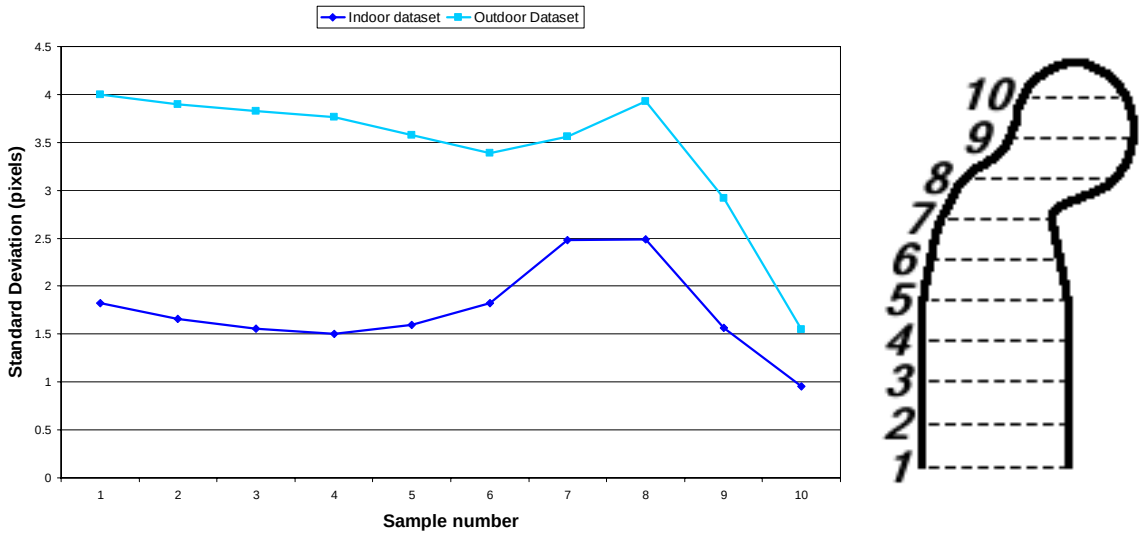


Figure 67: Upper body shape consistency

Figure 67 shows that variability in extracted upper body shape is highest around and below shoulder level. Consistency is highest at head level, but the discriminatory capability (Section 6.4) of these features is low. This may suggest that the head model is too simple, and consequently there is insufficient variation between subjects, or that there is simply little discriminative information in these head shape features. However, the research presented in [Veres 04] based on averaged silhouette features suggests that the head does carry useful discriminative information. Bearing this in mind, it is likely that either the head model is indeed too simple, or that the sampling method used to obtain head width features is insufficient and discriminatory capability is lost in the sampling process.

6.7. Conclusions on Performance

The impressions drawn from a visual analysis of the example sequences in Sections 6.2.1-6.2.3 support the superiority of a hybrid model adaptation process, reinforced by the recognition rates achieved in Section 6.5.

The global model adaptation approach is successful in extracting an approximation of the subject's shape and motion, but the limited precision of the model greatly reduces recognition capability. The global approach also limits the number of model parameters considered for adaptation, which contributes to a higher level of error. The advantage of this approach is that there are very few large errors, but instead small errors are distributed throughout the example sequences.

The local model adaptation approach has almost the opposite characteristics. In most frames of the example sequences shape and motion is extracted with a high degree of accuracy, but there are instances where the error in model configuration is quite large. These errors occur because the adaptation process is localised, and if the search does not begin in the right region of the model space, a good configuration cannot be determined.

The hybrid model adaptation approach solves these problems to some extent by applying both strategies in repeated succession. The performance analyses focus on this approach, demonstrating some of the failure modes encountered when processing the gait database. In general, complete failure of extraction occurs only when the initialisation process (Chapter 3) fails, usually because one or more of the assumptions listed in Section 1.4 has been violated. Other smaller errors occur when the global expectation process is

unable to generate an initialisation sufficiently close to the correct model configuration, with the result that a poor local configuration is accepted. This may indicate that the global model is too simple, and is unable to adequately represent the global characteristics of the subject, or that the local deviation of the subject from the model is larger than expected.

The inadequacy of the pre-processing algorithm is a major source of error. No colour information is employed, and the background estimation process is relatively naïve. The large number of shadows typical in the outdoor dataset is a particular problem, obscuring subject edges and introducing many spurious edges. Background estimation and pedestrian detection algorithms are covered in great detail in other areas of the literature, some of which are mentioned in Section 3.2.1. It is likely that performance could be greatly improved by incorporating this research into the pre-processing algorithm.

The hybrid model adaptation approach is relatively fast, but it is a long way from real-time performance. In addition, a global modelling strategy may require a time lag in processing to make use of data in neighbouring frames, perhaps of the order of a second. This should not be a problem for most biometric applications, which are unlikely to require much faster response times.

Analysis of the variance of the model features extracted revealed that body shape features contribute by far the most to discriminatory capability, and this conclusion is supported in the tests carried out in Section 6.5. Body shape parameters may be usable in their current form for some applications, for others it is likely that they would need to be normalised for clothing in some fashion (see Section 7.2). This may even be possible using additional sensing modalities, for example millimetre-wave or infrared imagery.

Recognition capability and consistency of extraction both suggest accurate and reliable model estimation, but this is not guaranteed. It is always possible to extract a model of the subject consistently badly. Though not a great concern for gait recognition applications, for the wider field of motion capture it is desirable to develop more authoritative measures of extraction performance.

Chapter 7. Conclusions and Future Work

7.1. Conclusions

This thesis examines the characteristics and applications of local and global modelling strategies in computer vision-based moving object analysis. It is shown that both have distinctive strengths and weaknesses, and that it is possible to combine the two strategies to attain improved performance in markerless motion capture.

New models are defined for the representation of human shape and motion, combining a geometric basis with deformable contours for efficient shape parameterisation. Prior knowledge of normal human anatomical proportions and gait motion is used to guide the model extraction process. An occlusion model provides additional guidance on the expected visibility of parts of the leg throughout the gait cycle, increasing the reliability of local search methods.

A simple global method for pedestrian detection and tracking is introduced, using temporal accumulation to estimate the mean shape of the subject. Periodicity information is derived by masking specific areas of the subject's legs, applying knowledge of the normal human gait pattern to predict the approximate motion of the subject's legs.

This initial model forms the basis on which a model adaptation process operates, generating a more accurate representation of the subject. Two basic approaches are introduced, employing a global or a local modelling strategy. The global approach is generally resistant to noise, as any missing data can be compensated for by data surrounding it in space and time. However, the model is highly constrained to restrict computational requirements and it is only possible to construct a crude approximation of the subject. The local approach on the other hand only considers one frame at a time, and a vastly reduced quantity of image data. As a result it is possible to generate a much more accurate representation of the subject. This approach suffers from the disadvantage that it is difficult to cope with missing data, as there is less surrounding data to draw on. As a result, some models extracted are very poor representations of the subject.

A novel hybrid model adaptation approach is introduced, which combines the two basic strategies in an attempt to eliminate their respective weaknesses, while retaining their strengths. Local adaptation is employed to find the best model configuration nearby the

initialisation, and the global strategy is limited to deciding which of these configurations are reliable, and generating a new initialisation from these local models. These processes are repeated in an iterative process, until the global model generated from the data reaches a stable point, or a maximum number of iterations are reached.

The performance of model extraction is estimated through visual analysis, recognition capability and statistically measured consistency. The hybrid model adaptation approach achieves a CCR of 98.6% on data captured within the laboratory under controlled conditions, which compares favourably with other published approaches. The recognition rate is reduced to 87.1% under outdoor conditions where there is greater variability, and is the first approach to tackle this dataset. These results were achieved on a large database of 115 subjects, with around 20 gait sequences captured per subject for each condition.

One of the advantages of a model-based approach is that each parameter is directly related to known subject properties, such as their leg shape or the degree to which their upper body rises and falls during gait. ANOVA performed on the feature vectors extracted from the gait database indicates that a great deal of recognition capability is derived from shape information. This is of some concern, as body shape may change drastically according to clothing worn. Recognition based only on dynamic parameters results in greatly reduced performance. However, it should be possible to normalise most of these parameters for covariate factors, and use them to some degree for recognition. It is also important to note that there are certain applications where this is not an important issue, for example tracking a particular person among multiple cameras, using their gait and whole-body appearance to confirm identity.

This thesis demonstrates that it is possible to apply global and local evidence gathering algorithms to a large database of walking people, and achieve a high level of accuracy in recognition using body shape and characteristic gait motion.

7.2. Future Work

Three general areas may be identified as the primary directions for future research:

1. **Generalisation of approach to less constrained capture conditions**
2. **Examination of covariate effects**
3. **Generalisation of approach to arbitrary human motion**

The first area is a practical issue for the implementation of smart surveillance systems. Although capture conditions can be controlled to some extent in entry control systems and in the placement of security cameras, generally it is not possible to guarantee a particular viewpoint on the subject. The approach presented within this thesis employs a 2D model, fitted to subjects walking fronto-parallel to the camera view plane. This approach could certainly be generalised to other viewpoints, but it is likely that a 3D model would be necessary. Additional research is required to extend the approach to a 3D space, and to cope with the resulting increase in the number of model parameters required. It is also desirable to integrate this approach with existing research on pedestrian tracking algorithms, which are capable of dealing with the crowded scenes that may be encountered in surveillance imagery.

The second area relates to the effect of changing variables such as clothing, footwear and loading, changes over extended periods of time, and even changes due to mood or illness. Appendix C includes a preliminary analysis for the SDB, which confirms that covariate factors significantly affect the discriminatory capability of model features. Although some research has been conducted into this area (see Section 1.3.2), current research has been mostly limited to appearance-based approaches, and little work has been done in the area of normalising gait descriptions for these effects. However, we should bear in mind that other biometrics have similar factors to consider. Facial appearance for example will vary according to expression, presence of glasses, hairstyle and son on. There are also blind people who do not have usable iris patterns, manual labourers without usable fingerprints, and children who are unable to stay still long enough for biometric samples to be taken. This is not a problem unique to gait.

The final area of research is in extending this approach, which is currently based on the analysis of walking subjects, to analysis of arbitrary human motion. It may be possible to measure mean motion patterns for other activities such as in sports or dancing, which could replace the mean gait patterns. For completely general motion extraction it is necessary to consider the case where no prior motion model is available, and one must be constructed from available image data. This is a difficult problem, and one that is unlikely to be solved in the near future. However, it is in this area that the widest class of potential applications become available, and is our eventual goal.

Appendix A:
Pre-processing Results

The following figures illustrate the typical edge data retrieved from sequences in the indoor and outdoor datasets. Each sequence is sampled at a rate of one frame in ten, corresponding to the example sequences provided in Sections 6.2.1-6.2.3. The pre-processing method used to obtain this data is described in Section 3.2.1.

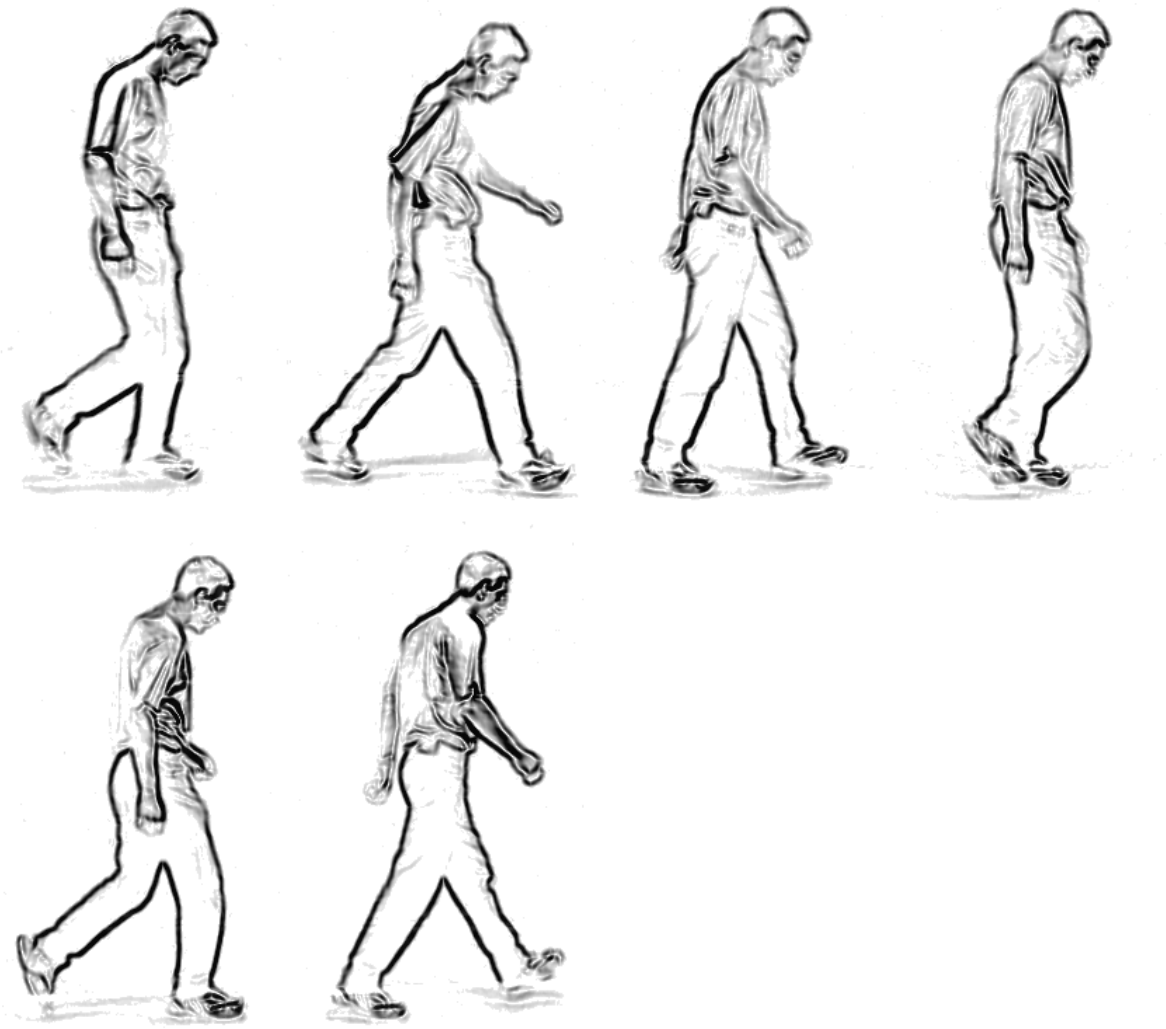


Figure 68: Pre-processed image data, sequence '008a013s00R'

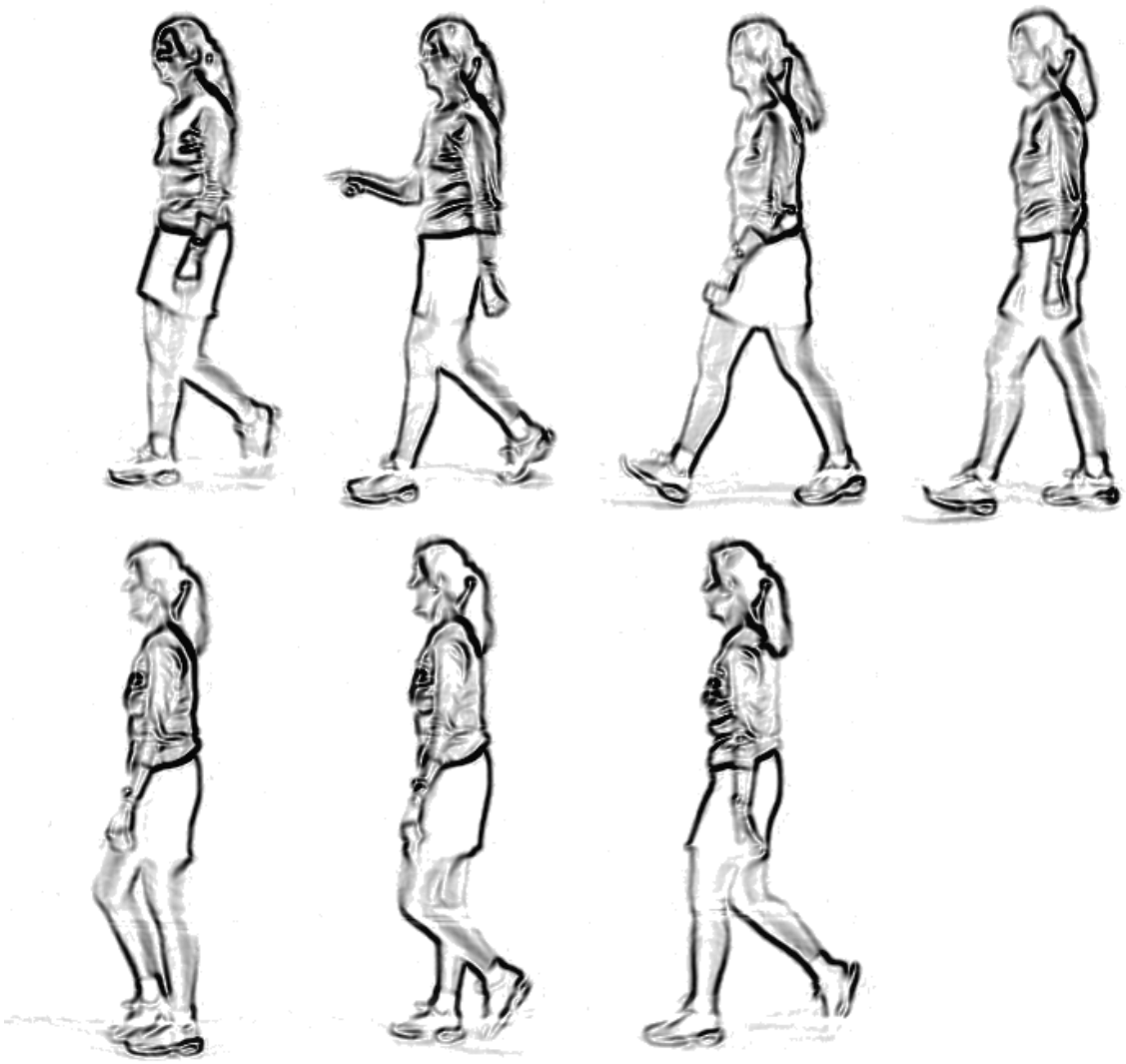


Figure 69: Pre-processed image data, sequence ‘012a033s00L’

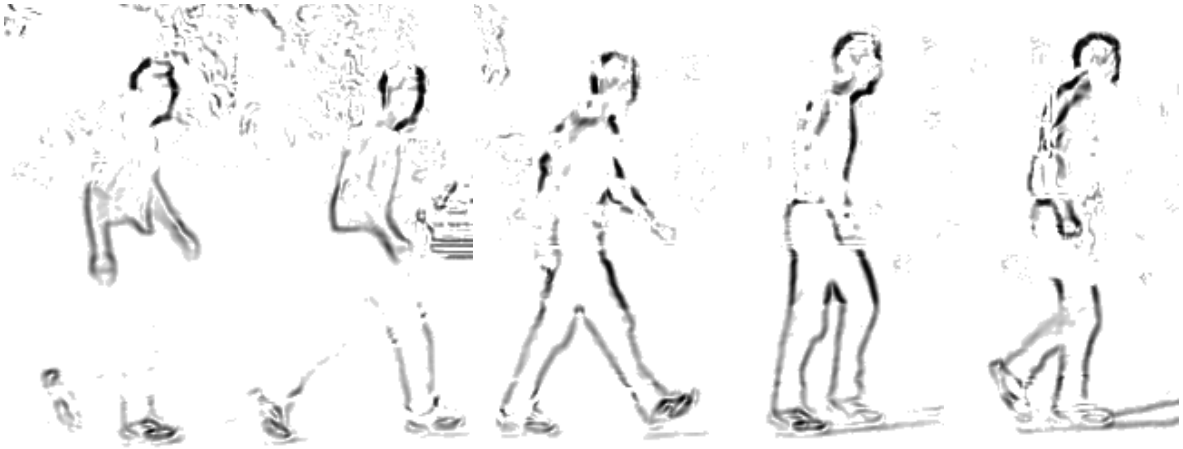


Figure 70: Pre-processed image data, sequence ‘008e013s00R’

Figure 71: Pre-processed image data, sequence ‘012e033s00L’

Appendix B:
Normalisation for Periodicity Analysis

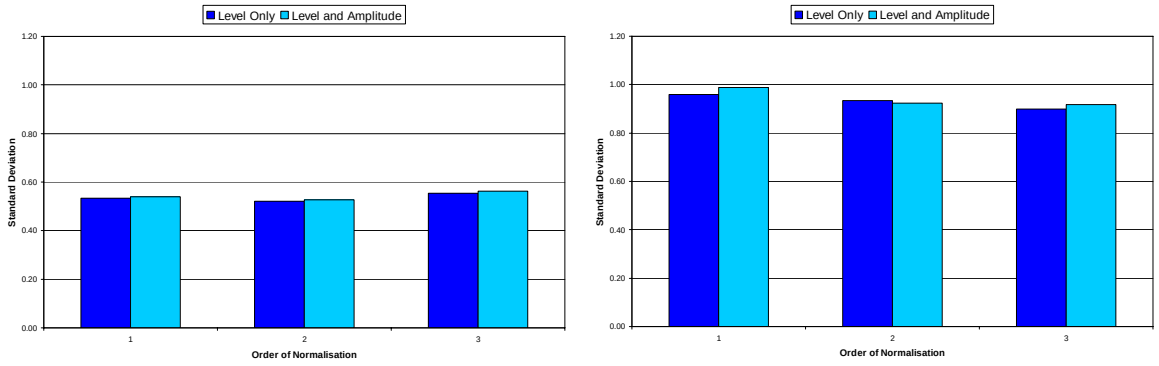
Section 3.2.2 describes a normalisation operation using low-order polynomials to model and remove noise from an underlying sinusoidal sequence:

$$S_N = \frac{S_M - p_1(S_M)}{p_2(\,|S_M - p_1(S_M)|\,)} \tag{13}$$

Where S_N is the normalised sequence, S_M is the measured sequence and $p_1(x)$ and $p_2(x)$ denote the best polynomial fit to sequence x, computed by least-squares linear regression [Fox 97]. The order of the polynomials p_1 and p_2 control the degree of normalisation applied. However, the utility of the second normalisation term p_2 is unclear. To investigate this component, periodicity extraction performance is measured indirectly through the within-subject standard deviation of period (see Section 3.2.2):

$$MSTDV = \frac{1}{NSub}\sum_i stdv(i) \tag{18}$$

Where $MSTDV$ is the mean within-subject standard deviation in extracted gait period, $NSub$ is the total number of subjects, and s is the standard deviation in period for subject i. Figure 72 compares performance when the normalisation process corrects level only (order of $p_2 = 0$), and when both level and amplitude variations are removed (order of p_2 = order of p_1). These results demonstrate that the inclusion of p_2 in the normalisation operation in all cases results in only a marginal difference, and is usually counter-productive.



(a) Indoor dataset (b) Outdoor Dataset

Figure 72: Consistency of period extraction

Appendix C:
Full ANOVA Results

The following tables list the ANOVA *f*-statistic computed for each feature extracted from the LDB using the global, local and hybrid model adaptation approaches. Note that the first two approaches do not determine *y*-motion amplitude, and so the *f*-statistic of this feature is listed as zero. The global model adaptation approach extracts only geometric shape features, so the contour width features are similarly listed as zero.

A preliminary analysis is also included for the SDB (Table 8), which comprises 3177 sequences from 10 subjects walking under indoor capture conditions. Unlike the LDB, this database includes covariate factors such as different walking speeds, footwear, clothing and loading for each subject (see Section 1.4). Hybrid model adaptation was used to generate the feature vectors for this dataset. Compared with the features extracted from the indoor dataset of the LDB, it is clear that the discriminative capability of most of the model features is severely reduced on this dataset. No attempt has yet been made to normalise extracted features for covariate effects, so this result is not surprising.

The discriminatory capability of some ankle rotation features improve on this dataset, suggesting that these features may become more important when different forms of footwear or gait speeds are considered. Some upper body shape features also improve on this dataset, though the reason for this is unclear. The majority of features however have reduced discriminatory capability. Gait speed and frequency in particular are almost useless for this dataset, as each subject varies their walking speed.

Research by [Tanawongsuwan 03] and [Yam 04] suggests that it may be possible to normalise gait dynamics for different walking speeds. Different sensing modalities such as millimetre-wave or infrared imagery may enable clothing-independent estimation of shape parameters. Further research (Section 7.2) may also reveal other methods for removing covariate effects from the feature vector.

Table 5: ANOVA for features extracted using global model adaptation

x	Feature Name	F-statistic		x	Feature Name	F-statistic	
		Indoor	Outdoor			Indoor	Outdoor
0	Hip Rotation Θ(00)	21.2	11.1	42	Ankle Rotation Θ(12)	2.5	2.9
1	Hip Rotation Θ(01)	14.7	8.4	43	Ankle Rotation Θ(13)	2.2	3.6
2	Hip Rotation Θ(02)	7.6	6.2	44	Ankle Rotation Θ(14)	2.2	3.4
3	Hip Rotation Θ(03)	9.2	7.6	45	Hip Rotation Mean	17.7	9.3
4	Hip Rotation Θ(04)	11.8	6.6	46	Hip Rotation Amplitude	14.6	8.8
5	Hip Rotation Θ(05)	10.0	5.4	47	Knee Rotation Mean	13.6	10.8
6	Hip Rotation Θ(06)	8.5	4.8	48	Knee Rotation Amplitude	22.9	9.2
7	Hip Rotation Θ(07)	9.0	4.9	49	Ankle Rotation Mean	4.2	2.3
8	Hip Rotation Θ(08)	12.2	8.0	50	Ankle Rotation Amplitude	2.3	2.4
9	Hip Rotation Θ(09)	14.5	11.4	51	Gait Speed	102.8	23.1
10	Hip Rotation Θ(10)	13.4	9.4	52	Gait Frequency ω	146.4	84.2
11	Hip Rotation Θ(11)	10.4	8.2	53	Y Motion Amplitude A_y	0.0	0.0
12	Hip Rotation Θ(12)	11.3	8.5	54	Torso Height TW	17.1	17.7
13	Hip Rotation Θ(13)	12.7	11.0	55	Torso Width TW	42.8	16.4
14	Hip Rotation Θ(14)	14.2	11.9	56	Head Height HH	14.4	6.3
15	Knee Rotation Θ(00)	22.2	7.9	57	Head Width HW	24.0	9.6
16	Knee Rotation Θ(01)	5.1	8.1	58	Head dx HDX	64.5	20.4
17	Knee Rotation Θ(02)	4.1	6.4	59	Head dy HDY	16.2	5.3
18	Knee Rotation Θ(03)	3.6	5.6	60	Hip Width LWH	55.2	18.5
19	Knee Rotation Θ(04)	5.3	5.5	61	Knee Width LWKU	69.5	32.2
20	Knee Rotation Θ(05)	4.5	4.2	62	Knee Width LWKL	203.7	43.3
21	Knee Rotation Θ(06)	4.4	4.3	63	Ankle Width LWA	188.7	36.5
22	Knee Rotation Θ(07)	5.5	5.3	64	Leg Width 00	0.0	0.0
23	Knee Rotation Θ(08)	11.5	6.6	65	Leg Width 01	0.0	0.0
24	Knee Rotation Θ(09)	18.4	9.9	66	Leg Width 02	0.0	0.0
25	Knee Rotation Θ(10)	19.3	10.2	67	Leg Width 03	0.0	0.0
26	Knee Rotation Θ(11)	15.0	9.2	68	Leg Width 04	0.0	0.0
27	Knee Rotation Θ(12)	9.2	7.0	69	Leg Width 05	0.0	0.0
28	Knee Rotation Θ(13)	8.6	6.2	70	Leg Width 06	0.0	0.0
29	Knee Rotation Θ(14)	14.8	8.3	71	Leg Width 07	0.0	0.0
30	Ankle Rotation Θ(00)	2.9	2.4	72	Leg Width 08	0.0	0.0
31	Ankle Rotation Θ(01)	3.4	2.3	73	Leg Width 09	0.0	0.0
32	Ankle Rotation Θ(02)	3.3	2.7	74	Upper Body Width 00	0.0	0.0
33	Ankle Rotation Θ(03)	3.1	2.9	75	Upper Body Width 01	0.0	0.0
34	Ankle Rotation Θ(04)	3.3	3.1	76	Upper Body Width 02	0.0	0.0
35	Ankle Rotation Θ(05)	3.7	2.4	77	Upper Body Width 03	0.0	0.0
36	Ankle Rotation Θ(06)	3.3	1.5	78	Upper Body Width 04	0.0	0.0
37	Ankle Rotation Θ(07)	3.0	2.0	79	Upper Body Width 05	0.0	0.0
38	Ankle Rotation Θ(08)	2.6	2.8	80	Upper Body Width 06	0.0	0.0
39	Ankle Rotation Θ(09)	2.2	2.8	81	Upper Body Width 07	0.0	0.0
40	Ankle Rotation Θ(10)	1.9	2.0	82	Upper Body Width 08	0.0	0.0
41	Ankle Rotation Θ(11)	2.0	1.9	83	Upper Body Width 09	0.0	0.0

Table 6: ANOVA for features extracted using local model adaptation

x	Feature Name	f-statistic		x	Feature Name	f-statistic	
		Indoor	Outdoor			Indoor	Outdoor
0	Hip Rotation Θ(00)	45.5	17.7	42	Ankle Rotation Θ(12)	3.1	4.6
1	Hip Rotation Θ(01)	38.1	21.5	43	Ankle Rotation Θ(13)	2.7	5.2
2	Hip Rotation Θ(02)	29.5	19.3	44	Ankle Rotation Θ(14)	2.2	5.7
3	Hip Rotation Θ(03)	29.9	18.5	45	Hip Rotation Mean	48.3	25.5
4	Hip Rotation Θ(04)	31.3	16.6	46	Hip Rotation Amplitude	23.4	16.2
5	Hip Rotation Θ(05)	28.8	15.6	47	Knee Rotation Mean	50.6	37.2
6	Hip Rotation Θ(06)	22.5	15.3	48	Knee Rotation Amplitude	56.7	37.0
7	Hip Rotation Θ(07)	14.3	18.6	49	Ankle Rotation Mean	4.4	5.8
8	Hip Rotation Θ(08)	19.4	20.2	50	Ankle Rotation Amplitude	2.0	4.0
9	Hip Rotation Θ(09)	25.6	20.9	51	Gait Speed	102.8	23.1
10	Hip Rotation Θ(10)	27.0	22.9	52	Gait Frequency ω	146.4	84.2
11	Hip Rotation Θ(11)	27.9	11.2	53	Y Motion Amplitude A_y	0.0	0.0
12	Hip Rotation Θ(12)	38.9	14.5	54	Torso Height TW	17.1	17.7
13	Hip Rotation Θ(13)	39.2	13.9	55	Torso Width TW	42.8	16.4
14	Hip Rotation Θ(14)	41.2	13.5	56	Head Height HH	14.4	6.3
15	Knee Rotation Θ(00)	50.1	40.6	57	Head Width HW	24.0	9.6
16	Knee Rotation Θ(01)	37.9	45.0	58	Head dx HDX	64.5	20.4
17	Knee Rotation Θ(02)	42.4	48.3	59	Head dy HDY	16.2	5.3
18	Knee Rotation Θ(03)	50.7	51.4	60	Hip Width LWH	58.0	20.3
19	Knee Rotation Θ(04)	41.6	42.5	61	Knee Width LWKU	71.5	31.6
20	Knee Rotation Θ(05)	31.0	22.1	62	Knee Width LWKL	202.1	44.5
21	Knee Rotation Θ(06)	22.9	22.9	63	Ankle Width LWA	202.0	36.7
22	Knee Rotation Θ(07)	20.0	27.3	64	Leg Width 00	65.2	18.4
23	Knee Rotation Θ(08)	28.3	23.8	65	Leg Width 01	81.3	27.2
24	Knee Rotation Θ(09)	54.1	24.1	66	Leg Width 02	100.2	36.1
25	Knee Rotation Θ(10)	63.7	25.8	67	Leg Width 03	136.2	47.8
26	Knee Rotation Θ(11)	37.1	16.5	68	Leg Width 04	172.1	59.0
27	Knee Rotation Θ(12)	28.7	9.4	69	Leg Width 05	186.1	56.2
28	Knee Rotation Θ(13)	34.3	16.0	70	Leg Width 06	202.9	52.4
29	Knee Rotation Θ(14)	42.6	26.9	71	Leg Width 07	304.5	54.0
30	Ankle Rotation Θ(00)	2.7	5.6	72	Leg Width 08	358.2	52.6
31	Ankle Rotation Θ(01)	3.8	5.0	73	Leg Width 09	206.2	44.6
32	Ankle Rotation Θ(02)	3.9	4.1	74	Upper Body Width 00	89.0	14.8
33	Ankle Rotation Θ(03)	3.5	3.2	75	Upper Body Width 01	80.7	16.9
34	Ankle Rotation Θ(04)	3.8	3.3	76	Upper Body Width 02	77.8	17.5
35	Ankle Rotation Θ(05)	4.2	3.0	77	Upper Body Width 03	80.0	19.3
36	Ankle Rotation Θ(06)	3.7	2.3	78	Upper Body Width 04	83.2	19.3
37	Ankle Rotation Θ(07)	2.7	2.3	79	Upper Body Width 05	59.6	19.1
38	Ankle Rotation Θ(08)	1.8	3.2	80	Upper Body Width 06	48.0	14.1
39	Ankle Rotation Θ(09)	1.5	3.9	81	Upper Body Width 07	28.9	9.7
40	Ankle Rotation Θ(10)	1.4	3.6	82	Upper Body Width 08	39.7	11.2
41	Ankle Rotation Θ(11)	1.7	3.9	83	Upper Body Width 09	27.7	8.2

Table 7: ANOVA for features extracted using hybrid model adaptation

x	Feature Name	f-statistic		x	Feature Name	f-statistic	
		Indoor	Outdoor			Indoor	Outdoor
0	Hip Rotation Θ(00)	40.3	5.8	42	Ankle Rotation Θ(12)	11.0	3.4
1	Hip Rotation Θ(01)	50.1	7.7	43	Ankle Rotation Θ(13)	7.0	1.9
2	Hip Rotation Θ(02)	46.7	12.6	44	Ankle Rotation Θ(14)	2.8	1.8
3	Hip Rotation Θ(03)	38.5	23.0	45	Hip Rotation Mean	35.7	8.6
4	Hip Rotation Θ(04)	25.7	19.9	46	Hip Rotation Amplitude	45.5	17.6
5	Hip Rotation Θ(05)	23.7	10.9	47	Knee Rotation Mean	31.2	7.5
6	Hip Rotation Θ(06)	31.5	14.9	48	Knee Rotation Amplitude	54.0	18.0
7	Hip Rotation Θ(07)	37.1	16.9	49	Ankle Rotation Mean	19.6	1.9
8	Hip Rotation Θ(08)	32.6	17.3	50	Ankle Rotation Amplitude	1.6	6.8
9	Hip Rotation Θ(09)	28.1	17.8	51	Gait Speed	102.8	21.6
10	Hip Rotation Θ(10)	18.3	15.9	52	Gait Frequency ω	146.4	84.2
11	Hip Rotation Θ(11)	24.9	12.1	53	Y Motion Amplitude A_y	87.3	24.9
12	Hip Rotation Θ(12)	38.1	9.0	54	Torso Height TW	17.1	17.7
13	Hip Rotation Θ(13)	45.2	6.9	55	Torso Width TW	42.8	16.4
14	Hip Rotation Θ(14)	40.8	5.6	56	Head Height HH	14.4	6.3
15	Knee Rotation Θ(00)	44.7	11.2	57	Head Width HW	24.0	9.6
16	Knee Rotation Θ(01)	33.9	20.4	58	Head dx HDX	64.5	20.4
17	Knee Rotation Θ(02)	31.4	19.3	59	Head dy HDY	16.2	5.3
18	Knee Rotation Θ(03)	34.0	16.2	60	Hip Width LWH	58.8	20.5
19	Knee Rotation Θ(04)	30.6	12.8	61	Knee Width LWKU	73.2	32.4
20	Knee Rotation Θ(05)	20.0	7.7	62	Knee Width LWKL	202.6	42.4
21	Knee Rotation Θ(06)	24.8	8.3	63	Ankle Width LWA	203.1	35.9
22	Knee Rotation Θ(07)	29.9	7.9	64	Leg Width 00	116.8	32.9
23	Knee Rotation Θ(08)	41.9	6.6	65	Leg Width 01	122.0	40.0
24	Knee Rotation Θ(09)	47.7	5.1	66	Leg Width 02	109.1	49.7
25	Knee Rotation Θ(10)	40.3	4.6	67	Leg Width 03	135.5	60.7
26	Knee Rotation Θ(11)	31.7	5.4	68	Leg Width 04	256.5	64.0
27	Knee Rotation Θ(12)	26.9	7.7	69	Leg Width 05	285.8	67.1
28	Knee Rotation Θ(13)	41.5	19.4	70	Leg Width 06	280.2	62.6
29	Knee Rotation Θ(14)	47.6	14.2	71	Leg Width 07	293.4	56.4
30	Ankle Rotation Θ(00)	3.2	1.9	72	Leg Width 08	406.6	62.8
31	Ankle Rotation Θ(01)	4.6	1.9	73	Leg Width 09	330.5	59.3
32	Ankle Rotation Θ(02)	10.3	1.9	74	Upper Body Width 00	96.9	16.8
33	Ankle Rotation Θ(03)	3.5	2.2	75	Upper Body Width 01	105.4	17.9
34	Ankle Rotation Θ(04)	1.3	2.3	76	Upper Body Width 02	97.0	20.6
35	Ankle Rotation Θ(05)	2.0	2.4	77	Upper Body Width 03	98.5	22.2
36	Ankle Rotation Θ(06)	7.6	2.4	78	Upper Body Width 04	104.0	23.1
37	Ankle Rotation Θ(07)	16.2	2.6	79	Upper Body Width 05	66.8	21.1
38	Ankle Rotation Θ(08)	5.6	3.0	80	Upper Body Width 06	44.7	15.2
39	Ankle Rotation Θ(09)	3.7	4.1	81	Upper Body Width 07	28.0	10.7
40	Ankle Rotation Θ(10)	4.3	4.8	82	Upper Body Width 08	25.8	9.1
41	Ankle Rotation Θ(11)	12.4	3.3	83	Upper Body Width 09	25.9	8.3

Table 8: ANOVA for features extracted using hybrid model adaptation (SDB)

x	Feature Name	f-statistic		x	Feature Name	f-statistic	
		Indoor	SDB			Indoor	SDB
0	Hip Rotation Θ(00)	40.3	13.4	42	Ankle Rotation Θ(12)	11.0	0.4
1	Hip Rotation Θ(01)	50.1	18.9	43	Ankle Rotation Θ(13)	7.0	3.8
2	Hip Rotation Θ(02)	46.7	25.5	44	Ankle Rotation Θ(14)	2.8	9.1
3	Hip Rotation Θ(03)	38.5	19.3	45	Hip Rotation Mean	35.7	1.8
4	Hip Rotation Θ(04)	25.7	10.6	46	Hip Rotation Amplitude	45.5	0.8
5	Hip Rotation Θ(05)	23.7	9.5	47	Knee Rotation Mean	31.2	3.3
6	Hip Rotation Θ(06)	31.5	15.5	48	Knee Rotation Amplitude	54.0	11.7
7	Hip Rotation Θ(07)	37.1	17.0	49	Ankle Rotation Mean	19.6	7.7
8	Hip Rotation Θ(08)	32.6	15.5	50	Ankle Rotation Amplitude	1.6	2.8
9	Hip Rotation Θ(09)	28.1	7.9	51	Gait Speed	102.8	0.6
10	Hip Rotation Θ(10)	18.3	6.8	52	Gait Frequency ω	146.4	5.1
11	Hip Rotation Θ(11)	24.9	6.8	53	Y Motion Amplitude A_y	87.3	29.5
12	Hip Rotation Θ(12)	38.1	10.2	54	Torso Height TW	17.1	6.0
13	Hip Rotation Θ(13)	45.2	17.7	55	Torso Width TW	42.8	38.5
14	Hip Rotation Θ(14)	40.8	15.9	56	Head Height HH	14.4	4.8
15	Knee Rotation Θ(00)	44.7	27.4	57	Head Width HW	24.0	0.9
16	Knee Rotation Θ(01)	33.9	24.3	58	Head dx HDX	64.5	13.8
17	Knee Rotation Θ(02)	31.4	5.6	59	Head dy HDY	16.2	14.3
18	Knee Rotation Θ(03)	34.0	7.7	60	Hip Width LWH	58.8	26.4
19	Knee Rotation Θ(04)	30.6	10.1	61	Knee Width LWKU	73.2	8.5
20	Knee Rotation Θ(05)	20.0	10.1	62	Knee Width LWKL	202.6	7.0
21	Knee Rotation Θ(06)	24.8	15.2	63	Ankle Width LWA	203.1	8.9
22	Knee Rotation Θ(07)	29.9	17.3	64	Leg Width 00	116.8	26.3
23	Knee Rotation Θ(08)	41.9	16.3	65	Leg Width 01	122.0	21.0
24	Knee Rotation Θ(09)	47.7	21.5	66	Leg Width 02	109.1	11.0
25	Knee Rotation Θ(10)	40.3	26.5	67	Leg Width 03	135.5	33.7
26	Knee Rotation Θ(11)	31.7	25.4	68	Leg Width 04	256.5	77.0
27	Knee Rotation Θ(12)	26.9	3.2	69	Leg Width 05	285.8	100.7
28	Knee Rotation Θ(13)	41.5	4.3	70	Leg Width 06	280.2	30.7
29	Knee Rotation Θ(14)	47.6	8.6	71	Leg Width 07	293.4	18.3
30	Ankle Rotation Θ(00)	3.2	0.5	72	Leg Width 08	406.6	15.4
31	Ankle Rotation Θ(01)	4.6	1.3	73	Leg Width 09	330.5	12.6
32	Ankle Rotation Θ(02)	10.3	1.9	74	Upper Body Width 00	96.9	23.8
33	Ankle Rotation Θ(03)	3.5	3.5	75	Upper Body Width 01	105.4	39.8
34	Ankle Rotation Θ(04)	1.3	0.9	76	Upper Body Width 02	97.0	43.4
35	Ankle Rotation Θ(05)	2.0	13.4	77	Upper Body Width 03	98.5	52.4
36	Ankle Rotation Θ(06)	7.6	18.9	78	Upper Body Width 04	104.0	63.9
37	Ankle Rotation Θ(07)	16.2	25.5	79	Upper Body Width 05	66.8	98.9
38	Ankle Rotation Θ(08)	5.6	19.3	80	Upper Body Width 06	44.7	164.5
39	Ankle Rotation Θ(09)	3.7	10.6	81	Upper Body Width 07	28.0	7.3
40	Ankle Rotation Θ(10)	4.3	9.5	82	Upper Body Width 08	25.8	5.3
41	Ankle Rotation Θ(11)	12.4	15.5	83	Upper Body Width 09	25.9	4.1

References

[Aggarwal 99] J K Aggarwal and Q Cai. "Human Motion Analysis: A Review." *Computer Vision and Image Understanding*, **73**(3):428-440, 1999.

[Al-Mazeed 03] A H Al-Mazeed, M S Nixon and S R Gunn. "Fusing Complementary Operators to Enhance Foreground/Background Segmentation." *Proc. British Machine Vision Conference*, 501-510, 2003.

[Angeloni 94] C Angeloni, P O Riley and E D Krebs. "Frequency Content of Whole Body Gait Kinematic Data." *IEEE Trans. Rehabilitation Engineering*, **2**(1):40-46, 1994.

[Ayyappa 97] E Ayyappa. "Normal Human Locomotion, Part 1: Basic Concepts and Terminology." *Journal of Prosthetics and Orthotics*, **9**(1):10-17, 1997.

[Baumberg 94] A Baumberg and D Hogg. "Learning Flexible Models from Image Sequences." *Proc. European Conference on Computer Vision*, 299-308, 1994.

[Bazin 05] A I Bazin, L Middleton and M S Nixon. "Probabilistic Fusion of Gait Features for Biometric Verification." *Proc. Information Fusion*, 2005.

[BenAbdelkader 02] C BenAbdelkader, R Cutler and L Davis. "Stride and Cadence as a Biometric in Automatic Person Identification and Verification." *Proc. Automatic Face and Gesture Recognition*, 372-377, 2002.

[BenAbdelkader04] C BenAbdelkader, R Cutler and L Davis. "Gait Recognition Using Image Self-Similarity." *Applied Signal Processing*, **4**:1-14, 2004.

[Bregler 04] C Bregler, J Malik and K Pullen. "Twist Based Acquisition and Tracking of Animal and Human Kinematics." *International Journal of Computer Vision*, **56**(3):179-194, 2004.

[Cheung 05a] G K M Cheung, S Baker and T Kanade. “Shape-From-Silhouette Across Time Part I: Theory and Algorithms.” *International Journal of Computer Vision*, **62**(3):221-247, 2005.

[Cheung 05b] G K M Cheung, S Baker and T Kanade. “Shape-From-Silhouette Across Time Part II: Applications to Human Modeling and Markerless Motion Tracking.” *International Journal of Computer Vision*, **63**(3):225-245, 2005.

[Collins 02] R T Collins, R Gross and J Shi. “Silhouette-based Human Identification from Body Shape and Gait.” *Proc. Automatic Face and Gesture Recognition*, 351-356, 2002.

[Cootes 92] T J Cootes, C J Taylor, D H Cooper and J Graham. “Training Models of Shapes From Sets of Examples.” *Proc. British Machine Vision Conference*, 9-18, 1992.

[Cunado 03] D Cunado, M S Nixon and J N Carter. “Automatic Extraction and Description of Human Gait Models for Recognition Purposes.” *Computer Vision and Image Understanding*, **90**(1):1-41, 2003.

[Cutler 00] R Cutler and L Davis. “Robust Real-Time Periodic Motion Detection, Analysis, and Applications.” *IEEE Trans. Pattern Analysis and Machine Intelligence*, **22**(8):781-796, 2000.

[Davison 01] A J Davison, J Deutscher and I D Reid. “Markerless Motion Capture of Complex Full-Body Movement for Character Animation.” *Proc. Computer Animation and Simulation*, 3-14, 2001.

[Dempster 77] A P Dempster, N M Laird and D B Rubin. “Maximum-likelihood from incomplete data via the EM algorithm.” *Journal of the Royal Statistical Society B*, **39**:1-38, 1977.

[Dontcheva 03] M Dontcheva, G Yngve and Z Popovic. “Layered Acting for Character Animation.” *ACM Trans. Graphics*, **22**(3):409-416, 2003.

[Drummond 02] T Drummond and R Cipolla. "Real-Time Visual Tracking of Complex Structures." *IEEE Trans. Pattern Analysis and Machine Intelligence*, **24**(7):932-946, 2002.

[Elgammal 02] A Elgammal, R Duraiswami, D Harwood and L S Davis. "Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance." *Proceedings of the IEEE*, **90**(7):1151–1163, 2002.

[Foster 03] J P Foster, M S Nixon and A Prügel-Bennett. "Automatic Gait Recognition using Area-Based Metrics." *Pattern Recognition Letters*, **24**(14):2489-2497, 2003.

[Fox 97] J Fox. "Applied Regression Analysis, Linear Models and Related Methods." *Sage Publications*, 1997.

[Gavrila 96] D Gavrila and L Davis. "3-D Model-based Tracking of Humans in Action: a Multi-view Approach." *Proc. Computer Vision and Pattern Recognition*, 73-80, 1996.

[Gavrila 99] D M Gavrila. "The Visual Analysis of Human Movement: A Survey." *Computer Vision and Image Understanding*, **73**(1):82-98, 1999.

[Han 04] J Han and B Bhanu. "Statistical Feature Fusion for Gait-Based Human Recognition." *Proc. Computer Vision and Pattern Recognition*, 842-847, 2004.

[Haritaoglu 00] I Haritaoglu, D Harwood and L S Davis. "W^4: Real-Time Surveillance of People and Their Activities." *IEEE Trans. Pattern Analysis and Machine Intelligence*, **22**(8):809-830, 2000.

[Hu 04] W Hu, T Tan, L Wang and S Maybank. "A Survey on Visual Surveillance of Object Motion and Behaviours." *IEEE Trans. Systems, Man and Cybernetics*, **34**(3):334-352, 2004.

[Hayfron-Acquah 03] J B Hayfron-Acquah, M S Nixon and J N Carter. "Automatic Gait Recognition by Symmetry Analysis." *Pattern Recognition Letters*, **24**(13):2175-2183, 2003.

[Illingworth 88] J Illingworth and J Kittler. "A Survey of the Hough Transform." *Computer Vision, Graphics and Image Processing*, **44**:87-116, 1988.

[Inman 81] V T Inman, H J Ralston and F Todd. "Human Walking." *Williams and Wilkins, Baltimore*, 1981.

[Isard 98] M Isard and A Blake. "CONDESATION – Conditional Density Propagation for Visual Tracking." *International Journal of Computer Vision*, **29**(1):5-28, 1998.

[Johannson 73] G Johannson. "Visual perception of biological motion and a model for its analysis." *Perception and Psychophysics*, **14**:201-211, 1973.

[Kale 04a] A Kale, A K RoyChowdhury and R Chellappa. "Fusion of Gait and Face for Human Identification." *Proc. International Conference on Acoustics, Speech and Signal Processing*, 901-904, 2004.

[Kale 04b] A Kale, A Sundaresan, A N Rajagopalan, N P Cuntoor, A K Roy-Chowdhury, V Krüger and R Chellappa. "Identification of Humans Using Gait." *IEEE Trans. Image Processing*, **13**(9):1163-1173.

[Kass 87] M Kass, A Witkin and D Terzopoulos. "Snakes: Active Contour Models." *International Journal of Computer Vision,* **1**(4):321-331, 1987.

[Kochanek 84] D Kochanek and R Bartels. "Interpolating Splines with Local Tension, Continuity and Bias Control." *Computer Graphics*, **18**(3):33-41, 1984.

[Lappas 02] P Lappas, J N Carter and R I Damper. "Robust Evidence-based Object Tracking." *Pattern Recognition Letters*, **23**:253-260, 2002.

[Larssen 04] A T Larssen. "Physical Computing – Representations of Human Movement in Human-Computer Interactions." *Proc. Asia-Pacific Conference on Human-Computer Interaction*, 661-665, 2004.

[Lee 02] L Lee and W E L Grimson. "Gait Analysis for Recognition and Classification." *Proc. Automatic Face and Gesture Recognition*, 155-162, 2002.

[Lee 04] C S Lee and A Elgammal. "Gait Style and Gait Content: Bilinear Models for Gait Recognition Using Gait Re-sampling." *Proc. Automatic Face and Gesture Recognition*, 147-152, 2004.

[Little 98] J Little and J Boyd. "Recognizing People by Their Gait: The Shape of Motion." *Videre*, **1**(2):2-32, 1998.

[Liu 04] Z Liu, L Malave, A Osuntugun, P Sudhakar and S Sarkar. "Towards Understanding the Limits of Gait Recognition", *Proc. SPIE Defense and Security Symposium: Biometric Technology for Human Identification*, 195-205, 2004.

[MacCormick 00] J MacCormick and A Blake. "A Probabilistic Exclusion Principle for Tracking Multiple Objects." *International Journal of Computer Vision*, **39**:57-71, 2000.

[McLachlan 92] G J McLachlan. "Discriminant Analysis and Statistical Pattern Recognition." *John Wiley*, 1992.

[Meyer 98] D Meyer, J Posl and H Niemann. "Gait Classification with HMMs for Trajectories of Body Parts Extracted by Mixture Densities." *Proc. British Machine Vision Conference*, 459-468, 1998.

[Moeslund 01] T B Moeslund and E Granum. "A Survey of Computer Vision-Based Human Motion Capture." *Computer Vision and Image Understanding*, **81**(3):231-268, 2001.

[Mowbray 04] S D Mowbray and M S Nixon. "Extraction and Recognition of Periodically Deforming Objects by Continuous, Spatio-temporal Shape Description." *Proc. Computer Vision and Pattern Recognition*, 895-901, 2004.

[Murray 64] M P Murray, A B Drought and R C Kory. "Walking Patterns of Normal Men." *Journal of Bone and Joint Surgery*, **46**(A):335-360, 1964.

[Nash 97] J M Nash, J N Carter and M S Nixon. "Dynamic Feature Extraction via the Velocity Hough Transform." *Pattern Recognition Letters*, **18**:1035–1047, 1997.

[Ning 04a] H Ning, T Tan, L Wang and W Hu. "People Tracking Based on Motion Model and Motion Constraints with Automatic Initialization." *Pattern Recognition*, **37**:1423-1440, 2004.

[Ning 04b] H Ning, T Tan, L Wang and W Hu. "Kinematics-based Tracking of Human Walking in Monocular Video Sequences." *Image and Vision Computing*, **22**:429-441, 2004.

[Nixon 99] M S Nixon, J N Carter, D Cunado, P S Huang and S V Stevenage, "Automatic gait recognition." Biometrics: Personal Identification in a Networked Society, *Kluwer Academic Publishing*, **11**:231-250, 1999.

[Nixon 03] M S Nixon, J N Carter, M G Grant, L G Gordon and J B Hayfron-Acquah. "Automatic Recognition by Gait: Progress and Prospects." *Sensor Review*, **23**(4):323-331, 2003.

[Novak 87] C L Novak and S A Shafer. "Color Edge Detection." *Proc. DARPA Image Understanding Workshop*, 35-37, 1987.

[Perrin 01] D P Perrin and C E Smith. "Rethinking Classical Internal Forces for Active Contour Models." *Proc. Computer Vision and Pattern Recognition*, 615-620, 2001.

[Perry 92] J Perry. "Gait Analysis: Normal and Pathological Function." *Slack Incorporated,* 1992.

[Phillips 02] P J Phillips, S Sarkar, I Robledo, P Grother and K Bowyer. "The Gait Identification Challenge Problem: Data Sets and Baseline Algorithm." *Proc. International Conference on Pattern Recognition*, 385-388, 2002.

[Plänkers 03] R Plänkers and P Fua. “Articulated Soft Objects for Multiview Shape and Motion Capture.” *IEEE Trans. Pattern Analysis and Machine Intelligence*, **25**(9):1182-1187, 2003.

[Prati 01] A Prati, R Cucchiara, I Mikic and M M Trivedi. “Analysis and Detection of Shadows in Video Streams: A Comparative Evaluation.” *Proc. Computer Vision and Pattern Recognition*, 571-576, 2001.

[Ripley 96] B Ripley. “Pattern Recognition and Neural Networks.” *Cambridge University Press*, 1996.

[Rosin 91] P L Rosin and T Ellis. “Detecting and Classifying Intruders in Image Sequences.” *Proc. British Machine Vision Conference*, 293-300, 1991.

[Rosin 95] P L Rosin and T Ellis. “Image Difference Threshold Strategies and Shadow Detection.” *Proc. British Machine Vision Conference*, 347-356, 1995.

[Sarkar 05] S Sarkar, P J Phillips, Z Liu, I R Vega, P Grother and K Bowyer. “The Human ID Gait Challenge Problem: Data Sets, Performance and Analysis.” *IEEE Trans. Pattern Analysis and Machine Intelligence*, **27**(2):162-177, 2005.

[Shakhnarovich 02] G Shakhnarovich and T Darrell. “On Probabilistic Combination of Face and Gait Cues for Identification.” *Proc. Automatic Face and Gesture Recognition*, 176-181, 2002.

[Shutler 02] J D Shutler, M G Grant, M S Nixon and J N Carter. “On a Large Sequence Based Human Gait Database.” *Proc. Recent Advances in Soft Computing*, 66-71, 2002.

[Shutler 06] J D Shutler and M S Nixon. “Zernike Velocity Moments for Sequence-Based Description of Moving Features.” *Image and Vision Computing*, **24**:343-356, 2006.

[Sidenbladh 02] H Sidenbladh, M J Black and L Sigal. “Implicit Probabilistic Models of Human Motion for Synthesis and Tracking.” *Proc. European Conference on Computer Vision*, 784-800, 2002.

[Sonka 99] M Sonka, V Hlavac and R Boyle. “Image Processing, Analysis and Machine Vision (2nd Edition).” *PWS Publishing*, 1999.

[Sony 05] Sony Computer Entertainment Inc. http://www.eyetoy.com

[Stauffer 00] C Stauffer and W Grimson. “Learning Patterns of Activity Using Real-time Tracking.” *IEEE Trans. Pattern Analysis and Machine Intelligence*, **22**(8):747–757, 2000.

[Stevenage 99] S V Stevenage, M S Nixon and K Vince. “Visual Analysis of Gait as a Cue to Identity.” *Applied Cognitive Psychology*, **13**(6):513-526, 1999.

[Tanawongsuwan 03] R Tanawongsuwan and Aaron Bobick. “Modelling the Effects of Walking Speed on Appearance-Based Gait Recognition.” *Proc. Computer Vision and Pattern Recognition*, 783-790, 2003.

[Tolliver 03] D Tolliver and R T Collins. “Gait Shape Estimation for Identification.” *Proc. Audio- and Video-Based Biometric Person Authentication*, 734-742, 2003.

[Urtasun 04a] R Urtasun and P Fua. “3D Human Body Tracking using Deterministic Temporal Motion Models.” *Proc. European Conference on Computer Vision*, 92-107, 2004.

[Urtasun 04b] R Urtasun and P Fua. “3D Tracking for Gait Characterization and Recognition.” *Proc. Automatic Face and Gesture Recognition*, 17-22, 2004.

[Veeraraghavan 04] A Veeraraghavan, A R Chowdhury and R Chellappa. “Role of Shape and Kinematics in Human Movement Analysis.” *Proc. Computer Vision and Pattern Recognition*, 730-737, 2004.

[Vegas 03] I R Vega and S Sarkar. “Statistical Motion Model Based on the Change of Feature Relationships: Human Gait-Based Recognition.” *IEEE Trans. Pattern Analysis and Machine Intelligence*, **25**(10):1323-1328, 2003.

[Veres 04] G V Veres, L Gordon, J N Carter, M S Nixon. "What Image Information is Important in Silhouette-Based Gait Recognition?" *Proc. Computer Vision and Pattern Recognition*, 776-782, 2004.

[Veres 05] G V Veres, M S Nixon, L Middleton and J N Carter. "Fusion of Dynamic and Static Features for Gait Recognition over Time." *Proc. International Conference on Information Fusion*, 2005.

[Wagg 03] D K Wagg and M S Nixon. "Model-Based Gait Enrolment in Real-World Imagery." *Proc. Multimodal User Authentication*, 189-195, 2003.

[Wagg 04a] D K Wagg and M S Nixon. "On Automated Model-Based Gait Extraction and Analysis." *Proc. Automatic Face and Gesture Recognition*, 11-16, 2004.

[Wagg 04b] D K Wagg and M S Nixon. "Automated Markerless Extraction of Walking People Using Deformable Contour Models." *Computer Animation and Virtual Worlds*, **15**(3-4):399-406, 2004. Given as an oral presentation at *Computer Animation and Social Agents*, 2004.

[Wang 03a] L Wang, W Hu and T Tan. "Recent Developments in Human Motion Analysis." *Pattern Recognition*, **36**(3):585-601, 2003.

[Wang 03b] L Wang, T Tan, H Ning and W Hu. "Silhouette Analysis-Based Gait Recognition for Human Identification." *IEEE Trans. Pattern Analysis and Machine Intelligence*, **25**(12):1505-1518, 2003.

[Wang 04] L Wang, H Ning, T Tan and W Hu. "Fusion of Static and Dynamic Body Biometrics for Gait Recognition." *IEEE Trans. Circuits and Systems for Video Technology*, **14**(2):149-158, 2004.

[Whittle 99] M W Whittle and D Levine. "Three-dimensional Relationships between the Movements of the Pelvis and Lumbar Spine during Normal Gait." *Human Movement Science*, **18**:681-692, 1999.

[Williams 92] D J Williams and M Shah. "A Fast Algorithm for Active Contours and Curvature Estimation." *Computer Vision, Graphics and Image Processing: Image Understanding*, **55**(1):14-26, 1992.

[Winter 90] D A Winter. "Biomechanics and Motor Control of Human Movement (2nd Edition)." *John Wiley and Sons*, 1990.

[Winter 91] D A Winter. "The Biomechanics and Motor Control of Human Gait: Normal, Elderly and Pathological." *University of Waterloo press, Ontario*. 1991.

[Wu 83] C F J Wu. "On the Convergence Properties of the EM Algorithm." *Annals of Statistics*, **11**(1):95-103, 1983.

[Yam 04] C Yam, M S Nixon and J N Carter. "Automated Person Recognition by Walking and Running via Model-Based Approaches." *Pattern Recognition*, **37**(5):1057-1072, 2004.

[Yang 92] Y H Yang and M D Levine. "The Background Primal Sketch: An Approach for Tracking Moving Objects." *Machine Vision Applications*, **5**:17–34, 1992.

[Yoo 03] J Yoo and M S Nixon. "On Laboratory Gait analysis via Computer Vision." *Proc. Biologically Inspired Machine Vision, Theory and Application*, 109-113, 2003.

[Zhang 04] J Zhang, R Collins, Y Liu. "Representation and Matching of Articulated Shapes." *Proc. Computer Vision and Pattern Recognition*, 342-349, 2004.

[Zhao 04] T Zhao and R Nevatia. "Tracking Multiple Humans in Complex Situations." *IEEE Trans. Pattern Analysis and Machine Intelligence*, **26**(9):1208-1221, 2004.

[Zhou 05] X Zhou, B Bhanu and J Han. "Human Recognition at a Distance in Video by Integrating Face Profile and Gait." *Proc. Audio- and Video-based Biometric Person Authentication*, 533-543, 2005.