UNIVERSITY OF SOUTHAMPTON

Modelling and Extracting Periodically Deforming Objects by Continuous, Spatio-temporal Shape Description.

by

Stuart David Mowbray

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

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

July 2008

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

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

Doctor of Philosophy

by Stuart David Mowbray

This thesis proposes a new model for describing spatio-temporally deforming objects. Through a novel use of Fourier descriptors, it is shown how arbitrary shape description can be extended to include spatio-temporal shape deformation. It is further demonstrated that these new spatio-temporal Fourier descriptors have the ability to be used as the basis for both the recognition and extraction of deforming objects. Application of this new recognition technique to human gait sequences demonstrates recognition rates of over 86% for individual human subjects, signifying that these descriptors possess unique descriptive properties. Based upon the new spatio-temporal Fourier descriptor model, a new technique for the detection and extraction of deforming shapes from an image sequence is presented through a new variant of the Hough transform (the Continuous Deformable Hough Transform) that utilises spatio-temporal shape correlation within an evidence-gathering context. This new technique demonstrates excellent success rates and tolerance to noise, correctly extracting human subjects in image sequences corrupted with noise levels of up to 80%. The technique is also tested extensively using real-world data, thus demonstrating its worth in a modern-day computer vision system. Both the spatio-temporal Fourier descriptor model, the Continuous Deformable Hough Transform, and aspects of their application are fully discussed throughout the thesis, along with ideas and suggestions for future research and development.

Contents

Acknowledgements **viii**

1 Context and contributions **1**

1.1 Allied research areas . . . 1

1.2 Application domain . . . 2

1.3 Scope of this thesis . . . 3

1.4 Thesis structure . . . 4

1.5 Publications resulting from this work . . . 6

2 Introduction **7**

2.1 Motivation . . . 7

2.2 Shape description and modelling . . . 8

2.2.1 Area-based metrics . . . 9

2.2.2 Perimeter-based metrics . . . 11

2.3 Shape detection . . . 13

2.4 Biometrics . . . 16

2.4.1 Gait as a biometric . . . 16

2.5 Summary and aims . . . 17

3 Modelling deformable objects using spatio-temporal Fourier descriptors **19**

3.1 Modelling two-dimensional shapes . . . 19

3.1.1 Elliptic Fourier descriptors . . . 20

3.1.2 Invariance . . . 21

3.2 Modelling periodically deforming shapes . . . 25

3.2.1 A short note on frequency content and boundary length . . . 26

3.2.2 Motion model normalisation . . . 27

3.2.3 Shape-size invariance . . . 28

3.3 Discussion . . . 30

4 Gait recognition using spatio-temporal Fourier descriptors **31**

4.1 Forming a database of shape-sequences . . . 31

4.2 Calculation of spatio-temporal Fourier descriptors . . . 33

4.3 Feature Selection . . . 34

4.4 Classification and initial recognition testing . . . 36

4.4.1 The effects of feature selection . . . 37

4.4.2 The effect of distance or low spatial resolutions on performance . . . 40

4.5 Summary and discussion . . . 41

5 The Continuous Deformable Hough Transform 43

5.1 Locating deformable objects 43

5.2 An introduction to Hough transforms 44

5.3 Foundation theory 45

5.3.1 The Velocity and Continuous Velocity Hough Transforms 45

5.3.2 Shape model dimensionality 46

5.3.3 Motion model dimensionality 48

5.4 Continuous Deformable Hough Transform theory 48

5.4.1 The CDHT kernel model 49

5.4.2 Considerations when working with discrete signals 53

5.4.3 The CDHT voting process 56

5.4.4 Pseudo-code algorithm 57

5.5 Dimensionality 58

5.6 Summary 58

6 Application of the CDHT: from theory to practice 59

6.1 Data and pre-processing 59

6.2 Generating a generic gait kernel 61

6.3 Results 61

6.4 Discussion 62

6.5 Summary 63

7 Noise performance testing of the CDHT 67

7.1 Choosing an appropriate noise model 69

7.1.1 Problems with additive Gaussian noise and binary data 69

7.1.2 A binary noise model 70

7.1.3 Testing 71

7.2 Results and discussion 73

7.3 Summary 76

8 Real-world performance testing of the CDHT 77

8.1 Test data 77

8.2 Raw real-world data testing 78

8.2.1 Analysis of results 80

8.2.2 Summary 85

8.3 Pre-processed real-world data testing 86

8.3.1 Background subtraction and edge data generation 86

8.4 Analysis of results 88

8.5 Summary 90

9 Occlusion testing 91

9.1 Test data 92

9.2 Analysis of results 93

9.3 Discussion and summary 98

10 Future work and conclusions 100

10.1 Future work 100

10.1.1 Application-specific work 100

10.1.2 Technique-specific work and performance enhancements 102
10.2 Summary and conclusions 105

A Results from the image sequence used in Chapter 6 107

B Detailed real-world raw image data test results 113

C Detailed real-world pre-processed image data test results 127

References 141

List of Figures

2.1 Eadweard Muybridge's studies of human and animal motion. 8
2.2 Example of a binary image and associated zeroth- and first- Cartesian moments. 10
2.3 Example of a binary image and associated horizontal and vertical projections. 11
2.4 4- and 8-connected chain code coding systems. 12
2.5 A shape and its associated 8-connected chain code. 12
2.6 Discrete shape representation and transformation. 13
2.7 Fourier-based shape representation and transformation. 14
2.8 Parameterisation of a straight line. 15

3.1 Cartesian plane to complex plane mapping. 20
3.2 Two-dimensional shape reconstruction from Fourier descriptors using increasing numbers of harmonic components. 21
3.3 Example of how a shape is reconstructed from rotating elliptic phasors. . 22
3.4 Magnitude of semi-major and semi-minor elliptic axes for a rotating phasor. 22
3.5 Basic two-dimensional shapes. 24
3.6 Example shape sequence of a walking human. 29
3.7 Shifted magnitude plot of spatio-temporal Fourier descriptors for Figure 3.6. 29

4.1 Silhouette extraction. 32
4.2 Clustering of different subjects using three Fourier descriptors: Subject 1 ∘; Subject 2 △; Subject 3 □; Subject 4 ×. 35
4.3 Examples of class separation and associated Bhattacharyya distance metrics. 36
4.4 The effects of varying Bhattacharyya distance thresholds on classification performance. 38
4.5 Number of features used during classification vs CCR% (k=3). 39
4.6 Varying image resolutions, scaled by subject height. 41

5.1 The Velocity Hough Transform (VHT) for a circle. 47
5.2 Example of a normalised shape sequence descriptors. 49
5.3 Using zero-padding in the frequency domain to interpolate in the time domain. 54

6.1 Raw test data and associated edge map. 60
6.2 Superimposed generic kernel, scaled and translated by the extracted parameters. 64

6.3 Views of a two-dimensional cross-section of the accumulator. 65
6.4 Side-on views of the accumulator. 66

7.1 Low-resolution ground-truth image sequence. 68
7.2 Probability distribution function for potential noisy pixel value for a white pixel. 71
7.3 Frame 16 with varying levels of binary noise added. 72
7.4 Frame 16 with varying levels of clipped and binarised Gaussian noise added. 72
7.5 Comparison of the noise performance for the CDHT using the binary noise model and the Clipped Gaussian noise model. 73
7.6 Comparison of the noise performance for the CDHT and the CVHT, both using the Clipped Gaussian noise model. 74
7.7 Accumulators for varying noise levels. 75

8.1 Original and edge-detected real-world data for Subject 14. 79
8.2 Resulting Shape Sequence Template overlaid on the original image data for Subject 14. 81
8.3 Accumulator for the image sequence 'Subject 14'. 82
8.4 Partial results for Subject 25, showing the subject walking into partial occlusion. 84
8.5 Original and Canny edge map for Subject 31, showing failure to detect edges of the lower torso due to poor contrast. 84
8.6 Cross-section of the accumulator space for Sequence 42, showing a vote-wise near miss. 85
8.7 Background subtraction process. 87
8.8 Accumulator for the image sequence 'Subject 14' using pre-processed data. 89

9.1 Occlusion limits for one gait cycle. 92
9.2 Occlusion data. 94
9.3 Results for the occlusion tests using varying MHD thresholds. 96
9.4 Box and whisker plots of the Euclidean distances at varying occlusion levels 97

10.1 Correlation of maximum bounding box width and height parameters. . . . 103
10.2 Correlation of maximum bounding box height and velocity scaling parameters, showing the lower and upper bounds (95% limits) of male (blue) and female (red) subject height and stride length. 104

List of Tables

3.1 Invariant elliptic Fourier descriptors. 24

4.1 Initial results of k-nearest neighbour classification 37
4.2 Best result of k-nearest neighbour classification 40
4.3 Results of k-nearest neighbour classification (k=3) for varying resolutions 41

6.1 Extracted parameters for the test image sequence. 62

9.1 Hit rate at varying occlusion levels for results falling within 1 pixel or frame of the baseline data 98

Acknowledgements

I would like to thank my supervisor, Professor Mark Nixon, for his advice and encouragement; my fellow ISIS research group members for their advice and entertainment; my family for their emotional and financial support; my friends for stopping me from going mad and helping me relax when things haven't been going quite to plan; and my girlfriend Anna for her discussions on clinical gait (thus proving that it's always prudent to have a medic close by) and her continued love and support, whether things were going to plan or not! Thanks should also be given to my uncle, Professor Austin Tate, for providing much-needed inspiration in my younger years. There are undoubtedly people who I have forgotten to mention, but to anyone who has helped me/entertained me/cheered me up (delete as applicable!) over the recent years, a big thank you!

I also gratefully acknowledge partial support by the European Research Office of the US Army under Contract No. N68171-01-C-9002.

Chapter 1

Context and contributions

1.1 Allied research areas

Fourier descriptors have been applied successfully to many shape recognition problems and provide a method of analytically representing a shape's contour. Many advantages are gained from having a shape description in an analytic form. If the shape is scaled or rotated, for example, then the transformations can be simply applied to the analytic form, which is continuous, and discretisation errors are therefore kept to a minimum. Further to this, being Fourier-based, Fourier descriptors allow access to any frequency-based properties that the shape may possess. For example, small details of a shape are contained in the higher frequencies of the Fourier descriptors, while a more general definition of the shape is contained in the lower frequencies. This also means that the general description of the shape is very compact, with very high frequencies usually containing little information about the actual shape itself. The use of Fourier descriptors to describe rigid shapes is well established. The use of Fourier descriptors to model deforming, non-rigid shape sequences however, is something that has not previously been researched or demonstrated, and it is this new modelling technique, along with its applications which this thesis presents. The ability to analytically represent a whole non-rigid shape sequence and to have access to the spatio-temporal frequency components of the shape as it deforms renders this model extremely flexible. This is demonstrated by the fact that the model is not only easily generalised to aid in shape detection, but is also suitable for discriminating between very similar shapes (for example, the shape of different human subjects).

In a modern computer vision system, a large part of the problem of being able to successfully interpret a scene is having the ability to identify and extract parameters (regarding pose and orientation, for example) of specific objects. Due to this, this area of computer vision has always been a naturally active area. To humans this is a task that we often do without thinking; it is a cognitive process that we have developed naturally since birth [4]. However, when viewing a scene, objects are often occluded by other objects, or subject to lighting variations, which make it very hard for automatic recognition systems to successfully identify them. Evidence-gathering techniques (or voting algorithms), such as the Hough transform and its successors [2, 5, 16, 19, 31] produce a statistical measure for the likelihood of an object with a given set of parameters being present in a scene, thus identifying a best-fit model of the object in question. Over the years, variants of the Hough transform have been developed, derived from various basic object models, with some even making an allowance for the fact that objects can move through an animated scene. However, none of these variants have considered objects that deform, of which there are many – particularly in the natural world where the majority of animals deform in some predictable way as they move. The lack of development in this area is perhaps due to a lack of suitable models for such objects and the fact that processing whole sequences of images, rather than just static scenes, only became a feasible research area in more recent years, as increasing computer processing power has permitted it within realistic time-scales. As previously mentioned, one of the aims of this thesis is to demonstrate a new model for such periodically deforming shapes, and as such it seems only natural to then use this model as the basis for a new evidence-gathering technique, which is also presented later in the thesis.

1.2 Application domain

One of the most intuitive application domains for the new shape deformation model is automatic human gait recognition. Human gait is a relatively new biometric, which aims to recognise people by the way they walk. Gait itself has long been recognised as a very individual characteristic – a fact that even Shakespeare recognised, as demonstrated in Julius Caesar (Act 1, Scene 3): "Tis Cinna; I do know him by his gait". Gait analysis has received considerable attention in the medical field and is well known as being one of the most complex of all human motions (and therefore very individual). Gait can also be monitored and analysed in a non-invasive manner, indeed it is this ability to observe gait non-invasively that makes it such a desirable biometric. In recent years, research into gait as a biometric has escalated and its suitability as a biometric has been established. The Defence Advanced Research Projects Agency (DARPA) recently funded a number of institutions to take part in a research program into Human Identification at a Distance. The participating institutions focusing specifically on gait were: the University of Southampton; the Massachusetts Institute of Technology; Georgia Technical Research Institute; the University of South Florida; University of Maryland; Carnegie Mellon University; and the National Institute for Standards in Technology (NIST). With human gait being such an active research area, it becomes a natural area of application for the techniques developed throughout this thesis.

1.3 Scope of this thesis

This thesis proposes a new method of modelling periodically deforming shapes. Based around the concept of Fourier descriptors, a new spatio-temporal Fourier descriptor is developed which is capable of capturing periodic shape deformation, such as that demonstrated by many animals, including humans. Being Fourier-based, this model provides many inherited advantages, such as being continuous in nature and providing access to the spatio-temporal frequencies of an animated shape sequence.

Derived from this model, a new evidence-gathering technique, based around the Hough transform, is also presented. This new technique, named the Continuous Deformable Hough Transform (CDHT), demonstrates all the same robust features of the original Hough transform, such as being very tolerant to noise and occlusion. Further to this, as the CDHT's kernel is based around the previously developed spatio-temporal Fourier descriptor model, it also inherits all the benefits of continuous shape representation, such as the ability to undergo affine transformations with minimal risk of discretisation errors occurring.

Both the model and the Continuous Deformable Hough Transform developed as part of this thesis are tested through their application to human gait recognition and detection, respectively. Specific features of the model are used to test the discriminability of the model's parameters, and a 'generic' model is used as the basis for the detection of walking humans in video footage.

The purpose of applying these new techniques to human gait is not necessarily to develop a fool-proof gait recognition system, but to test the discriminability of the features of the model and the extraction capabilities of the CDHT. Therefore, throughout the thesis, the concept of human gait is considered in only a simplified form, where the gait of a subject is measured from an almost camera-parallel (side-on) perspective, thus appearing in two-dimensions. In an ideal scenario for these tests, the camera would be precisely plane-parallel to the subject, and thus gait should appear to be perfectly periodic (within the limitations of the subject's capability to perfectly reproduce their movements during each gait cycle). In reality, however, the data used during this thesis was gathered using a fixed camera observing each subject walking by. This will inevitably cause slight errors in the periodicity of the gait cycle, but the skewing of the camera viewpoint introduced here is considered negligible for the purposes of the work contained in this thesis. Indeed, it would be very difficult in the real-world application of gait recognition to observe every moving subject in a plane-parallel manner, and therefore the results demonstrated in this thesis are more likely to describe the performance of the techniques described when applied in the real-world, rather than in a perfect environment.

1.4 Thesis structure

The overall structure of this thesis is split into three conceptual sections. The first section introduces the background for the research presented throughout the remainder of the thesis, and lays down the theoretical and formal definitions for the new spatio-temporal Fourier descriptor model used to describe periodically deforming shapes.

The second section shows how the parameters of the new spatio-temporal Fourier descriptor model can be used to discriminate and classify temporally deforming shape sequences. In this thesis, this is demonstrated through the application of the technique to automatic human gait recognition, although it should be noted that this new model could be applied to any periodically deforming objects or shapes.

The development of a new evidence-gathering mechanism, the Continuous Deformable Hough Transform (CDHT), is detailed and demonstrated in the third section of the thesis. The CDHT uses the spatio-temporal Fourier descriptor model as the underlying kernel model for detecting periodically deforming objects. Again, human gait is the application domain here, and the thesis demonstrates how the CDHT can be used to successfully detect and extract the correct model parameters for a range of human subjects in both artificial and real-world image sequences. This section also details the performance testing measures carried out to demonstrate the CDHT's ability to cope with noise and occlusion.

Chapter 2

This chapter introduces and gives background information relating to the areas of general shape description, Fourier descriptors, gait recognition, and shape detection. Other studies related to these areas are discussed and motivations for the remainder of the thesis are developed.

Chapter 3

This chapter describes and formally defines the new spatio-temporal Fourier descriptor model. Spatial and temporal normalisation procedures are also discussed here, and examples of spatio-temporal Fourier descriptors are demonstrated.

Chapter 4

This chapter demonstrates the ability of the new Fourier descriptor model to discriminate between shape sequences. This discriminability is tested using portions of the University of Southampton's Large Gait Database [38]. During the testing 115 subjects and 1062 gait sequences are used, resulting in a Correct Classification Rate of 86.2%.

Chapter 5

Building on the new Fourier descriptor model, Chapter 5 introduces the theory behind the Continuous Deformable Hough Transform (CDHT). Spatio-temporal and velocity scaling mechanisms are described and discussed, and pseudo-code for the CDHT is presented.

Chapter 6

Chapter 6 takes the CDHT from theory into practice, describing the application of the CDHT to a real image sequence. Considerations to the practical application of the CDHT to real image sequences are also discussed here.

Chapter 7

This chapter describes simulated noise tests conducted on the CDHT. Varying noise levels from 0% – 100% are used, with the CDHT exhibiting excellent tolerance to noise, even under extreme conditions.

Chapter 8

Chapter 8 describes and demonstrates 'real-world' testing on the CDHT, with outdoor data from the University of Southampton's Large Gait Database being used for the test data. The data is used in its 'raw' format, to estimate the CDHT's performance on such data, and also in a pre-processed form using basic image processing techniques, in order to test the CDHT's performance as part of a realistic modern-day computer vision system.

Chapter 9

In this chapter, the effects of occlusion on the overall performance of the CDHT are examined, using the results of the real-world data tests from Chapter 8 as a baseline measure of performance for comparison.

Chapter 10

This chapter concludes the thesis with suggestions for further research areas, and a summary and critical discussion of the work.

1.5 Publications resulting from this work

1. Mowbray, S. D. and Nixon, M. S. (2003), Automatic Gait Recognition via Fourier Descriptors of Deformable Objects, in Proceedings Audio- and Video-based Biometric Person Authentication 2003 (AVBPA 2003), pages 566-573, Guildford, UK, 2003.

2. Mowbray, S. D. and Nixon, M. S. (2004), Extraction and Recognition of Periodically Deforming Objects by Continuous, Spatio-temporal Shape Description, in Proceedings IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2004 (CVPR 2004), volume 2: pages 895-901, Washington DC, 2004.

Chapter 2

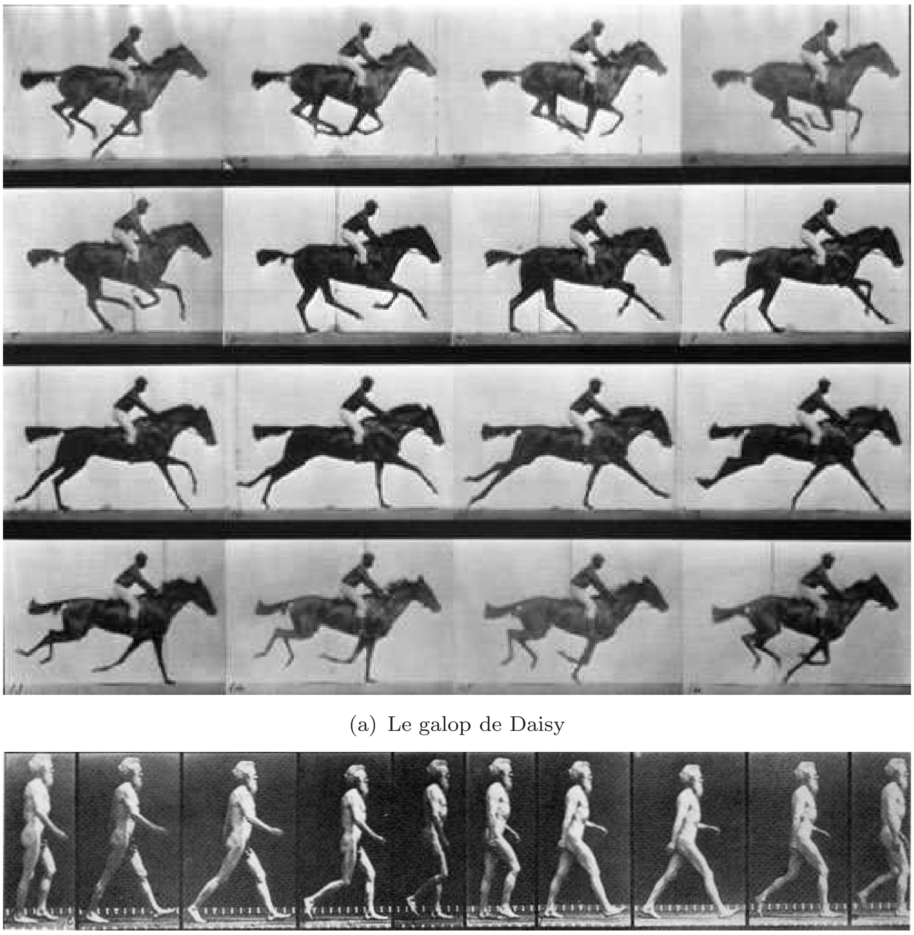
Introduction

2.1 Motivation

Many living objects deform in some way as they progress through their daily lives. A large majority of these deformations, specifically in animals, are periodic in nature. One only has to consider the shape formed by an object such as a walking human, or a cantering horse to observe this periodic deformation. Eadweard Muybridge, a pioneer in the photography of moving images, studied many such forms of periodic motion in humans and animals (Figures 2.1(a) and 2.1(b)), and his work is considered by many to represent the beginning of the science of biomechanics. It is the aim of this thesis to develop techniques to describe and detect periodically deforming and moving objects such as these.

The vast majority of previous research in the computer vision community has concentrated on the detection, description, or recognition of apparently rigid and stationary shapes. This was mostly, no doubt, due to the lack of computer processing power and memory to enable work on more complicated shapes and image sequences. Recently, as computer power has increased and digital video has become more and more widespread, we have seen the inclusion of motion in many aspects of computer vision. However, although these 'spatio-temporal' techniques have included motion, the vast majority do not take into account that objects can, and often do, deform as they move.

Detecting any form of shape or object, either rigid or non-rigid, static or moving, first requires that the shape in question can be described mathematically in some way. This could be an image of the shape, with the mathematical model simply being a list of pixel locations and values, but more often, and more preferably, this will be an analytical 'model' of some form or another. Providing one can do this, techniques then need to be employed to take an image or model of a new scene and efficiently and robustly search for new occurrences of this shape. The aim of modelling a shape mathematically before



(a) Le galop de Daisy

(b) Sequence of Eadweard Muybridge Walking

Figure 2.1: Eadweard Muybridge's studies of human and animal motion.

detection is to capture the 'essence' of the shape, so that when detection takes place the detection algorithm isn't confused by external factors, such a lighting or occlusion.

2.2 Shape description and modelling

To describe a shape in a way that is unambiguous, but also general enough to capture only the true characteristics of the shape is a problem that has been studied since the emergence of computer vision, and before this, in cognitive psychology. Although, as humans, we are very good at 'seeing' and recognising everyday objects, one only has to think of an object such as a table or chair to realise that, when asked to describe it unambiguously, we often fail miserably. One person may describe a table as having "four legs and a flat top", but this description may also describe a bed, a chest of drawers, a stool, or a number of other objects. Even in most modern object recognition systems, it is often the responsibility of human operators to first provide a description of the objects of interest. Given a layperson's apparent lack of ability in providing these object descriptions unambiguously, it is little wonder that some automatic object recognition systems fail. This is the reason that a good model of an object is essential.

In the field of computer vision, two major schools of research have emerged over the holistic description of shapes, these being perimeter-based and area-based. Perimeter-based shape descriptors consider a shape to be defined by its internal and external boundaries, which often manifest themselves as edges in images. Conversely, area-based descriptors consider shapes to be defined by the properties of the actual space they cover.

Perception of both the perimeter of a shape and the area it covers are subject to error, especially in cases of occlusion. Typically in the case of occlusion, however, the occluding object obscures more information pertaining to the area of the shape than the perimeter.

2.2.1 Area-based metrics

Perhaps one of the most commonly used area-based shape metric is that of statistical moments. Low-order statistical moments are encountered in everyday statistics, such as the zeroth- and first-order moments, which are used to calculate the *mass* (or *area* when applied to images) and the *mean* (or *center of mass*) of an object respectively.

When expressed in two-dimensions simple Cartesian moments are expressed as

$$m_{pq} = \int\limits_{\mathbb{R}} \int\limits_{\mathbb{R}} x^p y^q f(x,y) dx dy \tag{2.1}$$

where $\mathbb{R}$ is a set of real numbers. For a discrete image P_{xy}, this becomes

$$m_{pq} = \sum_{x \in \mathbb{R}} \sum_{y \in \mathbb{R}} x^p y^q P_{xy} \tag{2.2}$$

where (x, y) forms a Cartesian pixel location.

Often, moments are applied to binary images where pixels in foreground objects are denoted by a 1 and the background is denoted by 0s. In a simple example of a binary image containing one object, the total area of the shape therefore becomes

$$\begin{aligned} A &= m_{00} \\ &= \sum_{x \in \mathbb{R}} \sum_{y \in \mathbb{R}} x^0 y^0 P_{xy} \\ &= \sum_{x \in \mathbb{R}} \sum_{y \in \mathbb{R}} P_{xy} \end{aligned} \tag{2.3}$$

and the Centre of Mass of the object is found by normalising the two first-order moments by m_{00}, as shown in Equation 2.4.

$$\begin{aligned} \bar{x} &= \frac{m_{10}}{m_{00}} & \bar{y} &= \frac{m_{01}}{m_{00}} \\ \bar{x} &= \frac{\sum_{x\in\mathbb{R}}\sum_{y\in\mathbb{R}} xP_{xy}}{A} & \bar{y} &= \frac{\sum_{x\in\mathbb{R}}\sum_{y\in\mathbb{R}} yP_{xy}}{A} \end{aligned} \tag{2.4}$$

An example of these basic Cartesian moments applied to a simple binary image can be seen in Figure 2.2.

$y\backslash x$	0	1	2	3	4	5	6	7	8	9
0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	1	0	0	0	0
2	0	0	0	1	1	1	1	0	0	0
3	0	0	0	1	1	1	1	0	0	0
4	0	0	0	1	0	0	1	0	0	0
5	0	0	0	1	0	0	1	0	0	0
6	0	0	1	1	1	1	1	1	0	0
7	0	0	1	1	1	1	1	1	0	0
8	0	0	0	1	1	1	1	0	0	0
9	0	0	0	0	0	0	0	0	0	0

Area: 30, Center of Mass: $(x = 4.5,\ y = 5.0)$

FIGURE 2.2: Example of a binary image and associated zeroth- and first- Cartesian moments.

Many variations of moments exist, such as centralised moments, which are invariant to translation, and Hu moments, which are invariant to translation, rotation, and scaling. More recently, Shutler [37] extended centralised moments to model shapes in motion. These moments, known as velocity moments, demonstrate how a shape's statistical description can be improved by utilising temporal correlation; a concept which is demonstrated further in this thesis. Recently, Foster [10] also used area-based masks to measure the dynamics of the changing area of a shape through time as a means of discriminating between human subjects.

Another often used area-based metric that is worth mentioning is that of projection. In their simplest form, projections for binary images consist of a summation of each column and each row of the data to produce individual horizontal and vertical projections. Again, this is demonstrated in Figure 2.3 using the same binary image that was used in the previous Cartesian moment demonstration.

Projections are reasonably crude metrics for recording information about a shape's area, and are inherently affected by changes in the shape's orientation or size. However, they are very quick to calculate when these parameters are known, and as such, this technique might lend itself well to areas of computer vision such as Optical Character Recognition (OCR), or any other other application domain where speed and simplicity in a constrained environment are required.

$y\backslash x$	0	1	2	3	4	5	6	7	8	9	H-Proj
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	1	1	0	0	0	0	2
2	0	0	0	1	1	1	1	0	0	0	4
3	0	0	0	1	1	1	1	0	0	0	4
4	0	0	0	1	0	0	1	0	0	0	2
5	0	0	0	1	0	0	1	0	0	0	2
6	0	0	1	1	1	1	1	1	0	0	6
7	0	0	1	1	1	1	1	1	0	0	6
8	0	0	0	1	1	1	1	0	0	0	4
9	0	0	0	0	0	0	0	0	0	0	0
V-Proj	0	0	2	7	6	6	7	2	0	0	

Figure 2.3: Example of a binary image and associated horizontal and vertical projections.

Although interesting subject matters in themselves, the metrics and methods of area-based shape description will not be discussed further in this thesis, as the thesis' central research area is derived from the use of perimeter-based metrics; area-based metrics are presented here merely to juxtapose the two techniques for the reader.

2.2.2 Perimeter-based metrics

We now move on to the area of shape description which this thesis directly addresses and extends – that of perimeter-based shape description. In contrast to area-based shape description metrics, which aim to capture properties relating to the space which a shape covers, perimeter-based metrics use the boundaries of this space.

Early attempts at perimeter description resulted in the Chain codes [12]. Chain codes represent the perimeter of an object by coding a numerical value representing the relative direction needed to move from one edge point to another, typically continuing around the shape's boundary until the start point is reached. One could think of this as analogous to the American 'block' system used when giving directions, for example, "Walk two blocks north and 3 blocks east", which could be represented as NNEEE in this specific chain code notation. The number of values used to encode a direction change is dependent upon whether the boundary points around the shape are 4- or 8-connected, as described in Figure 2.4.

Chain codes represent a *complete* boundary description, in that a full reconstruction of the shape's boundary can be reproduced from the chain code, but are inherently

	2	
1	·	3
	4	

2	3	4
1	·	5
8	7	6

FIGURE 2.4: 4- and 8-connected chain code coding systems.

difficult to work with in situations where invariance to scale and rotation are needed. An example of a basic shape and its associated 8-connected Chain Code can be seen in Figure 2.5.

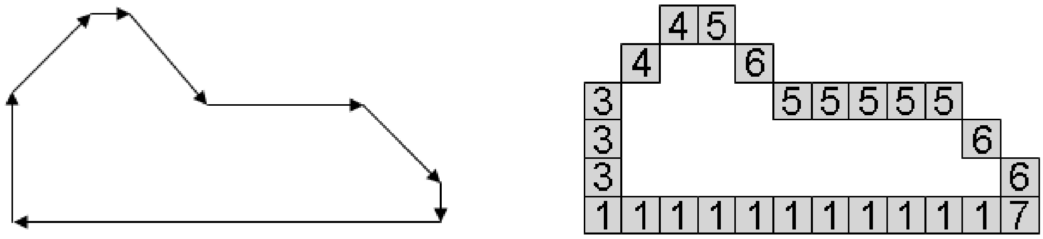
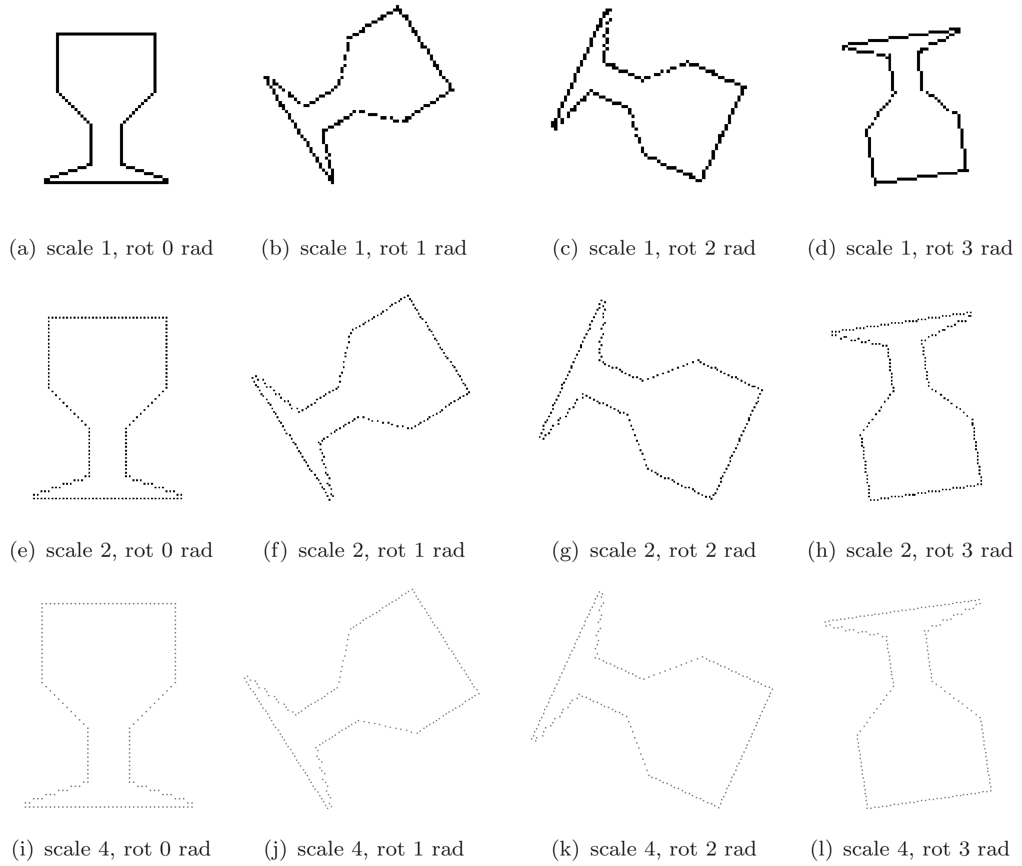


FIGURE 2.5: A shape and its associated 8-connected chain code.

More recently, Fourier descriptors have been used with great success to model shapes and object boundaries. Fourier descriptors transform a space-domain representation of the boundary pixels (a set of pixel points) into the frequency domain, allowing any periodic properties of a shape's perimeter to be analysed. Low frequencies are associated with macroscopic features of the shape's perimeter, while higher frequencies are associated with more microscopic, sharp changes in the perimeter, such as corners.

Fourier descriptors have the advantage that they are reasonably easy to be made invariant to scale, rotation, and start point. Furthermore, as the Fourier descriptor representation of a shape's boundary defines an analytic curve, scale and rotation transformations can be performed in the frequency domain, avoiding many possible discretisation errors associated with such transformations when performed in the time domain. This can clearly be seen when comparing Figure 2.6 with Figure 2.7. Figure 2.6 depicts a discrete method of representing a shapes boundary (a bitmap), whereas 2.7 depicts a continuous, Fourier-based shape description, both of which are rotated and scaled. It is easy to see that any rotation using a discrete boundary representation can cause undesirable discretisation effects. Similarly, the bitmap in Figure 2.6 consists purely of a set of discrete sample points, and when scaled these points 'explode' out from the origin of the scaling operation and the shape boundary lacks any continuity. Conversely, the rotated and scaled boundaries shown in Figure 2.7 demonstrate how a continuous model can be transformed with very little distortion due to discretisation, with the only significant errors being made when sampling takes place to convert this continuous curve back into the discrete world of computer graphics for display purposes.


(a) scale 1, rot 0 rad (b) scale 1, rot 1 rad (c) scale 1, rot 2 rad (d) scale 1, rot 3 rad

(e) scale 2, rot 0 rad (f) scale 2, rot 1 rad (g) scale 2, rot 2 rad (h) scale 2, rot 3 rad

(i) scale 4, rot 0 rad (j) scale 4, rot 1 rad (k) scale 4, rot 2 rad (l) scale 4, rot 3 rad

FIGURE 2.6: Discrete shape representation and transformation.

Many studies in computer vision have used Fourier descriptors successfully to model shape boundaries [15, 25, 29, 35]. Aguado et al [2] was the first to combine Fourier descriptors and the Generalised Hough Transform to exploit the power of continuous shape representation and the robustness of the Hough transform to detect static shapes.

2.3 Shape detection

Being able to describe a shape through whatever means, whether area-based or perimeter-based, naturally leads to the desire to detect that shape when presented with a new scene. A number of forms of shape detection, some of which are now almost de-facto standards, have been used for a number of years, two of which are template matching and the Hough transform and its successors [5, 19].

Template matching involves correlating a kernel (a simple representation of the shape being sought) with the possible feature points (typically edges pixels) in a new scene.

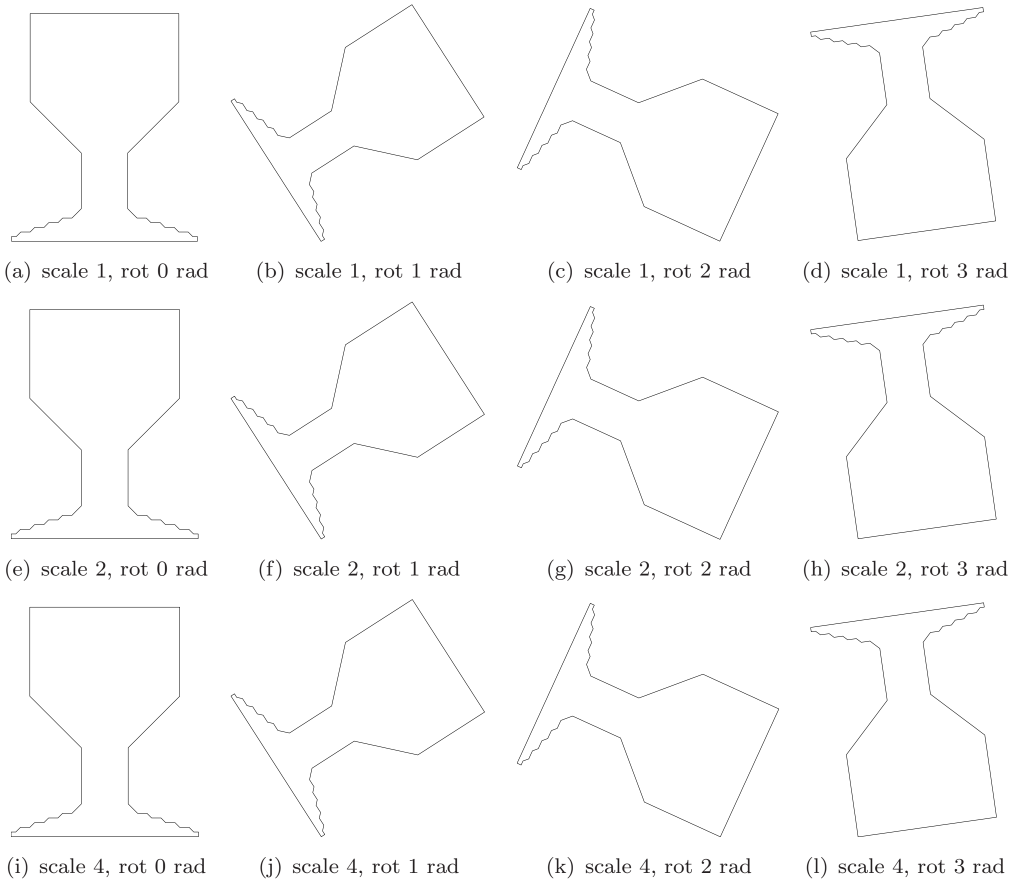
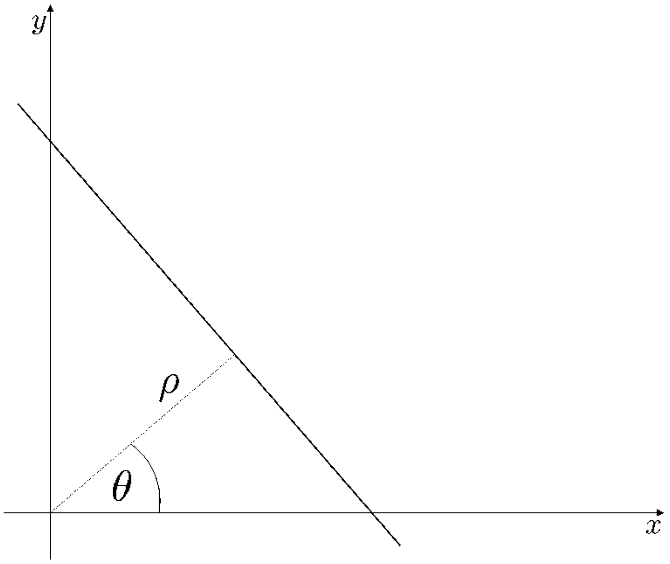

(a) scale 1, rot 0 rad (b) scale 1, rot 1 rad (c) scale 1, rot 2 rad (d) scale 1, rot 3 rad
(e) scale 2, rot 0 rad (f) scale 2, rot 1 rad (g) scale 2, rot 2 rad (h) scale 2, rot 3 rad
(i) scale 4, rot 0 rad (j) scale 4, rot 1 rad (k) scale 4, rot 2 rad (l) scale 4, rot 3 rad

FIGURE 2.7: Fourier-based shape representation and transformation.

The computational complexity of template matching is governed by the dimensionality and size of this scene and the number of feature points in the kernel, as a correlation coefficient for each possible feature point in both the kernel and the test feature-space has to be calculated – the kernel literally has to be tested in every single position to obtain the maximum correlation coefficient, which hopefully then defines the location of the shape being sought. It is this high computational demand that often limits the application of template matching, with more constrained and efficient shape detection methods being favoured instead.

One of these more efficient shape detection methods is the Hough transform, which uses a technique known as evidence-gathering. Evidence-gathering techniques work by allowing any given feature point in the test space (e.g. a binary edge image) to cast a number of 'votes' for the correct parameters (location, size, orientation, etc.) of the shape being sought. By this process, each feature point suggests 'evidence', in the form of these votes, for a possible shape with a certain set of parameters. The votes are cast into an accumulator space (also known as a Hough space), the dimensionality of which is equal to the number of unknown parameters of the shape being sought.

An example of the original form of the Hough transform – that for straight lines, is given by first examining the parameterisation for a straight line, as shown in Figure 2.8.



$$\rho = x \cos \theta + y \sin \theta$$

FIGURE 2.8: Parameterisation of a straight line.

Given that two of the parameters in this parameterisation, x and y, are known for each feature point in any given scene, the accumulator for the Hough transform for a straight line is two-dimensional, with the free (unknown) parameters being ρ and θ – the distance between the line and the origin, and the angle of the line to the origin respectively. Similar Hough transforms can be developed for any shape that can be described parametrically, such as a circle, which is expressed by Equation 2.5, and therefore has a three-dimensional accumulator, parameterised by a, b, and r, where $(a,\ b)$ are the coordinates of the center of the circle, and r is the radius.

$$(x - a)^2 + (y - b)^2 = r^2 \tag{2.5}$$

When all votes have been cast for each feature point, the accumulator space is searched for either a global maximum of votes, if only one shape is being sought, or a number of local maxima, if more than one shape is being sought. The unknown parameters of the shapes found then correspond to the location of these maxima in the accumulator.

A number of advancements have been made since the original formulation of the Hough transform for straight lines, probably the most notable of which is the Generalised Hough Transform [5], which parameterises arbitrary shapes by forming a table (known as a 'R-table') of polar vector mappings from a shape's edge points to a given reference point (usually its center of mass).

The Velocity Hough Transform (VHT) [31] is a more recent development that extends the concept of the Hough transform by including velocity as a free parameter, allowing the detection of shapes moving across an image sequence. The commonplace use of video over the past decade has undoubtedly contributed to the need and practical feasibility for the development of such techniques. Limited computer processing power ensured that detection of objects moving through image sequences was something that could not have even been conceived of at the time of Hough's original work.

Even more recently, Grant [16] developed the Continuous Velocity Hough Transform (CVHT). The CVHT further extended Nash's Velocity Hough Transform by not only allowing a shape to be tracked across an image sequence, but also by allowing arbitrary shapes to be used. Furthermore, Grant's work incorporated Aguado et al's approach to modelling the shape continuously – a fact that dramatically reduces discretisation errors when testing for the presence of a shape at differing scales or rotations, as previously seen. A variant of the CVHT also featured the inclusion of a non-linear motion model, allowing for arbitrary (but previously known) patterns of motion to be detected.

2.4 Biometrics

The use of biometrics to uniquely identify a person has grown rapidly over the years. From the use of fingerprints, to the identification of humans through analysis of iris patterns, biometric-based techniques are fast becoming commonplace in a variety of application domains. Biometrics have been used most notably in the areas of security and surveillance, where reliable verification of individuals, without the risk of forgery, is essential. Usual methods of personal identification either involve lengthy enrollment and calculation procedures in order to gain accuracy (such as DNA testing), or are fast but unreliable (such as a PIN code system). Biometric-based methods, however, have the potential to provide the end user with an excellent combination of speed and security.

2.4.1 Gait as a biometric

Recently, a significant amount of attention has been devoted to the use of human gait patterns as a biometric, and to the analysis of human motion in general. Studying this motion, however, is a difficult task. Studies of human gait have been carried out in many disciplines, such as psychology, biomechanics, and various medical fields. However, these studies have usually required human interaction, such as manual labelling of the data or the use of markers. The aim of computer vision-based analysis of human gait therefore, is to automatically recover and describe human gait accurately and with minimal human intervention. Several models have been proposed for the description of the human body and also for the description of human gait [1, 14, 41]. Human gait has many advantages over other biometrics, but perhaps its most notable advantages are that it is non-invasive, offering a means to verify identity without a subject's active participation, and that it offers a means of offering recognition at a distance – something that many other biometrics are not capable of.

Many descriptors of human gait are kinematic in nature, relying on geometric descriptions of the various body parts and mathematical modelling of the transformations which describe their movement. A number of kinematic features exist, such as the angles of various body parts through time, their velocity, and their acceleration. Other studies have taken a more statistical approach to produce a unique gait descriptor. These have included the use of Principal Component Analysis (for dimensionality compression) combined with Canonical Analysis (for classification) [20], the use of velocity moments [39], and more recently the use of Hidden Markov Models [24]. Other approaches have included the use of area-based features [11], the analysis of 'frieze' patterns [27], and the use of static parameters, such as height and stride-length [23]. Research has also being carried out into describing human motion by analysing the bilateral symmetry inherent in human gait [7, 18].

Most popular models of human gait used in computer vision have taken one of two forms. One approach has been to model the pendular movement of the human thigh using simple harmonic motion [9], while the other approach models the thigh and lower leg movement as coupled oscillators [42]. These models, however, represent only sections of the human body and do so by finding a best fit of the model to the image data.

Many other approaches have been taken to the recognition of humans from shape statistics and gait models, and good reviews of current state-of-the-art techniques can be found in [33] and [34].

2.5 Summary and aims

This thesis presents a new and original model for representing periodically-deforming objects through the novel use of spatio-temporal Fourier descriptors. By utilising this spatio-temporal Fourier descriptor model, a biometric can be obtained which is capable of discriminating between human subjects. Further to this, it will be shown how a normalised version of this model can be used to form a basis for a new variant of the Hough transform (the CDHT), which is capable of extracting moving, deformable objects (such as walking human subjects) from an image sequence. This new Hough transform technique not only proves to inherit many desirable features from the standard Hough transform, such as its ability to deal with noise and occlusion, but also provides a continuous, spatio-temporal shape representation to deal with the effects of discretisation, and has the computational benefit of having a parameter space with a fixed number of dimensions.

Throughout this thesis, use of the spatio-temporal Fourier descriptor model and the CDHT are demonstrated through their application to human gait. More specifically, this will be shown through demonstrating their application to automatic gait recognition and to the extraction of walking humans from image sequences.

Chapter 3

Modelling deformable objects using spatio-temporal Fourier descriptors

This chapter discusses the concept of extending Fourier descriptors to include temporal information, rather than just spatial information as used previously. Spatio-temporal Fourier descriptors offer a continuous method of representing an entire sequence of a periodically deforming and moving shape in a complete and compact form. In order to explain the concept of using Fourier descriptors to model sequences of shapes fully, we will first examine the background theory of Fourier descriptors for two-dimensional shapes, then progress to develop and explain the theory for spatio-temporal Fourier descriptors.

3.1 Modelling two-dimensional shapes

If a shape's boundary forms a closed curve, then this curve c – a function of arc-length, l, with total arc-length L, can be considered to be periodic such that

$$c(l + L) = c(l) \tag{3.1}$$

Two-dimensional shapes are usually represented on a Cartesian plane, where the x and y coordinates of the shape are a function of the shape's arc-length. If one defines a mapping from the Cartesian plane to a complex plane, however, then it is possible to represent a shape's boundary as a complex function of arc-length

$$c(l) = x(l) + j.y(l) \tag{3.2}$$

A diagrammatic example of this mapping can be seen in Figure 3.1.

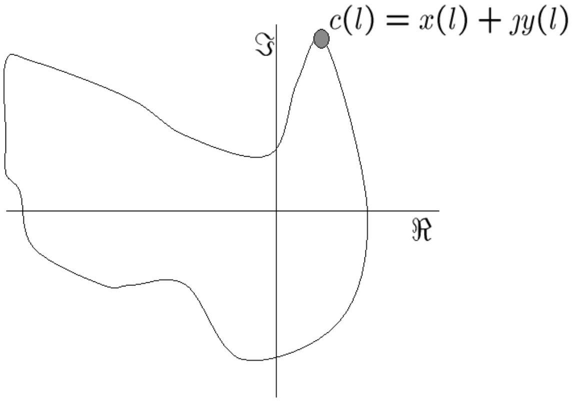


Figure 3.1: Cartesian plane to complex plane mapping.

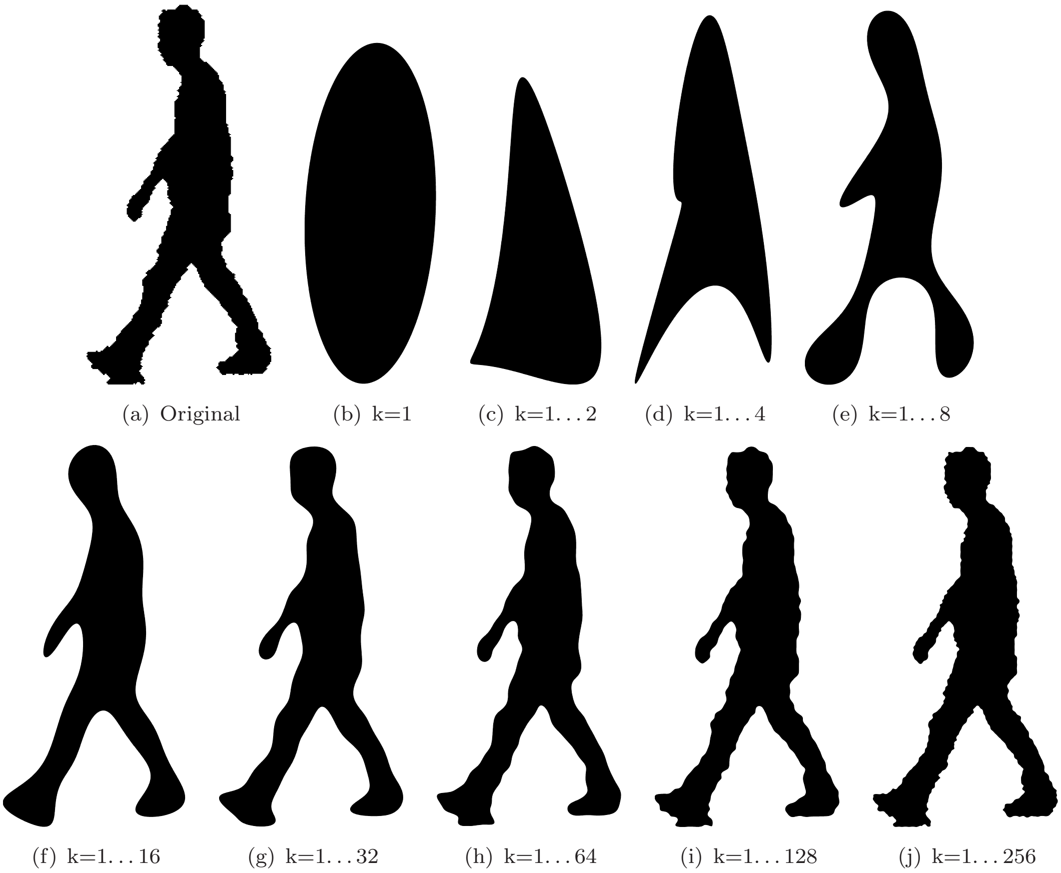
3.1.1 Elliptic Fourier descriptors

Due to the periodicity of $c(l)$ it is possible to represent the shape's boundary using a Fourier series, with the coefficients of the series, a_{xk}, b_{xk}, a_{yk}, and b_{yk}, being the Fourier descriptors of the shape

$$\begin{aligned} c(l) = \quad & \tfrac{a_{x0}}{2} + \int_{k=1}^{\infty} \left\{ a_{xk} \cos\left(\tfrac{kl2\pi}{L}\right) + b_{xk} \sin\left(\tfrac{kl2\pi}{L}\right) \right\} + \\ & j\left(\tfrac{a_{y0}}{2} + \int_{k=1}^{\infty} \left\{ a_{yk} \cos\left(\tfrac{kl2\pi}{L}\right) + b_{yk} \sin\left(\tfrac{kl2\pi}{L}\right) \right\} \right) \end{aligned} \tag{3.3}$$

The coefficients of Equation 3.3 describe the boundary of the curve $c(l)$ in increasing detail as k increases, with the components when $k = 0$ (the DC component) being closely related to the center of mass of the shape enclosed by $c(l)$. Figure 3.2 shows how the inverse function of Equation 3.3 can be used to reconstruct an original shape using various subsets of the increasing values of k, from just a single component ($k = 1$) to the first 256 components ($k = 1 \ldots 256$). It can be seen that at $k = 1$ the reconstructed shape consists simply of an ellipse, with the inclusion of increasingly higher harmonic components revealing more and more detail until, in Figure 3.2(j), the boundary of the shape is virtually complete at the resolution shown.

For each harmonic in a set of these 'elliptic' Fourier descriptors, the Fourier coefficients a_{xk}, b_{xk}, a_{yk}, and b_{yk} can be interpreted as a rotating elliptic phasor (a rotating vector), as l changes, thus giving the name "elliptic Fourier descriptor". Using this interpretation, each elliptic phasor orbits around the previous harmonic's elliptic locus with a speed that


(a) Original (b) k=1 (c) k=1. . . 2 (d) k=1. . . 4 (e) k=1. . . 8

(f) k=1. . . 16 (g) k=1. . . 32 (h) k=1. . . 64 (i) k=1. . . 128 (j) k=1. . . 256

Figure 3.2: Two-dimensional shape reconstruction from Fourier descriptors using increasing numbers of harmonic components.

is proportional to its harmonic value k. The original shape is then reconstructed from a summation of these rotating phasors, as seen in Figure 3.3.

3.1.2 Invariance

Descriptors that are invariant to rotation and start point can be obtained by considering the lengths of the semi-major and semi-minor axes of each harmonic phasor's elliptic locus. The dimensions of these axes are given by [32] as the Euclidean distance between the origin and the coordinate pairs (a_{xk}, a_{yk}) and (b_{xk}, b_{yk}) for the semi-major and semi-minor axes respectively, as shown in Figure 3.4.

However, these measures only hold true when the phasor's start point is aligned with the semi-major axis of the elliptic locus. This can be readily seen by studying the equations for the x and y projections of the elliptic locus in Cartesian space when $l = 0$, as shown by Equation 3.4.

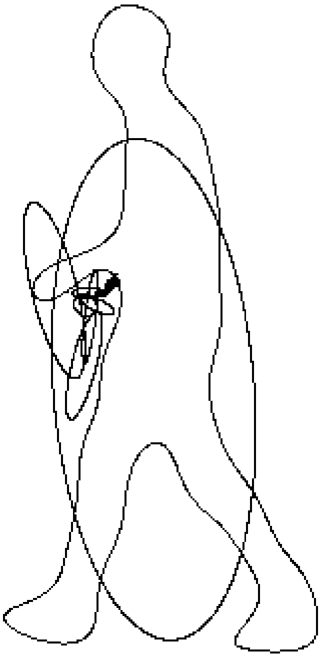


Figure 3.3: Example of how a shape is reconstructed from rotating elliptic phasors.

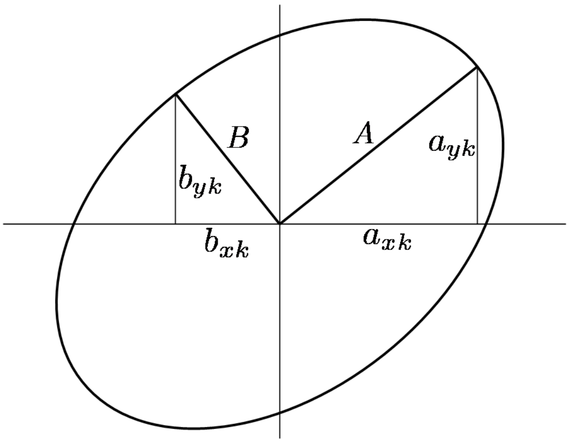


Figure 3.4: Magnitude of semi-major and semi-minor elliptic axes for a rotating phasor.

$$\begin{aligned} & x_k(0) = a_{xk}\cos(0) + b_{xk}\sin(0) \\ & y_k(0) = a_{yk}\cos(0) + b_{yk}\sin(0) \\ \Rightarrow \quad & x_k(0) = a_{xk} \\ & y_k(0) = a_{yk} \end{aligned} \tag{3.4}$$

Equation 3.4 demonstrates that the only time for which the locus is defined exactly by the coefficients a_{xk} and a_{yk} is at the start point of the elliptic phasor, when $l = 0$. This will not always be the case, and so for this measure of axes length to hold true, a rotation operator must be applied to all four Fourier coefficients, as discussed in [25], to align the start point of the phasor with the semi-major axis of the harmonic's elliptic locus.

$$\begin{bmatrix} a'_{xk} & a'_{yk} \\ b'_{xk} & b'_{yk} \end{bmatrix} = \begin{bmatrix} \cos\theta_k & \sin\theta_k \\ -\sin\theta_k & \cos\theta_k \end{bmatrix} \begin{bmatrix} a_{xk} & a_{yk} \\ b_{xk} & b_{yk} \end{bmatrix} \tag{3.5}$$

where θ_k is the angular rotation of the semi-major axis of the k^{th} harmonic's elliptic locus. This is discussed further in [32].

Another, perhaps simpler method for determining the location (and subsequently the magnitude) of the semi-major axis and semi-minor axis can be demonstrated by finding the two points of inflexion of a distance function defined from the centre of the ellipse to its locus, given by

$$\begin{aligned} & dist_k(l) = \sqrt{(dist_{xk}(l)^2 + dist_{yk}(l)^2)} \\ \text{where} & \\ & dist_{xk}(l) = a_{xk}\cos\left(\tfrac{kl2\pi}{L}\right) + b_{xk}\sin\left(\tfrac{kl2\pi}{L}\right) \\ \text{and} & \\ & dist_{yk}(l) = a_{yk}\cos\left(\tfrac{kl2\pi}{L}\right) + b_{yk}\sin\left(\tfrac{kl2\pi}{L}\right) \end{aligned} \tag{3.6}$$

These two points of inflexion (one a local maximum and one a local minimum) can be found by setting the derivative of $dist_k(l)$ to 0 and solving for l. The first solution which satisfies this is given at distance

$$sol_{1k} = \frac{L}{4\pi}\tan^{-1}\left[\frac{2(a_{xk}b_{xk} + a_{yk}b_{yk})}{a_{xk}^2 + a_{yk}^2 - b_{xk}^2 - b_{yk}^2}\right] \tag{3.7}$$

with the second solution being most readily found by noticing that the semi-major and semi-minor axes are orthogonal

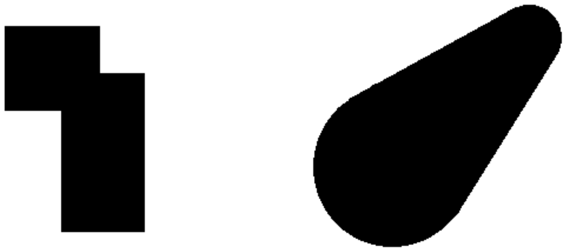
$$sol_{2k} = sol_{1k} - \frac{L}{4} \tag{3.8}$$

This gives the solution for a point on the ellipse located $\frac{\pi}{2}$ radians around its contour from the point specified by sol_{1k}.

Substituting sol_{1k} and sol_{2k} into Equation 3.6 gives the length of the semi-major and semi-minor axes of the k^{th} harmonic ellipse, which can be summed to give a descriptor for this harmonic which is invariant to both rotation and start point. Invariance to scale can be easily obtained by normalising each harmonic's descriptor with respect to the magnitude of the first harmonic's descriptor, and invariance to translation can be obtained by discarding the descriptors for $k = 0$, which describes the shape's center of mass. Thus, descriptors that are invariant to start point, translation, rotation, and scale can be defined as

$$d_k = \frac{dist_k(sol_{1k})+dist_k(sol_{2k})}{dist_k(sol_{11})+dist_k(sol_{21})} \qquad k \in \mathbb{Z}+ : k \leqslant \lfloor \tfrac{L}{2} \rfloor \tag{3.9}$$

Invariant elliptic Fourier descriptors for the two different shapes featured in Figure 3.5 are shown in Table 3.1. These Fourier descriptors, calculated using the method above, are shown for the original shape, the shape after a rotation transformation, and after both a rotation and a scaling transformation.



(a) Shape 1

(b) Shape 2

FIGURE 3.5: Basic two-dimensional shapes.

Shape	Transformation	1^{st} Harm	2^{nd} Harm	3^{rd} Harm	4^{th} Harm
1	Original	1.000	0.149	0.135	0.081
	Rotation	1.000	0.149	0.133	0.079
	Rotation & Scaling	1.000	0.146	0.135	0.080
2	Original	1.000	0.102	0.083	0.021
	Rotation	1.000	0.103	0.083	0.021
	Rotation & Scaling	1.000	0.101	0.086	0.021

TABLE 3.1: Invariant elliptic Fourier descriptors.

It is clear to see that the descriptors for Shape 1, even after several affine transformations, are very similar, and quite different from those of Shape 2. The slight variations in the actual values of the descriptors can be put down to discretisation errors. This shows the uniqueness of the descriptors in describing a shape and also the invariant properties they possess. It is worth noting that the first harmonic's descriptor always equals unity – this is due to the scale normalisation process described earlier.

Making elliptic Fourier descriptors totally invariant in this way is a good method for analysing basic properties of two-dimensional shapes, but this invariance comes at a cost: the descriptors lose their ability to reconstruct the shape and therefore lose much of their potential descriptive power (in this case they lose information relating to the rotation, start point, true scale, and eccentricity of the elliptic phasors). Due to this, when we examine the application of elliptic Fourier descriptors to modelling periodically deforming shapes in the next section, a different approach to invariance will be used, which treats a sequence of shapes as a whole, rather than just individual two-dimensional shapes.

3.2 Modelling periodically deforming shapes

If we now consider a shape which deforms between points 0 and T in time, indexed by t, we can model the periodicity of the deforming shape s, at arc-length index l as

$$s(t,\ l) = s(t + T,\ l) \tag{3.10}$$

Given this representation, it is possible to model the whole periodically deforming boundary of a shape in this shape sequence as a two-dimensional complex Fourier series

$$s(t,\ l) = \int\limits_{k_t=0}^{\infty}\int\limits_{k_l=0}^{\infty} \hat{s}(k_t,\ k_l)e^{j2\pi(\frac{t.k_t}{T}+\frac{l.k_l}{L})} \tag{3.11}$$

given here in complex form for brevity. The Fourier coefficients of this series then characterises one whole period of the shape's movement.

If we consider the discrete case for a periodically deforming and moving shape then T becomes equivalent to the number of images in one period of motion, L is the length of the boundary of the object, and the Fourier coefficients, $\hat{s}$, can be represented by a discrete two-dimensional complex Fourier series

$$\hat{s}(k_t,\ k_l) = \frac{1}{T.L}\sum_{t=0}^{T}\sum_{l=0}^{L} s(t,\ l)e^{-j2\pi(\frac{k_t.t}{T}+\frac{k_l.l}{L})} \tag{3.12}$$

with these spatio-temporal Fourier descriptors then describing the way the spatial frequency components of each frame vary over time.

3.2.1 A short note on frequency content and boundary length

As each time-domain signal[1] (a boundary signal of each frame of the sequence) is of variable length, and therefore of variable potential frequency content, an unequal number of spatial frequencies will be available for each shape in a given sequence. It is therefore necessary to impose a limit on the number of spatial frequencies that are calculated in order to aid in the calculation of the spatio-temporal descriptors. To ensure that the Nyquist sampling criterion is satisfied, this frequency limit must be no larger than half of the length of the shape in the image sequence with the shortest boundary length. A maximum frequency limit is therefore defined as

$$k_{max} = \left\lfloor \frac{\min\limits_{t}(L_t)-1}{2} \right\rfloor \qquad t \in 0 \ldots T-1 \tag{3.13}$$

where L_t is the complex time-domain signal length of the t^{th} shape in the sequence. It should be noted that any truncation of the spatial frequency information, such as that possibly encountered here, may result in a loss of information relating to small details in the shape's boundary. In practice, however, and certainly in the case here, loss of high order harmonics should not prove to be a hindrance, as the true lengths of the time-domain signals do not vary by a large amount, resulting in only the loss of only 'extreme' high-order harmonics. These harmonics are likely to be either noise in the original signal, or extremely insignificant harmonics, which are very probably contaminated by noise.

Due to this variation in boundary length, the spatial coefficients for each shape in a given image sequence are calculated and then band-limited during the calculation of the spatio-temporal Fourier descriptors, so that each signal contains the same number of spatial harmonics, as follows

$$\hat{s}(k_t,\ k_l) = \sum_{t=0}^{T} c_t^*(k_l) e^{-j2\pi(\frac{k_t.t}{T})} \tag{3.14}$$

where

$$c_{k_t}^*(k) = \begin{cases} \frac{c_{k_t}(k)}{2.k_{max}+1} & k \in 0 \ldots k_{max} \\ \\ \frac{c_{k_t}(k+L_t-2.k_{max}-1)}{2.k_{max}+1} & k \in k_{max}+1 \ldots 2.k_{max} \end{cases} \tag{3.15}$$

and

$$c_{k_t}(k) = \sum_{l=0}^{L} s(k_t,\ l) e^{-j2\pi(\frac{k_t.l}{L})} \tag{3.16}$$

[1] Occasionally, especially when referring to core Fourier theory, the term time-domain will be used to describe the Euclidean description of a shape's boundary as a function of time. However, as we are dealing with a two-dimensional shape contour deforming over time, this could become confusing for the reader. When this is the case, attempts will be made to disambiguate the two concepts.

The scaling factor encountered in Equation 3.15 ensures that the energy of the signal is conserved when restricting the number of harmonics in the signal, thus ensuring that Parseval's theorem [8], specifying that the energy in both the time and frequency domains is equal, is still satisfied.

For clarity, the band-limiting function shown here will be omitted from further formulae, but it should be noted that in practice this step is necessary to ensure that the spatio-temporal Fourier descriptors can be calculated correctly. Band-limiting the spatial frequency domain in this way is equivalent to resampling each curve, $c(l)$, so that each contains the same number of samples. In this case, however, band-limiting in the frequency domain is preferred to resampling in the time domain in order to eliminate discretisation errors when resampling.

3.2.2 Motion model normalisation

A moving shape's position changes over time. The amount of change in spatial position, however, is not always directly related to the shape's size: a tall person walking slowly may take shorter strides, and will therefore travel less distance, than a short person taking longer strides, perhaps due to being hurried. It is therefore desirable to normalise this motion pattern independently of the shape's size.

When a shape moves it causes the spatial DC components of Equation 3.12, which describe the shape's center of mass, to vary, with movement corresponding to changes in the real and imaginary parts of the DC component for movements in the x-axis and y-axis respectively. Due to the separability property of the Fourier transform, the spatial Fourier DC components can be calculated and then normalised with respect to the whole image sequence before the temporal calculation of the spatio-temporal Fourier descriptors has occurred, thus creating a motion model which is shape-size invariant. During this normalisation process the shape's motion model is translated so that the center of mass of the first shape in the sequence, at $t = 0$, is located at the origin $(0, 0)$. The motion model is then scaled so that the total Euclidean distance that the shape travels within one period from the origin is unity.

$$\hat{s}(k_t,\ k_l) = \frac{1}{T}\sum_{t=0}^{T}\left[motion_{norm}\left(\frac{1}{L}\sum_{l=0}^{L} s(t,\ l)e^{-j2\pi\frac{k_l.l}{L}}\right)\right]e^{-j2\pi\frac{k_t.t}{T}} \tag{3.17}$$

where

$$
\begin{aligned}
motion_{norm}(f(t,\ 0)) = \quad & \tfrac{\Re(f(t,\ 0)) - \Re(f(0,\ 0))}{\sqrt{\Re(f(T-1,\ 0))^2 + \Im(f(T-1,\ 0))^2}} + \qquad t \in 0 \ldots T-1 \\
& j.\tfrac{\Im(f(t,\ 0)) - \Im(f(0,\ 0))}{\sqrt{\Re(f(T-1,\ 0))^2 + \Im(f(T-1,\ 0))^2}}
\end{aligned}
\tag{3.18}
$$

3.2.3 Shape-size invariance

Making the descriptors shape-size invariant is important if we are to neglect the effect of scale in the shape sequence model (which can be affected by camera-to-subject distance, for example). Normalisation of the sequence with regard to shape-size is performed with respect to the largest spatial dimension within the whole shape sequence (in the case of the human subjects used in this thesis, this corresponds to the largest perceived height of the subject through a gait sequence). This normalisation process is performed not only to ensure that the aspect ratio of the shape remains correct, but also to ensure that any relative changes to the shape's overall size within the sequence remain intact. Again, this takes place in the spatial frequency domain, before full formulation of the spatio-temporal descriptors, thus ensuring that discretisation errors that would normally affect scaling in the time domain do not occur.

The shape-size normalization process is shown below in Equations 3.19 and 3.20

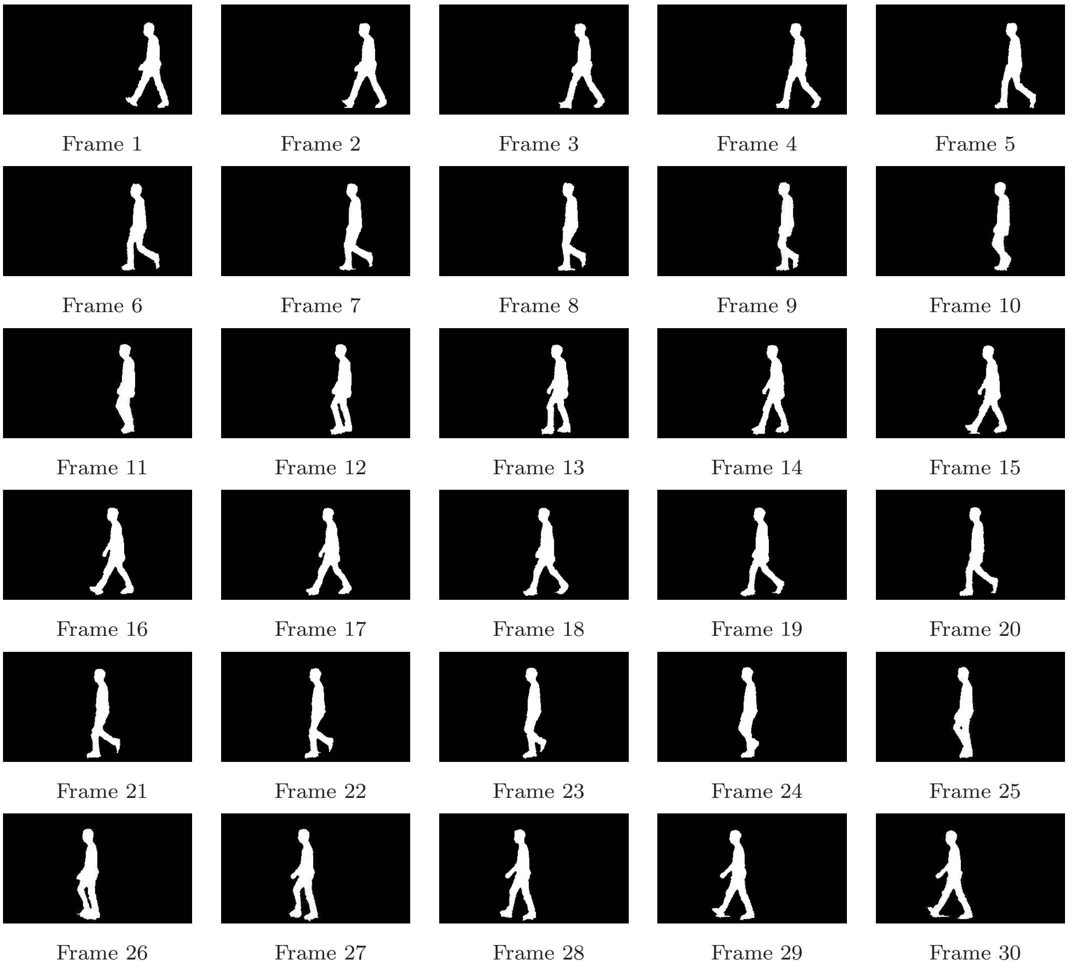
$$
\hat{s}(k_t,\ k_l) = \frac{1}{T} \sum_{t=0}^{T} \left[size_{norm} \left(\frac{1}{L} \sum_{l=0}^{L} s(t,\ l) e^{-j2\pi \frac{k_l.l}{L}} \right) \right] e^{-j2\pi \frac{k_t.t}{T}} \tag{3.19}
$$

where

$$
\begin{aligned}
& size_{norm}(f(t,\ k_l)) = \tfrac{f(t,\ k_l)}{\max\limits_{t}(\max\limits_{k_l}(\Im(s(t,\ k_l))) - \min\limits_{k_l}(\Im(s(t,\ k_l))))} \\
& k_l \in 1 \ldots L-1,\ t \in 0 \ldots T-1
\end{aligned}
\tag{3.20}
$$

A graphical example of a shape sequence and the magnitude plot of a subset of its associated normalised spatio-temporal Fourier descriptors are given in Figures 3.6 and 3.7 respectively. Note that for convenience and clarity the spatio-temporal DC component has been removed in Figure 3.7 and the descriptors have been rearranged so that the harmonics run from negative frequencies, through DC, to positive frequencies.

Several artefacts from the segmentation process can be seen in Figure 3.6, such as those exhibited in Frame 29 and Frame 30, where shadows cast on the floor in the original image sequence have erroneously been included in the extracted silhouette. These will obviously affect the resulting Fourier descriptors, but since small artefacts such as these represent only a small level of detail in the time-domain signal, they shouldn't affect the lower range of harmonics (which represent the more general, overall form of the shape's boundary) significantly.



Frame 1 Frame 2 Frame 3 Frame 4 Frame 5

Frame 6 Frame 7 Frame 8 Frame 9 Frame 10

Frame 11 Frame 12 Frame 13 Frame 14 Frame 15

Frame 16 Frame 17 Frame 18 Frame 19 Frame 20

Frame 21 Frame 22 Frame 23 Frame 24 Frame 25

Frame 26 Frame 27 Frame 28 Frame 29 Frame 30

FIGURE 3.6: Example shape sequence of a walking human.

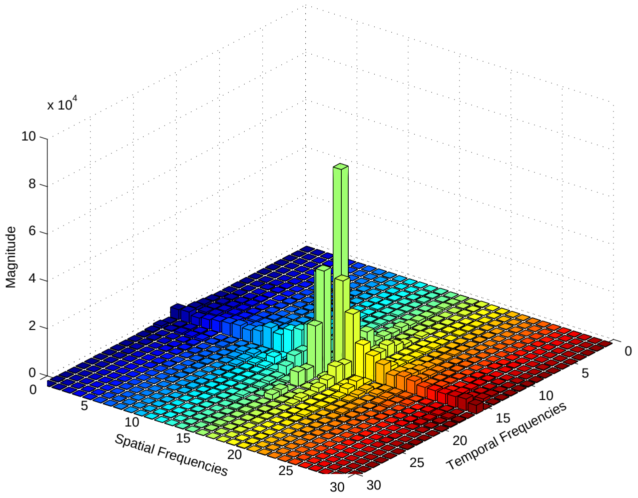


FIGURE 3.7: Shifted magnitude plot of spatio-temporal Fourier descriptors for Figure 3.6.

.

It can be seen in Figure 3.7 that temporal frequencies drop to near zero at a much quicker rate than the spatial frequencies. The rate at which the spatial frequencies drop to zero will be determined by the complexity of the shape in question, and the rate at which the temporal frequencies drop to zero will be determined by the complexity of the temporal deformation of the shape. Medical studies have shown that human gait has a maximum temporal frequency of around 5Hz [3], the Fourier descriptors shown in Figure 3.7 certainly fit in with this observation for the gait sequence shown in Figure 3.6.

3.3 Discussion

This chapter has shown a new and novel use of Fourier descriptors to model periodically deforming objects. The descriptive power of these descriptors will now be examined in Chapter 4, with an application to human gait recognition, before their use as the basis for a new variant of the Hough transform, specifically designed to extract parameters for periodically deforming objects, is examined and developed in subsequent chapters.

Chapter 4

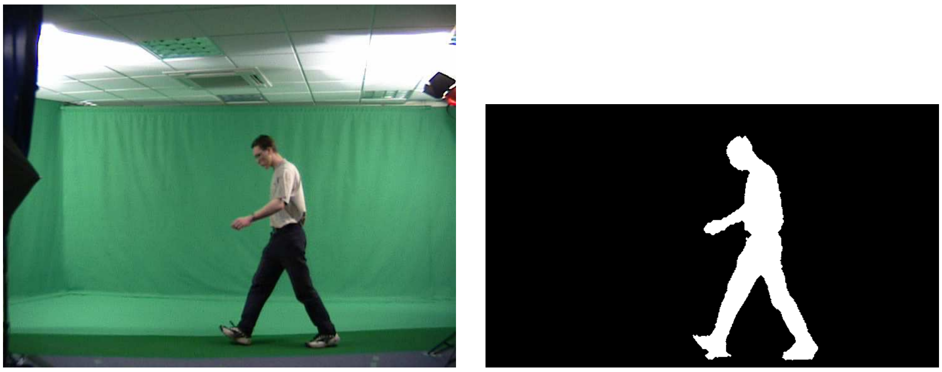
Gait recognition using spatio-temporal Fourier descriptors

The Fourier descriptor model described in the previous chapter provide a means of capturing spatio-temporal information which is not only complete (if phase is included), but also very compact. The majority of the information relating to a deforming shape is contained within the lower frequencies of the descriptors, with higher frequencies containing only minor details, which are also much more susceptible to noise.

Access to the spatio-temporal frequency domain of the shape sequence provides a convenient method of analysing fundamental structural properties of a deforming shape, and as such, the spatio-temporal Fourier descriptors provide a good theoretical basis for discriminating between differing deformable objects.

4.1 Forming a database of shape-sequences

To test the recognition capabilities of this new model, spatio-temporal Fourier descriptors were calculated for each subject in the Large Gait Database at the University of Southampton [38]. This database consists of 115 subjects and 1062 sequences (only right-to-left walking sequences were used in this study). In order to extract the boundary for each shape, background extraction (via chroma-keying) was first performed on each image in a particular sequence. Each image was then thresholded to produce a silhouette (see Figure 4.1). The boundaries from each subject were then extracted by following the boundary of each silhouette, starting from the top of the head (thus having a consistent start point), to produce a complex boundary signal, from which the spatio-temporal Fourier descriptors were calculated.



(a) Original Image (b) Extracted Silhouette

Figure 4.1: Silhouette extraction.

The image sequences were captured under laboratory conditions so as to provide accurate segmentation of the subject. Obviously, if these techniques were to be applied to other data then suitable background subtraction/object extraction techniques should be applied to provide an accurate segmentation of the objects of interest prior to proceeding with the calculation of the spatio-temporal Fourier descriptors.

The subject extraction process consists of four major stages: Chroma-key extraction; image cropping; connected components analysis; and morphological filtering. The first two of these stages can be considered to be background subtraction stages, while the latter two stages 'clean' the binary image to remove any unwanted artefacts. A more detailed description of this process is given below.

Chroma-key extraction

The footage from the Southampton database was filmed against a green-screen background. This green hue was chosen as it would conflict minimally with the clothing of the subjects being filmed (subjects were not allowed to wear green clothing).

After filming, chroma-key technology was used to perform background extraction on the image sequence as follows

1. A conversion was made between the RGB (Red, Green, Blue) colour-space to the HSV (Hue, Saturation, Value) colour-space.
2. A gamut of the HSV colour-space was defined which encompassed an optimal variation of the background HSV values.
3. All the pixels which had values within the selected HSV region were set to 0 (i.e. the background), while all other pixels were set to 1 (i.e. the foreground), to produce a binary image. This process was repeated, from stage 1, for each image in the image sequence.

Image cropping

Some regions of the images (such as the ceiling and lights) did not contain any green hue and as such were assigned to foreground pixels during the chroma-keying process. Since these regions were consistent throughout the image sequences they could be defined and then cropped from each image.

Connected components analysis

After cropping each image, analysis of the "connected components" was performed to produce several labelled areas. The largest of these areas was assumed to be the human subject and all other areas were assumed to be noise with respect to the image content and assigned pixel values of 0, rendering them as part of the background.

Morphological filtering

The previous processing stages ensures that a connected area representing the human silhouette is found. However, 'holes' may exist in this binary silhouette. A morphological 'closing' operation is performed to eliminate this

1. Dilation is performed to expand the borders of the foreground image, thus filling in any small holes within the silhouette.
2. The image is eroded to remove the pixels added onto the outer border of the silhouette. Note that the inner portion of the silhouette will not change here, hence preserving any filled in holes.

Both the dilation and the erosion were performed using a 3×3 square structuring element.

4.2 Calculation of spatio-temporal Fourier descriptors

Using the silhouettes produced by the methods described above, spatio-temporal Fourier descriptors were produced using the techniques described in Chapter 3. Measures were taken to ensure that the start point for each image in the image sequences was kept consistent. This was achieved by scanning for the start point of the boundary of each subject from the top-left of the image. Since all subjects were walking upright and right-to-left, this ensured that the start point for each boundary was located at the left-most pixel at the top of the head. The effects of rotation were also not considered in this study, as all subjects were walking on a horizontal plane. Following the production of the boundary signals, the Fourier descriptors were then calculated.

4.3 Feature Selection

The resulting descriptors contain a large number of spatio-temporal harmonics, even though the majority of the information about the subject's gait is contained in the relatively small subset of these. In order to use these descriptors for classification purposes it was therefore necessary to reduce the number of descriptors, or 'features' for each subject. This was necessary for two reasons, firstly to extract only the descriptors that would be useful for classification, and secondly to reduce the dimensionality of the feature space – ensuring feasible classification speeds.

The primary aim of feature selection in this case is to minimise intra-subject variance and maximise inter-subject variance, in order to increase the Correct Classification Rate (CCR). Successfully achieving this will result in good clustering in the feature space, such as that seen in Figure 4.2, which shows clustering of 4 different subjects using the three Fourier descriptors which gave greatest class separation. Although only 3 descriptors are used in Figure 4.2 (due to the limitation of only being able to display a three-dimensional plot) many more descriptors were used in classification, and any subjects which overlap or lay close in such a three-dimensional feature space should be separated by the addition of further dimensions found through the feature selection algorithm explained below.

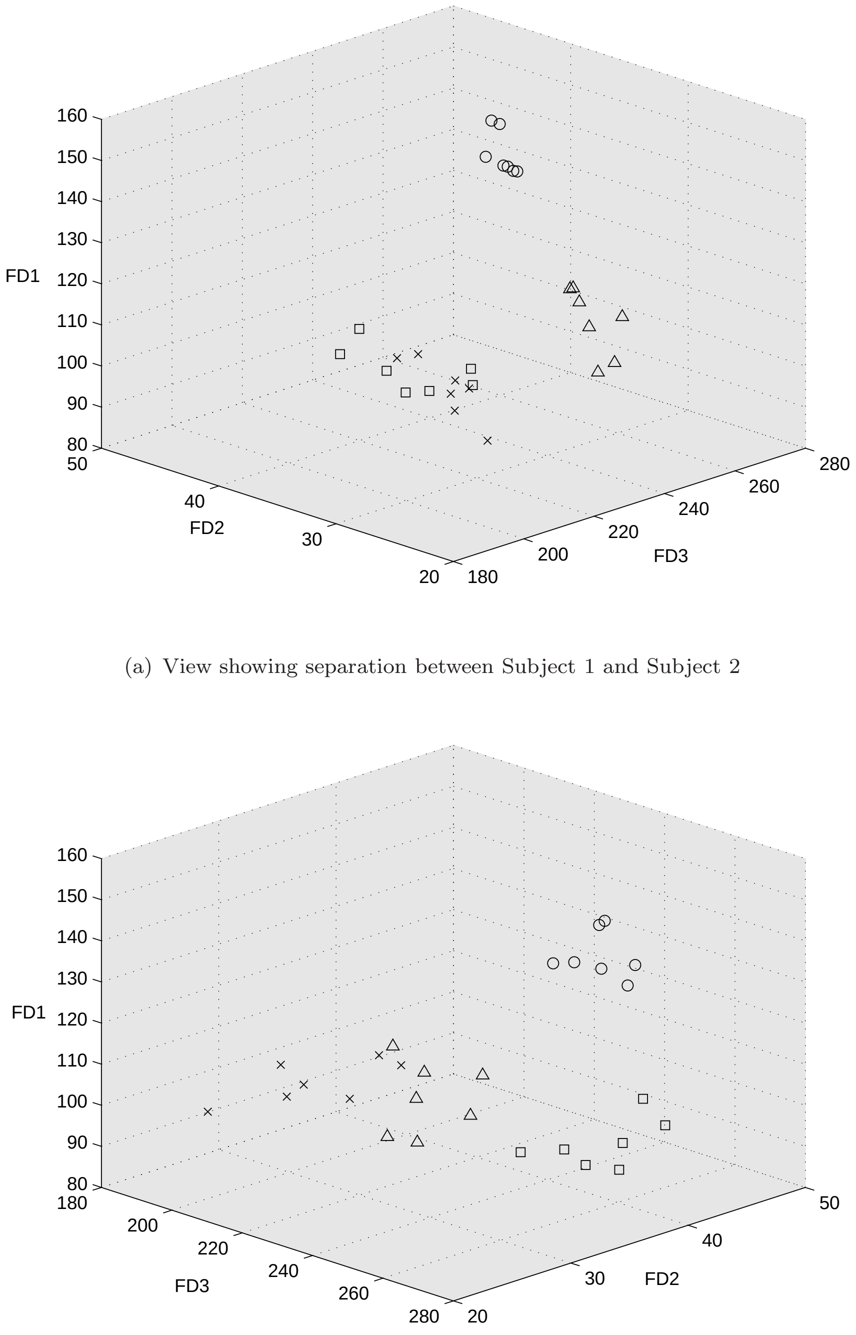
To obtain a measure of inter-subject variance one can use a variation of the Bhattacharyya distance metric to measure inter-class separation due to mean-difference with respect to the class covariances [13] – a method also used by Yam [41]. In practical terms this metric gives a measure of the distance between class means weighted by the amount of overlap between the two classes due to sample variation from the mean. Figure 4.3(a) demonstrates two classes with a relatively low Bhattacharyya distance from each other, with many samples from each class encroaching into the space occupied by the other. In contrast, Figure 4.3(b) demonstrates two classes with a relatively high Bhattacharyya distance from each other, with a well-defined space between the samples of each class being observed; these are classes with good separation.

The separation between the two classes a and b, for a given feature, is given by

$$S_{a,b} = [m_a - m_b] \left[\frac{\sum_{\mathbf{a}} + \sum_{\mathbf{b}}}{2} \right]^{-1} [m_a - m_b]^T \tag{4.1}$$

where m_a is the class mean and $\sum_a$ is the covariance matrix of class a, with equivalent terms for class b.

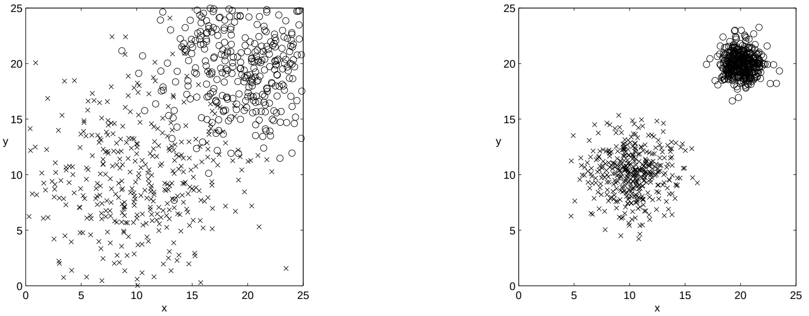
To gain a measure of a feature's ability to separate classes successfully a mean value of S was determined for each feature as



(a) View showing separation between Subject 1 and Subject 2

(b) View showing separation between Subject 3 and Subject 4

Figure 4.2: Clustering of different subjects using three Fourier descriptors: Subject 1 ○; Subject 2 △; Subject 3 □; Subject 4 ×.



(a) Classes with a low Bhattacharyya distance: Class 1: ×, N(10, 2); Class 2: ◦, N(20, 1)

(b) Classes with a high Bhattacharyya distance: Class 1: ×, N(10, 5); Class 2: ◦, N(20, 4)

Figure 4.3: Examples of class separation and associated Bhattacharyya distance metrics.

$$\bar{S} = \frac{1}{N^2} \sum_{a=1}^{N} \sum_{b=1}^{N} S_{a,b} \tag{4.2}$$

where N is the number of subjects for the given data set. $\bar{S}$ is then proportional to the class separability measure of the given feature, with larger values of $\bar{S}$ implying good class separability.

It should be noted that the feature selection process used here was chosen for reasons of simplicity, but other more sophisticated feature selection algorithms could have been employed. In particular, the ANOVA (Analysis Of Variance) technique [37], used to test for significant class mean differences, would be particularly well suited to this. Principal Component Analysis (PCA) could also be used as a means of reducing the dimensionality of the data set prior to classification.

4.4 Classification and initial recognition testing

Initial testing was performed using twenty features selected using the method described above as having the highest class separability values.

Classification was performed using a K-nearest neighbour classifier and cross-validated with the leave-one-out rule. This classifier assigns a test subject to be the same class as that of the modal class of the k^{th} nearest neighbouring subjects to it. If no modal class is found, then the test subject is assigned to the class of the nearest neighbouring subject. The distance between classes is measured by the Euclidean distance, ED, given by

$$ED = \sqrt{\sum_{n=0}^{N-1} (\mathbf{x_n} - \mathbf{y_n})^2} \tag{4.3}$$

where N is the dimensionality of the feature set, and $\mathbf{x_n}$ and $\mathbf{y_n}$ are the values of the n^{th} feature of the samples $\mathbf{x}$ and $\mathbf{y}$ respectively.

The classification results for $k = 1$ and $k = 3$ are shown in Table 4.1.

TABLE 4.1: Initial results of k-nearest neighbour classification

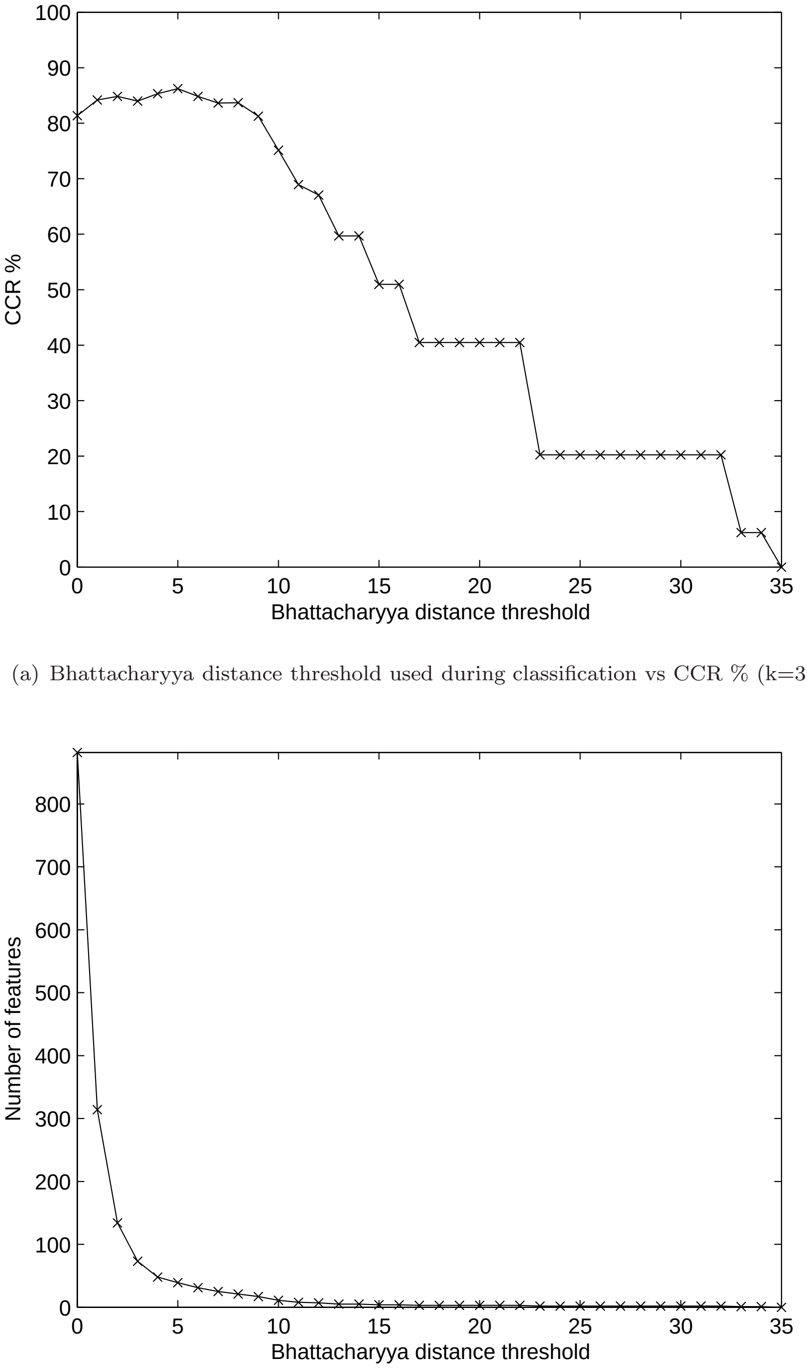
Database	k	CCR(%)
SOTON LARGE	1	82.5
SOTON LARGE	3	83.7

4.4.1 The effects of feature selection

The size of the features set used for classification for the initial recognition test was determined arbitrarily, by choosing the twenty features that had the highest average Bhattacharyya distance score, thus hopefully separating the subjects well in the feature space. The number of features used, however, will affect classification performance: use too few 'good' features (features that should increase class separability) and the feature space will not be separated out enough; use so many features that 'bad' features start to become included and classification performance could be compromised by spreading well defined clusters in the feature space. The appropriate number of features used when classifying is therefore an important factor.

To determine the optimal number of features to use, the recognition test described above was conducted using varying numbers of features[1]. Figure 4.4(a) shows the Correct Classification Rate (CCR) for features chosen by applying a threshold to the Bhattacharyya distance scores, using only features which had a distance score greater than this threshold. On examination of Figure 4.4(b), which shows the asymptotic relationship between distance threshold and number of features selected, we can see that using a threshold of zero yields classification with all features, and close to 900 features are selected. As this threshold is increased, the number of features selected for classification drops off rapidly until, at a distance threshold of 5, just 39 features are used. It is at this distance threshold of 5 that the greatest CCR is found, 86.2%, as can be seen in Figure 4.4(a). From this point, the number of features selected by applying a greater threshold tails off dramatically (as does the CCR), until at a distance threshold value of 35, no features are selected.

[1] It is worth noting that feature selection is dependent upon the data set used, and the features selected for these tests may not show similar results with different data sets.



(a) Bhattacharyya distance threshold used during classification vs CCR % (k=3)

(b) Bhattacharyya distance threshold vs number of features selected)

FIGURE 4.4: The effects of varying Bhattacharyya distance thresholds on classification performance.

In order to examine the relationship between the number of features selected and CCR further we can examine Figure 4.5, which shows this relationship with the number of features plotted logarithmically for ease of viewing. It can be seen from this that selecting a handful of the best features yield recognition rates significantly higher than that which could occur by chance, with 4 features yielding a CCR of around 50% and 5 features yielding a CCR of just under 60%. Increasing the number of features past this point gives smaller, but significant increases in performance, with an increase of 5 features to 10 giving a 15% performance increase, continuing further until at 39 features we reach our greatest CCR of 86.2%. Adding more features past this point generally causes the performance to tail-off, and it can be seen that using many more features actually causes the performance to degrade slightly.

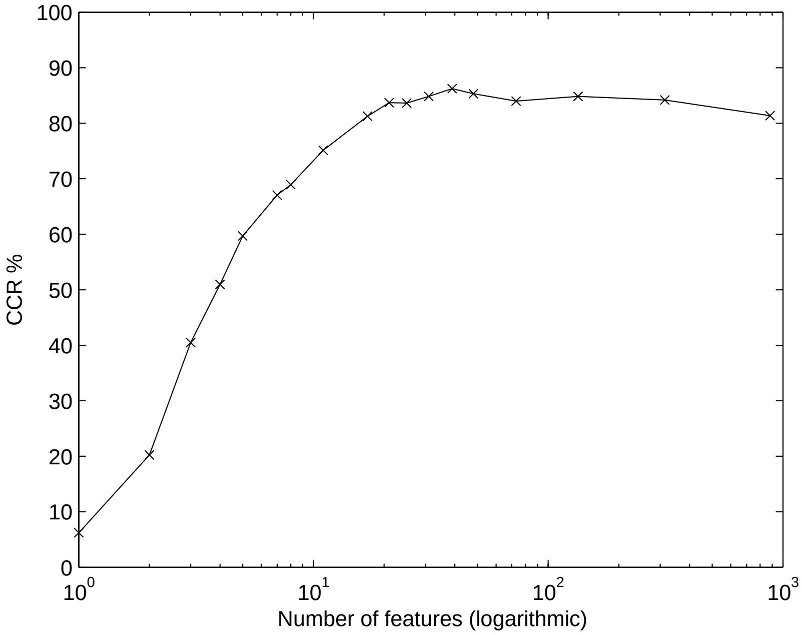


FIGURE 4.5: Number of features used during classification vs CCR% (k=3).

From these results, it can be inferred that any features with a Bhattacharyya distance score of 5 or greater can be classed as 'good' features, whilst any with a lesser score cause breaking down of the clustering in the feature space, and can be classed as 'bad' features.

The best classification rates for the University of Southampton's Large Gait Database using both $k = 1$ and $k = 3$ are shown in Table 4.2.

As can be seen, the selected spatio-temporal Fourier descriptors show a good ability at being able to discriminate between human subjects. These results compare favourably with other studies using the same database, with results of 82.9% and 71.2% for $k = 1$ and $k = 3$ respectively being reported by [10]. Hayfron-Acquah [17] reported recognition

Table 4.2: Best result of k-nearest neighbour classification

Database	k	CCR(%)
SOTON LARGE	1	84.5
SOTON LARGE	3	86.2

rates on the same database using a technique based around spatio-temporal symmetry of 92.8% for $k = 1$ and 86.0% for $k = 3$. Thus showing an improved recognition rate over the technique discussed here for $k = 1$, but not for $k = 3$, with the consistency between k=1 and k=3 suggesting that this technique produces features that demonstrate better clustering in the feature space.

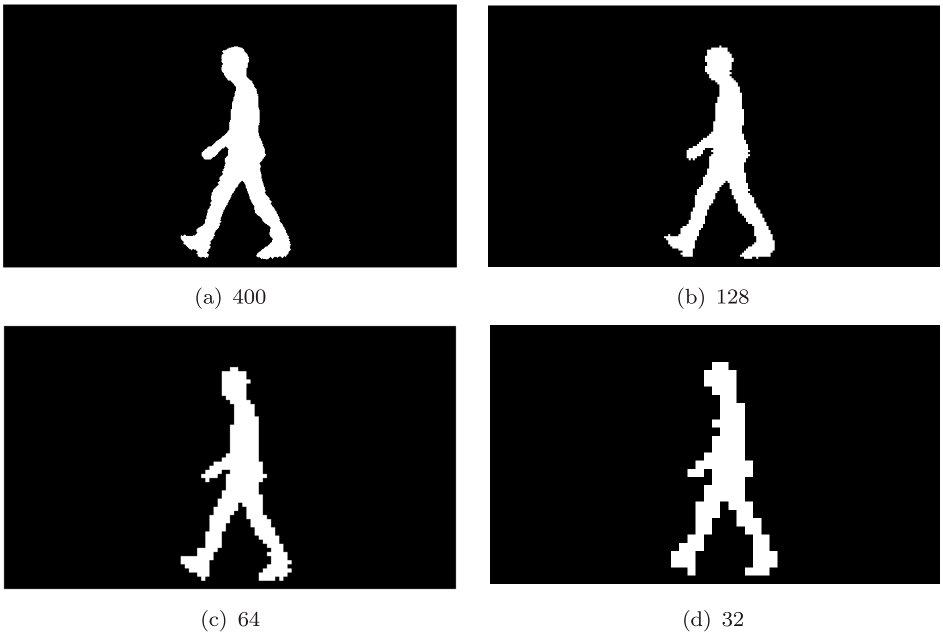
4.4.2 The effect of distance or low spatial resolutions on performance

As mentioned in Chapter 1, one of the main advantages of human gait as a biometric is that it can be captured non-invasively and at a distance. It was therefore deemed prudent to evaluate the performance and robustness of using Fourier descriptors for the recognition of deformable objects at varying spatial resolutions, to simulate the effect of distance.

The effect of distance was simulated by decreasing the spatial resolution of each image in a given image sequence. The original images were subsampled so that the minimum height of the subjects were 128, 64, and 32 pixels respectively (see Figure 4.6) and test feature sets were produced using the same method as the previous recognition test. The original full resolution feature set for each subject was then used for k-nearest neighbour classification. It proved difficult to extract boundaries from sequences with a subject height of less than 32 pixels and so, for the purposes of this test, 32 pixels can be considered to be the lowest limit at which successful classification can be performed. With this said, a subject height of 32 pixels mimics a significant camera-to-subject distance, and could be considered a real challenge for any recognition system.

The results of classification at these resolutions, which are shown in table 4.3, show that the image resolution can be relatively small without a great loss of resolution.

The fact that such a loss of spatial resolution results in only a small drop in CCR can be accounted for by the fact that the majority of the information in a given Fourier descriptor is contained in the low-level descriptors and that for a spatial boundary of length N we can obtain up to $\frac{N}{2}$ descriptors. Therefore, so long as we have a sufficiently large value of N, an adequate number of descriptors can be obtained for recognition. Discretisation errors introduced at such low resolutions, however, mean that errors will be introduced into the descriptors, resulting in the drop of CCR seen here.


(a) 400

(b) 128

(c) 64

(d) 32

Figure 4.6: Varying image resolutions, scaled by subject height.

Table 4.3: Results of k-nearest neighbour classification (k=3) for varying resolutions

Image Height	CCR(%)
400	86.2
128	84.5
64	85.7
32	82.4

During the tests presented here, the highest spatial frequency used for recognition was 10Hz (using the same feature set as was previously selected to give the highest CCR), thus requiring a minimum boundary length of 21 pixels for each boundary signal. This requirement is more than fulfilled, even at the low resolutions used here.

4.5 Summary and discussion

In this chapter we have put the theory of spatio-temporal Fourier descriptors, as discussed in the previous chapter, into practice and have formulated descriptors for 1062 image sequences split between 115 subjects. After calculation of the descriptors, feature selection was performed by calculating the average Bhattacharyya distance between classes for a given feature, with those features that gave the greatest Bhattacharyya distance, and therefore the greatest average class-separability potential, being regarded as the best features. Classification was then performed using a K-nearest neighbour classifier.

The results presented in this chapter demonstrate that the spatio-temporal Fourier descriptor model holds a great amount of descriptive power. As Shutler noted [37], a shape's statistical description greatly improves with the inclusion of temporal information. Recognition rates of over 86% on a large database of subjects are an encouraging indication that spatio-temporal Fourier descriptors uniquely describe not only a shape's spatial description, but also any temporal deformations applied to that shape.

It is clear that gait as a biometric shows promise for recognition of subjects at a distance, and the use of Fourier descriptors provides a means of capturing and modelling the salient information required for the discrimination of different subjects at a variety of spatial resolutions.

Chapter 5

The Continuous Deformable Hough Transform

5.1 Locating deformable objects

In Chapter 4 it was shown how spatio-temporal Fourier descriptors could be used to discriminate between human subjects, or more generally between deformable moving objects. A necessary stage in the extraction of the boundary of a moving shape, however, is that a clean segmentation of the shape is performed. During the tests carried out in the previous chapter segmentation was performed via chroma-keying, but this is a very limited form of segmentation, working only in laboratory conditions, or in situations where a strong contrast in hue or intensity between the object and the background exists.

Segmentation could be made much easier if the subject or object in question could first be located and tracked somehow through a given image sequence – knowing the position and pose of the an object can form a good initialisation point for object extraction. Most current techniques for tracking work on a frame-by-frame basis, identifying an object in one frame and tracking it through successive ones. These techniques, however, often rely on the presumption that the shape is constant throughout the image sequence and that the shape's movement and location is initialised. This is not always the case, and human gait gives a good counterexample to this, where the general boundary of the subject deforms and self-occludes as the subject moves with variable velocity. Further to this, humans demonstrate variation in the way they walk or move. For example, some people of similar height and stature walk faster or slower than others by taking larger or shorter strides.

Evidence-gathering techniques have previously been used to detect static rigid shapes [2, 5, 19], and more recently to detect moving rigid shapes [16, 31] without initialisation or the need for a strong shape definition. Indeed, evidence-gathering techniques provide a robust and reliable method of determining the location and other unknown parameters of a shape, and usually demonstrate good tolerance to noise and occlusion. Shape deformation, which occurs frequently in the real world, has never previously been extracted using an evidence-gathering approach, but periodically deforming shapes seem an excellent candidate for such an approach as, through the use of the new spatio-temporal Fourier descriptor model, shape deformation can now be defined parametrically.

In this chapter we develop a new spatio-temporal form of the Hough transform – the most common form of evidence-gathering mechanism, which is capable of locating deformable shapes (in both space and time) in a given image sequence, and which demonstrates excellent tolerance to noise.

5.2 An introduction to Hough transforms

The Hough transform [19] has been used for many years as a form of perimeter-based object identification, utilising an evidence-gathering approach. The Hough transform works by 'voting' for unknown (or 'free') parameters in a parametric shape model. Typically the Hough transform identifies the location and size of a shape, but any number of unknown parameters can be estimated using the technique.

The original Hough transform was used to detect straight lines in edge-detected images, but since then variants have been developed to detect other parametric shapes, such as circles and ellipses. A more generalised shape detection algorithm, based on the Hough transform, was also developed by Ballard [5] in the form of the Generalised Hough Transform, which allowed the detection of arbitrary two-dimensional shapes. The concept of parameterising arbitrary shapes for extraction was developed further by Aguado et al [2] who used a continuous shape model, based around elliptic Fourier descriptors. Again, this was used as the basis for extracting two-dimensional shapes.

Recent research has further developed the Hough transform to include linear or parametrically modelled motion [31], and even more recently to include arbitrary motion and a continuous shape model [16].

A limitation with all of the previous approaches to detecting moving objects via the Hough transform is that they all assume a fixed shape model, whereas in the real world many shapes, particularly the shapes of organic objects, are anything but fixed, and often deform in a periodic manner. To overcome this, a new Hough transform is proposed and developed, known as the Continuous Deformable Hough Transform (CDHT). The CDHT uses a continuous, spatio-temporal shape model based around the spatio-temporal Fourier descriptors developed in Chapter 3 to detect periodically deforming, moving objects (e.g. humans and animals) within a dynamically changing scene. This new Hough Transform's originality lies in the fact that it deals with the problem of detecting deformable, non-rigid, objects, whereas previous Hough transforms have only detected rigid, fixed-shaped objects.

5.3 Foundation theory

Before the CDHT is discussed in full, other spatio-temporal evidence-gathering techniques will be described that form a basis for the work presented through the rest of this chapter.

5.3.1 The Velocity and Continuous Velocity Hough Transforms

Many image sequences contain a significant amount of temporal correlation, a fact which is frequently utilised by video compression techniques, such as the MPEG series. Very few machine vision algorithms take advantage of this fact however. The Velocity Hough Transform (VHT), developed by Nash [31], was the first evidence-gathering technique to use this temporal correlation to extract the optimal parameters of a linearly moving shape (a conic section). Simple extensions to the VHT, made by Grant [16], made it possible to extract any shape and motion combination, given that both were able to be modelled parametrically.

The VHT was originally developed as an extension of the Hough Transform for circles by adding a linear velocity parameter as a free parameter into the formulae used to cast votes into the accumulator space.

The original Hough transform for circles uses the following formulae to cast votes

$$\begin{aligned} a_x &= c_x + r.\cos(\theta) \\ a_y &= c_y + r.\sin(\theta) \end{aligned} \tag{5.1}$$

where a_x and a_y are the coordinates of the vote to be cast, c_x and c_y are the center coordinates of the circle to be cast into the accumulator (consequently, also the feature pixels from the test image sequence), and r and θ are the polar parameters of the circle in question.

The VHT extends these formulae to include velocity as follows

$$\begin{aligned} a_x &= c_x + r.\cos(\theta) + v_x.t \\ a_y &= c_y + r.\sin(\theta) + v_y.t \end{aligned} \tag{5.2}$$

where v_x and v_y describe the linear velocity of the circle and t represents a time reference relative to the start of the object's motion ($t = 0$).

During the voting process, the time reference, t, and the x-axis and y-axis velocities, v_x and v_y, are used to determine the location of the center coordinates of the circle in the initial frame of motion, at $t = 0$, therefore focusing all votes for the correct circle onto one set of coordinates in the accumulator. Thus, the parameters voted for when using the VHT are the center coordinates $(a_x,\ a_y)$ of the circle at its initial time reference ($t = 0$) with a linear velocity described by $(v_x,\ v_y)$.

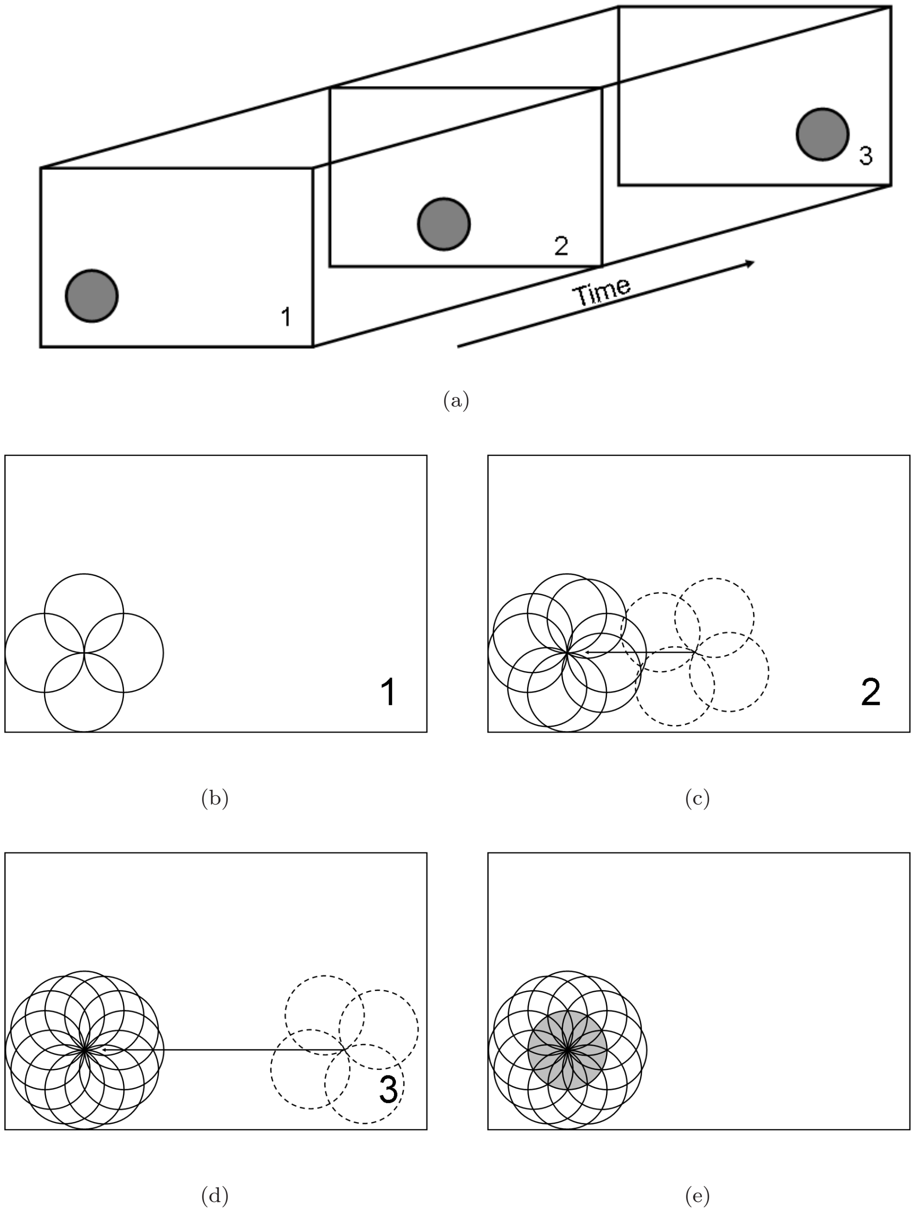
As an example to illustrate this, one could imagine a circle moving with a velocity of $(v_x = 3,\ v_y = 2)$, so that at the frame at time $t = 2$ the center of the circle would have moved by $(\Delta x = 6,\ \Delta y = 4)$. This information can then be used to back-project any votes cast to the center coordinates of the circle at $t = 0$, thus forming one single maximum peak in the accumulator for the correctly identified moving circle.

A graphical example of the voting algorithm for the VHT for a circle is given in Figure 5.1. Figure 5.1(a) shows the original sequence of three frames, while Figures 5.1(b) to 5.1(d) show the accumulator for this image sequence for a correctly identified set of parameters after processing each frame of Figure 5.1(a). The votes depicted by dashed circles in frames 2 and 3 of the accumulator sequence show the original position of the loci of votes before the velocity information is taken into account, causing the votes to be back-projected to time $t = 0$ where all votes for each frame coincide to locate the correct parameters for the moving circle.

Much in the same way that Aguado et al [2] had extended the Generalised Hough Transform (GHT) to use a continuous shape model, the VHT was extended by Grant to create the Continuous Velocity Hough Transform (CVHT) [16]. The CVHT uses a Fourier descriptor parameterisation of the moving shape being sought, in effect replacing the R-table of the GHT, and as such provides an analytic and continuous shape model, ensuring that discretisation errors, usually associated with applying affine transformations to discrete shape models, can be minimised. However, as was the case with the VHT, the CVHT is only able to extract rigid moving shapes.

5.3.2 Shape model dimensionality

The VHT inherits many features from the original Hough transform, including, unfortunately, the "curse of dimensionality", which gets worse as the number of free parameters in the shape's kernel increases. The Generalised Hough transform (GHT) goes quite some way to solving this problem for the case of static, rigid shapes, but suffers from discretisation problems, especially when searching for shapes at much larger scales than the original shape model. A solution to this was to derive a continuous shape model



(a)

(b)

(c)

(d)

(e)

Figure 5.1: The Velocity Hough Transform (VHT) for a circle.

using Fourier descriptors [2] and use this to derive the vote coordinates in place of the GHT's 'R-Table'.

The CVHT combined the VHT and this continuous variant of the GHT, thus providing a Hough transform for extracting arbitrary moving objects using a parameter space with a fixed dimensionality.

5.3.3 Motion model dimensionality

The original VHT used a very basic motion model, assuming linear motion and a constant velocity scaling factor. The motion model itself is built into the kernel of the VHT (Equation 5.2), thus changing the motion model requires a change to the kernel itself. If a more complex motion model is added into the kernel of the VHT, however, the dimensionality of the accumulator space then increases proportionally with number of free parameters in the new motion model. Coupled with the fact that the parametric complexity of the shape model can be large for all but the simplest of geometric shapes, the VHT can soon become an impractical evidence-gathering technique for all but the most basic of shape and motion models.

The issue of variable motion is addressed to an extent by a variant of the CVHT that uses a variable-parameter motion model, termed a "motion template". These motion templates assume that the target object's path of motion is known *a priori* and that, as such, they can be modelled using a known motion model. As with the perimeter of the shape, these motion templates were modelled continuously using Fourier descriptors, and were allowed to be modified by four free parameters to control spatio-temporal scaling, temporal phase, and spatial orientation.

5.4 Continuous Deformable Hough Transform theory

All variations of the Hough transform to date have been designed purely with the detection of rigid, non-deformable objects in mind, either static or moving. As previously mentioned, however, many objects in the real world are not rigid and deform in some predictable, periodic manner as they move.

One specific application of the CVHT was the detection of moving human subjects, but the very nature of the CVHT restricted it to finding rigid objects only, meaning that only parts of the body which were ostensibly rigid could be found using this approach.

A new variant of the Hough transform is now introduced: the Continuous Deformable Hough Transform (CDHT), which is designed to solve the problem of being able to detect moving, deformable shapes in image sequences. As with the CVHT, the CDHT is based around the use of Fourier descriptors, so as to avoid discretisation problems, but whereas the CVHT uses Fourier descriptors to model a rigid shape and its associated motion model, the CDHT uses the spatio-temporal Fourier descriptor model developed earlier in this thesis to model a periodically deforming and moving shape as a single entity.

5.4.1 The CDHT kernel model

If we represent a spatio-temporal shape sequence as a set of complex Fourier descriptor, $\hat{s}(k_t,\ k_l)$, as described in Chapter 3, then a kernel can be defined which casts votes in the accumulator space for each feature point (i.e. each edge pixel), given a newly presented image sequence.

The basis for the CDHT kernel is a normalised set of shape sequence descriptors, derived from Equation 3.19 – a set of complex spatio-temporal Fourier descriptors of the shape we wish to find. The shape sequence descriptors are normalised so that the shape's maximum height and width with respect to the whole shape sequence are set to unity. One drawback to this approach to normalisation is that the aspect ratio of the shape is lost. There are reasons for this, however, particularly when the technique is applied to finding humans, as humans themselves do not have a fixed aspect ratio, and thus allowing non-uniform scaling of both width and height (and therefore justifying normalisation of both) is desirable.

The motion model of the shape sequence descriptors, provided by the spatial DC components, is also normalised, independently of shape size, so that the linear distance that the shape travels (in its direction of motion) falls in the range $[0, 1)$ over one period of deformation, with the shape's initial center of mass at the start of the shape sequence being located at $(0, 0)$. An example of a reconstructed normalised shape sequence descriptor set is shown in Figure 5.2.

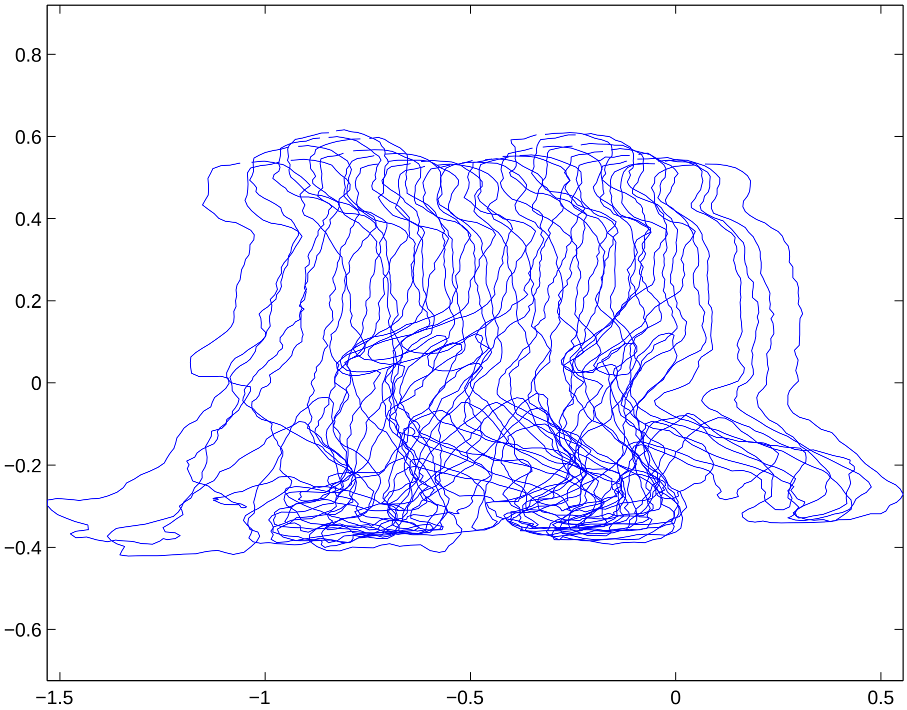


FIGURE 5.2: Example of a normalised shape sequence descriptors.

During the following formal definition of the kernel of the CDHT, the normalised spatio-temporal Fourier descriptor set, which we shall refer to as the Shape Sequence Template (SST), will be represented by the following equivalence, to aid clarity

$$SST(k_t,\ k_l) \equiv norm(\hat{s}(k_t,\ k_l)) \tag{5.3}$$

where *norm* is the normalisation function discussed above and $\hat{s}(k_t,\ k_l)$ is the set of spatio-temporal Fourier descriptors, as described in Chapter 3.

To aid in the definition of the CDHT's kernel, the following domains are also declared at this point

D_p	All pixels in an image.
D_f	All frames in an image sequence.

In order to produce loci of votes for given scales and translations (both spatially and temporally) the SST needs to be transformed. These transformations occur during the reconstruction phase of the CDHT's kernel, but before the signals are fully transformed back into the time domain. Temporal transformations take place in the spatio-temporal frequency domain, and spatial transformations take place in the spatial frequency domain.

Temporal scaling

Since time is one-dimensional, if we wish to scale the SST temporally to extend the period of deformation over a greater period of time, we simply have to multiply each spatio-temporal Fourier descriptor by a scalar factor. This increases the overall power of the spatial frequency deformation described by the spatio-temporal Fourier descriptors in the SST, in effect normalising the energy contained within this spatial frequency deformation when observed over a greater period of time

$$SST(k_t, k_l) = t_s SST(k_t, k_l) \tag{5.4}$$

In practice, as discretised frames of spatial data (i.e. shape boundary signals sampled equally through time) are used to test the CDHT, this scaling mechanism is always used in conjunction with temporal interpolation during the reconstruction of the CDHT's kernel, as discussed later.

Velocity Scaling

Before we proceed any further, we will define a new term: a Shape Template (ST), which represents an individual frame of the spatio-temporal Fourier descriptor model in the spatial frequency domain. The full set of STs can therefore be calculated as

$$ST(t, k_l) = \int_{k_t=0}^{K_T} SST(k_t,\ k_l) e^{j2\pi \frac{t.k_t}{K_T}} dk_l \tag{5.5}$$

With the SST now a collection of STs in the spatial frequency domain, the motion pattern over one period of shape deformation is defined by the varying DC components of the STs through time, t.

The initial SST normalisation process, as described in Equation 5.3, ensured that the motion pattern was uniformly scaled to be in the range (0, 1), therefore preserving the overall motion pattern of the shape in the sequence, but discarding any velocity information. This was a necessary step, since no assumptions should be made with regard to the relationship between the shape's size and its velocity, and as such, spatial size and velocity may need to be scaled independently. To give an example of why this is so, one could imagine two people of differing size and stature walking in a busy crowd. In this situation the two people may well be walking at very similar velocities, even though their physical sizes are very different.

Velocity scaling is a relatively simple process, requiring only scaling of the spatial DC components. Since scalar multiplication in the spatial frequency domain is analogous to spatial scaling by the same factor in the time domain, this is a simple process and merely involves multiplication of the DC components for each ST by a given velocity scaling factor. Velocity scaling is therefore defined as

$$ST(t,\ 0) = v_s.ST(t,\ 0) \quad t = (0,\ T] \tag{5.6}$$

Spatial Scaling

If it was sufficient to spatially scale each shape in the SST in a uniform manner then it would be possible to simply multiply every spatial Fourier descriptor for a given ST (excluding $k_l = 0$, which contains the DC-based motion model) by a scalar value. Non-uniform scaling, however, as required in the case of the CDHT, is not so simple. In this situation a problem arises due to the nature of the complex Fourier transform when using complex time-domain data. This leads to the spectra for both the spatial x and y projections, which are combined as a complex signal, being combined when transformed into the spatial frequency domain. In order to perform non-uniform scaling it is therefore necessary to decompose each complex ST into its respective x and y frequency spectrum, then scale each respectively by the associated x_s and y_s scaling factors. This decomposition operation ensures that the scaling remains in the spatial frequency domain, thus reducing the impact of discretisation errors, which may have been encountered had scaling been applied in the time domain.

The real part of a complex time-domain signal (in this case representing the x axis projection) produces a frequency spectrum with an even real part and an odd imaginary part. Conversely, the imaginary part of a complex time-domain signal (representing the y axis projection) produces a frequency spectrum with an odd real part and an even imaginary part. In the spatial frequency domain the spectra for the two projections are therefore combined, but are separable by an odd/even function decomposition.

This spectra decomposition is detailed below in Equation 5.7. For sake of clarity here we shall refer to the real part of frame t's spatial frequency domain (representing the spectrum of the x axis projection) as $X(t,\ k_s)$ and the imaginary part (representing the spectrum of the y axis projection) as $Y(t,\ k_s)$.

$$X(t,\ k_l) = \begin{cases} \frac{\Re(ST(t,\ k_l))+\Re(ST(t,\ -k_l))}{2} + j\frac{\Im(ST(t,\ k_l))-\Im(ST(t,\ -k_l))}{2} & k_l > 0,\ t = (1,\ T] \\ \Re(ST(t,\ 0)) & k_l = 0,\ t = (1,\ T] \end{cases}$$

$$Y(t,\ k_l) = \begin{cases} \frac{\Im(ST(t,\ k_l))+\Im(ST(t,\ -k_l))}{2} + j\frac{\Re(ST(t,\ k_l))-\Re(ST(t,\ -k_l))}{2} & k_l > 0,\ t = (1,\ T] \\ \Im(ST(t,\ 0)) & k_l = 0,\ t = (1,\ T] \end{cases} \tag{5.7}$$

Scaling is then achieved by multiplying each spectrum by its respective scaling factor, avoiding scaling the DC components, therefore being equivalent to scaling after translating the shape's center of mass to the origin.

$$\bar{X}(t,\ k_l) = \begin{cases} x_s X(t,\ k_l) & k_l > 0 \\ X(t,\ k_l) & k_l = 0 \end{cases}$$

$$\bar{Y}(t,\ k_l) = \begin{cases} y_s Y(t,\ k_l) & k_l > 0 \\ Y(t,\ k_l) & k_l = 0 \end{cases} \tag{5.8}$$

5.4.2 Considerations when working with discrete signals

When used with discrete images sequences, the CDHT will use a discretised accumulator space. Due to this, when transforming the SST into the time domain in order to form loci of votes in this accumulator, there eventually becomes a need to discretise the location of the votes with respect to both time (t) and spatial boundary length (l). If this discretisation is performed incorrectly, and samples become spaced too far apart from each other, then these loci of votes will appear unconnected, and only a sparse representation of votes may be formed. This effect can often also be seen when scaling regular discretised shapes using time-domain scaling, and was demonstrated in Figure 2.6 in Chapter 2.

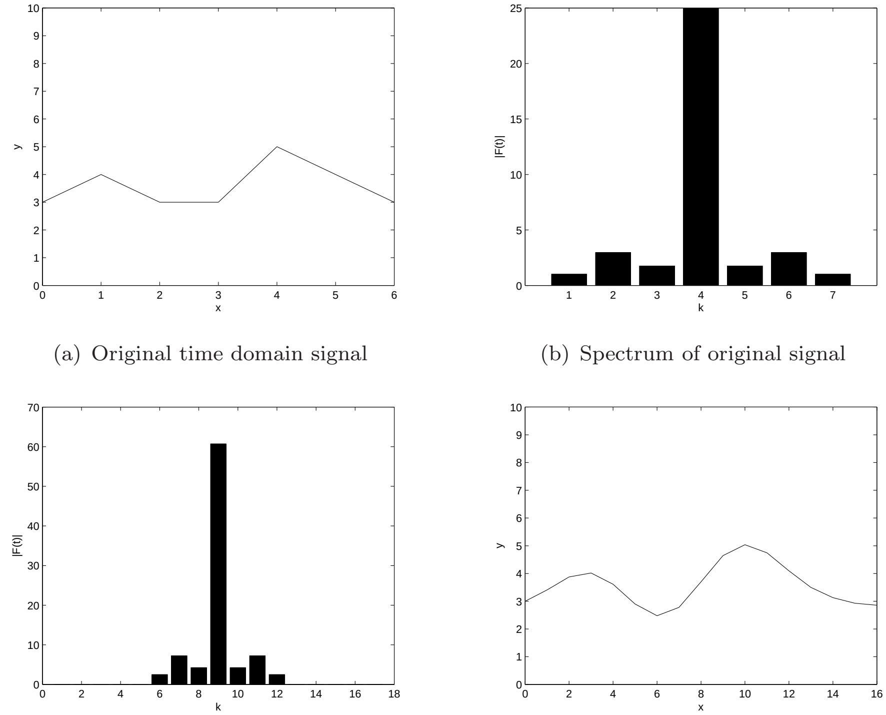
In order to resample this data into the appropriate number of samples for a given scaling factor, we draw on the fact that zero-padding in the frequency domain, leads to interpolation in the time domain [28]. Fourier interpolation in the frequency domain essentially involves zero-padding a signal's spectrum by appending zero-valued frequencies to the extremes of both the positive and negative frequency components. When transformed into the time domain, this results in the addition of evenly spaced interpolated sampling points. Although no further information or detail can be added to the time-domain signal through this method of interpolation, it does allow the time-domain signal to be spread over a larger number of perceived sampling points.

The basic principal of Fourier interpolation is demonstrated in Figure 5.3, where the time-domain signal in Figure 5.3(a), containing N samples, is first transformed into the frequency domain, shown in Figure 5.3(b), resulting in K frequency components. The spectrum of this original signal is then zero-padded (Figure 5.3(c)) to contain K_2 frequency samples and scaled by $\frac{N}{N_2}$, in order to conserve the energy of the signal, before being transformed back into the time domain, as shown in Figure 5.3(d). This reconstructed time-domain signal can then observed as the original time-domain signal interpolated over N_2 points.

Conversely, if the number of samples needed to rescale temporally is less than the number of temporal descriptors available, truncation of the descriptors must occur (removing an equal number negative and positive frequency-based descriptors). This will inevitably result in a loss of high-frequency detail, so care should be observed when performing this operation to ensure that an adequate representation of the time-domain signal is not lost.

Temporal interpolation

Temporal interpolation dictates the number of samples in time over which one period of the shape sequence exists. This is always used with a time scaling parameter, t_s, as was described previously.



(a) Original time domain signal

(b) Spectrum of original signal

(c) Zero-padded spectrum

(d) Interpolated time domain signal

Figure 5.3: Using zero-padding in the frequency domain to interpolate in the time domain.

When temporal scaling is required, Fourier interpolation is applied to the temporal dimension of the SST, which when transformed into the spatial frequency domain interpolates the periodic change in spatial frequencies over T_S frames, where T_S is the total number of desired samples (frames) in the new temporally scaled shape sequence. As time is one-dimensional, T_S will always equal the temporal scaling factor, t_s.

Spatial interpolation

As spatial scaling is two-dimensional and not non-uniform, calculating the number of samples over which to interpolate is more complicated, and there is not a direct (or a particularly simple) correspondence between the arc-length of a shape's boundary and the spatial scaling parameters x_s and y_s. The original boundary signals used in this thesis were sampled with a maximum Euclidean sampling distance of 1 pixel (due to 4-connected boundary sampling), resulting in a total arc-length of L for a given two-dimensional shape. Basic spatial scaling will simply increase the distance between these samples, not increase their number, resulting in an incomplete tracing of the loci of votes in the accumulator space of the CDHT. Given this, resampling is needed if we wish to maintain a maximum sampling distance of one pixel after spatial scaling has taken place. Again this is performed in the frequency domain in order to reduce any unnecessary discretisation errors.

The total arc-length of the perimeter of a scaled two-dimensional shape can be calculated by observing that the length of a shape's perimeter is the sum of all the Euclidean distances between an infinite number of samples around the shape's boundary. When considering this for the application of shapes formed from analytic two-dimensional curves, this can be calculated by integrating the first-order derivative of a Euclidean distance function, formed from the analytic descriptions for the x and y projections, for all values between 0 and 2π around the shape's perimeter. Thus giving the total Euclidean distance traversed around the arc-length of the shape's boundary

$$L_s = \int_0^{2\pi} \sqrt{x'(l)^2 + y'(l)^2} \tag{5.9}$$

Working in the spatial frequency domain, analytic descriptions for the scaled x and y projections are readily obtained from a Fourier series with coefficients defined by the spatial Fourier descriptors a_{xk}, b_{xk}, a_{yk}, and b_{yk} obtained from a given Shape Template

$$x(l) = \frac{a_{x0}}{2} + x_s \left[\sum_{k=1}^{K} a_{xk} \cos(kl) + b_{xk} \sin(kl) \right] \tag{5.10}$$

$$y(l) = \frac{a_{y0}}{2} + y_s \left[\sum_{k=1}^{K} a_{yk} \cos(kl) + b_{yk} \sin(kl) \right] \tag{5.11}$$

with the first-order derivatives of these scaled x and y projections being

$$x'(l) = \sum_{k=1}^{K} x_s k \left[-a_{xk} \sin(kl) + b_{xk} \cos(kl) \right] \tag{5.12}$$

$$y'(l) = \sum_{k=1}^{K} y_s k \left[-a_{yk} \sin(kl) + b_{yk} \cos(kl) \right] \tag{5.13}$$

The total arc-length, L_s, can then be found by substituting Equations 5.12 and 5.13 into Equation 5.9.

L_s then gives a minimum number of samples, and therefore determines the amount of zero-padding needed in the spatial frequency domain, in order to reconstruct a signal that has a unitary maximum inter-sample Euclidean distance.

The CDHT kernel model

With the previous transformations and considerations in mind, the CDHT kernel can now be defined as being the rescaled version of the SST, for the range of all scale values, transformed into the time domain.

$$\bar{\omega}(t,\ l,\ x_s,\ y_s,\ t_s,\ v_s) = \int\limits_{k_t=0}^{K_T-1} \left[\int\limits_{k_l=0}^{K_L-1} SST_{scaled}(k_t,\ k_s,\ x_s,\ y_s,\ t_s,\ v_s) e^{\frac{j2\pi k_l}{K_L}} dk_l \right] e^{\frac{j2\pi k_t t}{K_T}} dk_t \tag{5.14}$$

This kernel defines a scaled spatio-temporal locus of feature points for a set of given scale parameters, effectively replacing the 'R-table' used by the GHT.

5.4.3 The CDHT voting process

The votes to be cast into the accumulator are found by offsetting the CDHT's kernel from the coordinates of each feature point in a given test image sequence, IS, defined by

$$IS = \left\{ \bar{\lambda}(f, \mathbf{p}) \mid f \in D_f,\ \mathbf{p} \in D_p \right\} \tag{5.15}$$

where $\bar{\lambda}(f, \mathbf{p})$ is a function that defines the feature points $\mathbf{p}$, with coordinates (p_x, p_y), for each frame, f.

Given this, the accumulator vote function is defined as the x and y projections of the kernel model (the real and imaginary parts respectively) back-projected from each feature point, for a given set of scale and temporal offset parameters

$$A_{pf} = \left\{ \begin{array}{l} \bar{\lambda}(f, p_x) - \Re\left\{\bar{\omega}(f - t, l, x_s, y_s, t_s, v_s)\right\}, \\ \bar{\lambda}(f, p_y) - \Im\left\{\bar{\omega}(f - t, l, x_s, y_s, t_s, v_s)\right\} \end{array} \middle| \, t \in D_t, l \in D_l \right\}, f \in D_f, \mathbf{p} \in D_p \tag{5.16}$$

for values of $f \geq t$ (as time is a monotonically increasing function), where D_t is the domain of the possible temporal locations for the start of the shape sequence and D_l is the domain of the arc-length parameter of the given temporal sample's spatial boundary. A_{pf} then defines a set of vote coordinates, from which votes will be cast into the accumulator.

With the accumulator vote function defined, it is now necessary to define a matching function that will map the vote coordinates defined by A_{pf} into the accumulator space. This matching function determines if a point in the accumulator's parameter space $\overline{pf}$, should be incremented for a point, $\bar{a}$, in set A, and if so by how much it should be incremented. The simplest form of matching function simply increments a matched accumulator point by unity. Thus given as

$$M(\overline{pf}, \bar{a}) = \begin{cases} 1 & \overline{pf} = \bar{a} \\ 0 & \overline{pf} \neq \bar{a} \end{cases} \tag{5.17}$$

This matching function is then applied to A_{pf} for a range of parameter values, thus fully defining the continuous version of the CDHT as

$$\begin{aligned} &CDHT(\delta, t, x_s, y_s, t_s, v_s) = \\ &\quad \int\limits_{l} \int\limits_{t=0}^{F} \int\limits_{f=t}^{F} \int\limits_{p} M(\delta,\ t,\ \bar{\lambda}(p, f) - \bar{\omega}(f - t, \delta, x_s, y_s, t_s, v_s))\ dp\ df\ dt\ dl \end{aligned} \tag{5.18}$$

where δ is the spatial translation vector and t is the temporal translation (in frames) to the center of mass of the first shape in the shape sequence. In practice, this continuous parameter space is then sampled into a discrete parameter space, given by

$$\begin{aligned} &DCDHT(\delta, t, x_s, y_s, t_s, v_s) = \\ &\quad \sum_{l \in D_l} \sum_{t=0}^{F} \sum_{f=t}^{F} \sum_{p \in D_p} M(\delta,\ t,\ \bar{\lambda}(p, f) - \bar{\omega}(f - t, \delta, x_s, y_s, t_s, v_s)) \end{aligned} \tag{5.19}$$

with the global maxima of this space corresponding to the best matches of the reconstructed SST within the given image sequence.

5.4.4 Pseudo-code algorithm

A pseudo-code algorithm detailing the CDHT voting process is detailed below

```
For ($x_s = x_{smin}...x_{smax}$)
  For ($y_s = y_{smin}...y_{smax}$)
    For ($t_s = t_{smin}...t_{smax}$)
      For ($v_s = v_{smin}...v_{smax}$)
        Generate scaled kernel, $SST_{vst}$
        For ($l = L$)
          For ($f = 0...F$)
            For (frame $t...F$ in the image sequence)
              Generate time domain kernel, $\bar{\omega}(f-t,\ l,\ x_s,\ y_s,\ t_s,\ v_s)$
              For (each feature pixel ($p_x$, $p_y$))
                Generate vector $\delta(a_x,\ a_y)$ from $\bar{\omega}(f-t,\ l,\ x_s,\ y_s,\ t_s,\ v_s)$
                Generate the vote vector ($vote_x,\ vote_y$)=($p_x,\ p_y$)-$\delta(a_x,\ a_y)$
                DCDHT[($vote_x,\ vote_y$), $t,\ x_s,\ y_s,\ t_s,\ v_s$]++
```

5.5 Dimensionality

While computation time for the CDHT will obviously be dependent on the implementation of the algorithm and the available hardware, it is worth noting that the accumulator is a fixed seven-dimensional space, with the parameters being

δ	a two-dimensional spatial translation vector.
t	a temporal translation parameter.
x_s	a spatial scaling parameter along the x-axis.
y_s	a spatial scaling parameter along the y-axis.
t_s	a temporal scaling parameter.
v_s	a velocity scaling parameter (distance travelled in one period).

5.6 Summary

This chapter has presented the foundation theory for a new form of the Hough transform, the Continuous Deformable Hough Transform, which uses an evidence-gathering approach to find moving, periodically-deforming shapes. For each feature point in a newly presented image sequence, loci of votes are calculated from a scaled kernel model, derived from the spatio-temporal Fourier descriptor model presented in Chapter 3. These votes are then cast into an accumulator space. After all feature points are processed, the global maximum of the accumulator space indexes the best-fit parameters of the model of the shape in the image sequence in question. If more than one shape sequence is expected to be found in the image sequence then the parameters of the models for these will be indexed by a number of local maxima.

Chapter 6

Application of the CDHT: from theory to practice

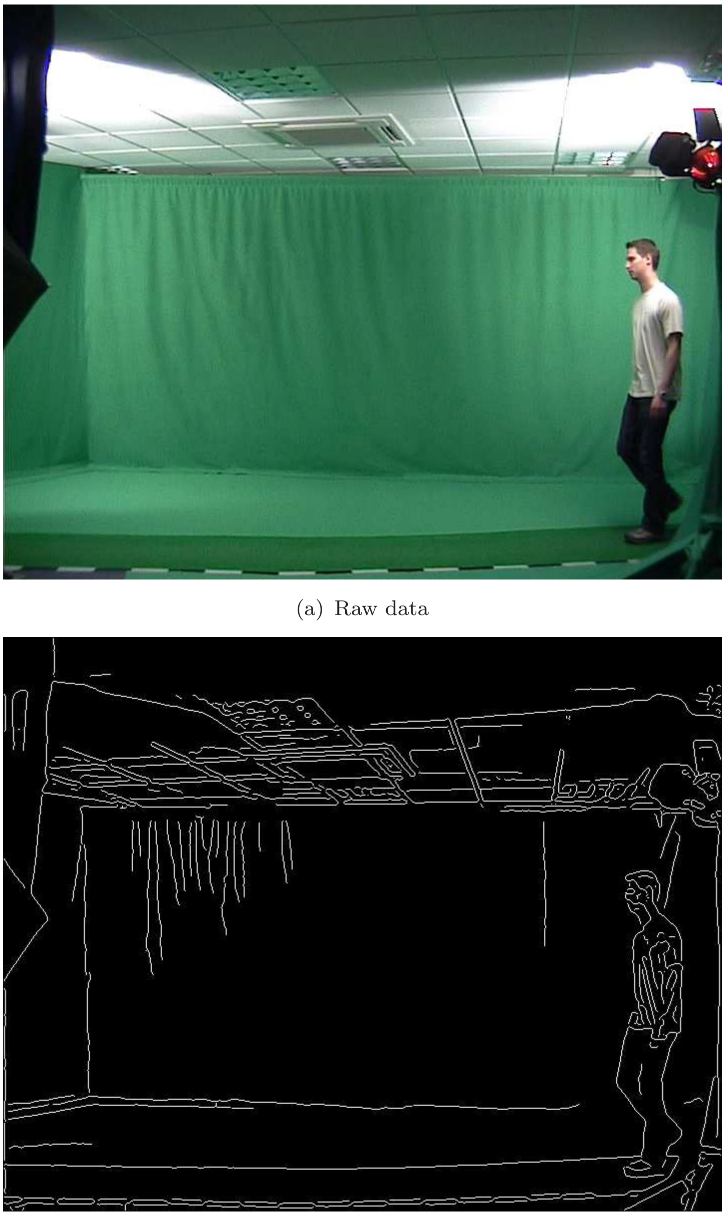
With the theory of the Continuous Deformable Hough Transform (CDHT) now defined, this chapter will now demonstrate how it can be applied to a real image sequence to successfully find the correct parameters for a moving and deforming object. Due to the constraints of time and processing power, we will perform limited testing on full-sized image sequences, but shall do so here in order to demonstrate the CDHT's ability to work on such data and to enable the reader to fully analyse the voting mechanisms and potential results obtained from using high resolution image sequences. Performance and robustness tests, which will be presented later in the thesis, use many image sequences and will therefore use spatially subsampled images.

6.1 Data and pre-processing

The data used to test the application of the CDHT is taken from the University of Southampton's Large Gait Database, consisting of PAL DV image sequences (720 x 576) at 25 frames per second. This usually equates to a typical gait cycle taking up between 20 and 30 frames of data – more than sufficient to capture the temporal deformation demonstrated in human gait, which is the primary subject matter.

Prior to being used as input into the CDHT algorithm, the data was processed by edge-detecting each image using the Canny operator. The parameters for the Canny operator were judged by eye to give a good balance between clean edges being detected in the subject and weak or noisy edges being detected in the background. A typical original frame and resulting edge map is shown in Figure 6.1.

The test was limited to searching for temporal scales (number of frames over which one full gait cycle is sampled) of between $20 \ldots 30$, in 1 frame increments, spatial scales of



(a) Raw data

(b) Canny detected edges

FIGURE 6.1: Raw test data and associated edge map.

between $x = 150 \ldots 170$ and $y = 280 \ldots 360$, in 5 pixel increments, and velocity scales of $230 \ldots 260$, also in increments of 5 pixels. While this is not an exhaustive set of free parameters,[1] it was deemed, by examining a number of test image sequences, to be a sufficient range to cover most of the human subjects in the Southampton database, and therefore provide a realistic set of search parameters.

[1] An exhaustive set of search parameters would require an infinite amount of computation time!

6.2 Generating a generic gait kernel

The CDHT was applied to the image sequence shown in Figure 6.1(b) using a generic gait kernel model. This kernel was generated by producing an averaged Fourier descriptor model using ten sequences generated from silhouette data from the Southampton Large Gait Database. This was produced in the same manner as described in Chapter 4. These test sequences were chosen manually with the criterion that they should be relatively clean extractions from the original data, and would therefore form a good representative model of human gait. An equal number of male and female subjects were chosen, so at to try to minimise any gender bias in the final kernel model.

Averaging many Fourier descriptor models to produce a single kernel model in this way can be thought of as analogous to the process of Fourier signal averaging – a technique used to extract clean signals from noisy data by calculating the mean spectra for many samples of the same signal. Averaging spectra in this way allows deviations from the true signal, usually caused by noise, to effectively become averaged out, leaving only the spectrum for the true signal remaining. However, in the case of the human gait models used here, the spectra for the models also consist of inter-subject variation, along with a degree of noise. Averaging different subjects' gait spectra (so long as they are normalised with respect to each other) will therefore result in this inter-subject variance being averaged out, along with the noise. This may not seem a desirable effect, but given that the aim in this instance is to generate a generic gait kernel model, this is exactly the desired result.

The generic gait kernel is therefore defined as

$$SST_{generic} = \frac{\sum_{n}^{N} SST_n}{N} \tag{6.1}$$

where N is the total number of subjects used to generate the kernel.

6.3 Results

The parameters extracted after applying the CDHT to Figure 6.1(b) are shown in Table 6.1, with the generic gait kernel, scaled and translated to these parameters shown superimposed over the original image data demonstrated in Figure 6.2 (For brevity, only selected frames are shown here. The full result sequence can be found in Appendix A).

Table 6.1: Extracted parameters for the test image sequence.

Parameter	Description	Value
x	Starting location in the x dimension	592
y	Starting location in the y dimension	422
t	Starting location in the time dimension	6
x_s	Spatial scaling in the x dimension	160
y_s	Spatial scaling in the y dimension	330
t_s	Temporal scaling (period of the sequence in frames)	25
v_s	Velocity scaling (distance the shape has moved)	255

6.4 Discussion

The results show that the CDHT has performed well in extracting the correct starting location and scale parameters (in both space and time) for the subject in question. At first sight, a number of frames in Figure 6.2 seem to show an ill fit to the kernel model, with various limbs being slightly outside of the model boundaries, or being slightly ahead or behind in time. However, human gait is very variable – a fact that was actually exploited in Chapter 4 in order to differentiate between and classify different subjects. As such, it should come as no surprise that as a generic gait model is being used, the actual data will show slight variations from this. Indeed, the very fact that there is a difference between the model and the data, yet the algorithm still detects the location and scales of the subject correctly, is testament to the robustness of the CDHT. It is probably worth noting that many applications involving variants of the Hough transform involve searching for well-defined rigid objects, and as such they are able to use a very tailored kernel model. As a results of this the model often perfectly fits the test data. In the tests presented here, however, the shape being sought is not well-defined, but is an organic, deformable object – a walking human, and as such a generic model must be used. Given this, it is very unlikely that the model and data will ever match perfectly; the aim is to extract the best-fit parameters of the subject, not segment it perfectly from its surroundings.

Figure 6.3 shows a two-dimensional cross-section of the accumulator space for this test sequence, with the cross-section being drawn from the free parameters x and y, and the third dimension showing the number of votes cast in the accumulator. Even though the input data (the edge detected sequence shown in Figure 6.1(b)) is reasonably noisy[2], the number of votes cast by this noise (best seen in Figure 6.3(b)) is relatively low when compared with the maximal peak in the accumulator, which locates the correct parameters for the subject. The prominence of this peak can be seen even more clearly

[2] Noise in this instance is considered to be the set of edge pixels in the input image that don't belong to the subject.

if one examines Figure 6.4, which shows the same cross-section of the accumulator from two orthogonal side-on views.

6.5 Summary

The transition from the theory to the practical application of the CDHT has been very successful. Using a relatively unprocessed input sequence, the CDHT coped with background noise well and correctly located the free parameters of the subject being sought. The temporal correlation inherent within the CDHT is, no doubt, a major contributing factor to its success here, as Grant [16] and Nash [31] also found with their previous incarnations of temporal Hough transforms. Perhaps what makes the CDHT even more robust, however, is the fact that shape deformation, not only shape movement, is correlated through time, further decreasing the probability of a “false positive” peak appearing in the accumulator. How robust the CDHT truly is, however, will only become apparent with further performance testing, and so it is this area that the thesis will next address.

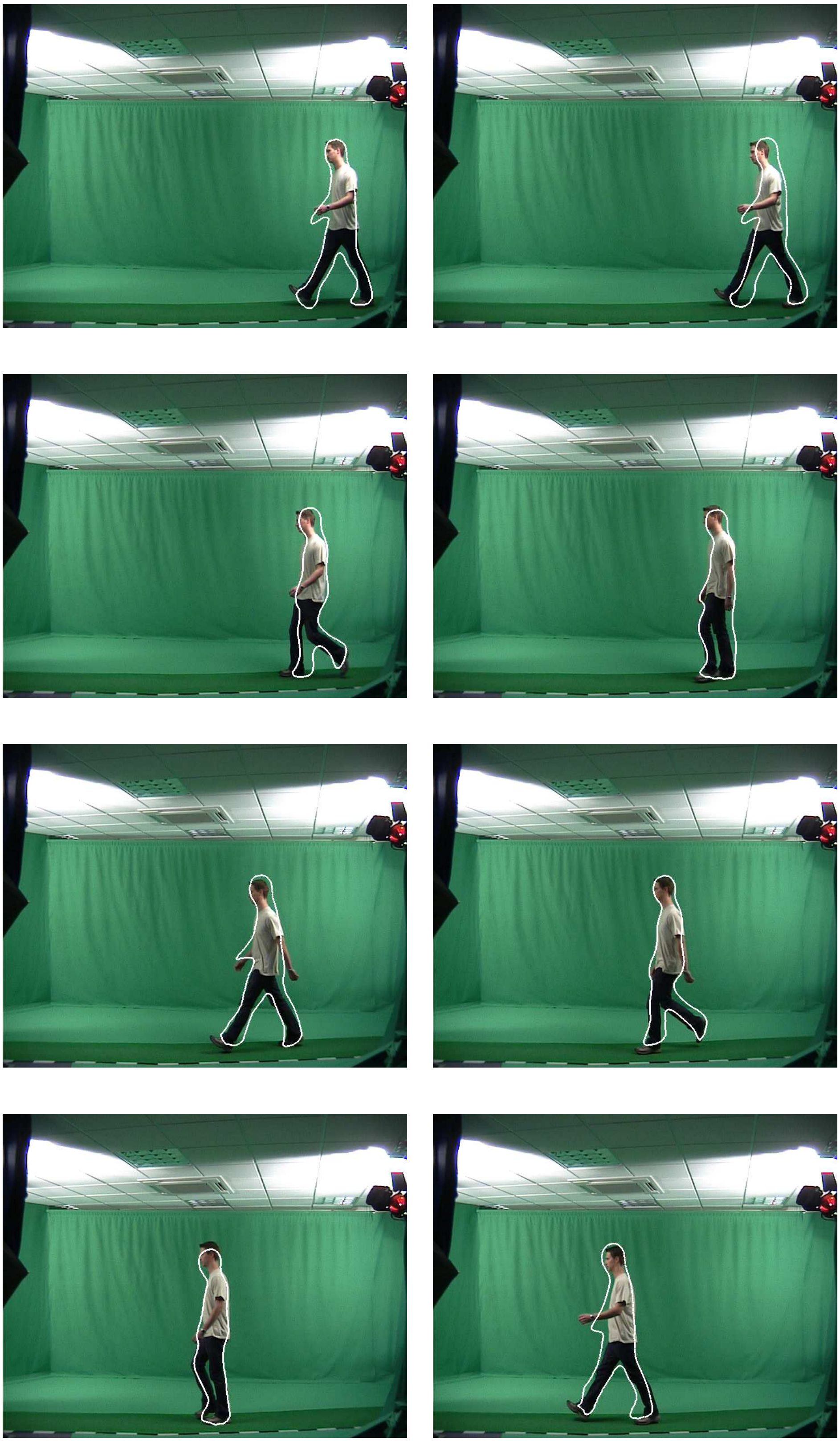
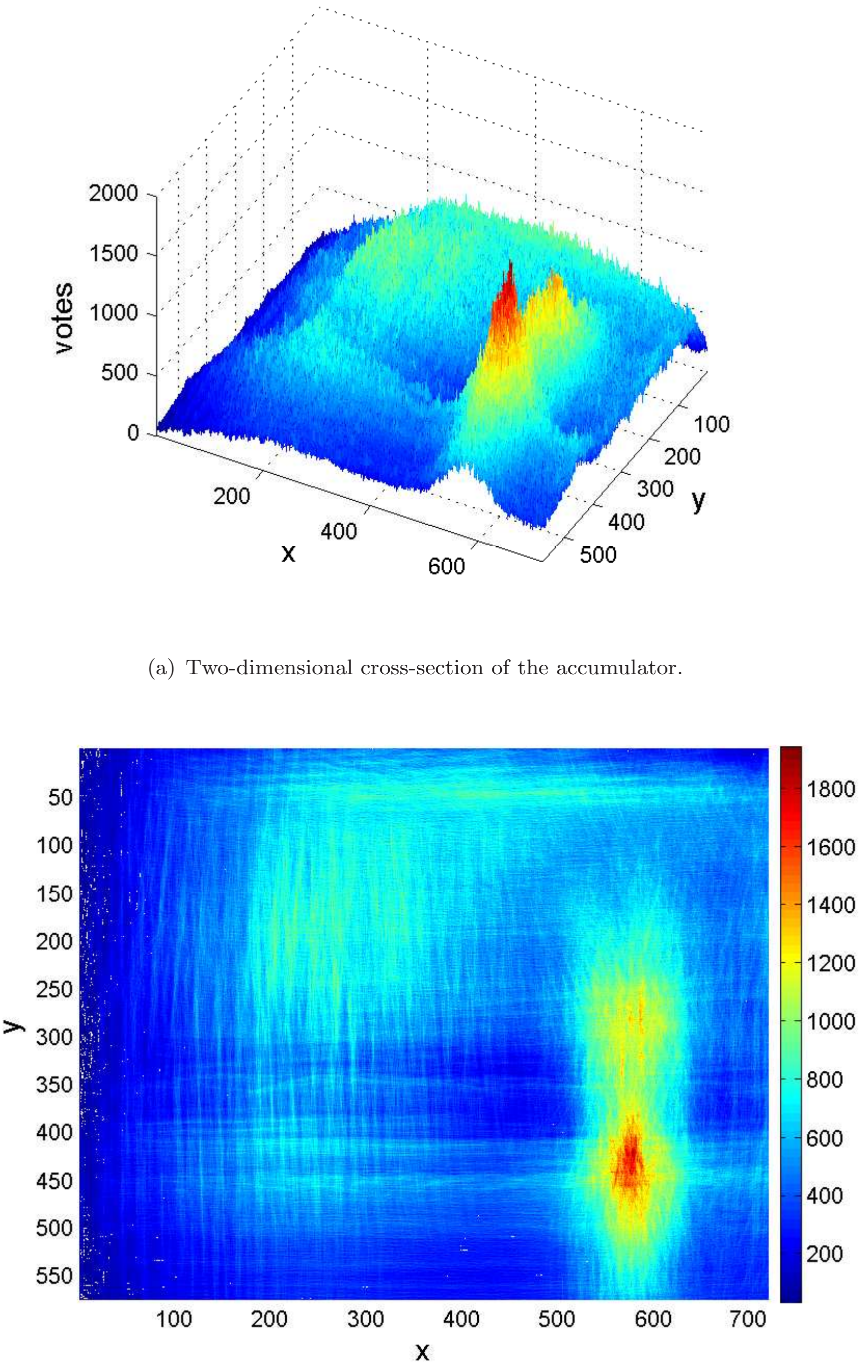


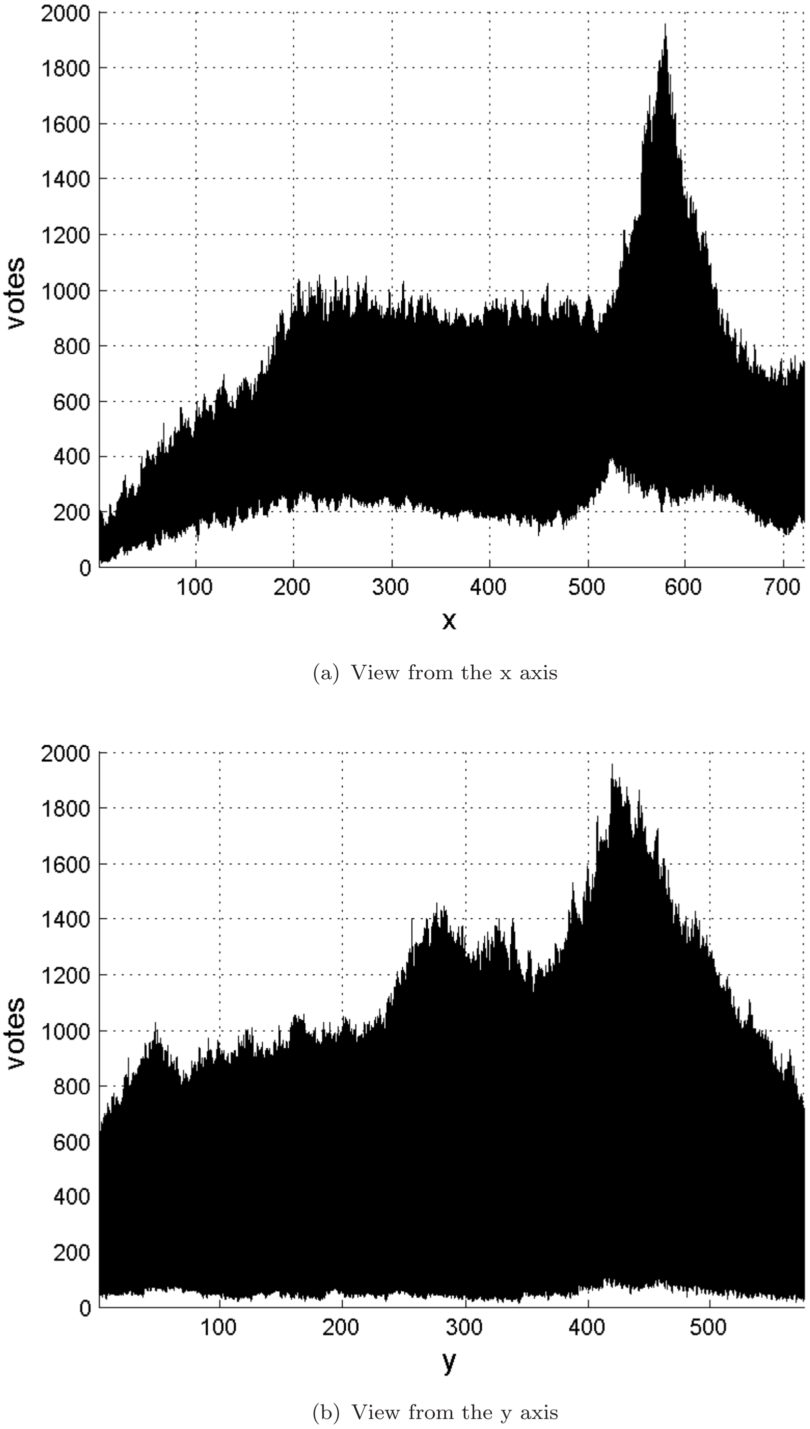
Figure 6.2: Superimposed generic kernel, scaled and translated by the extracted parameters.



(a) Two-dimensional cross-section of the accumulator.

(b) View from above.

Figure 6.3: Views of a two-dimensional cross-section of the accumulator.



(a) View from the x axis

(b) View from the y axis

FIGURE 6.4: Side-on views of the accumulator.

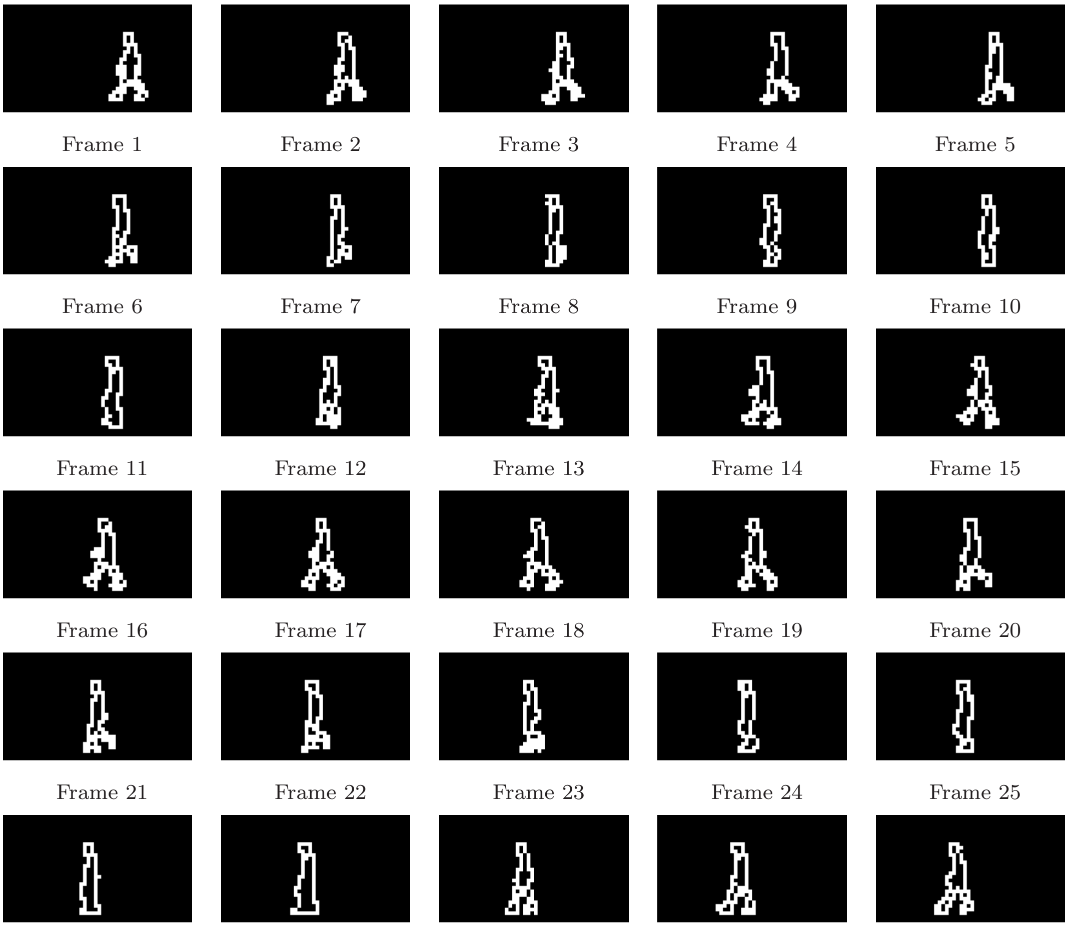
Chapter 7

Noise performance testing of the CDHT

The Hough transform and its derivatives are well known to be very robust to noise, due to their implicit evidence-gathering nature. Noise tends to lead to false feature points in the input data, and consequently false loci of votes in the accumulator space. Usually, however, any true feature points (i.e. edge pixels of the actual deformable shape being sought) remaining after the addition of noise will also cast loci of votes in the accumulator, which will then hopefully coincide to provide a maximal peak indexed by the correct parameters of the shape in question. As long as this peak is large enough so as not to be masked by the noise then the correct shape model parameters should be found. At some point of noise saturation, however, so many noisy votes may exist that the peak created by the shape being sought will become swamped with noisy votes, and at this point the CDHT will fail. Up until this point, a degradation in performance can be expected, and it is the aim of this chapter to examine this.

To test the degradation of the performance of the CDHT in the presence of additive noise, it is necessary to start from a known performance level – a "ground-truth" image sequence that causes the CDHT to perform optimally. In order to generate this ground-truth data set, the image sequence used in these noise performance tests was derived from the spatio-temporal Fourier descriptor model of the image sequence in Figure 3.6; it is based on a real-world example simply to provide a realistic shape and deformation sequence for the purposes of these tests. Using this model, a test sequence was reconstructed at a low resolution in order to make the computational demands of the performance tests detailed here feasible. An already segmented image sequence was used as the basis for generating the test image sequences in this case. Starting with 'clean' and human-validated edge data such as this is ideal for the purposes of performance testing, as the testing procedure wishes to address the effects of additive noise in isolation. Using data that needs to be edge-detected prior to testing in this case would introduce an additional performance variable – the performance of an edge-detector at removing noise and detecting true edges.

It was noted in Chapter 6 that the use of a generic CDHT kernel was necessary when using real-world data, as real-world data can vary a great deal. However, the area of interest here is how the CDHT's performance degrades when presented with noisy data added to a ground-truth data set. To produce this ground-truth data set, it is actually necessary to use an identical kernel model to the shape being sought, in order to force the CDHT to perform optimally. It is therefore valid, even desirable, to create this ground-truth data set and kernel model from the same data. This reconstructed ground-truth image sequence, to which noise will be added during performance testing, can be seen in Figure 7.1.



Frame 1 Frame 2 Frame 3 Frame 4 Frame 5

Frame 6 Frame 7 Frame 8 Frame 9 Frame 10

Frame 11 Frame 12 Frame 13 Frame 14 Frame 15

Frame 16 Frame 17 Frame 18 Frame 19 Frame 20

Frame 21 Frame 22 Frame 23 Frame 24 Frame 25

Frame 26 Frame 27 Frame 28 Frame 29 Frame 30

Figure 7.1: Low-resolution ground-truth image sequence.

7.1 Choosing an appropriate noise model

When testing the noise performance of the CVHT using binary images Grant [16] used a noise model comprising of a zero-mean Gaussian with a standard deviation of 3, using an additive ‘clipped’ form of Gaussian noise (discussed below). However, as Grant noted, adding Gaussian noise to binary image data can cause problems. These problems are now described in order to justify the use of a binary (salt and pepper) noise model, described in section 7.1.2, as the main noise model over a Gaussian noise model. The Gaussian noise model that Grant selected will, however, be used for further testing purposes, but only to form a basis for comparison with Grant’s earlier work.

7.1.1 Problems with additive Gaussian noise and binary data

When added to a binary image, Gaussian noise can produce quite unexpected results. This is due to the need to deal with “out of range” values when adding positive or negative noise values. If, for example, a positive noise value is added to a white pixel (with a value of 1), then a “brighter than white” pixel is produced, and a similar, but opposite condition exists with negative noise values and black pixels (with a value of 0). Obviously some mechanism needs to be put in place to deal with these out of range values.

Two mechanisms are available to deal with out of range pixel values: one can clip the values, as described in Equation 7.1

$$p = \begin{cases} 1 & p+n > 1 \\ 0 & p+n < 0 \\ p+n & p+n \in [0,\ 1] \end{cases} \tag{7.1}$$

where p and n are the pixel and noise values respectively, or one can make the pixel values ‘wrap-around’, as described in Equation 7.2

$$p = \begin{cases} 0 & p+n > 1 \\ 1 & p+n < 0 \\ p+n & p+n \in [0,\ 1] \end{cases} \tag{7.2}$$

However, both mechanisms have their faults. If out of range noise values are clipped, then pixels with additive noise have a tendency to retain their value, and conversely, if out of range values are made to wrap-around, then pixels have a slight tendency to become inverted. To demonstrate this one can imagine a white pixel (pixel value of 1) that has positive noise added to it. This pixel will now have a value of greater than 1 and in the case of a ‘clipped’ noise model would retain its value. Given a zero-mean

Gaussian noise model, the probability of this occurring is 0.5 (50% of the time). If negative noise is added to a white pixel, then the amount of noise added must be greater than any binarisation threshold that may be subsequently applied to the data (used to return the noisy image to a binary state and typically a value of 0.5), otherwise the pixel will again retain its value. The probability of this occurring will be dependent upon the standard deviation of the Gaussian noise model, but will be always be non-zero. A similar condition exists for black pixels (pixel value 0). Given these combined probabilities, it becomes obvious that a 'clipped' Gaussian noise model leaves pixels intact in more than 50% of cases; obviously an undesirable trait for a noise model.

Using a 'wrap-around' noise model one sees a similar, but opposite problem. In this model, a white pixel (pixel value 1) that has positive noise added would 'wrap-around' and become a black pixel (pixel value 0), with a similar condition existing for black pixels and the addition of negative noise. Given a Gaussian noise distribution, the probability of a pixel being inverted in this case is therefore 0.5. Again, however, there also exists a problem with binarisation of this model. If negative noise is added to a white pixel (or positive noise to a black pixel) then there is also a non-zero probability that this noisy pixel value will be great enough to exceed the binarisation threshold and also invert the pixel. The result of applying this noise model is therefore to invert pixels in more than 50% of cases; again, not a desirable trait, and one which could lead to an unnatural inversion of the majority of the image with the appropriate Gaussian noise model.

It might seem that both Gaussian noise models here are equally as flawed, but if one examines Figure 7.2, which shows the probability distribution function for a white pixel with zero-mean Gaussian noise added, it can be seen that this is not the case. Figure 7.2 shows that the probability of a white pixel remaining the same using the 'clipped' Gaussian model is $0.5+a$: the probability of either a positive noise value being added and clipped (0.5), or a negative noise value being added which doesn't push the noisy pixel value past the binarisation threshold (a). If one considers the 'wrap-around' Gaussian model, it can be seen that the probability of a pixel becoming inverted is $0.5 + b$: the probability of either a positive noise value being added, wrapping around and inverting the pixel (0.5), or a negative noise value being added which pushes the noisy pixel value past the binarisation threshold (b). Since the cumulative probability a is greater than the cumulative probability b for any given Gaussian, there is a lower probability of a pixel simply becoming inverted in the 'wrap-around' model than a pixel retaining the same value in the 'clipped' model.

7.1.2 A binary noise model

A good noise model for use on binary data should ensure that, after the addition of noise and undergoing any binarisation process, each pixel value has a 50% chance of being either a 1 or a 0. Given this, and the problems with using binarised additive Gaussian

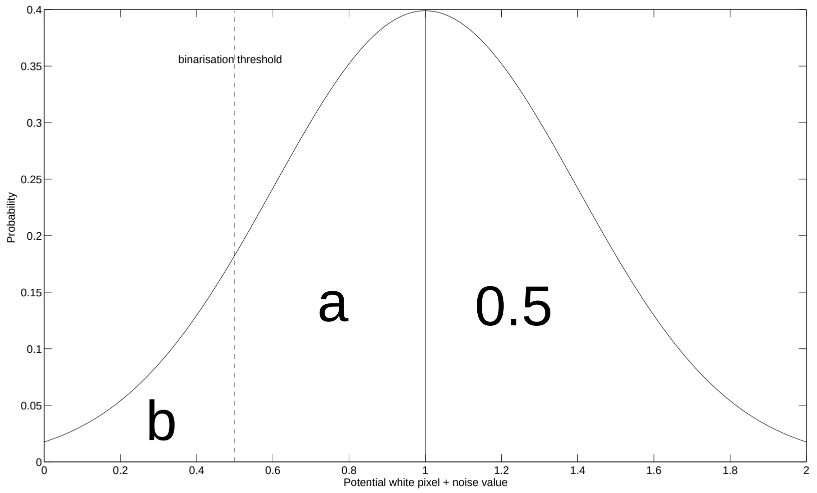


Figure 7.2: Probability distribution function for potential noisy pixel value for a white pixel.

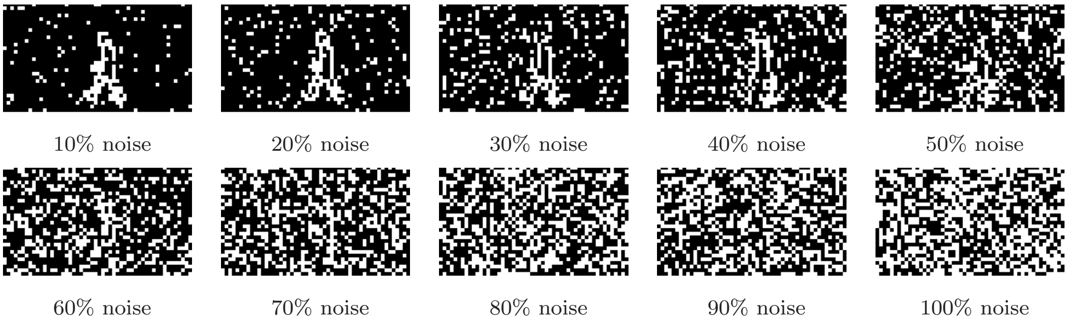
noise described above, a simple, but adequate alternative noise model was chosen that satisfies the aforementioned criterion for a good binary noise model and simply assigns each pixel effected an equal probability of being a 1 or a 0. This is used in conjunction with a uniform random variable controlling the amount of additive noise. For example, adding 10% noise in this way would change roughly 5% of the background pixels into edge pixels (a 1) and roughly 5% of the edge pixels into background pixels (a 0). At 100% noise, the image would contain a purely random distribution of edge and background pixels, with any true background or edge pixels remaining with equal probability, and only by chance. This noise level was also favoured by Shutler on a similar binary data set [37].

7.1.3 Testing

While the basic binary noise model described above is more applicable to binary data, noise testing was performed with both this model and Grant's 'clipped' Gaussian noise model to allow for comparison with the noise performance testing of the CVHT in [16]. The same image sequence used for testing the CVHT cannot be used in these tests, however, as the kernel used for Grant's previous noise testing was based around a rigid shape model, whereas the testing presented here uses a deforming shape model. The CDHT is more than capable of handling rigid shape kernels, just as the CVHT does, but as the aim of this thesis is to demonstrate its ability to extract deformable objects, to extract a rigid shape using the CDHT would undermine its abilities and not demonstrate its full potential.

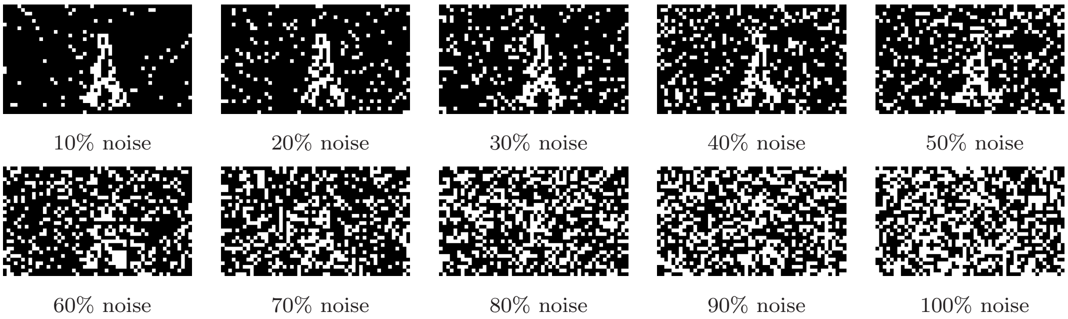
The CDHT was tested for its robustness to noise using 11 noise levels, from 0% – 100%. The varying noise levels (above 0%) used during testing, added to Frame 16 from the test sequence, are shown in Figure 7.3 using the basic binary noise model, and in Figure 7.4 using the clipped Gaussian model (post binarisation). The Gaussian noise model used a zero-mean Gaussian with a standard deviation of 3, as described in [16]. Using a standard deviation of 3 ensures that the problems associated with the clipped Gaussian noise model are minimal.

For each noise level, 50 test sequences were produced independently of each other from the ground-truth (0% noise) test sequence shown in Figure 7.1. These sequences were then used as input into the CDHT and the extracted parameters found were compared to the extracted parameters for the ground-truth test sequence in order to produce a measure of performance degradation. No tolerance to slight deviations in the extracted parameters was allowed in these tests, and therefore a set of extracted parameters could be counted as being incorrect, or a 'miss', even if there was only a slight difference from the ground-truth parameters.


10% noise 20% noise 30% noise 40% noise 50% noise

60% noise 70% noise 80% noise 90% noise 100% noise

Figure 7.3: Frame 16 with varying levels of binary noise added.


10% noise 20% noise 30% noise 40% noise 50% noise

60% noise 70% noise 80% noise 90% noise 100% noise

Figure 7.4: Frame 16 with varying levels of clipped and binarised Gaussian noise added.

7.2 Results and discussion

The result of the noise performance tests using both the binary noise model and the clipped Gaussian noise model can be seen in Figure 7.5. Figure 7.6 shows a comparison between the Continuous Deformable Hough transform and the Continuous Velocity Hough Transform, using a similar data set and the same noise model. In these results, the "hit rate" is defined as the percentage of correctly located subjects, in both space and time, for the 50 image sequences used in the test, where ground-truth location parameters for the subjects were taken directly from the model data used to generate the test image sequences.

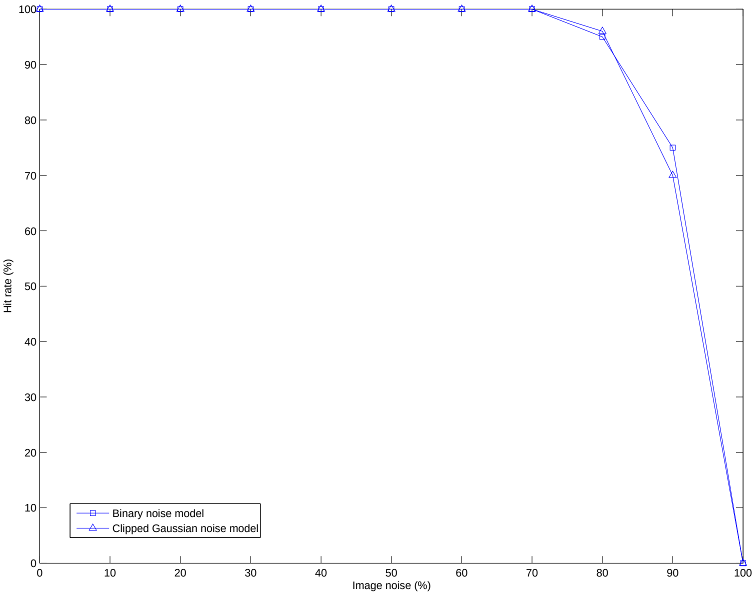


FIGURE 7.5: Comparison of the noise performance for the CDHT using the binary noise model and the Clipped Gaussian noise model.

These results, although on a reasonably basic data set, demonstrates that the CDHT has excellent tolerance to noise. It should be noted that to the human eye, the apparent motion of the human subject is not visible at noise levels of 60% and above in either noise model. The CDHT, however, still detects the correct parameters at noise levels of up to 70% with 100% accuracy. After this point, with both noise models, a drop in performance to an accuracy rate of around 96% at noise levels of 80% can be observed, followed by a fairly sharp drop-off to a 70% accuracy rate for the binary noise model, and 76% for the clipped Gaussian noise model at noise levels of 90%. Using both noise models, the CDHT failed to detect any sequences correctly at noise levels of 100%. This is understandable as no original edge data should have been present at this level.

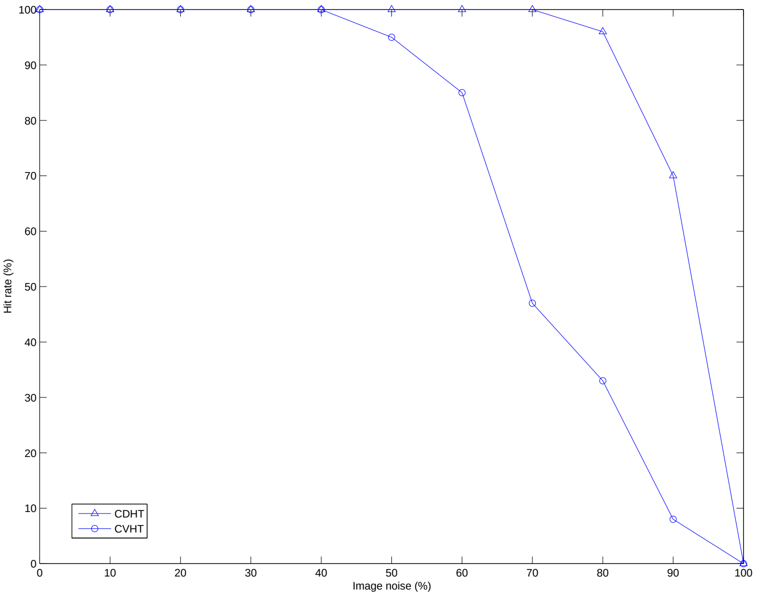
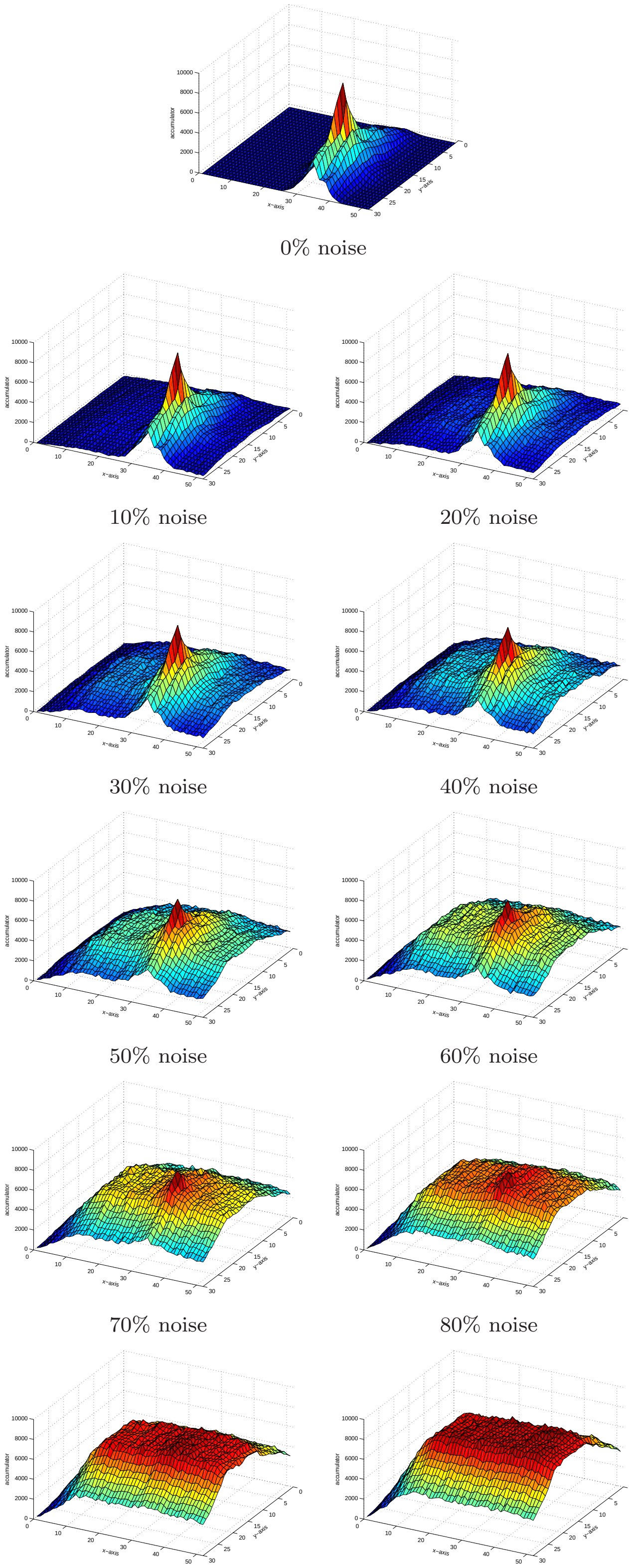


Figure 7.6: Comparison of the noise performance for the CDHT and the CVHT, both using the Clipped Gaussian noise model.

Figure 7.7 shows cross-sections of the accumulators, taken along the spatial x and y dimensions, for typical noisy image sequences from 10% - 100% respectively using the binary noise model. On examination of these accumulator cross-sections, it can be seen that there is a well-defined global maximum for each noise level up until the 80% noise level, from which point onwards the noise in the accumulator starts to mask the peak formed by the true parameters. Given that to the human eye, the image sequences containing above 60% noise are unrecognisable as that of a walking human, these are very encouraging results.

When comparing the CDHT with the CVHT it can be seen that the CDHT is much more accurate in extracting the correct parameters of the ground-truth shape sequence model in noise levels of between 40% and 90%, with the CDHT being able to tolerate a 30% increase in noise before any accuracy is lost. This can be accounted for by the fact that CDHT is designed to find deformable objects, where the shape of the object changes throughout the sequence, whereas the CVHT was designed to extract rigid moving objects. Given this, when applying the CDHT, each frame of data is less likely to be confused for another frame of data in the sequence at a different time, thus lessening the probability of confusion in the accumulator space. Combining this unique, varying shape description with the global nature of the CDHT to gather evidence over both space and time, ensures that corrupt data can be handled well. It should be noted that comparing the test results presented here with the test result presented by Grant



0% noise

10% noise

20% noise

30% noise

40% noise

50% noise

60% noise

70% noise

80% noise

90% noise

100% noise

Figure 7.7: Accumulators for varying noise levels.

in [16] is not exactly comparing like-for-like, as the data used in Grant's work consisted of a simple moving rigid object, and if presented with the same data, the CDHT would probably exhibit similar accuracy to the CVHT. A comparison between the CDHT and the CVHT is made here simply because the CVHT is the closest technique to the CDHT for modelling moving objects, and as such it is worth using as a baseline for the work presented here. The increase in accuracy at higher levels of noise owes as much to the uniqueness of the temporal deformation captured using the spatio-temporal kernel model as it does to the design of the CDHT itself. Nevertheless, the combination of these two powerful techniques has produced very impressive results.

7.3 Summary

This chapter has presented a set of 'best-case' noise performance tests for the CDHT. The results from these tests demonstrate that the CDHT has an excellent tolerance to noise. The ability to tolerate such extreme levels of noise is, no doubt, due to the global spatio-temporal evidence-gathering nature of the CDHT. If a frame of the input data is corrupted to the point where it is no longer recognisable, then the overall impact on the performance of the CDHT is not as great as it might be when using evidence-gathering techniques that do not take advantage of this temporal correlation.

The tests presented in this chapter have demonstrated the performance of the CDHT using pseudo-synthetic edge data, in that the edge data used is far 'cleaner' than the edge data that would be obtained from a real-world scene. With this in mind, the thesis will now proceed to examine the performance of the CDHT using real-world data, captured in a natural environment.

Chapter 8

Real-world performance testing of the CDHT

The results from Chapter 7 indicate that the CDHT works well when presented with a synthetic input, even in the presence of high levels of noise. When compared to the image sequence used in the previous chapter, however, edge-detected images sequences in the real world are rarely so well defined. Typically when looking at real-world data there is a particular subject, or subjects, of interest, such as walking humans in a busy scene. If a scene such as this is considered with regard to locating and extracting these subjects of interest, it soon becomes apparent that most of the image data can be considered to be 'noise' that simply obfuscates the problem. Further to this, certain aspects of image noise, such as Gaussian noise are usually always present in a recorded image, and even the best edge detection techniques will result in false edges being detected while true edges are not. Due to this, formal real-world performance testing was deemed necessary to prove the CDHT's worth outside of laboratory or synthesised conditions.

8.1 Test data

The University of Southampton's Large Gait Database contains image sequences filmed outdoors, in a natural environment. Fifty of these images sequences, chosen arbitrarily (the first 50 right-to-left walking sequence for each subject – Subjects 13-63[1]), were selected to be used as test data for real-world performance testing using the CDHT. Image sequences were taken for various subjects, but all from camera 'e' - "normal outside". From these image sequences one whole gait cycle and, wherever possible, 5 frames either side of this, were extracted. Due to computational time restraints, the image sequences were subsampled spatially by a factor of 2 along both axes and

[1] Due to incorrect labelling in the database used for testing, the image sequence for Subject 37 was omitted from testing, being replaced by the first image sequence for Subject 63.

scale parameters were also estimated and fixed for each sequence, therefore leaving the temporal start point and the two spatial start points as free parameters.

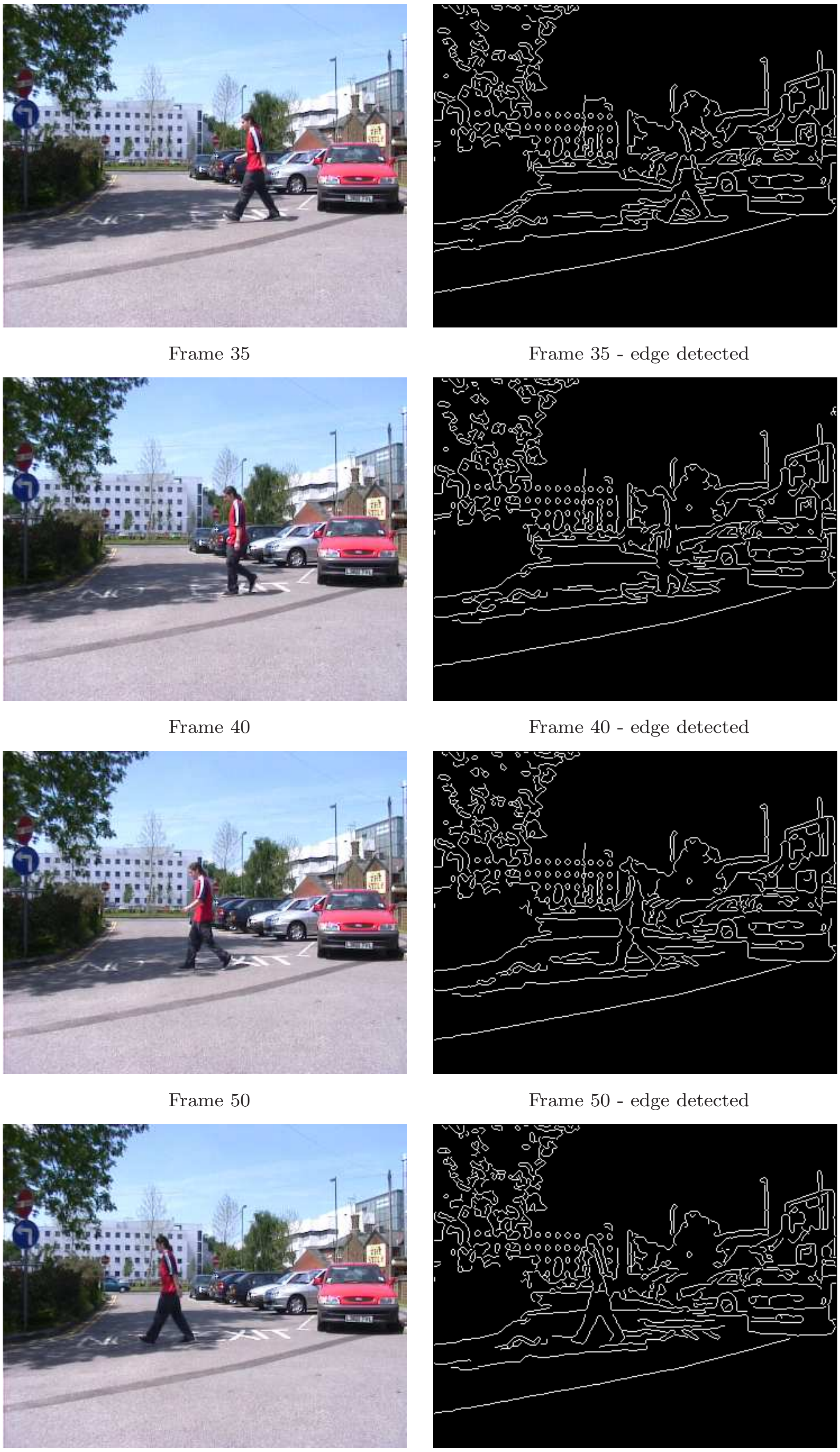
It should be noted that the data used in these tests differs significantly from the data first used to test the application of the CDHT to a real image sequence in Chapter 6. The data presented in Chapter 6 was collected under controlled laboratory conditions, and as such variables such as lighting and background noise were controlled tightly. Further to this, the data previously used for testing only ever contained one subject walking along a well defined path. In contrast, the data used in this chapter was collected outdoors, where the sun, clouds, and other transient objects cause variations in lighting, and occlusion and interruption to the subject's gait are often met due to people walking into or across the scene. Due to these factors, this data presents a far more significant challenge to any computer vision techniques, including this one, than the data used in Chapter 6.

It was decided that a two-tier testing structure would be applied to the real-world data: the first test to determine the CDHT's performance when presented with raw edge-detected image sequences; the second test to determine the CDHT's performance when presented with pre-processed image sequences. This decision was made to allow a full examination and comparison of the CDHT when applied to the real-world data – from a raw test of the CDHT, to a test more akin to how it might be applied in a real-life application environment (simple background subtraction is performed). Most applied computer vision systems pre-process the input data somewhat before presenting it to any image analysis algorithms, and these second set of tests aim to demonstrate the CDHT's performance under these conditions.

8.2 Raw real-world data testing

In order to determine a basic performance measure of the CDHT on the real-world data, the first tests were conducted using data that was processed as little as possible. The only pre-processing carried out prior to testing was a colour space conversion, to produce an intensity image, and edge-detection using the Canny operator (using parameters which were deemed optimal for the type of data in question. These parameters were consistent for each image sequence tested). Selected example frames from the image sequence for Subject 14 showing the original and edge-detected data prior to testing can be seen in Figure 8.1.

The edge data shown in shown in Figure 8.1 is far from perfect, and a substantial amount of noise exists in the background. In many frames, the edge detector only manages to extract partial boundaries for the walking human subject. This was often because the subject's clothing was of a similar intensity to much of the background and so didn't



Frame 35

Frame 35 - edge detected

Frame 40

Frame 40 - edge detected

Frame 50

Frame 50 - edge detected

Frame 60

Frame 60 - edge detected

FIGURE 8.1: Original and edge-detected real-world data for Subject 14.

produce a strong edge response. The binary edge data produced here was fed directly into the CDHT, which extracted a best-fit model for the free parameters of the generic gait kernel model for each image sequence.

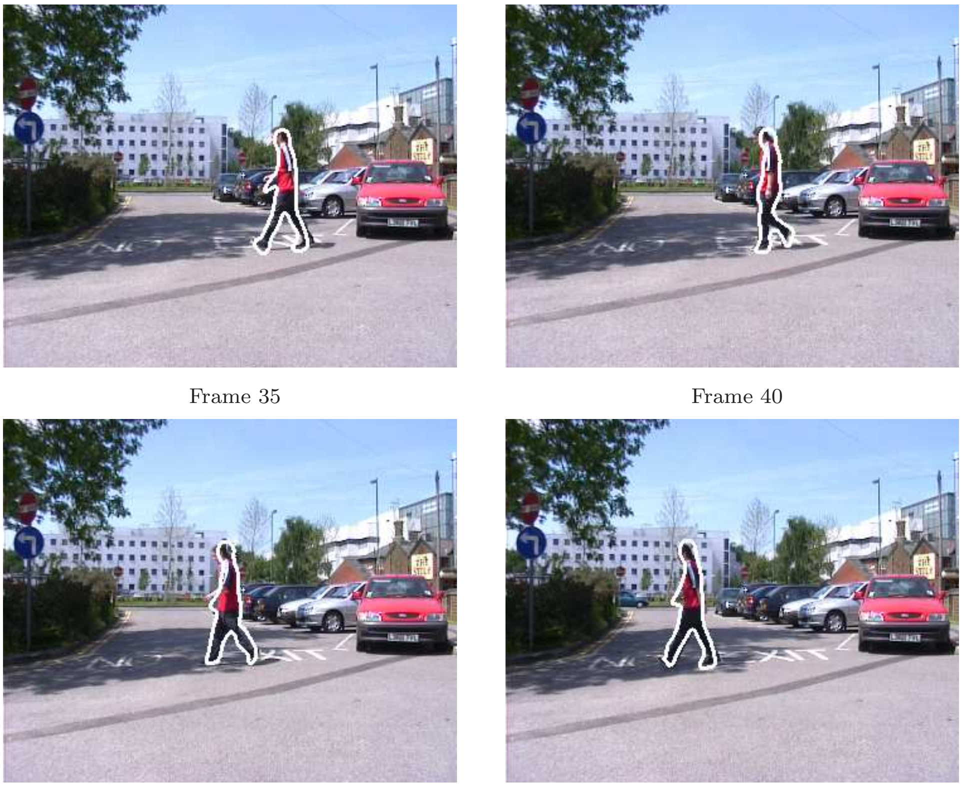
8.2.1 Analysis of results

When considering the success of the CDHT in the absence of any real ground-truth data, it was necessary to perceptually analyse whether the model seemed to 'fit' the subject in question. This approach relies on manually identifying cases when the model deviates from the true data significantly enough to warrant marking it as a 'miss'. This is obviously a subjective measure, but due to the subjective nature of vision and image interpretation as a subject on the whole, this method of verification was deemed suitable for this set of tests. Later, when we are testing the performance of the CDHT when subjected to image occlusion, some form of baseline result data will be necessary, as there will be a need for a benchmark performance score to compare the test results against. In these tests, however, this is not the case.

Due to the large quantities of image data involved in these tests, only a sample of the results will be presented in this chapter, but in order to verify the hit rates calculated by this method of analysis, the first and last frame from each gait sequence found as a result of these tests, with the shape model overlaid, can be found in Appendix B. The results from these tests actually demonstrate a mix of either exact or extremely close hits, or extreme misses, thus further strengthening the argument that the method of analysis used to verify the hit rates of these tests is sufficient for this particular set of results.

The hit rate for the whole of this particular data set was 82%, with 41 out of the 50 subjects being correctly located in both space and time. While these results may not seem as impressive as those for the synthetic data, even when the synthetic data was subjected to high levels of noise, this is an encouraging result. Given the overall amount of background edge data in this data set, along with the relatively poor edge responses at the boundary of the human subject in some images sequences, a slight drop in performance is quite understandable and, as will be shown, with very basic pre-processing a great increase in performance can be gained.

Another factor influencing these results is that the model data, which is the generic gait model detailed in Chapter 6, will not (unless by chance) exactly match the edge data supplied to the CDHT. Figure 8.2 shows the resulting kernel parameters, in the form of a reconstructed Shape Sequence Template, scaled and translated in space and time, overlaid on the same sample frames from Subject 14 as previously presented in Figure 8.1. From this, it can be seen that there is generally a very good correlation between the real-world gait sequence and the extracted model. However, as a generic gait model was used as the kernel for the CDHT, obvious deviations can be seen in the actual data. In particular, deviations can be seen around the feet in frames 35 and 60, where the posterior foot during the "toe off" points of the gait cycle extends past the boundaries of the model, and around the head, due to the subject walking with their head pointing down towards the ground. Deviations such as these can be observed in the results for many subjects, but as the fit to the model is otherwise very good, we consider these as hits.



Frame 35

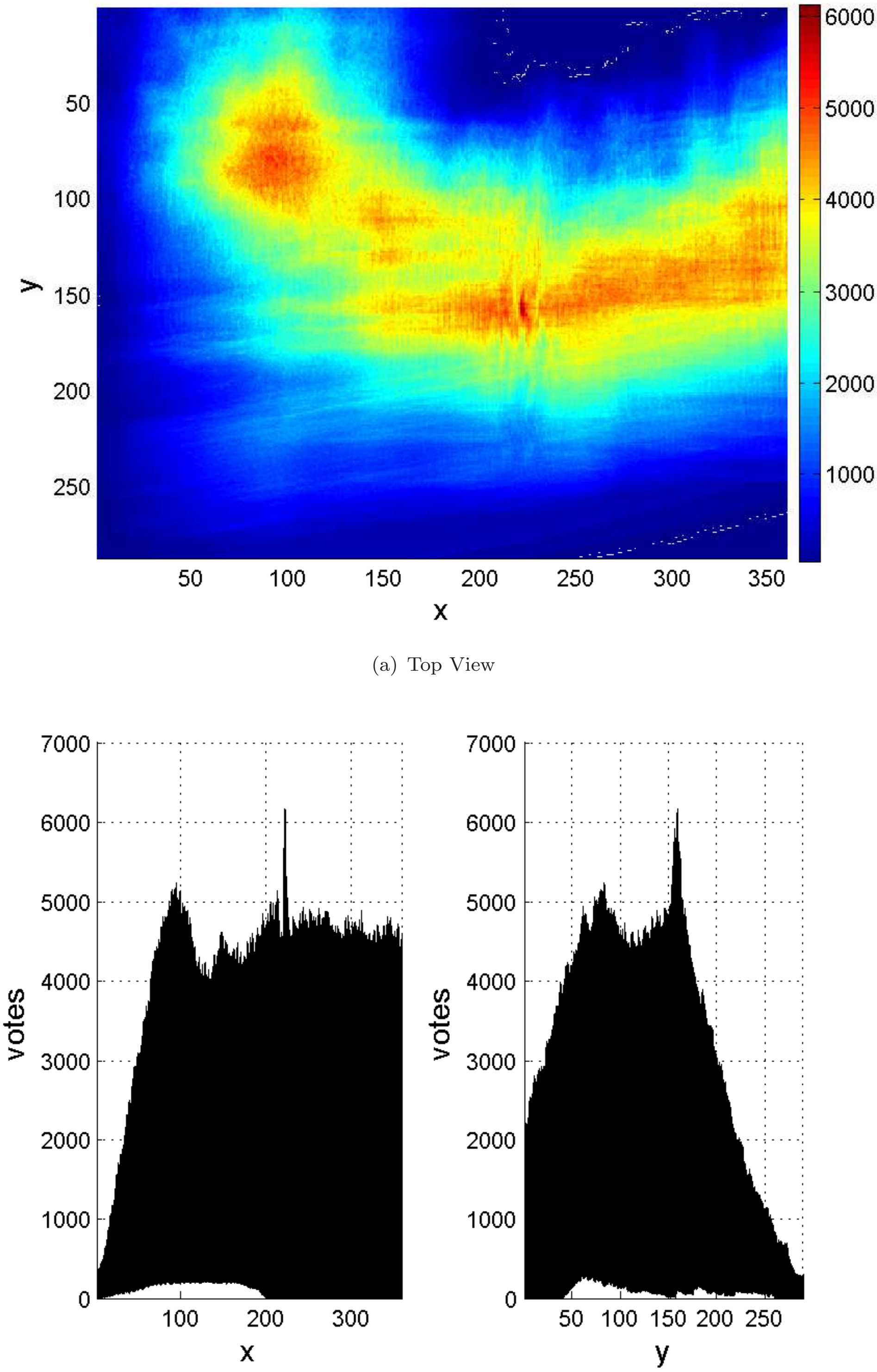
Frame 40

Frame 50

Frame 60

FIGURE 8.2: Resulting Shape Sequence Template overlaid on the original image data for Subject 14.

A cross-section of the resulting accumulator for the image sequence featured in Figure 8.2, taken at the parameters found by the CDHT, is shown in Figure 8.3. It can be seen from this that the peak of the accumulator is much less pronounced than in the synthetic data examples shown in Chapter 7, and there are a great many noisy votes cast for areas of the image sequence that contain background edge pixels. A large number of votes can be seen that correspond to areas in the original image sequence containing the tree towards the top left of the scene, and around the cars and buildings to the right, where there is a large amount of edge information present in the input data.



(a) Top View

(b) Views along the x and y axes

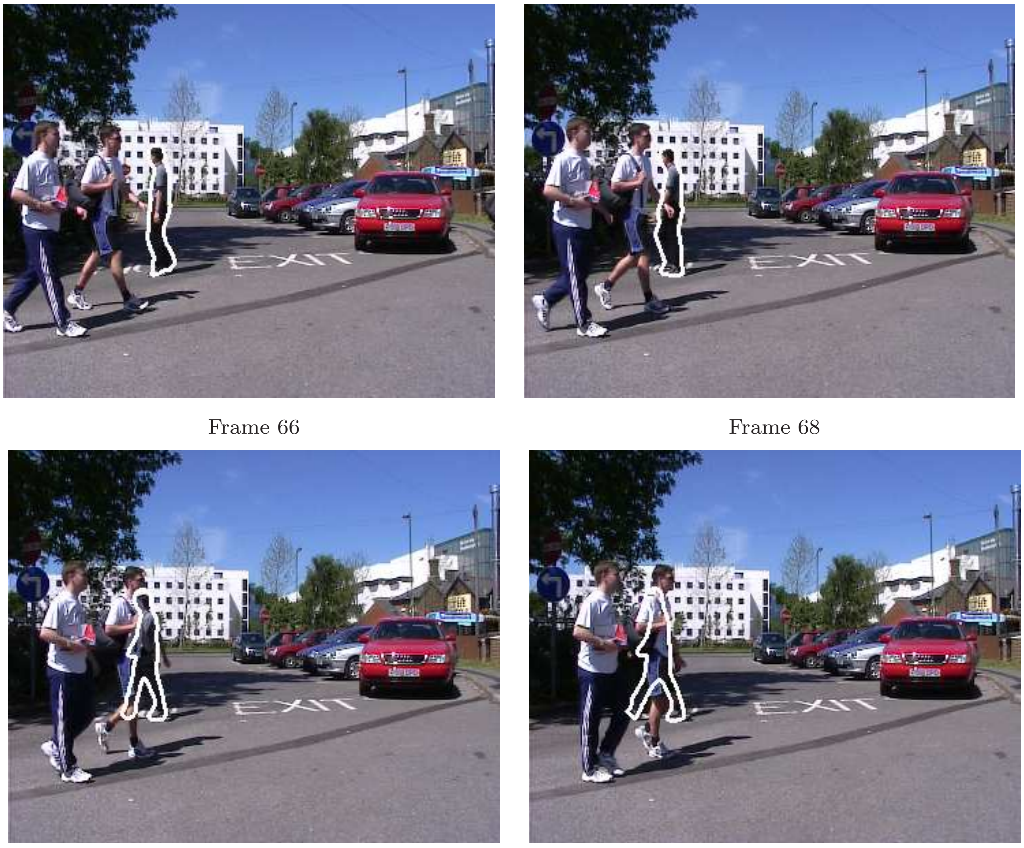
Figure 8.3: Accumulator for the image sequence 'Subject 14'.

Many image sequences in the test data set contained other human subjects walking into view, or cars driving past the road in the background. For some image sequences, the subject was obstructed from walking normally and their gait became erratic. For pathological examples of this behaviour, the CDHT fails to extract the correct parameters for the subject in question, but for other examples the CDHT still performs well. One particular test sequence which demonstrates the CDHT's resilience to temporal noise of this kind is that for Subject 25 (Figure 8.4), in which the subject can be seen walking behind a group of oncoming people, partially occluding the subject for part of the gait cycle. The CDHT still manages to correctly locate the correct parameters for this subject. Tolerance to complete frames of subject data being occluded in this way comes as no surprise, and the successful extraction of the correct parameters in this case is testament to the CDHT's spatio-temporal evidence-gathering nature – considering the image sequence as a whole, rather than just individual frames. An image sequence such as this would have caused purely spatial variants of the Hough Transform, such as the GHT, to break down completely. Frames of the result sequence for Subject 25, specifically showing the subject walking into this period of occlusion, are shown Figure 8.4[2].

Examination of the accumulator spaces and input data from these tests offers some insight into why the CDHT failed to extract the correct parameters for 9 of the images sequences tested. It has previously been mentioned that particularly noisy areas in the accumulator often relate to areas in the test scene containing trees and buildings, and, in failure cases, votes cast by these are a major contributing factor to the lack of success in extracting the correct model parameters. However, another major contributing factor is the quality of the edge data used as a source of input to the CDHT. In many of the failure cases there is often limited contrast between the intensity levels around the boundary of the subject and the background. Due to this, edge detection often fails to detect a number of true edges around the subject for many successive frames. An example of this can be seen in Figure 8.5, which shows an original frame of data and the edge-detected input to the CDHT for a frame in Sequence 31. Figure 8.5 demonstrates how poor contrast has affected the edge detection process, and the whole of the lower torso of the subject, which is in shadow, has been disregarded by the edge detector. Given that in instances such as these, the input data is only partially present, it is not hard to see why the CDHT failed to extract the correct model parameters, and this should be seen as a failure of the edge detector rather than the design of the CDHT itself.

Out of the 9 image sequences from which the CDHT failed to extract the correct model parameters, 7 demonstrated strong evidence to suggest that poor edge detection, particularly around the legs, was a major contributing factor in this failure (Sequences 22,

[2] It is worth noting that the tests presented here used a kernel model specifically designed to find subjects walking from right-to-left, and as such the people walking into the scene in the other direction, although causing significant amounts of edge data to be present in the input, would not have been detected by the CDHT during these tests.



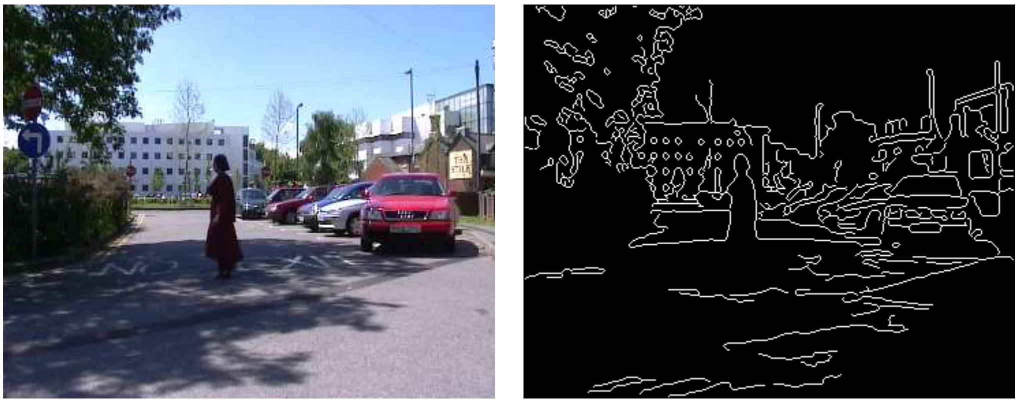
Frame 66

Frame 68

Frame 70

Frame 72

FIGURE 8.4: Partial results for Subject 25, showing the subject walking into partial occlusion.



Frame 59 – Original image

Frame 59 – Edge map

FIGURE 8.5: Original and Canny edge map for Subject 31, showing failure to detect edges of the lower torso due to poor contrast.

31, 32, 36, 41, 53, 59). Examination of the accumulator spaces for the remaining two incorrectly extracted sequences (42, 54), showed that the CDHT produced near misses (when considering the number of votes cast) to the actual true parameters, with a very high number of votes being cast into the accumulator at the indices that would have correctly located the human subject, but not quite a high enough number to create a peak sufficient to rise above that of the background noise. This can be seen by examining the cross-section of the accumulator space for Sequence 42, shown in Figure 8.6, in which a large peak can be seen at position ($x \approx 215,\ y \approx 160$) relating to the votes cast by subject of the image sequence. It can be seen that this peak, although very prominent, just fails to be masked by the noisy votes cast by the trees, centralised around position ($x \approx 100,\ y \approx 70$).

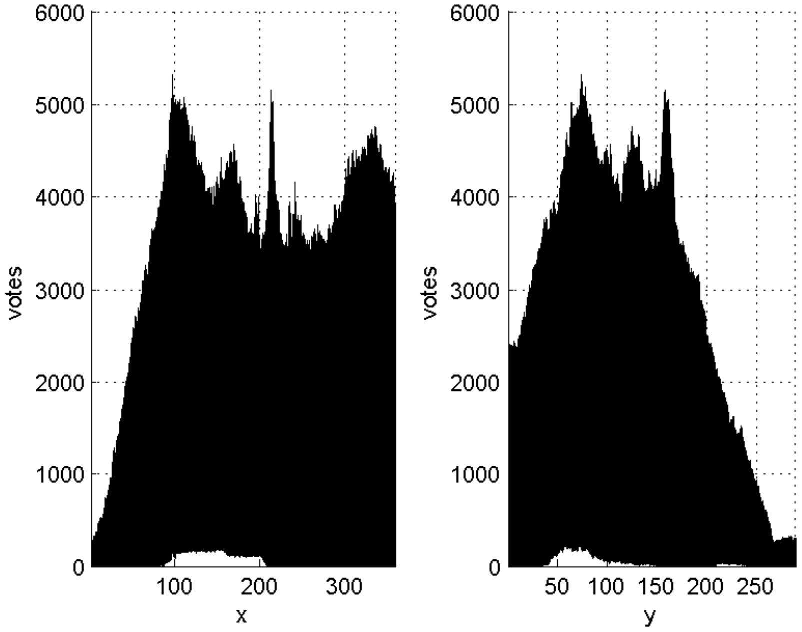


Figure 8.6: Cross-section of the accumulator space for Sequence 42, showing a vote-wise near miss.

8.2.2 Summary

The tests presented above show that the CDHT works well with what can be considered to be raw real-life data, and manages to deal with noise well if the majority of the feature points for the subject of an image sequence are present. On the whole, any lack of success when applying the CDHT to this raw real-life data can probably be attributed to a failure of the edge detector in detecting true edges, rather than a failure of the CDHT itself.

In practice, real-life image data is often pre-processed somewhat, often in an application-specific manner, and the CDHT's performance should benefit from this. To examine this further, and therefore give a true feel for the CDHT in a real-life computer vision environment, the next section will present tests performed using the same data set as above, but pre-processed using basic, generic image processing methods, which would be applicable to almost any application domain.

8.3 Pre-processed real-world data testing

When applying the CDHT in a realistic environment, one would usually pre-process the image data somewhat using heuristics related to the particular application domain, in order to remove unwanted edge data and noise. In the application domain of this thesis – finding walking humans, we are only concerned with moving objects and therefore a realistic and sensible first step in pre-processing the image data would seem to be the removal of any static features in the scene. In order to therefore test the performance of the CDHT in a more realistic environment, a simple background subtraction technique was applied to each image sequence prior to edge detection and being used as input into the CDHT.

8.3.1 Background subtraction and edge data generation

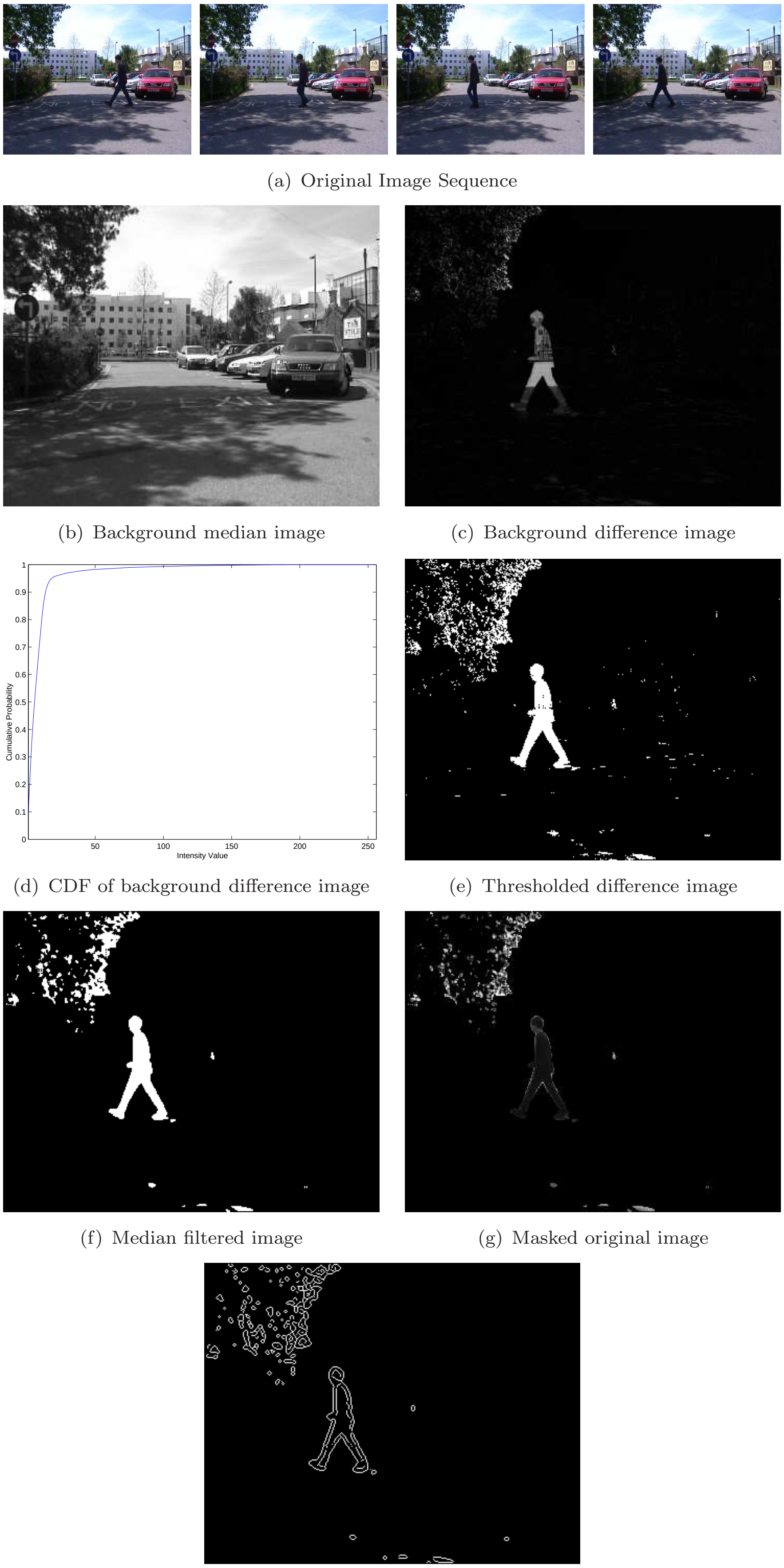
The simple background subtraction process used during these tests prior to edge-detection will now be detailed. The various stage of this process can be seen in Figure 8.7.

Before background subtraction, each image was transformed from a 3 channel RGB image to a single channel intensity image. After this, the background was calculated as the median, for each pixel using the set of all frames as

$$background(x, y) = \mathbf{median}(x, y, \mathbf{t}); \tag{8.1}$$

where x and y are the spatial pixel coordinates and $\mathbf{t}$ is a vector of time offsets for each frame ranging from $0..T-1$, with T being the length of the image sequence in frames. The median was used as a preferred method of obtaining an estimate of the average pixel intensity as it is less affected by statistical anomalies, such as those introduced when an object passes through a specific pixel region.

For each frame, a difference map was then formed by finding the absolute difference between the frame and the median background image. In theory, in creating this background map, differences between each frame and the median background should correspond to non-static (moving) pixels in the scene. In practice, however, if there is little difference in pixel intensity between the original image and the median background image then even non-static pixels may be removed.



(a) Original Image Sequence

(b) Background median image

(c) Background difference image

(d) CDF of background difference image

(e) Thresholded difference image

(f) Median filtered image

(g) Masked original image

(h) Canny edge-detected image

FIGURE 8.7: Background subtraction process.

In order to produce a binary image with which to mask the original frame data, a threshold level was calculated by forming a Cumulative Distribution Function (CDF) of the difference values from the difference image. Using this CDF, thresholding was then performed by setting the $0 - n^{th}$ percentile of grey levels to 0 and the $(n + 1)^{th} - 100th$ percentile of grey levels to 1. The images used in these test were thresholded up to the 85^{th} percentile, as this was judged to give optimal removal of background pixels whilst retaining true foreground pixels.

Due to natural image intensity variation, often caused by variable lighting, this thresholding process inevitably caused some true background pixels to be thresholded as foreground pixels (and unfortunately vice-versa). However, in the binary image produced from the last step, these background pixels tended to be very isolated and the majority of these could be removed by a simple median filtering operation. In the tests detailed here, this was performed using a 3x3 median filter.

To complete the background subtraction process, the median filtered binary image was used as a mask and applied to the original image data to extract the foreground pixels only. The resulting images were then edge-detected using the same method as the previous set of tests prior to being used as input directly into the CDHT.

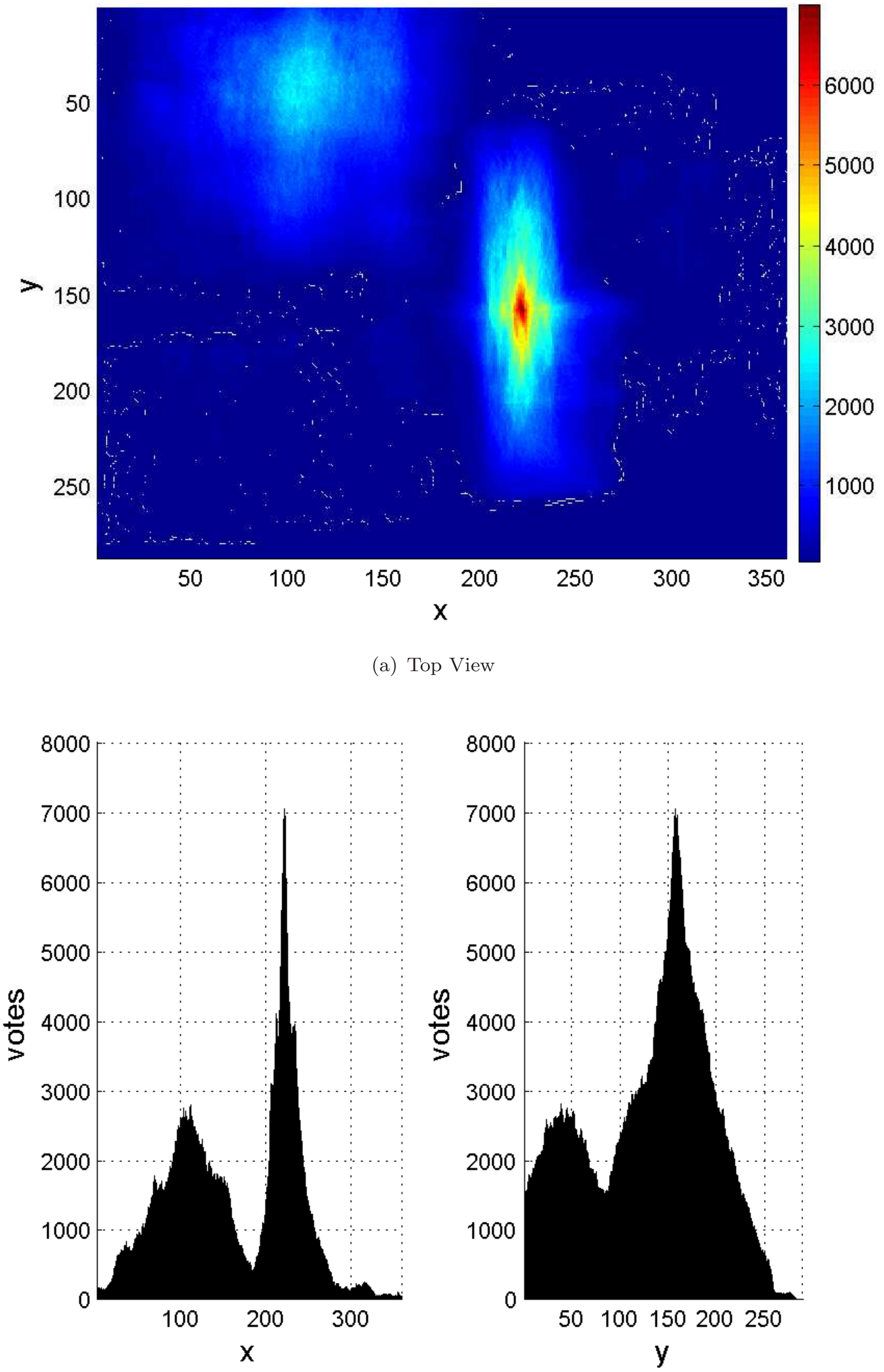
8.4 Analysis of results

This set of tests was conducted using the same 50 image sequences that was used during the previous real-world data tests, with the only difference being that pre-processing was carried out as described above. The hit rate for these tests was impressive, with the CDHT extracting 100% of the object's model parameters correctly. Given the hit rate of 82% for the previous 'raw' real-world data tests, this is a significant, but perhaps not surprising increase. Once again, due to the amount of image data involved in these tests, only a sample of the results will be presented in this section, with more detailed results being found in Appendix C.

In order to compare the results of these tests with the results from the previous real-world tests, which used raw edge-detected data only, we will again look at a cross-section of the accumulator for Subject 14, shown in Figure 8.8.

It can be seen from this cross-section that the peak of the accumulator is now much more prominent, and the area surrounding the maximal peak much cleaner, than that seen when raw edge-detected data was used. This is a result of much less background noise existing in the edge data prior to the application of the CDHT.

Although there is much less in the way of noise, a significant number of votes have still been cast that relate to the area occupied by the trees located towards the top-left of the original scene (which due to the effects of wind, demonstrated some temporal



(a) Top View

(b) Views along the x and y axes

FIGURE 8.8: Accumulator for the image sequence 'Subject 14' using pre-processed data.

variance). The noise created in the accumulator by this area, however, is much less using the pre-processed data, and the peak-to-noise ratio in this instance is much more pronounced, increasing the probability of the subject being correctly identified and not being confused by a local maximum in the background noise.

8.5 Summary

This chapter has presented a comprehensive set of performance tests using the CDHT to extract spatio-temporal model parameters from a challenging number of real-world data sequences. The results from these tests have shown that the CDHT is very successful when working with real-world data, even in the presence of significant amounts of noise, and in certain cases, partial occlusion, and a hit rate of 82% was obtained using raw edge-detected image sequences. Through the application of very basic pre-processing of the image sequences, in order to give a more accurate understanding of the CDHT's potential when used in a realistic computer vision system, this hit rate increased to 100%, thus proving that the CDHT is a viable tool for use in practical applications.

Chapter 9

Occlusion testing

Occlusion can cause problems to many computer vision algorithms. Metrics that one might normally associate with a shape, such as area, boundary, eccentricity, and colour distribution may not hold during periods of occlusion. As the Fourier descriptor model is a description of the shape as a whole, and not just one section of the boundary, regular Fourier descriptor models are usually vastly corrupted by occlusion. Indeed, any object detection or recognition technique that utilises the whole shape will be affected by occlusion. Given this, the need for robust computer vision techniques that can cope well under occlusion is great. It should be noted, however, that to hide part of an object from view is to hide important information relating to that object's description, and no vision system (including the human vision system) can be expected to perform well when significant proportions of an object of interest are occluded.

When compared to other shape detection techniques, evidence-gathering algorithms cope well with occlusion. This is largely due to the fact that, when determining whether an object defined by a specific set of parameters is present, each feature point is independent in providing 'evidence'. With Hough transform-based techniques, if a small portion of these edge points are obscured in some way, then the number of votes cast into the accumulator should only be reduced slightly if there are still many contributing edge points remaining. In the case of spatio-temporal evidence-gathering techniques, such as the Velocity Hough Transform or the Continuous Deformable Hough Transform, evidential feature points exist in a much greater quantity, with points existing through both space and time. Given this, one would expect a much greater tolerance to disruption or absence of features points. Even if whole frames of data are occluded, the other frames in the sequence should ensure that the object is still detected correctly.

The real-world data used in the previous chapter demonstrated that the CDHT performed very well at detecting periodic shape-sequences in complete image sequences. Often, however, image sequences in the real-world contain areas of occlusion, in which the view of the object(s) of interest are temporarily blocked by other foreground objects.

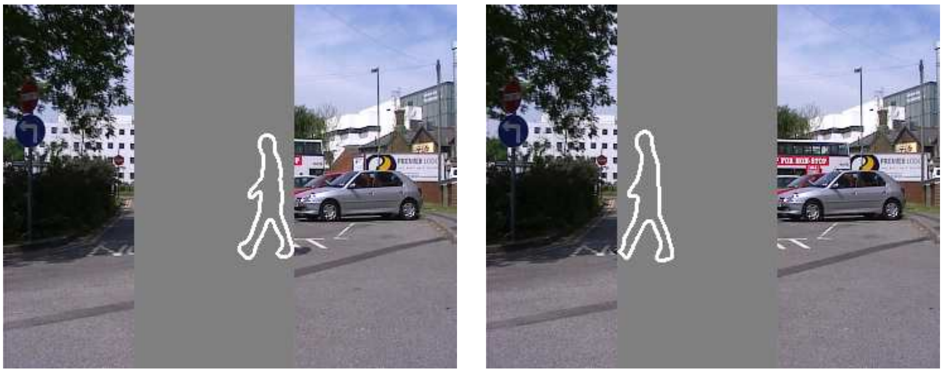
Indeed, an example of such a condition was described in the previous chapter, when two people crossed paths whilst walking. To claim that the CDHT is able to cope with large amounts of occlusion just based on this one example, however, would be foolish, and as such it was deemed prudent to examine the CDHT's ability to cope with occlusion using a set of well-defined tests.

9.1 Test data

To test for a degradation in performance when presented with occluded data, a baseline result set is needed. To test the CDHT's ability to handle occlusion, the results from the pre-processed real-world image set from the previous chapter was used as this result set. Since the CDHT achieved a hit rate of 100% when applied to this data set, it makes a good baseline result for testing any further changes to the data.

Occlusion was added as a further step to pre-processing using a static foreground object (variable width vertical bars), overlaid on top of the original image data in such a way that the object of interest was obscured. The amount of occlusion was varied between 0% and 100% of one period of the object's motion in 10% increments, in order to determine the level of occlusion at which the CDHT would fail to determine the correct model parameters for the shape sequence in question.

The process of creating the data for the occlusion tests was derived from the real-world test results from Chapter 8, Section 8.3 and the real-world test data itself, and minimum and maximum x-axis values from the reconstructed shape models found during real-world tests were used as the minimum and maximum values for the occlusion range of one gait cycle for a given gait sequence. These limits are demonstrated in Figure 9.1.



(a) Start of gait cycle

(b) End of gait cycle

Figure 9.1: Occlusion limits for one gait cycle.

To produce the actual test data, vertical blanking was added to the original image data using a mid grey-level intensity in order to simulate worst-case real-world occlusion.

As the tests performed here used the same pre-processing as those in Chapter 8, this vertical blanking section was effectively removed during the background subtraction process before edge-detection took place, leaving a zero-intensity period of occlusion in the resulting edge data. This is a desired effect, however, as any true static occluding object would be removed in much the same way. Examples of the occlusion levels used can be seen in Figure 9.2.

The tests were then repeated in the same manner as previous real-world data tests, using the CDHT to gather evidence and locate the best fit for the kernel model parameters.

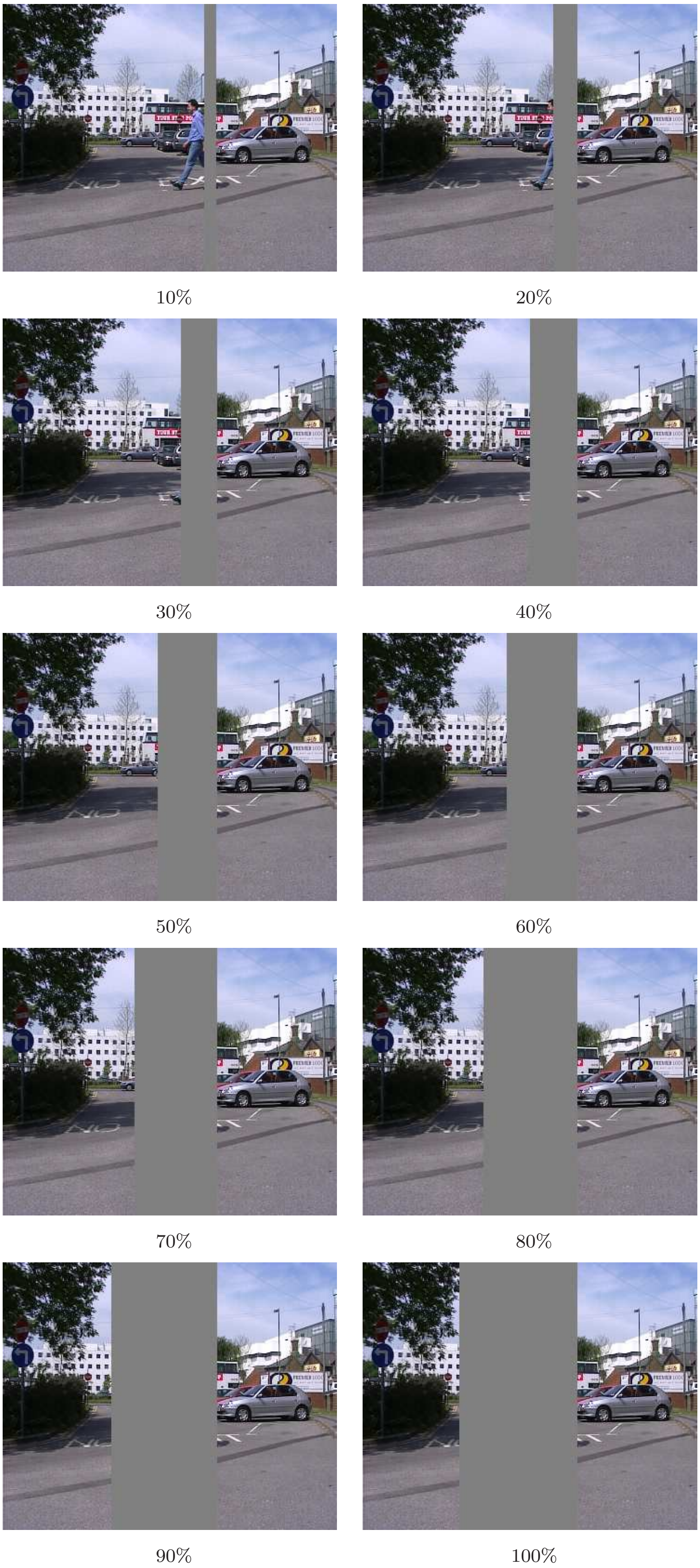
9.2 Analysis of results

Before analysing the results of these tests, it is prudent to consider factors that could affect the performance of the CDHT and cause spatio-temporal "near misses". Due to the nature of the Hough transform, as more noise is added to an image, or image sequence in this case, "peak spreading" can occur in the accumulator. This effect occurs as the edge pixels become corrupt and false edges occur around the true edge data. The true peak in the accumulator becomes clouded with noisy votes and more "near miss" votes are cast, effectively spreading and smoothing the true peak. Even votes cast by the shape itself may cause a slight spreading of the peak of the accumulator, as inconsistencies between the input data and the kernel of the Hough transform can introduce slight averaging errors. Further to this, the effects of discretisation can cause rounding to occur in the accumulator, with votes being cast either side of the true location of the peak.

Depending on the application domain of the CDHT, a near miss may actually be interpreted as a hit. Given a 360 x 288 image, such as those presented in all real-world tests in this thesis for example, locating the center of mass of a human subject spatially to within 2 or 3 pixels may be considered acceptable. For other applications this might be unacceptable.

With the factors described above in mind, it would be foolish to simply compare the results from these occlusion tests with the baseline data and look only for exact hits. As a result of this, no one measure of hit rate will be given for the results in this chapter. Alternatively, we will present the effective hit rate should a threshold determining what is considered a hit be applied to the results.

The measure for this threshold will be given by a simple Euclidean distance metric, which we will term the Maximum Hit Distance, or MHD, and takes three variables into account: the spatial x and y start point locations, and the temporal start point t, and is thus given by



10%

20%

30%

40%

50%

60%

70%

80%

90%

100%

FIGURE 9.2: Occlusion data.

$$MHD_{seq} = \sqrt{\Delta x_{seq}^2 + \Delta y_{seq}^2 + \Delta t_{seq}^2} \qquad (9.1)$$

where Δx_{seq}, Δy_{seq}, and Δt_{seq} are the respective distances of the differences between the extracted spatial and temporal start point parameters and the baseline data for the sequence seq, obtained from the real-world tests in Chapter 8.

A problem with calculating a distance metric threshold such as this is that distances in time and space are not easy to compare side-by-side. An error of 2 pixels spatially would probably be considered by most to be acceptable, whereas an error of 2 frames temporally would probably not. As such, the MHD should only be used as a rough guide to the proximity of the extracted parameters to the baseline parameters. With this said, the actual results from these tests showed that the margin of error when locating the spatio-temporal start point of the CDHT's kernel model was either within 2 pixels in either spatial direction or 1 frame temporally (a maximum theoretical MHD value of 3) of the baseline start point, or an extreme distance away (similar to the results seen in previous real-world tests), and as such all results could either be considered as hits or misses with some degree of certainty.

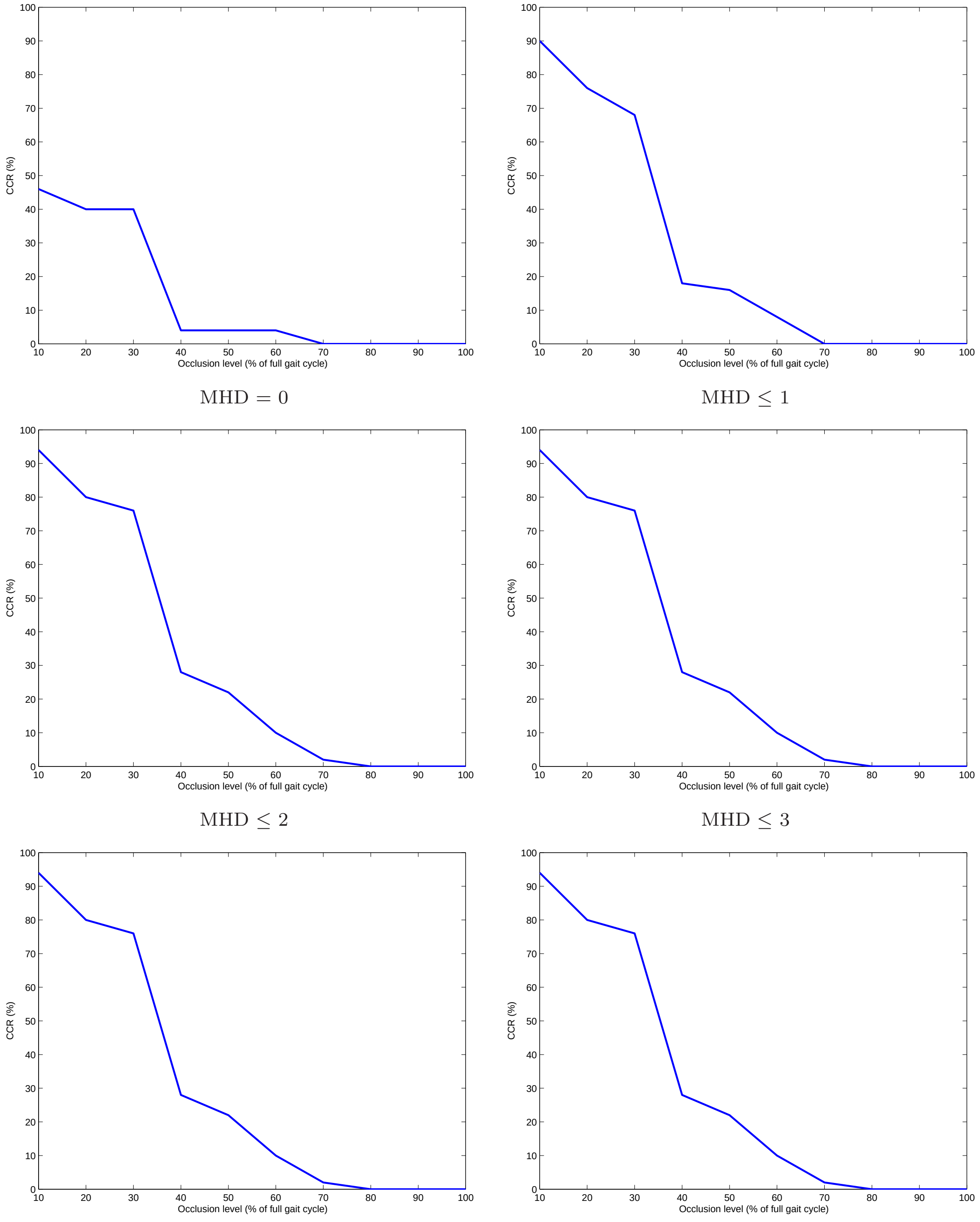
The results of these tests are shown in Figure 9.3. This figure details the hit rate for 50 sequences, using varying MHD threshold values from 0 to 5.

Figure 9.3 is worth taking some time to interpret. It can be seen that strong occlusion can cause a marked drop in performance, although this is not without obvious cause. Using a zero-tolerance Maximum Hit Distance (MHD) threshold of 0 yields a hit rate of only 46% for an occlusion level of 10%, with performance dropping off dramatically after an occlusion level of 30%, until at 70% no sequences are recognised as a hit.

It is interesting to note, however, that easing this threshold slightly, to a MHD threshold of 1, shows a large increase in hit rate (almost double) for most occlusion levels. At a MHD threshold of 1, hit rates of 90%, 74% and 68% are observed at occlusion levels of 10%, 20% and 30% respectively, with hit rates then dropping off again dramatically at an occlusion level of 40% to only 18%. Again, at this MHD threshold, the hit rate falls to 0% at a 70% occlusion level.

Increasing the MHD threshold to a value of 2 yields only a slight increases in hit rate at an occlusion level of 10% to 94%, but a noticeable increase at occlusion levels of 20% and 30% to 80% and 76% respectively. Again, past this point performance drops off dramatically, failing at 80%.

At MHD thresholds of between 3 and 5 – the upper limit of this test (as anything deviating this far from ground truth would definitely be considered a miss), no increase in hit rate is observed at any occlusion level. This can be explained if one examines Figure 9.4, which shows the distribution of the Euclidean distances between the baseline



MHD = 0

MHD ≤ 1

MHD ≤ 2

MHD ≤ 3

MHD ≤ 4

MHD ≤ 5

FIGURE 9.3: Results for the occlusion tests using varying MHD thresholds.

data parameters and the extracted parameters for each occlusion level as a box-and-whisker plot. It can be seen from this that occlusion levels of up to 30% produce results which all fall within a very small distance from the baseline results (a MHD of 0 - 2). Occlusion levels of greater than this tend to produce results which fall much greater distances away from the basline results. Indeed, if we examine the box-and-whisker plot for the occlusion level of 40% we see that the median distance from the baseline result is around 17. From this, one can clearly see why increasing the MHD threshold (thus increasing the distance from baseline results that can be classed as a 'hit') would have little effect on perceived performance between values of 2 and 5. The high hit rates at occlusion levels of between 10% and 30% can be explained due to the fact that, at these occlusion levels, the majority of the maxima in the accumulator space fall within a distance of between 0 and 2 of the baseline location – a very close, if not exact, hit. At an occlusion level of 40% and above, a much greater tolerance of what one considers a 'hit' is needed to see an increase in perceived performance.

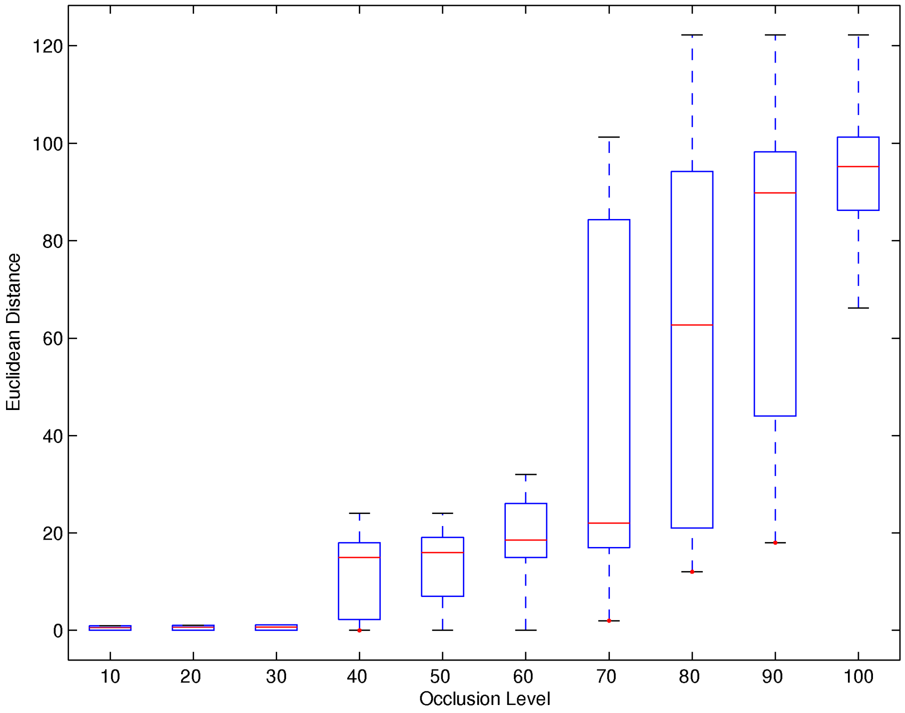


Figure 9.4: Box and whisker plots of the Euclidean distances at varying occlusion levels

To put this analysis in perspective, if we consider the percentage of the sequences, at varying levels of occlusion, that fall within 1 pixel, or 1 frame of the baseline parameters, we obtain the hit rates summarised in Table 9.1. At these resolutions, this threshold is a reasonable threshold to use, where extraction of the correct parameters with a 1 pixel or 1 frame tolerance level (equivalent to $\frac{1}{25}^{th}$ of a second in time at the frame rates used here) could easily be considered a correct hit. These results are drawn from the result

Table 9.1: Hit rate at varying occlusion levels for results falling within 1 pixel or frame of the baseline data

Occlusion level (%)	Hit rate (%)
10	94
20	80
30	76
40	28
50	22
60	10
70	2
80	0
90	0
100	0

set using a MHD of ≤ 2 (a combination of errors of 1 unit in all 3 dimensions gives a maximum distance of 1.73, which is below the MDA ≤ 2 threshold).

9.3 Discussion and summary

It would be wise at this point to compare our results with those from other spatio-temporal evidence gathering algorithms. The algorithm most similar to the CDHT is the CVHT. Using the CVHT, successful extractions were obtained at up to 70% occlusion levels [16]. However, this is perhaps not a fair comparison to make as the imagery used in these test was filmed under laboratory conditions, with fixed and constant lighting conditions. Due to this, the edge data used as an input into the CVHT for these tests was much cleaner, and importantly the subject was extracted very well when compared to the edge data used as an input for the CDHT. The real-world data set used for the CDHT was also very much more corrupted by noise due to non-static objects, with swaying trees, passing cars, and in extreme cases, multiple people crossing the field of view between the camera and subject. While this does not in any way discredit Grant's results for the CVHT, the result set for the occlusion tests using the CDHT do perhaps give a more accurate view of applying spatio-temporal evidence-gathering techniques in the outside world, where other forms of occlusion and noise are ever-present.

Grant also reported in [16] that testing on the CVHT concentrated specifically on the upper-torso of the subjects, as the the legs were more difficult to extract successfully using edge detection. For the tests presented here, the CDHT uses the whole boundary of the human subject, so includes edge data from these more problematic areas. One may naively think that using more of the edge data for the subject is a good thing, as more 'evidence' is gathered for the subject in question, but this is only holds true if the edge data is clean and intact. Using noisy sections of edge data may result in an overall drop of performance as near misses of votes cast in accumulator will simply cause peak-spreading in the accumulator and mask the true peak in more noise. While it is feasible for the CDHT to use a partial boundary model, such as that used in the tests on the CVHT, the CDHT's originality lies in its ability to detect objects that deform in some periodic way.

The additional noise that occurs in the real-world data set used for testing the CDHT, along with poor detection of the edges of the subject at times, is without doubt responsible for a marked drop in performance when compared with the results for the CVHT. The fact that the CDHT still correctly extracts the parameters of the subject 76% of the time under these conditions when the subject's gait cycle is occluded by up to 30%, is a very encouraging result, and it should be noted that appropriate heuristics, which could be applied to any variant of the Hough transform, would undoubtedly make the algorithm even more robust.

Chapter 10

Future work and conclusions

10.1 Future work

10.1.1 Application-specific work

Examining a signal in the frequency domain often allows defining characteristics of the signal to become apparent. In much the same way, the ability to analyse spatio-temporal gait patterns in the frequency domain should allow defining characteristics of a subject's gait to become apparent. In this thesis, this has already been demonstrated by way of applying the new spatio-temporal Fourier descriptor model to human gait recognition, where the defining characteristics of a specific human subject are extracted from more 'generic' human gait characteristics. These characteristic descriptors allowed a simple classifier (k-nearest neighbour) to discriminate between a large number of subjects with a high level of accuracy.

Although this thesis has concentrated on the detection and recognition of human subjects, the spatio-temporal Fourier descriptor model developed here is very generic and is applicable to any forms of periodically deforming objects. There are many fields, such as those allied to medicine and biology, in which periodic motion is present and the ability to model it concisely, and to a known level of specificity may be advantageous. In medicine, for example, various aspects of the cardio-vascular system exhibit periodic deformation as they respond to the beating of the heart, which in itself demonstrates strong regular periodic deformation.

Clinical gait analysis

The clinical analysis of gait (an area distinct from gait recognition), is another field in which the use of a spatio-temporal model of gait with variable generality, such as that presented in this thesis, could probably be put to good use. A large proportion of clinical gait analysis is carried out either by eye, or by analysis of marker-based imaging of subjects, with easily identifiable markers being placed over the subject's body at specific known points. The spatio-temporal movement and relationship of these markers are then used to form a model of the subject's gait, with this model then being examined by human analysts to search for abnormalities in gait.

Gait is often analysed for the two quite differing purposes of determining abnormalities and analysing sports performance. Pathological gait may reflect compensations for underlying problems and diseases, and certain abnormalities in a subject's gait cycle are often a good cue to these problems. A distinct example of pathological gait is demonstrated in the walk of somebody with Parkinson's disease. Often, this includes bradykinesic shuffling – a term used to describe the associated slow shuffling motion of a person with Parkinson's disease, which also often presents with a lack of arm swing, and a drooping of the head and shoulders [30]. Many such pathological gait disorders exist, relating to a number of medical conditions, and a list of these along with further discussion on pathological gait is given in [40].

In direct contrast to pathological gait analysis, sport scientists often use gait to examine athletes' performance. For example, being able to analyse the gait of a sprinter may highlight areas of weakness, which may then enable an appropriately trained physician to advise the sprinter as to how he or she could enhance their performance.

Areas such as these are obvious candidates for the application of an automatic gait analysis system based around techniques presented in this thesis.

Subject segmentation

Background subtraction techniques have arguably been a weak point in computer vision due to a number of technical reasons. This area is often a problematic one, as it is hard to define what the term 'background' means in all contexts. Often this is taken to be the static, or pseudo-static objects in a scene, such as grass, concrete structures, and buildings. When observing a scene, however, concentrating purely on this concept of a 'background' model is not always what may be required. Within a given scene, vision systems usually focus their attention on certain objects of interest, and prior knowledge of these objects could, in many cases, guide their segmentation. Often, especially in biometric and surveillance systems, these objects of interest are humans, who move in a periodic manner. The Continuous Deformable Hough Transform has proved to be a very robust method of detecting periodically, or near-periodically deforming objects, and given this, the extracted shape and deformation parameters would provide a very good basis for initialising an intelligent object segmentation system. Better segmentation of human subjects would inevitably improve automatic gait recognition techniques and would also allow more complicated behavioural analysis, as the details of a subject's movements could be more accurately extracted.

10.1.2 Technique-specific work and performance enhancements

The algorithms and techniques demonstrated throughout this thesis are shown and have been applied in a very 'raw' form. No heuristics have been used to help guide searching of the feature space nor has support been provided by prior knowledge of the application domain. Additionally, very basic classification and feature extraction techniques have purposely been used. This is with good reason. The concern of experimentation throughout this thesis has been in testing the spatio-temporal Fourier descriptor model and CDHT's baseline performance. While arguably better classification and feature extraction techniques could have been used, the aim of the experimentation was to prove and emphasise the model's descriptive uniqueness and the CDHT's robustness and generality, and not to demonstrate any form of optimised recognition or extraction. Nevertheless, certain parts of the algorithmic design could benefit greatly from various optimisation techniques or performance enhancements, and these are suggested for further work below.

Parameter space reduction via correlation

Studies have shown that normal human gait exists within a well defined set of limits, and that these parameters are often strongly correlated, especially for subjects within a similar age range [40]. Many of these correlations could be directly applied when using the CDHT to extract human gait parameters. Anatomical correlations would form a gamut of possible gait models in the CDHT's feature space, which in turn could be used to guide and speed up searching of the parameter space, due to the need to consider only a subset of all possible gait parameters.

If the parameters extracted for the correctly identified real-world image sequences are examined, it is easy to see that some are highly correlated, even if some variation exists. Figure 10.1 shows the correlation between the maximum bounding box width and height parameters (the spatial scaling parameters extracted via the CDHT). If we examine this figure a correlation can be seen between these two parameters. If we further analyse this data set, we can obtain a correlation coefficient for these two variables, which for these specific variables and data set is 0.99, demonstrating that the two variables are indeed very highly correlated. This is hardly surprising as subject height is well-known to be highly correlated to leg length, which in turn is highly correlated to the maximum width of the subject's stance when walking.

Figure 10.2 shows a similar correlation, for the same data set, between subject height and the distance that the center of mass of the subject has travelled. This figure also shows

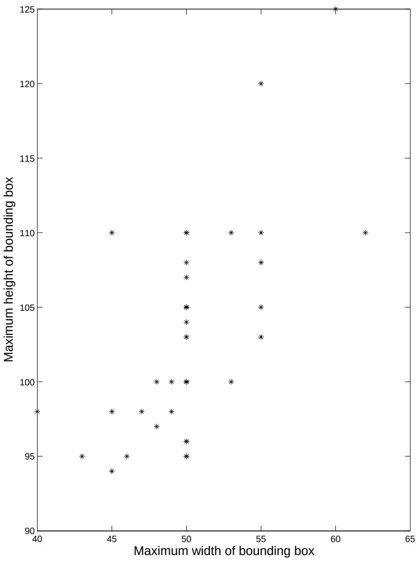


Figure 10.1: Correlation of maximum bounding box width and height parameters.

the 95% limits of both male and female standard (for British subjects) anthropometric measures for human subject height and stride length [36], where stride length roughly equates to the distance that the center of mass has travelled for the parameters extracted via the CDHT. Given that the test data consisted of both male and female subjects, it is not surprising that the majority of samples fall within the upper and lower bounds of both of these limits. The two samples actually falling outside these limits belong to Subjects 32 (107, 60) and 52 (108, 72). Upon examination of the data for these subjects, it can be seen that while Subject 54 is merely a slow walking individual, Subject 32 is actually prohibited from walking at normal speed due to their gait cycle being interrupted by passers-by (and is therefore a true anomaly). The correlation coefficient for these two variables is also 0.99, again suggesting a very high degree of correlation.

Given this high degree of correlation between certain parameters, the parameter space of the CDHT could be narrowed down significantly when searching for normal gait, using only the gamut of parameter values that covers the range for human gait. An alternative to this, that could be used to test for pathological gait as described above, would be to search for only parameters outside of this gamut (although this would obviously have to be bounded at some point).

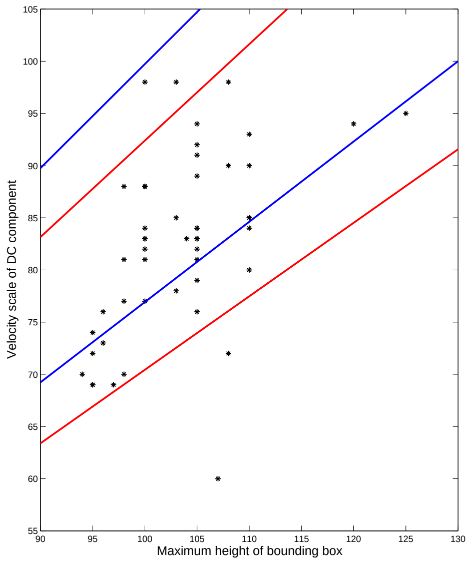


FIGURE 10.2: Correlation of maximum bounding box height and velocity scaling parameters, showing the lower and upper bounds (95% limits) of male (blue) and female (red) subject height and stride length.

Parallel processing and general heuristics

Although the CDHT's feature space is of a fixed dimensionality and does not change with object complexity (contrary to the case with the standard Hough transform), it still requires a considerable amount of processing time. This processing time is generally directly proportional to the number of potential feature points in a given image sequence (edge pixels). The processing required per feature point, is however, totally independent of the processing of other feature points and the algorithm would therefore lend itself well to a parallel or distributed method of computation. Theoretically, this could reduce computation time by a factor equal to the number of parallel nodes available to perform the processing. A good discussion on the implementation of parallel Hough transforms can be found in [26].

Pragmatic approaches to reducing computational requirements of the Hough transform could also be applied to the CDHT. For example, 'pyramidal' or multi-stage methods, such as the Hierarchical Hough Transform [6, 22] or the Adaptive Hough Transform [21] allow progressive refinement of the accumulator space to speed up computation time, whilst still retaining a high degree of accuracy.

As we have already seen in Chapter 8, Section 8.3 (Pre-processed real-world data testing), analytically removing feature points that will definitely not contribute to the meaningful interpretation of a scene can be one of the most effective methods of improving accuracy. In the case of the CDHT, where the subject of the search is usually moving, removing static features can greatly improve performance. In this thesis this was demonstrated through a very basic and crude form of background subtraction, but the removal of 'background' features often inadvertently leads to the accidental removal of genuine feature points belonging to the subject itself. More sophisticated methods of determining motion and removing non-salient features, such as incorporating the use of optical flow, could lead to a dramatic increase in performance and speed. Removing such features has both the combined benefit of decreasing processing, as there are less feature points to process, and also causing fewer false votes to be cast in the accumulator space, thus reducing the probability of any true maximal peaks being swamped by noise.

10.2 Summary and conclusions

The aims of this thesis have been two-fold. In the first instance, the aim was to test the feasibility of using spatio-temporal Fourier descriptors to model periodically deforming and moving shapes, and to test their discriminatory power. In this case, their discriminatory power was tested by applying the descriptors to the field of automatic gait recognition. A basic feature extraction technique, based around the Bhattacharya distance metric to measure inter-class separability, was used to find the spatio-temporal Fourier descriptors for human gait that would prove most useful in classification. Given that the aim was to test the basic discriminability of the descriptors alone, a simple classifier – a K-Nearest Neighbour classifier – was used for classification. The results were very promising, with Correct Classification Rates of 86.2% being achieved on a database of 115 subjects, totalling 1062 gait sequences. The effect of varying camera-to-subject distance on this recognition rate (simulated by sub-sampling in the spatial domain) was also tested, with only relatively small drops in performance being observed, especially when considering the drop in spatial resolution.

The second aim of this thesis was to combine the descriptive power of spatio-temporal Fourier descriptors with the robustness of the Hough transform, resulting in an new advanced algorithm for the detection and tracking of deformable moving objects – the Continuous Deformable Hough Transform (CDHT). Throughout the thesis, the CDHT has demonstrated excellent potential at detecting and extracting deformable objects as they move through an image sequence. When tested with simulated data, even at noise levels upwards of 70%, the CDHT demonstrated remarkable tolerance to noise, only failing totally with a noise level of 100% noise, which is to be expected as the image sequence at this point consists purely of noise.

On real-world data sets, the CDHT demonstrated its robustness under real-life conditions. Using data obtained outdoors, in which changing lighting conditions and shadows cause large amounts of image noise and variability, the CDHT still performed well, with hit rates of 82% being observed on an arbitrary sample of 50 test sequences using data that had merely been edge detected, and not ‘cleaned’ in any other way. Removing static features from these image sequences using very basic image processing techniques resulted in a hit rate of 100% on the same data set, thus proving the CDHT’s worth in a practical modern-day computer vision system. When worst-case synthetic occlusion was added to these image sequences, the CDHT demonstrated its ability to deal with partial occlusion, with 76% of the model parameters being extracted correctly (i.e. falling within a Euclidean distance of 1 pixel or frame from the true start point) at an occlusion level of 30%.

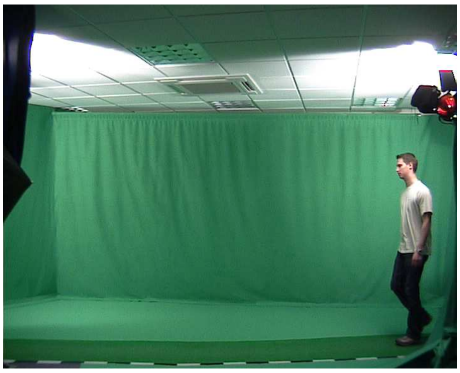
The new spatio-temporal Fourier descriptor model presented in this thesis provides a compact and complete way of representing moving periodic shape deformation. Due to these descriptors being Fourier-based, both macroscopic and microscopic details of the shape’s boundary and deformation can be analysed with ease. Further to this, spectral decomposition of both the shape and deformation ensures that the resulting descriptors form a good basis for discriminating between a number of deformable objects.

The CDHT effectively provides the user with a model of the most likely model parameters (location and scale, in both spatial axes, and the temporal axis) of a given deforming object within a dynamically changing scene. The robustness and effectiveness of the CDHT owes much to the spatio-temporal Fourier descriptor model described in Chapter 3, and to the fact that the model can be used to describe the shape and spatio-temporal deformation of an object in both a general form and in detail separately. The technique also benefits greatly from the CDHT’s spatio-temporal evidence-gathering nature, considering a period of shape deformation as a whole, rather than just individual frames. This combination of a powerful shape description model and a robust shape extraction mechanism ensures that the CDHT is an effective tool for use in the field of computer vision, and one which can offer many advantages as the field moves more and more into the realms of video processing.

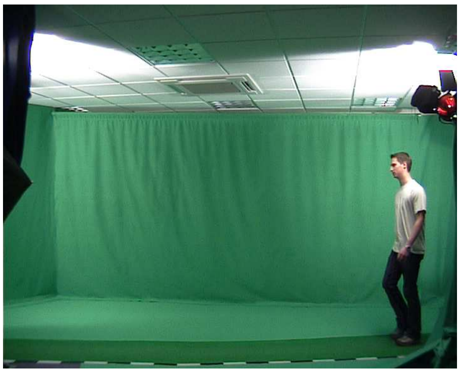
Appendix A

Results from the image sequence used in Chapter 6

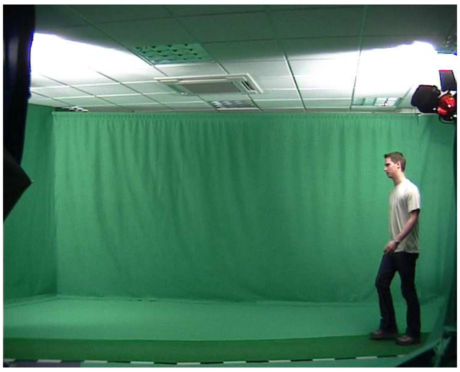
The following images display the complete results from the application of the CDHT to a real-world image sequence, as described in Chapter 6. These images are PAL DV frames (720 x 576, 25 fps), with the Shape-Sequence Template spatio-temporally rescaled and translated to the parameters extracted by the CDHT, overlaid on the original image sequence.



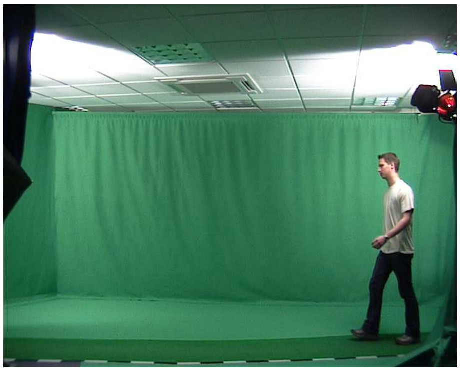
Frame 1



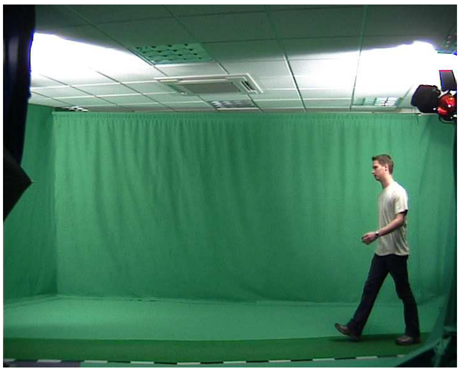
Frame 2



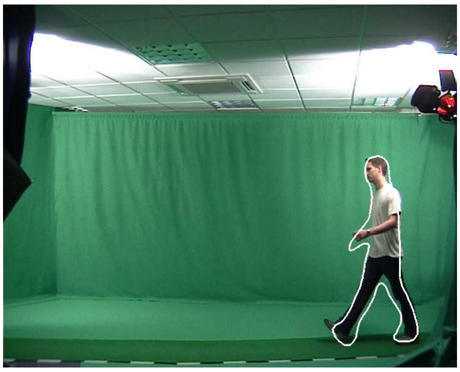
Frame 3



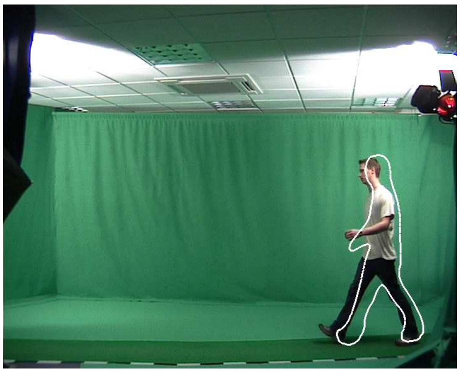
Frame 4



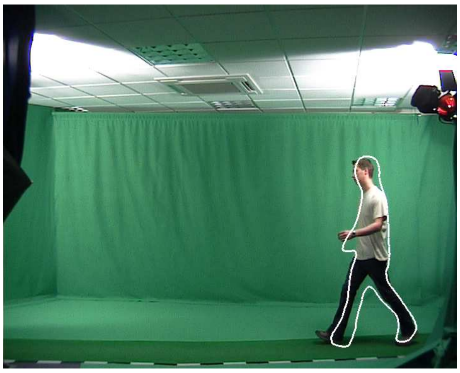
Frame 5



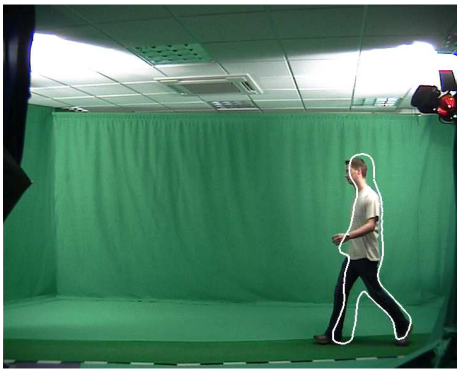
Frame 6



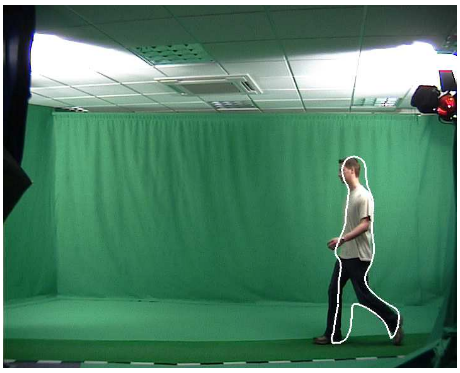
Frame 7



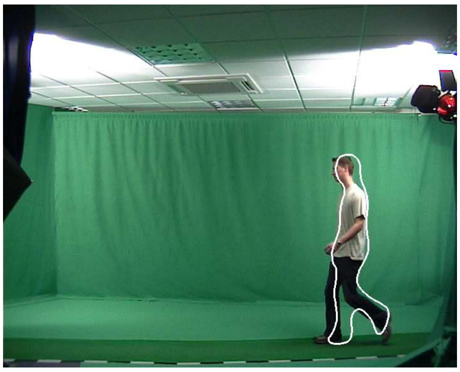
Frame 8



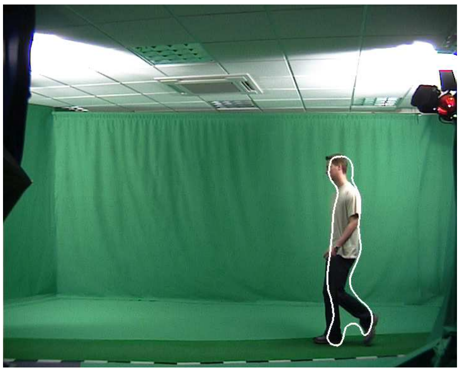
Frame 9



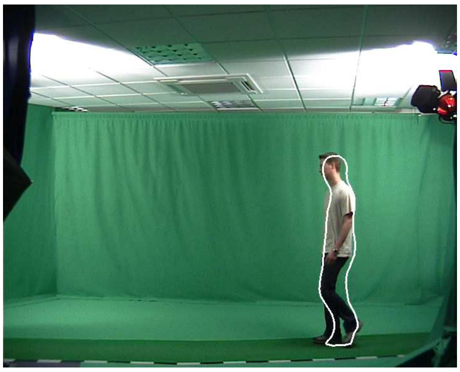
Frame 10



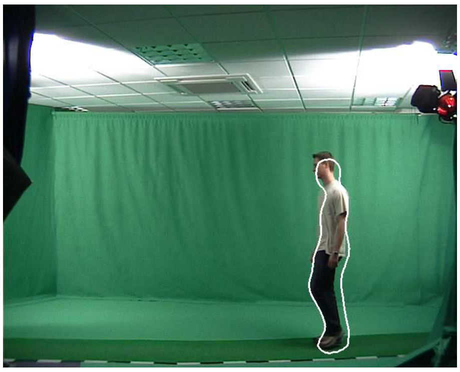
Frame 11



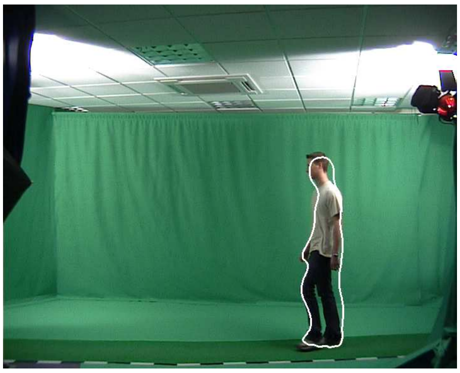
Frame 12



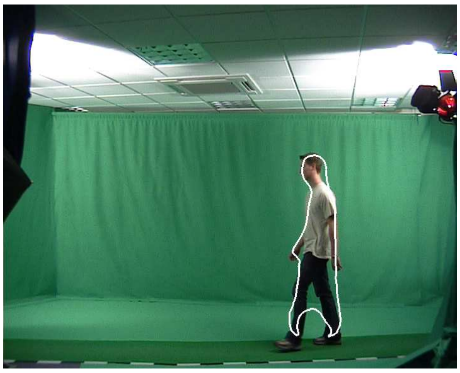
Frame 13



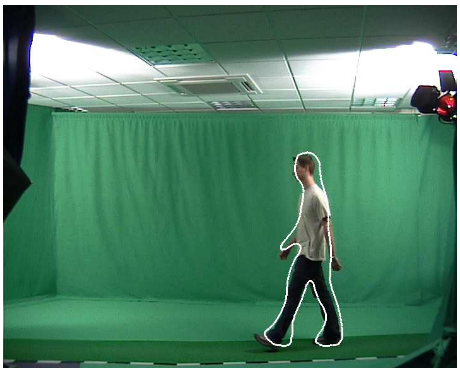
Frame 14



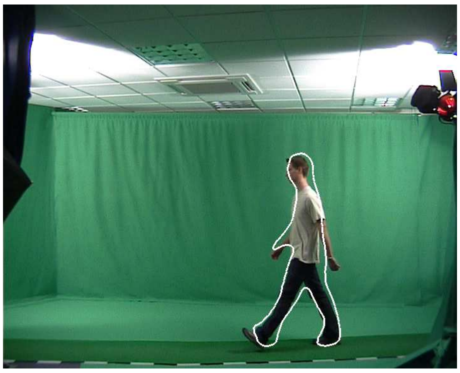
Frame 15



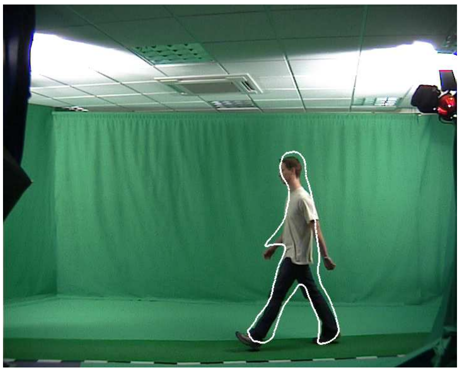
Frame 16



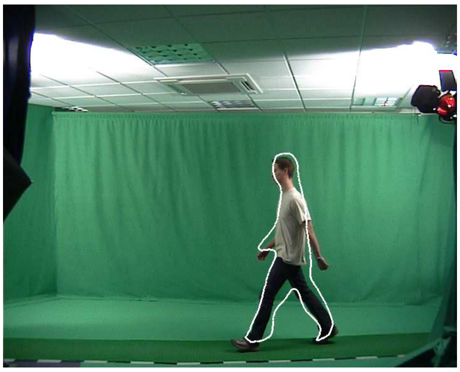
Frame 17



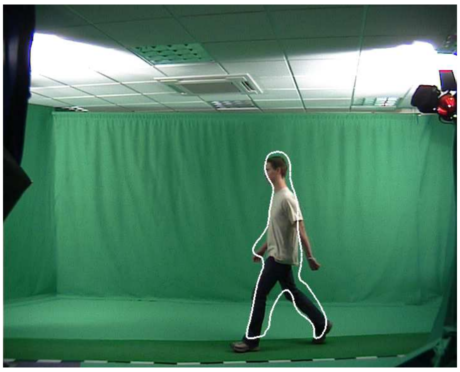
Frame 18



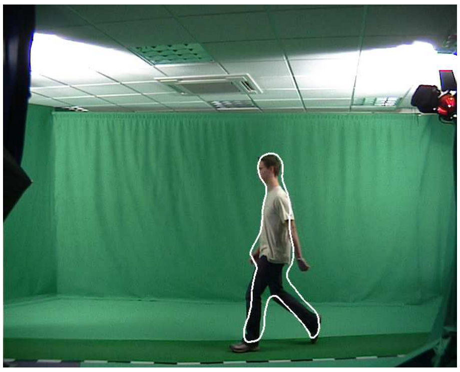
Frame 19



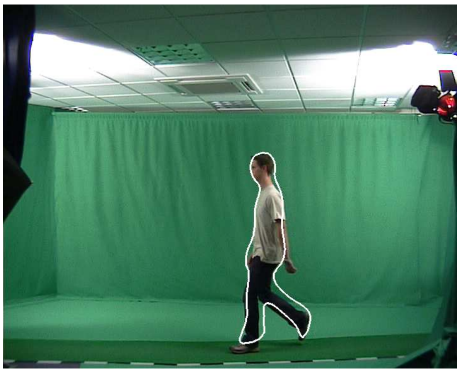
Frame 20



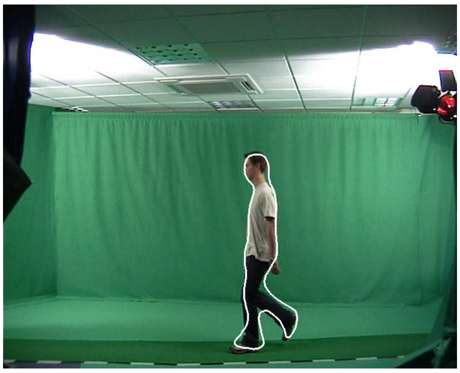
Frame 21



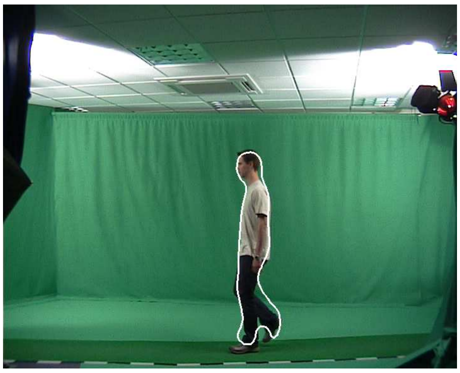
Frame 22



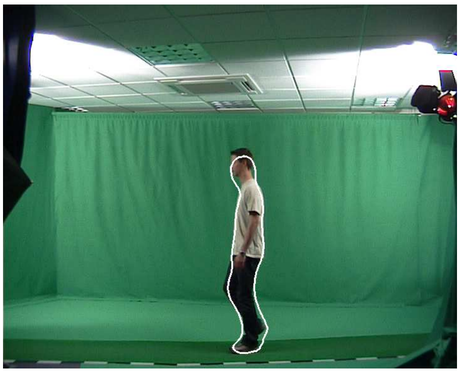
Frame 23



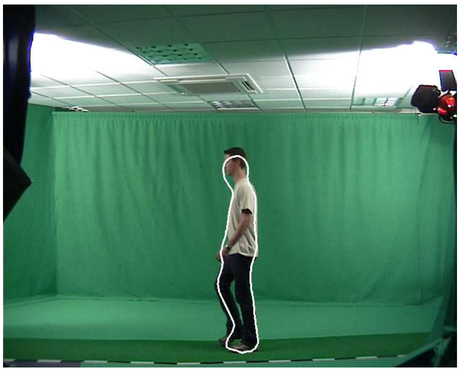
Frame 24



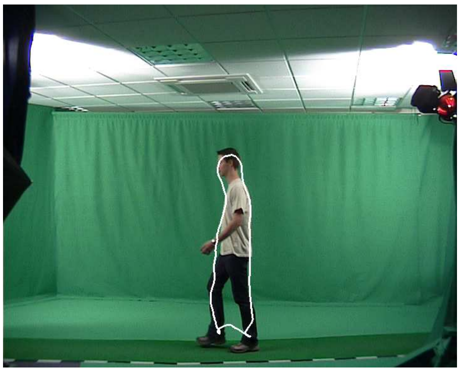
Frame 25



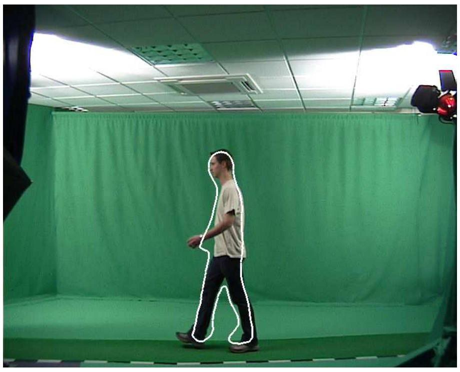
Frame 26



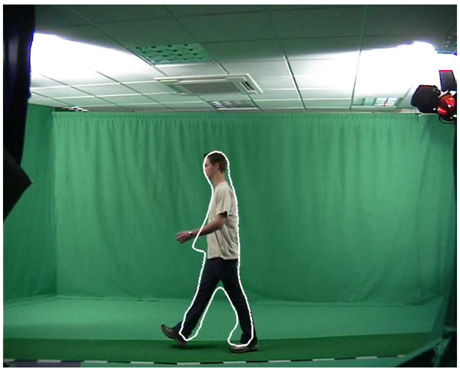
Frame 27



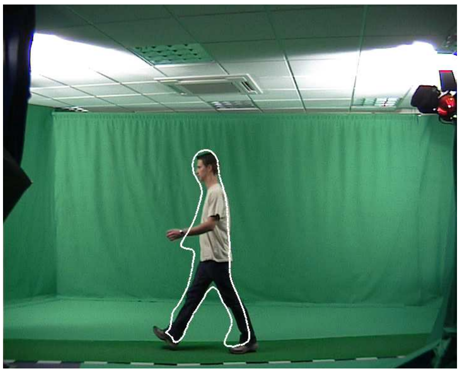
Frame 28



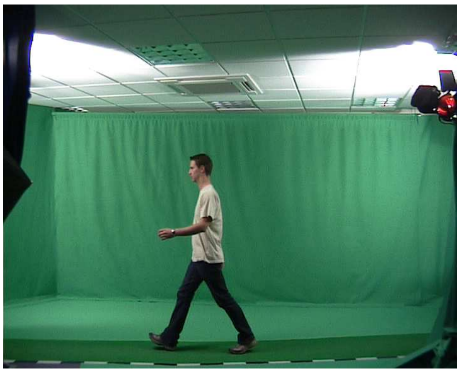
Frame 29



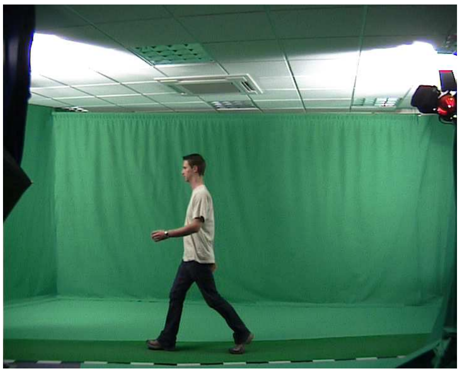
Frame 30



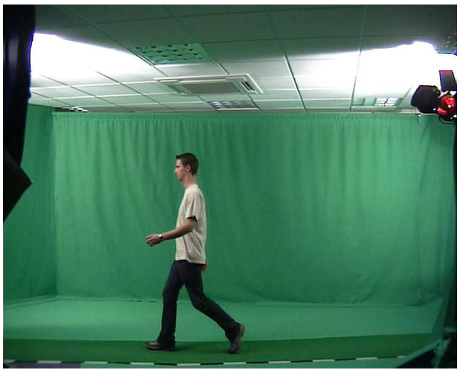
Frame 31



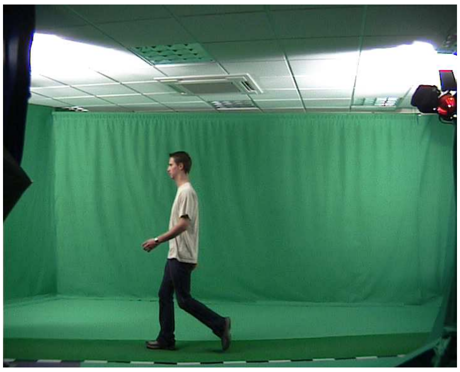
Frame 32



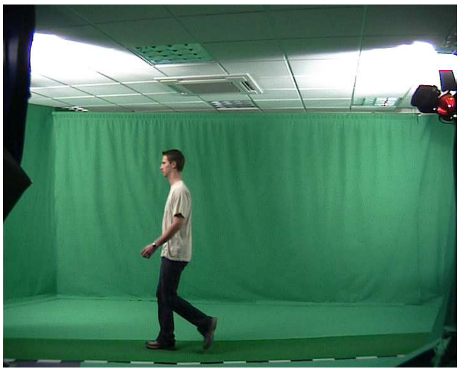
Frame 33



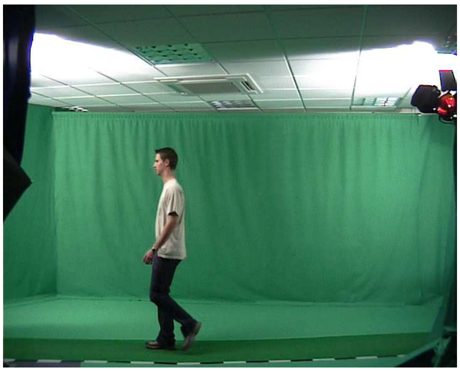
Frame 34



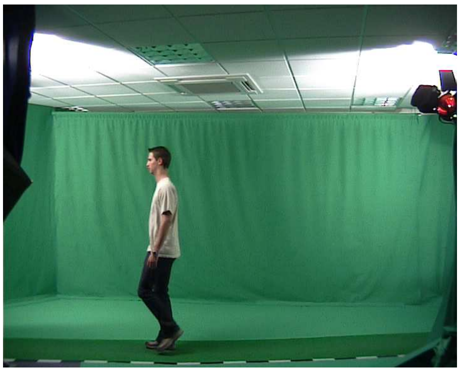
Frame 35



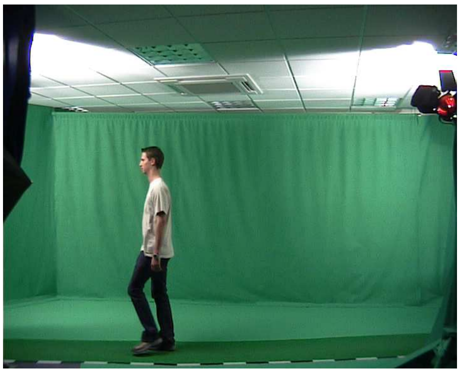
Frame 36



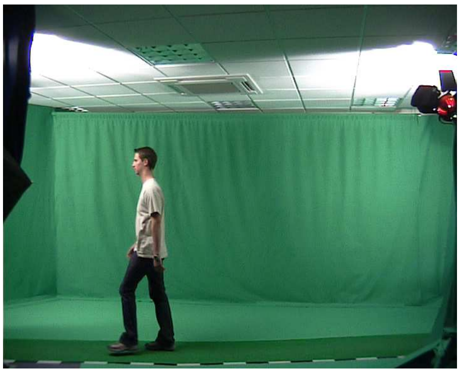
Frame 37



Frame 38



Frame 39

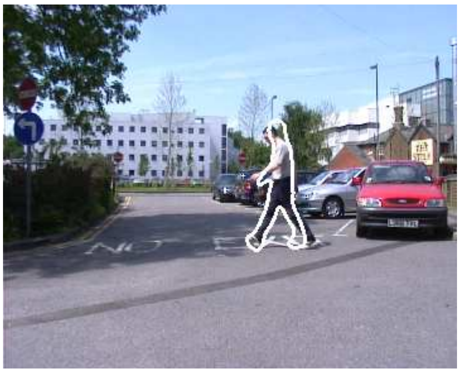


Frame 40

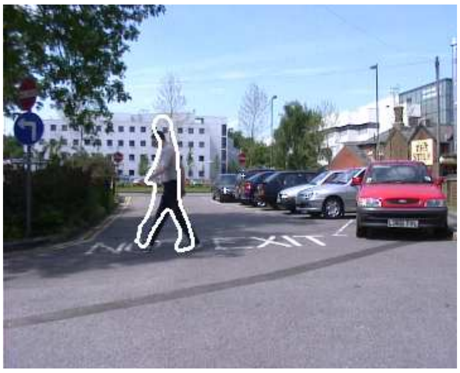
Appendix B

Detailed real-world raw image data test results

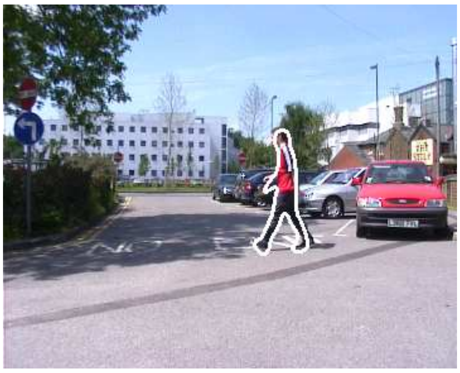
The following images display the first and last frames of each gait cycle identified using the CDHT from the real-world raw edge-detected data, with the reconstructed Shape Sequence Template, spatio-temporally rescaled and translated to the parameters extracted by the CDHT, are overlaid on the original image sequence. The sequence for Subject 37 was omitted from testing due to it being incorrectly marked in the University of Southampton's Large Gait Database as being a Right-To-Left sequence, when it was actually a Left-To-Right sequence. This was replaced with a sequence from Subject 63.



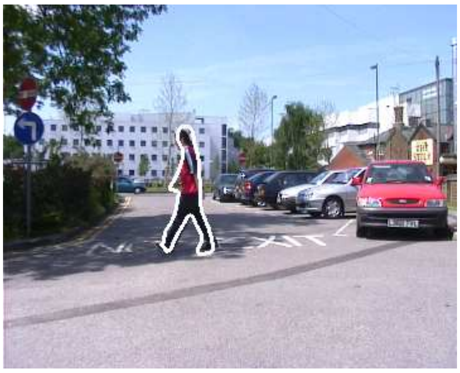
Subject 13: First frame



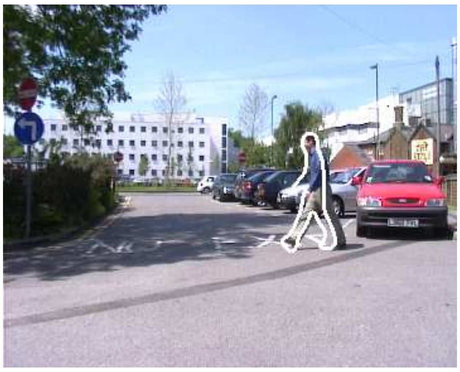
Subject 13: Last frame



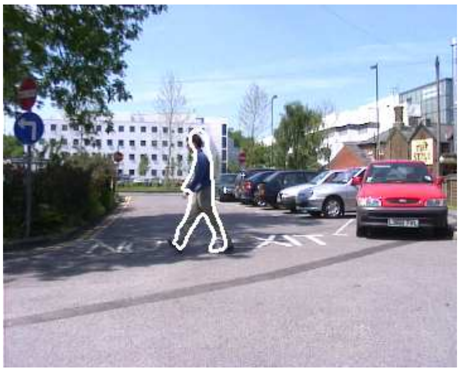
Subject 14: First frame



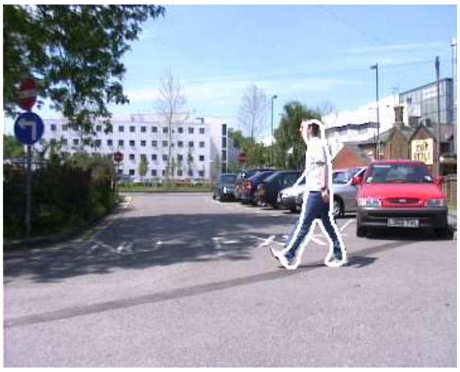
Subject 14: Last frame



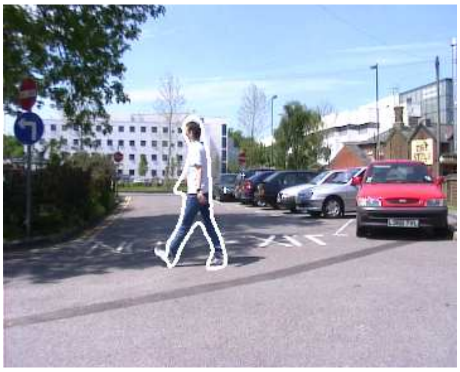
Subject 15: First frame



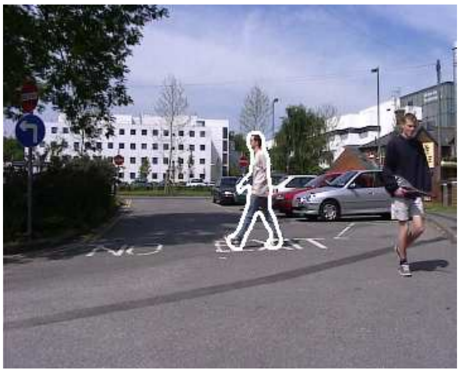
Subject 15: Last frame



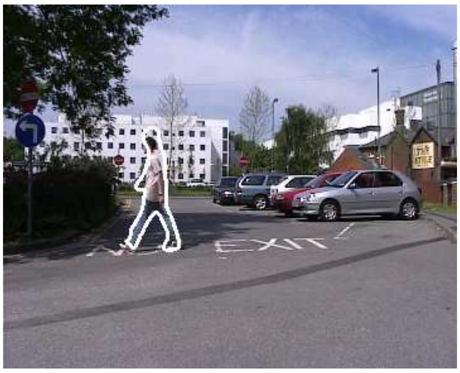
Subject 16: First frame



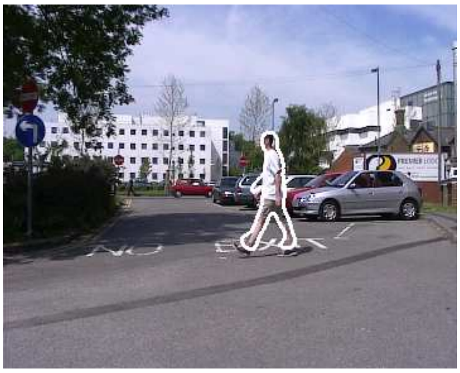
Subject 16: Last frame



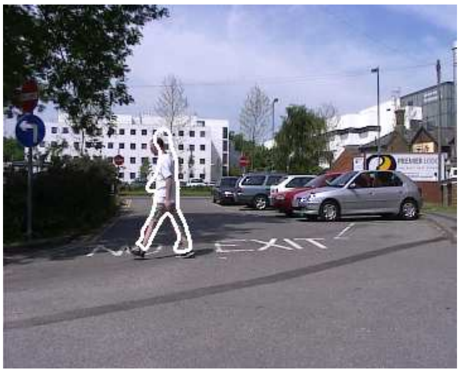
Subject 17: First frame



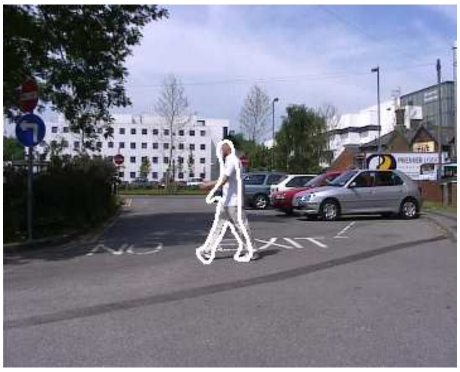
Subject 17: Last frame



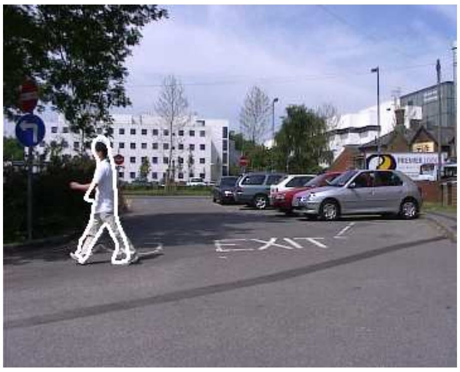
Subject 18: First frame



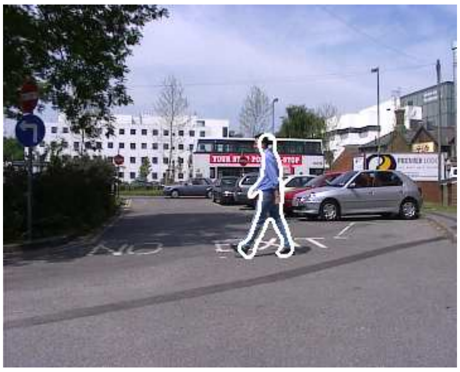
Subject 18: Last frame



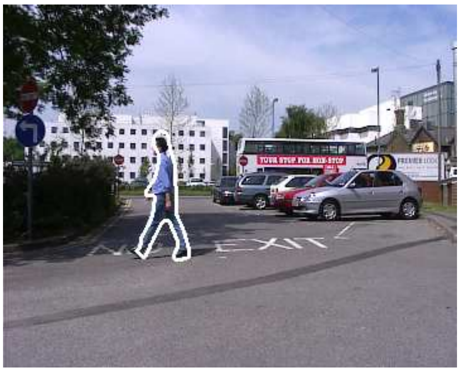
Subject 19: First frame



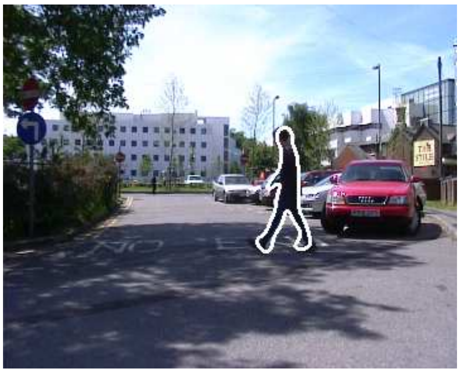
Subject 19: Last frame



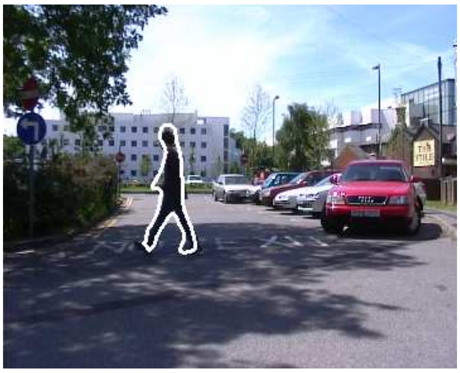
Subject 20: First frame



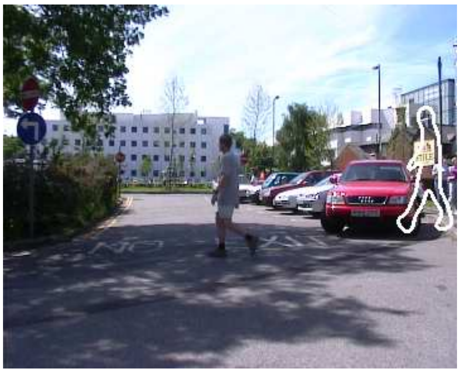
Subject 20: Last frame



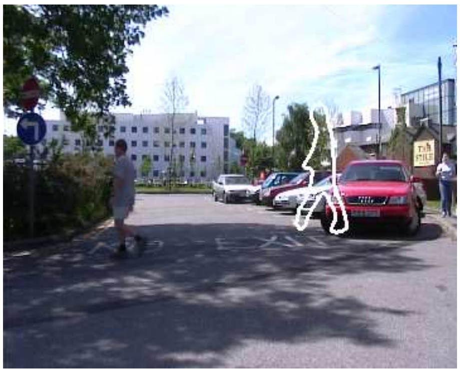
Subject 21: First frame



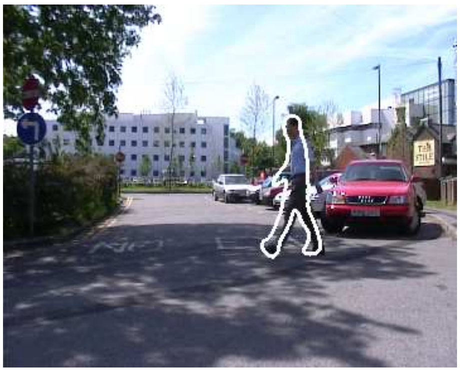
Subject 21: Last frame



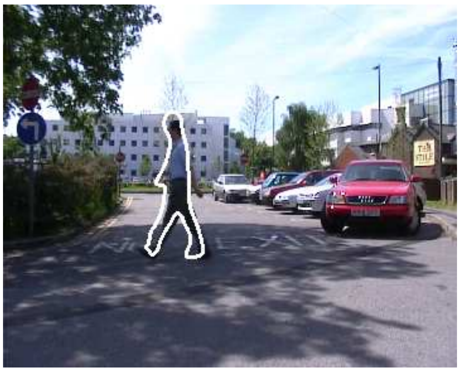
Subject 22: First frame



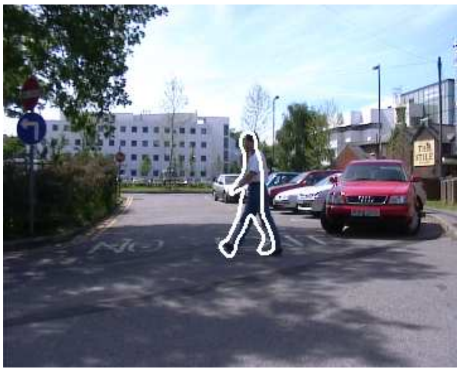
Subject 22: Last frame



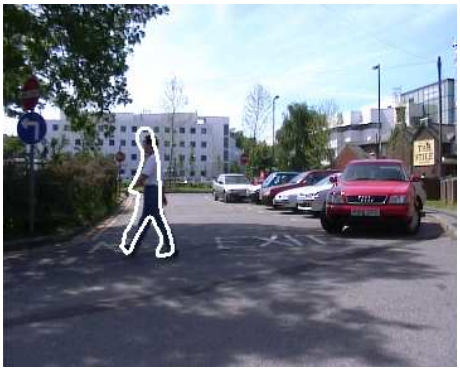
Subject 23: First frame



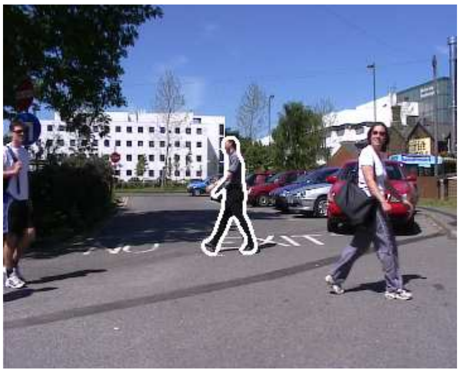
Subject 23: Last frame



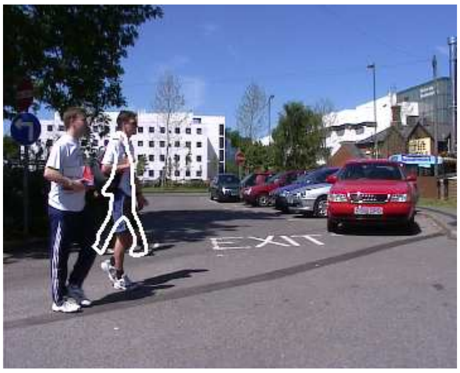
Subject 24: First frame



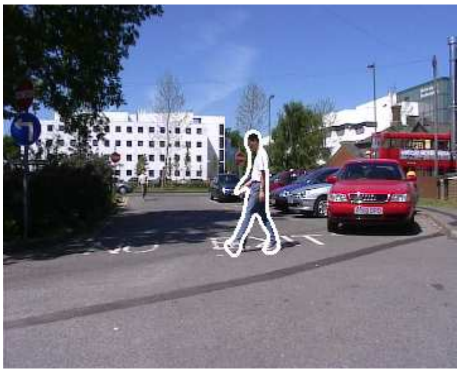
Subject 24: Last frame



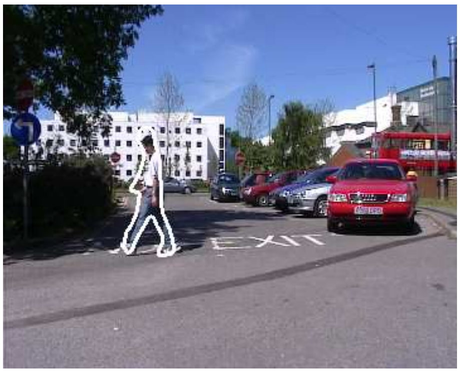
Subject 25: First frame



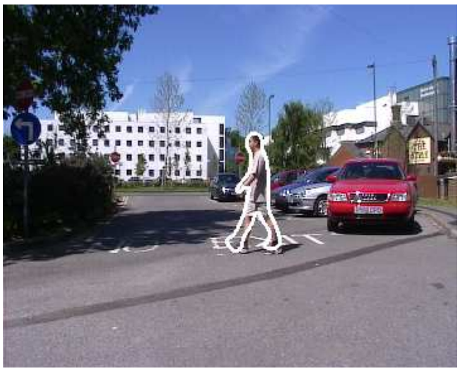
Subject 25: Last frame



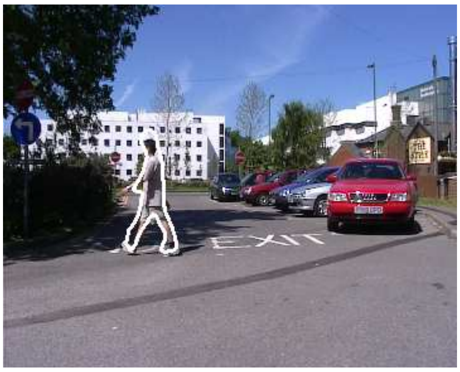
Subject 26: First frame



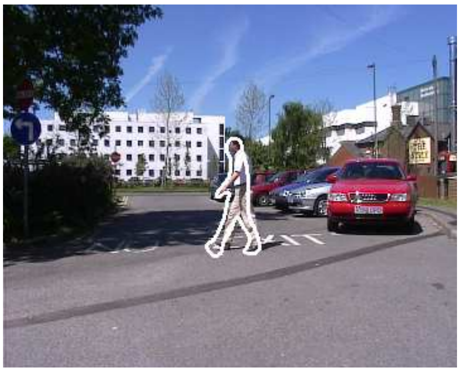
Subject 26: Last frame



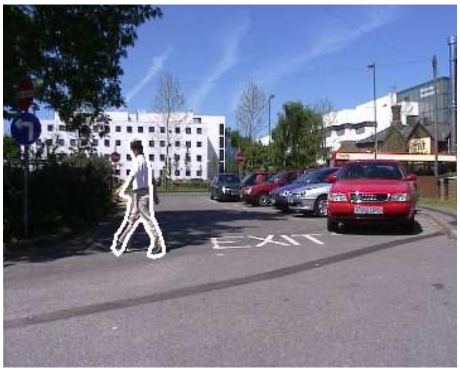
Subject 27: First frame



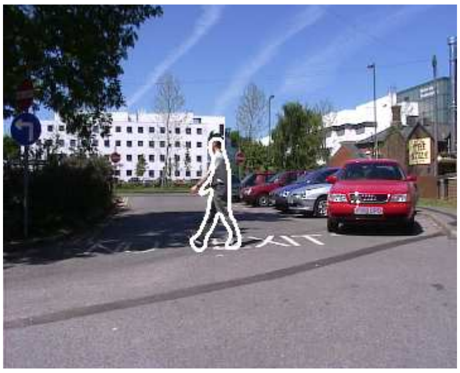
Subject 27: Last frame



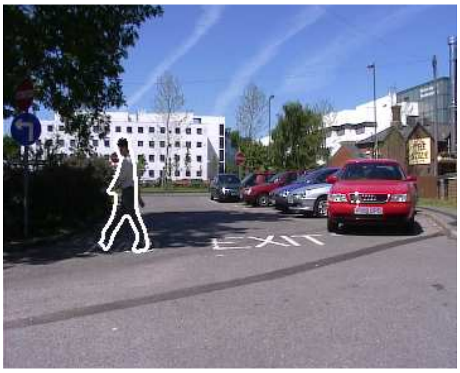
Subject 28: First frame



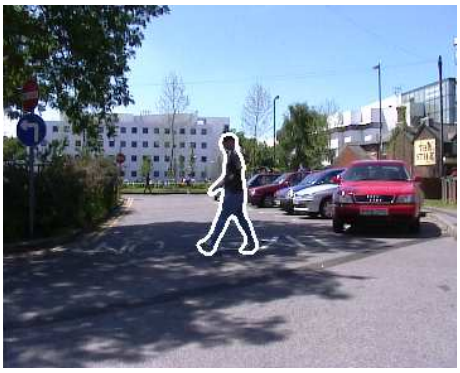
Subject 28: Last frame



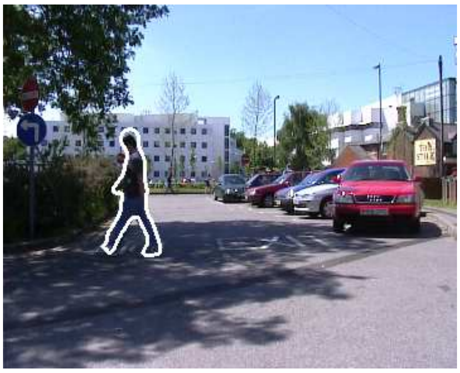
Subject 29: First frame



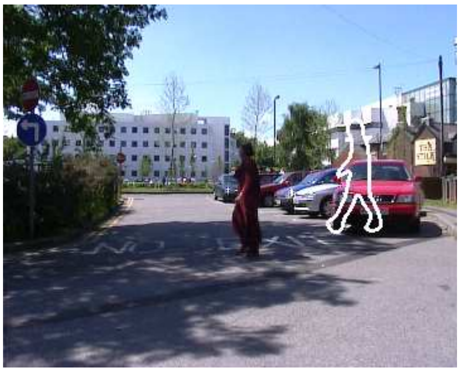
Subject 29: Last frame



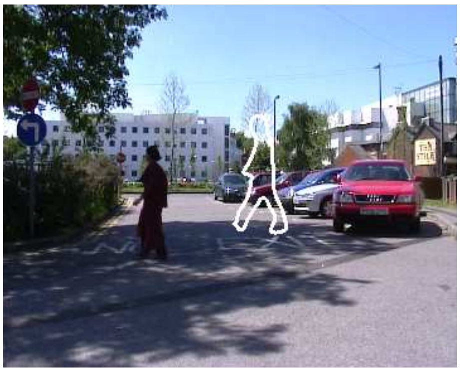
Subject 30: First frame



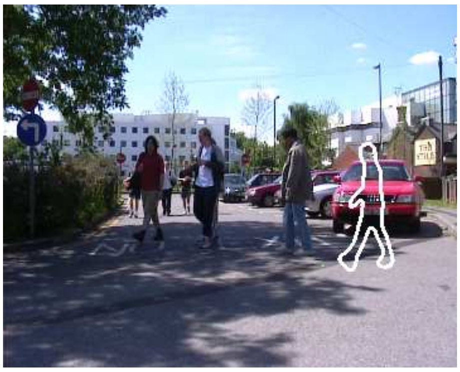
Subject 30: Last frame



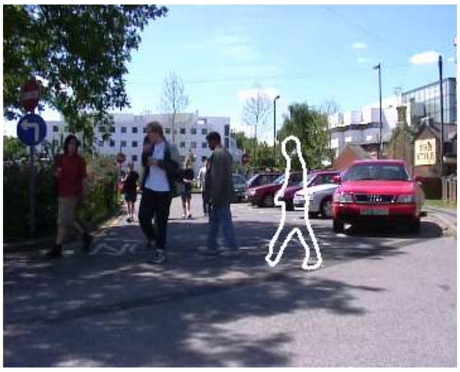
Subject 31: First frame



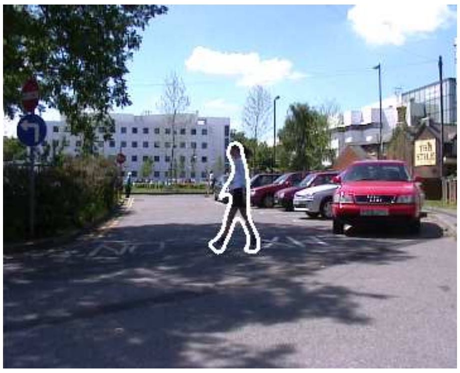
Subject 31: Last frame



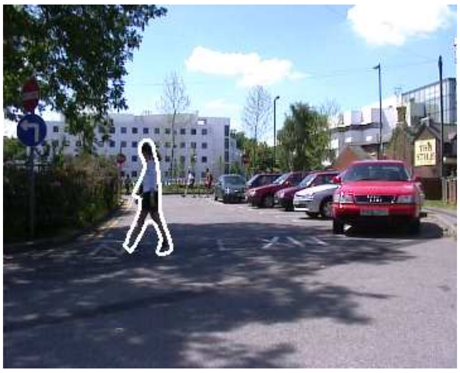
Subject 32: First frame



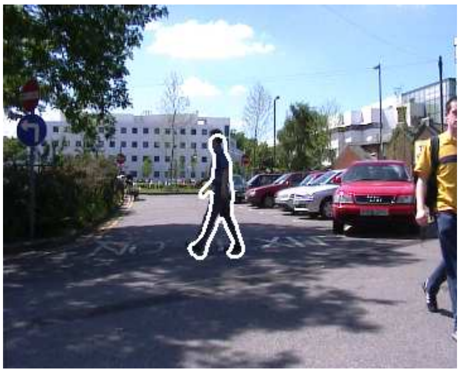
Subject 32: Last frame



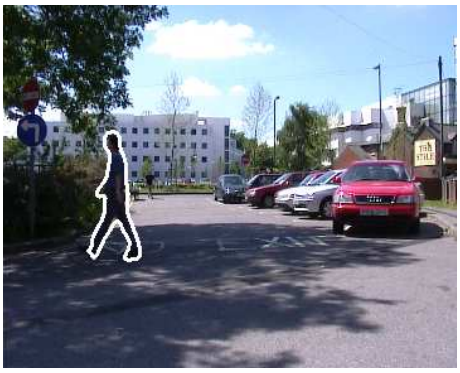
Subject 33: First frame



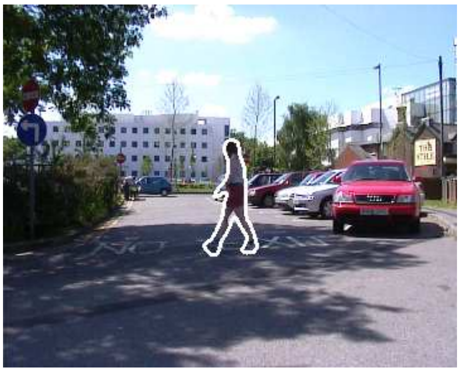
Subject 33: Last frame



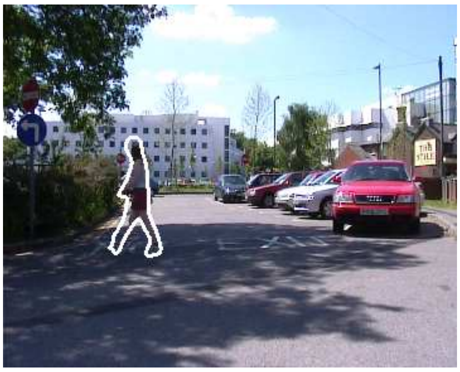
Subject 34: First frame



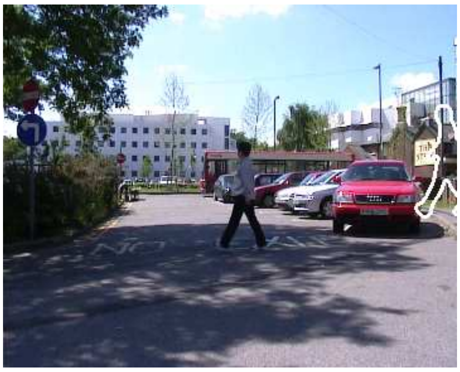
Subject 34: Last frame



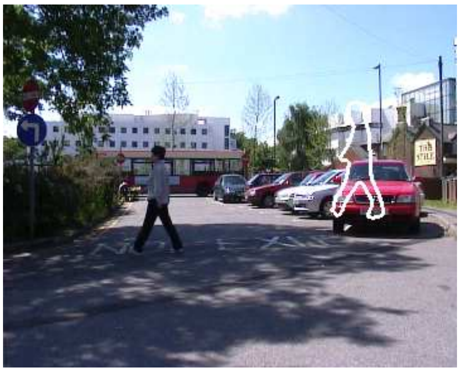
Subject 35: First frame



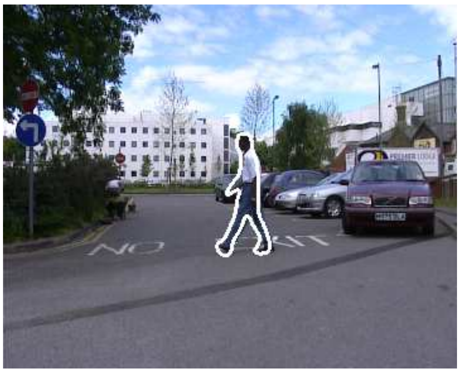
Subject 35: Last frame



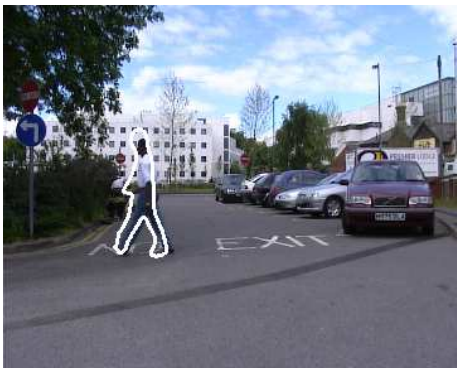
Subject 36: First frame



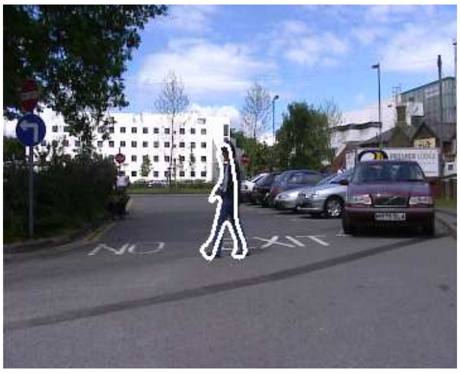
Subject 36: Last frame



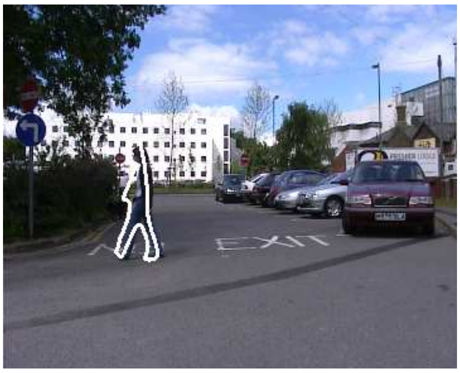
Subject 38: First frame



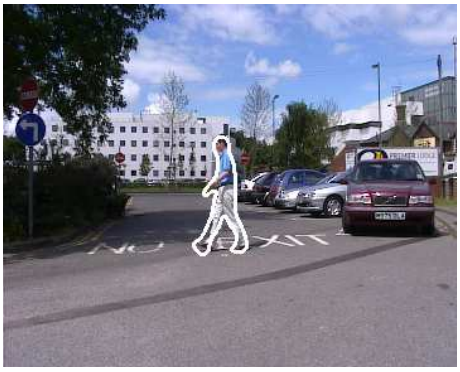
Subject 38: Last frame



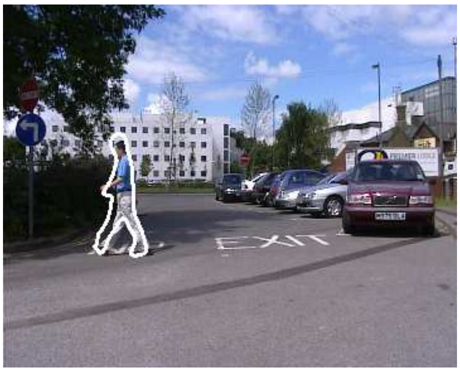
Subject 39: First frame



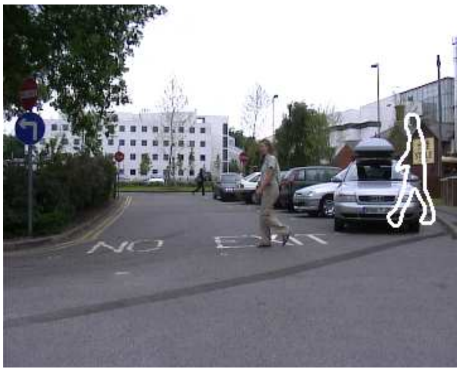
Subject 39: Last frame



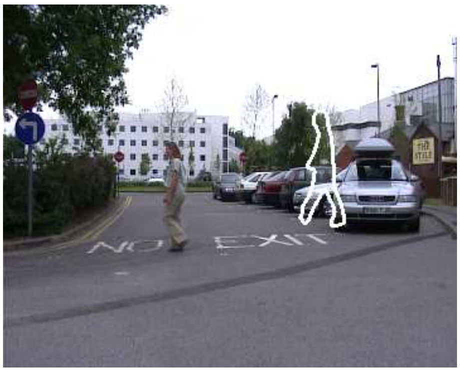
Subject 40: First frame



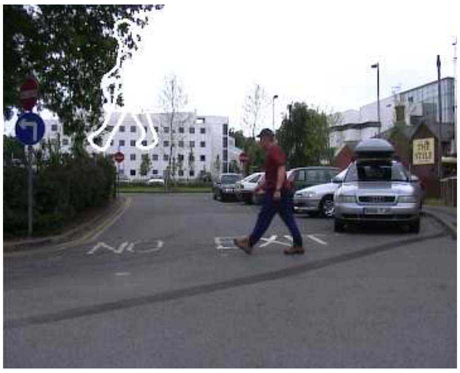
Subject 40: Last frame



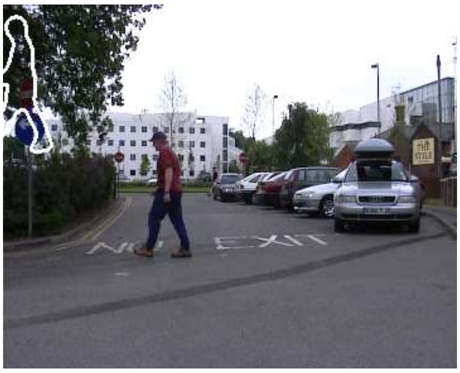
Subject 41: First frame



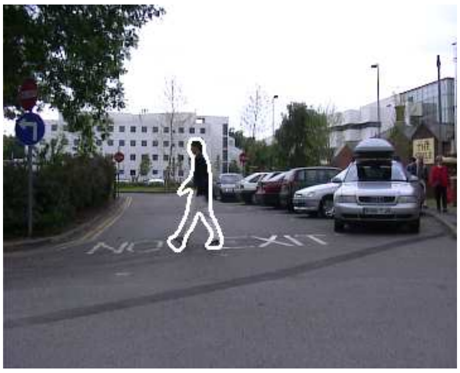
Subject 41: Last frame



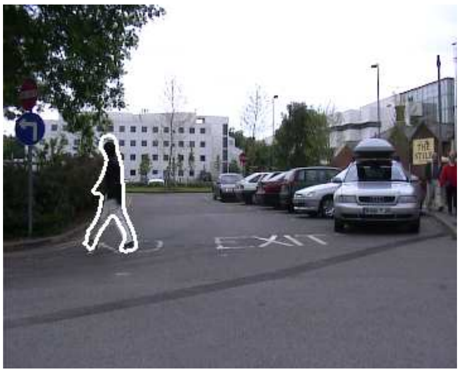
Subject 42: First frame



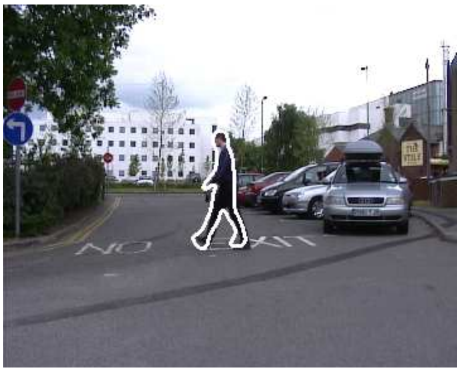
Subject 42: Last frame



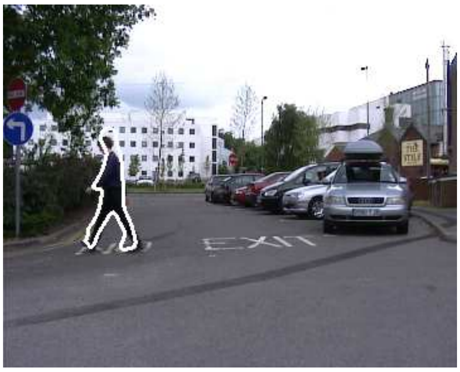
Subject 43: First frame



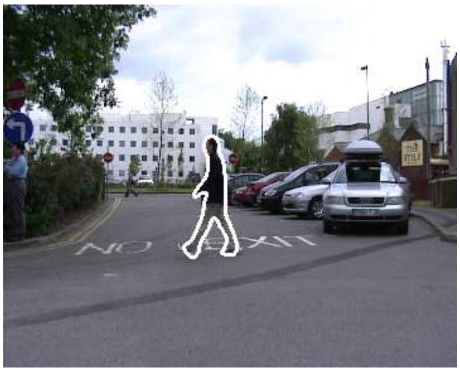
Subject 43: Last frame



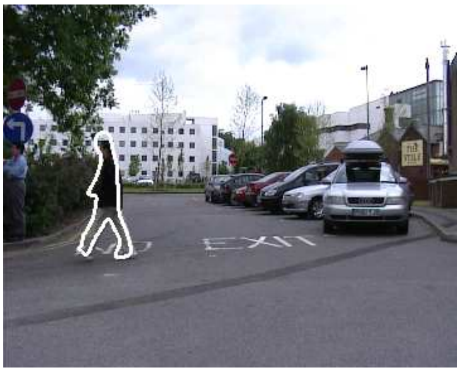
Subject 44: First frame



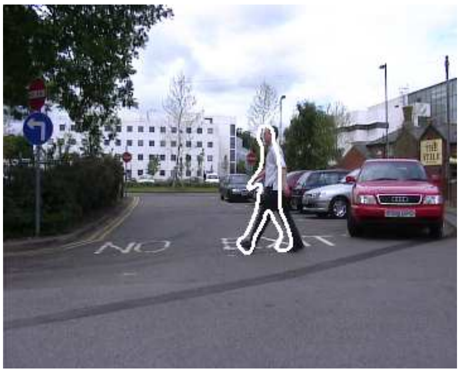
Subject 44: Last frame



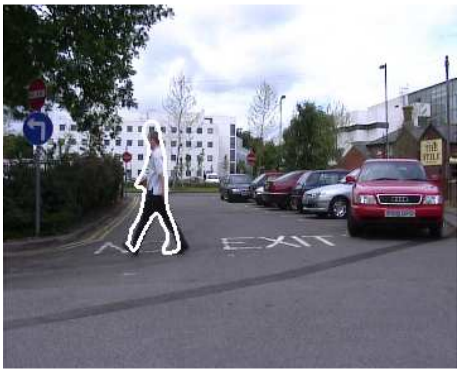
Subject 45: First frame



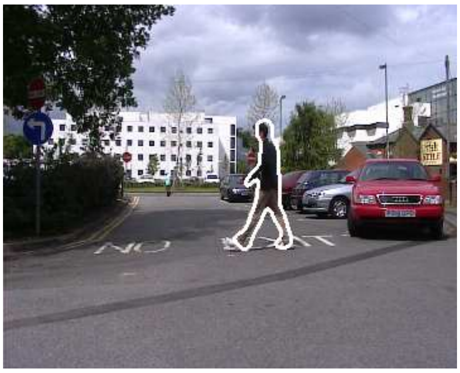
Subject 45: Last frame



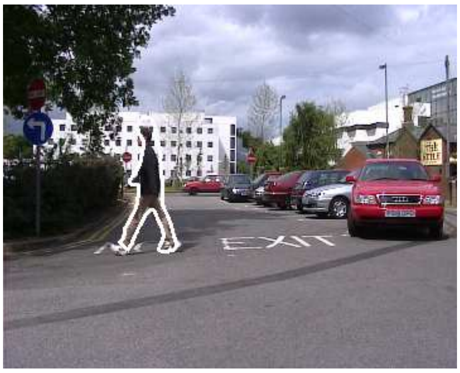
Subject 46: First frame



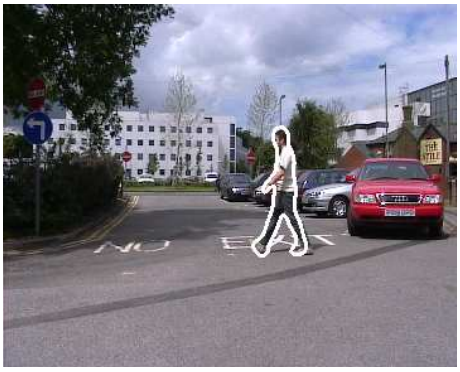
Subject 46: Last frame



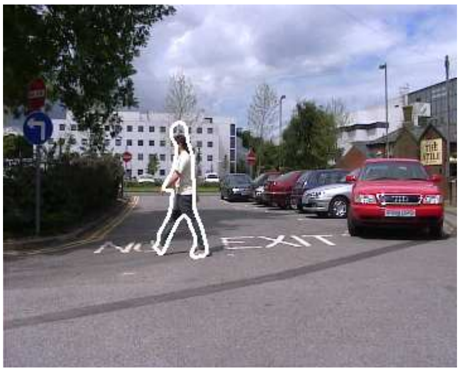
Subject 47: First frame



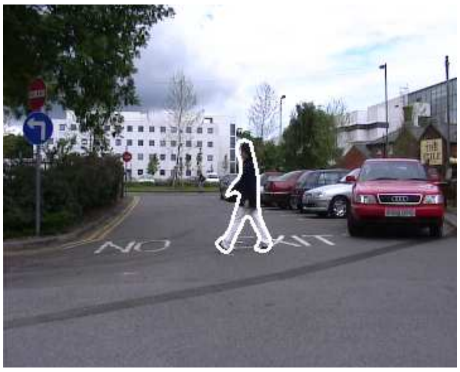
Subject 47: Last frame



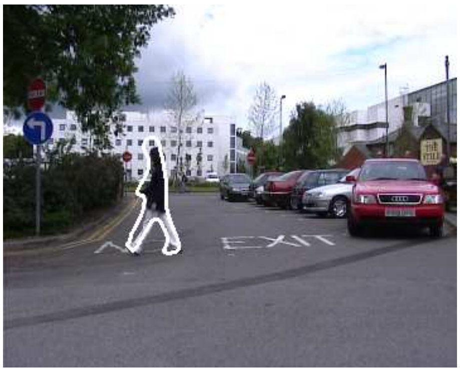
Subject 48: First frame



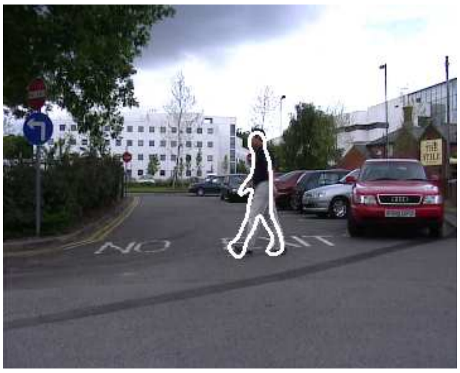
Subject 48: Last frame



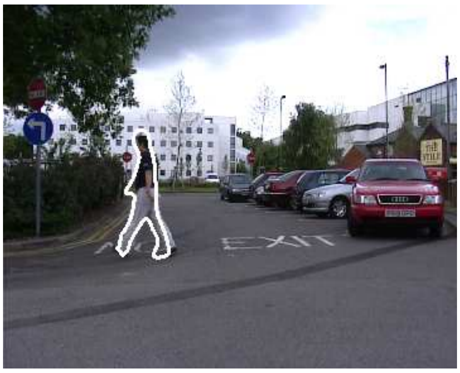
Subject 49: First frame



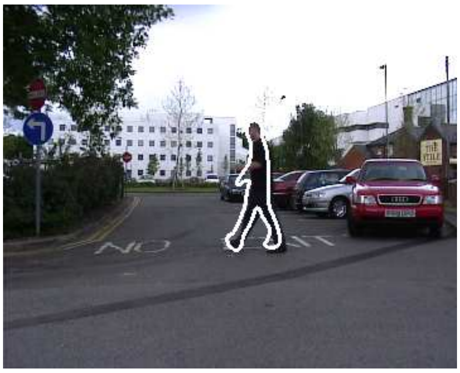
Subject 49: Last frame



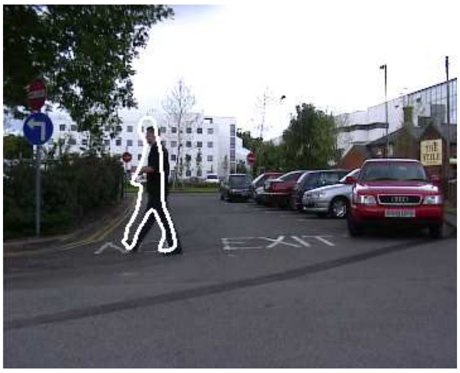
Subject 50: First frame



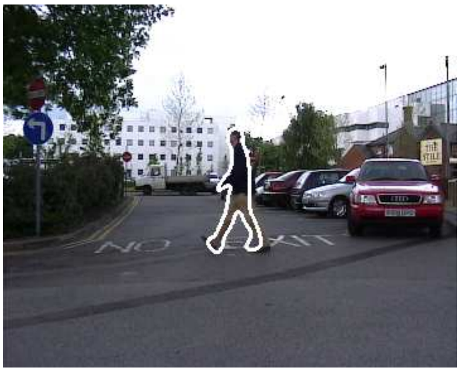
Subject 50: Last frame



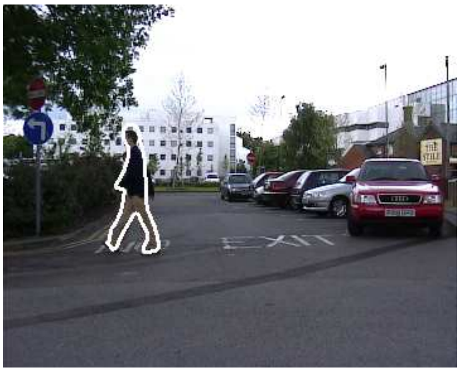
Subject 51: First frame



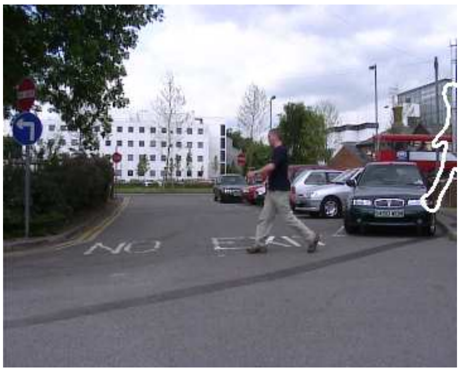
Subject 51: Last frame



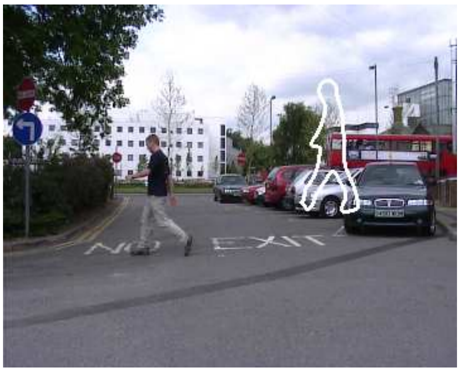
Subject 52: First frame



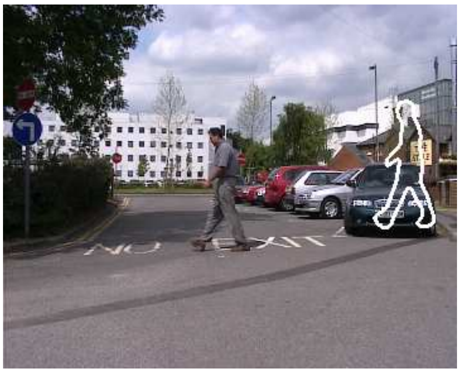
Subject 52: Last frame



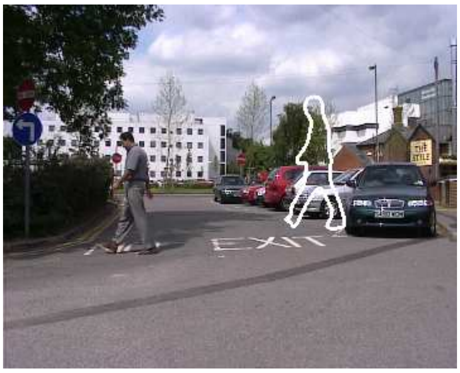
Subject 53: First frame



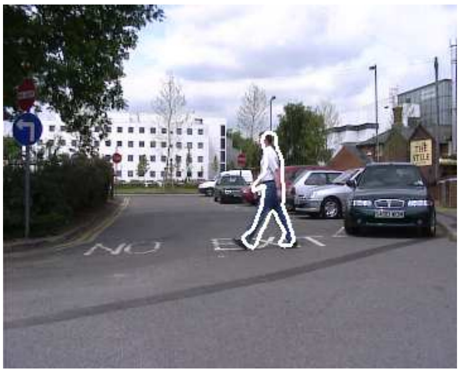
Subject 53: Last frame



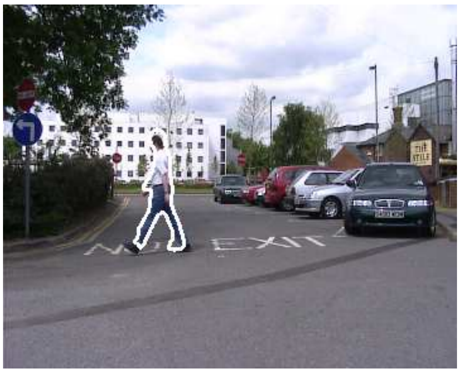
Subject 54: First frame



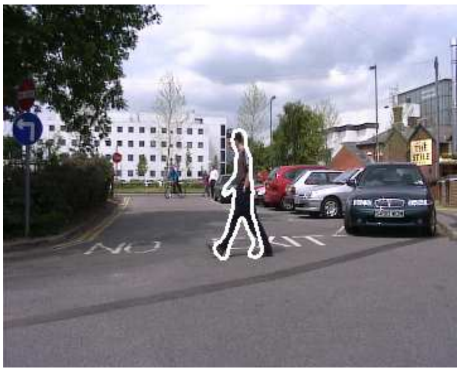
Subject 54: Last frame



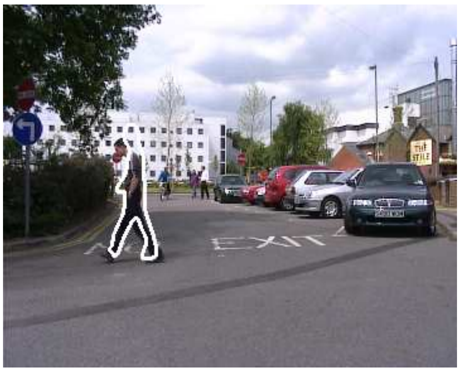
Subject 55: First frame



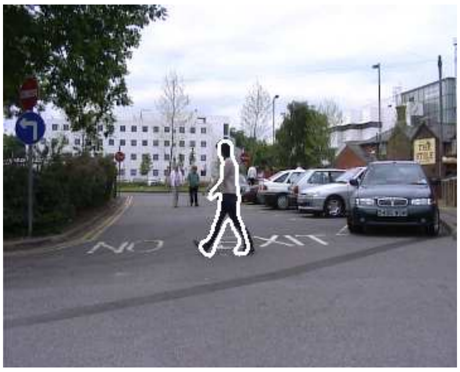
Subject 55: Last frame



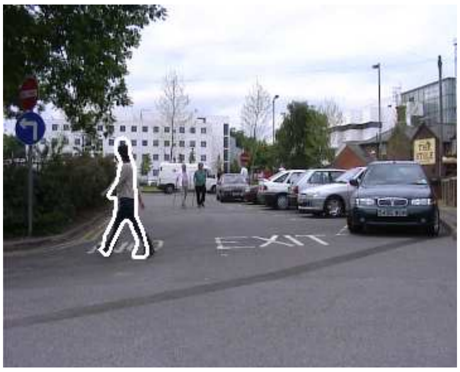
Subject 56: First frame



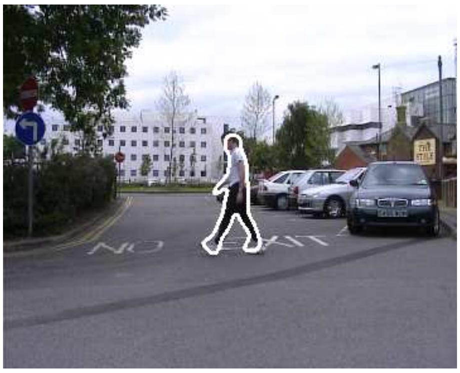
Subject 56: Last frame



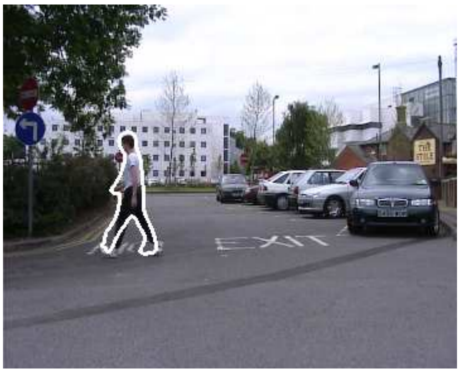
Subject 57: First frame



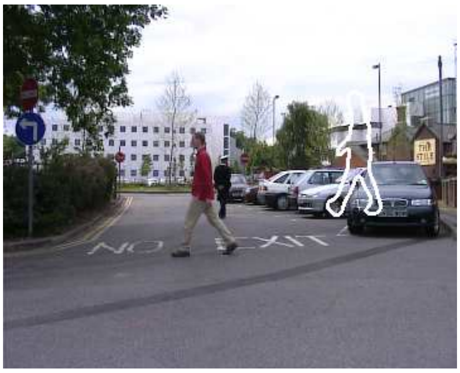
Subject 57: Last frame



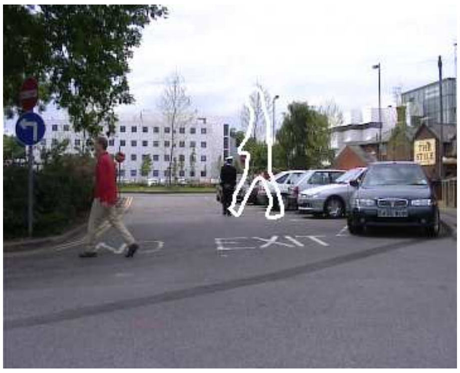
Subject 58: First frame



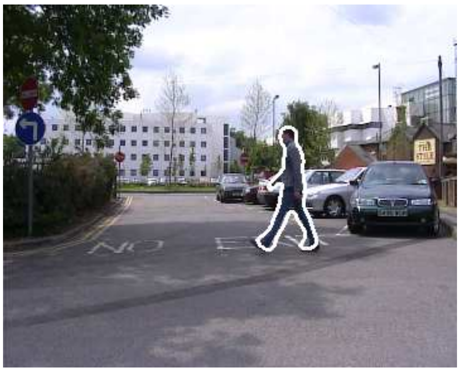
Subject 58: Last frame



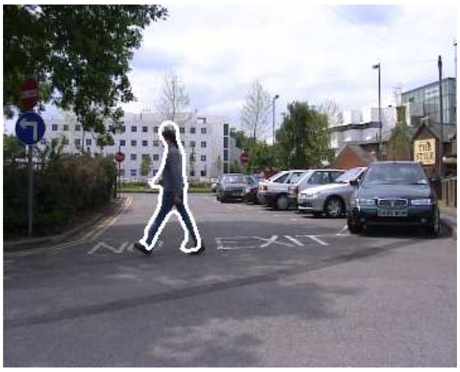
Subject 59: First frame



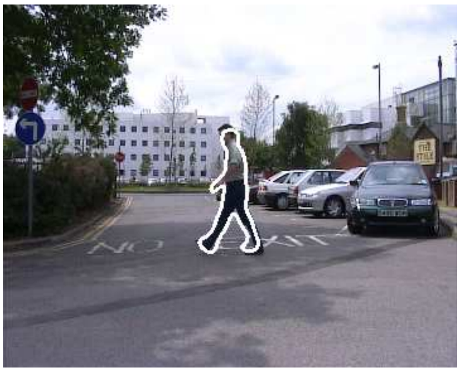
Subject 59: Last frame



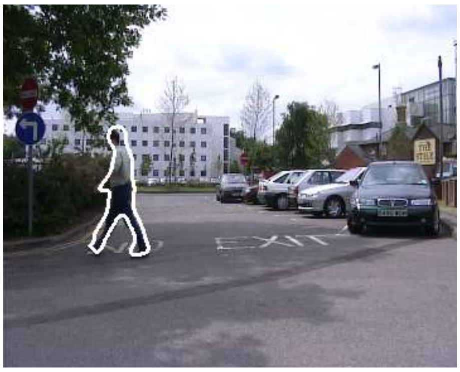
Subject 60: First frame



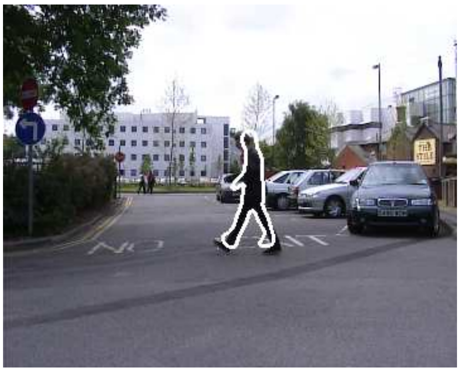
Subject 60: Last frame



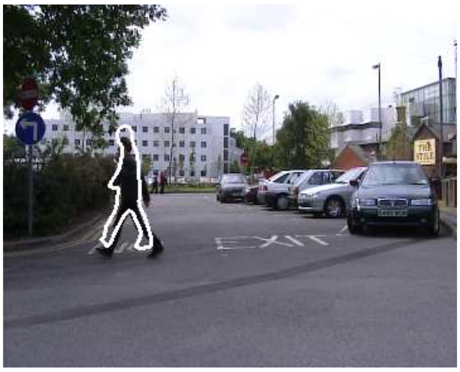
Subject 61: First frame



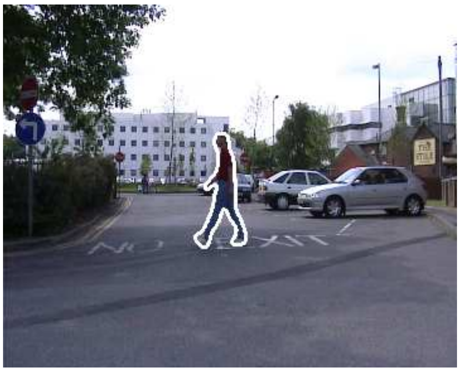
Subject 61: Last frame



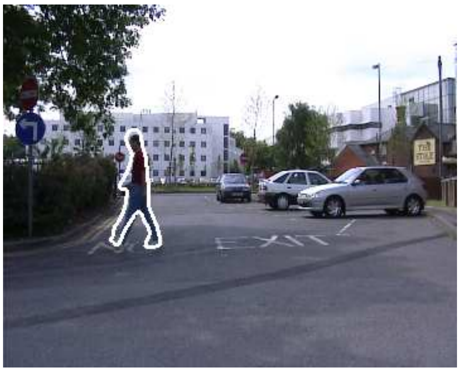
Subject 62: First frame



Subject 62: Last frame



Subject 63: First frame

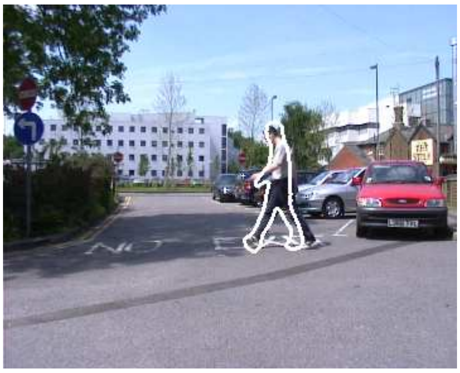


Subject 63: Last frame

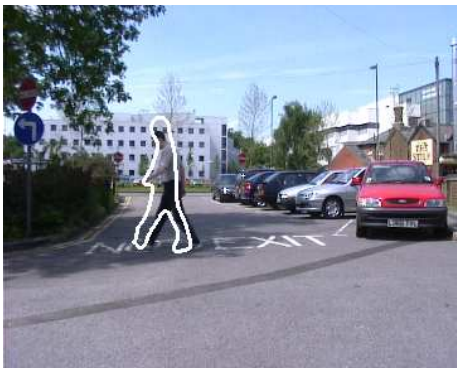
Appendix C

Detailed real-world pre-processed image data test results

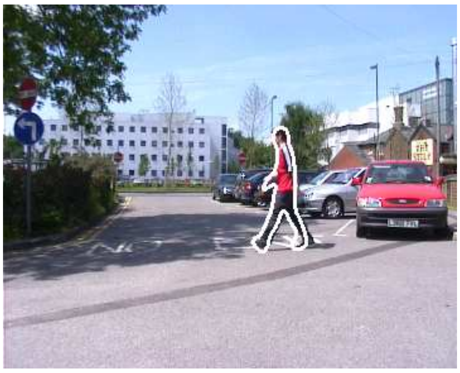
The following images display the first and last frames of each gait cycle identified using the CDHT from the real-world pre-processed image data, with the reconstructed Shape Sequence Template, spatio-temporally rescaled and translated to the parameters extracted by the CDHT, are overlaid on the original image sequence. The sequence for Subject 37 was omitted from testing due to it being incorrectly marked in the University of Southampton's Large Gait Database as being a Right-To-Left sequence, when it was actually a Left-To-Right sequence. This was replaced with a sequence from Subject 63.



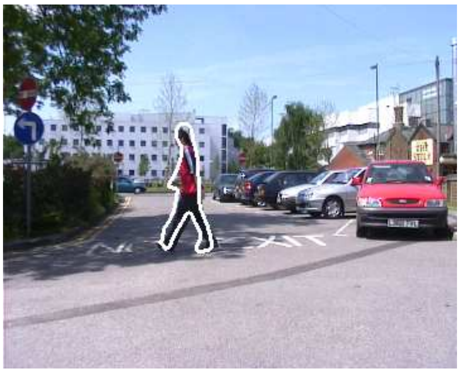
Subject 13: First frame



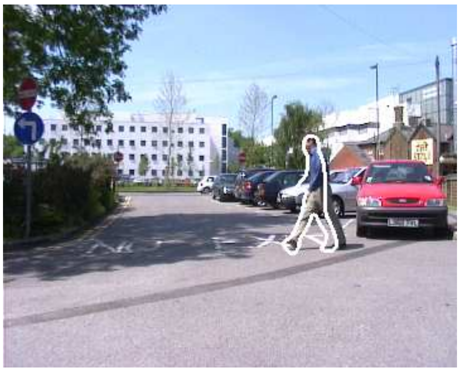
Subject 13: Last frame



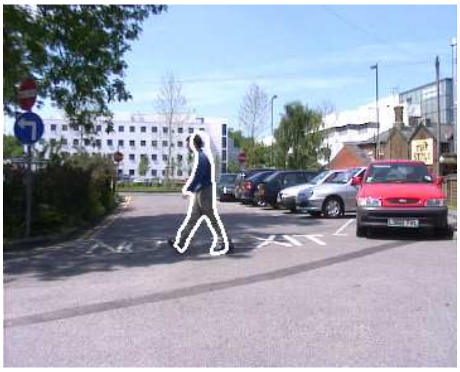
Subject 14: First frame



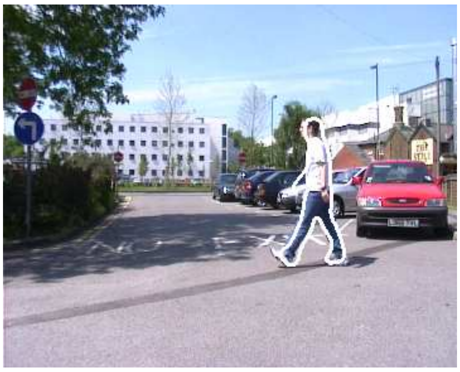
Subject 14: Last frame



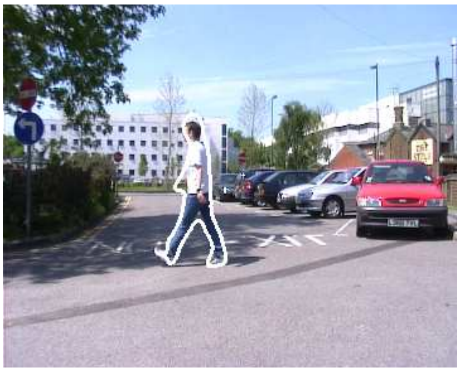
Subject 15: First frame



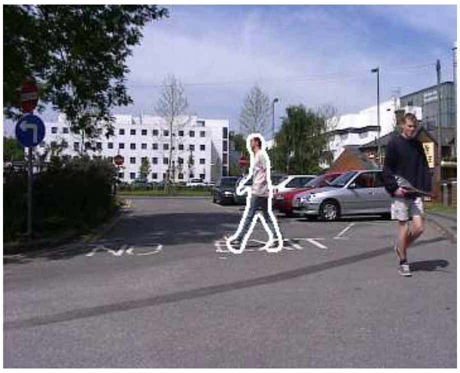
Subject 15: Last frame



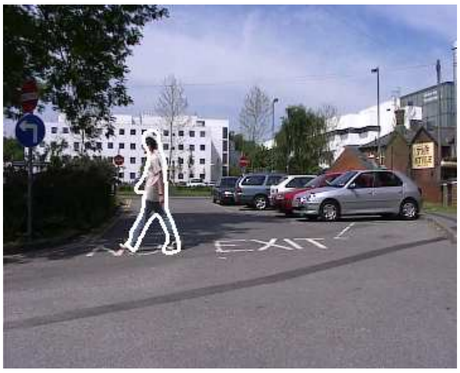
Subject 16: First frame



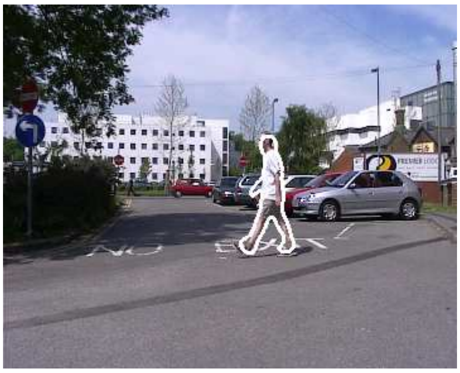
Subject 16: Last frame



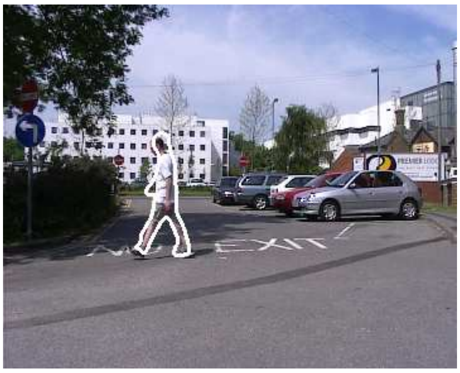
Subject 17: First frame



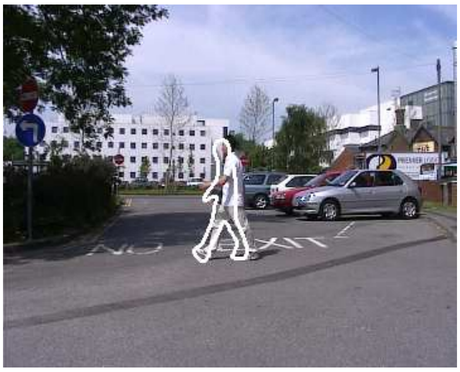
Subject 17: Last frame



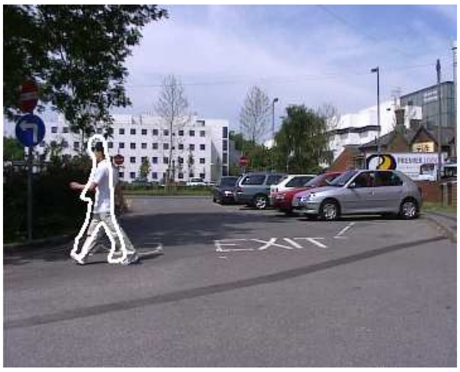
Subject 18: First frame



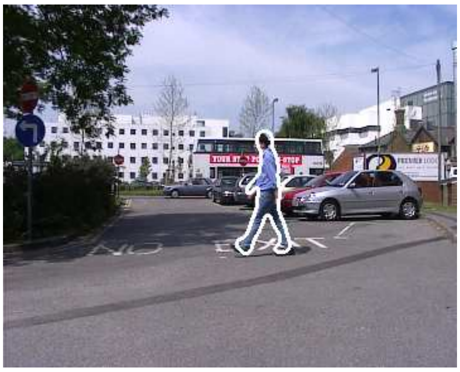
Subject 18: Last frame



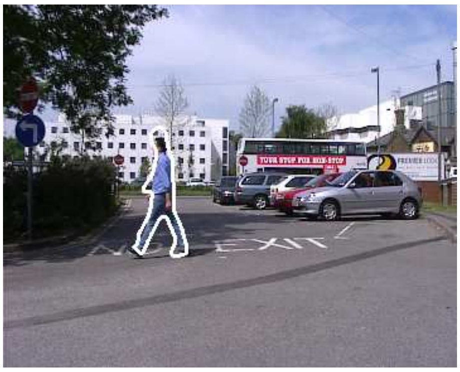
Subject 19: First frame



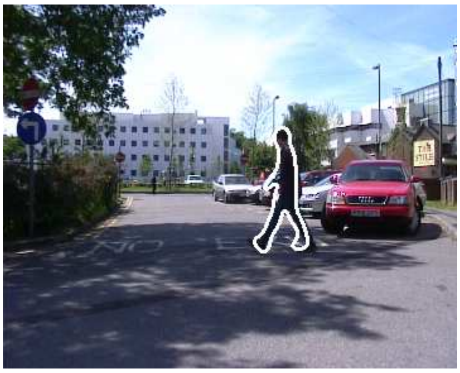
Subject 19: Last frame



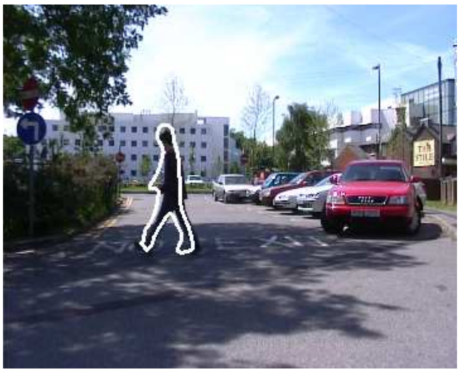
Subject 20: First frame



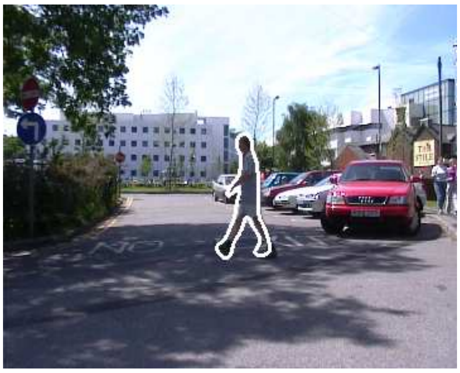
Subject 20: Last frame



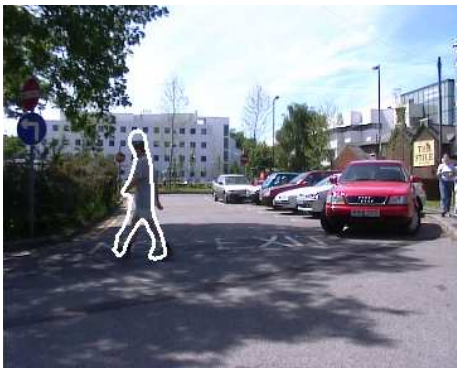
Subject 21: First frame



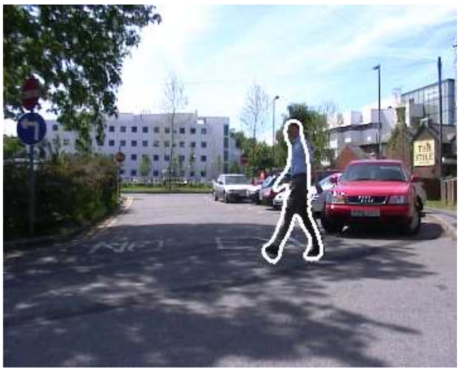
Subject 21: Last frame



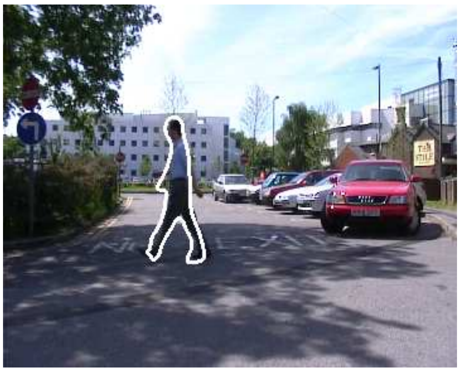
Subject 22: First frame



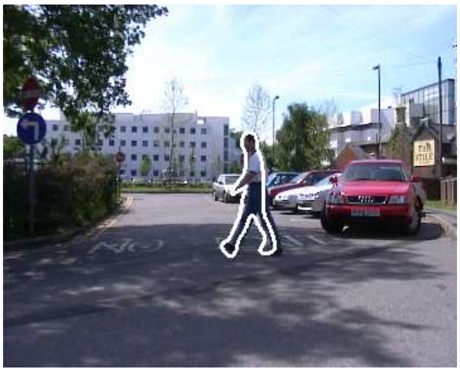
Subject 22: Last frame



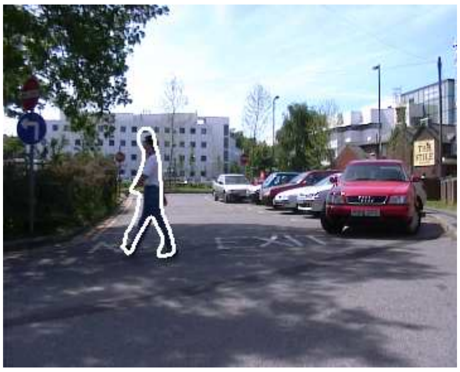
Subject 23: First frame



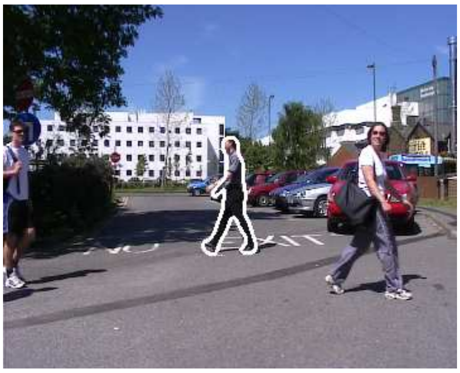
Subject 23: Last frame



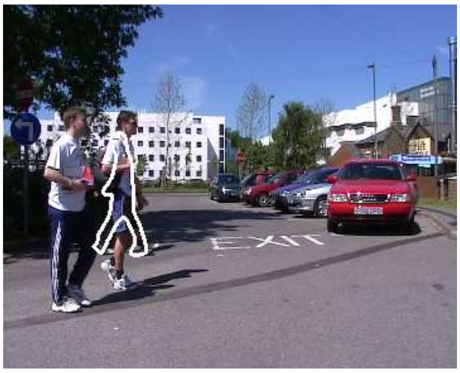
Subject 24: First frame



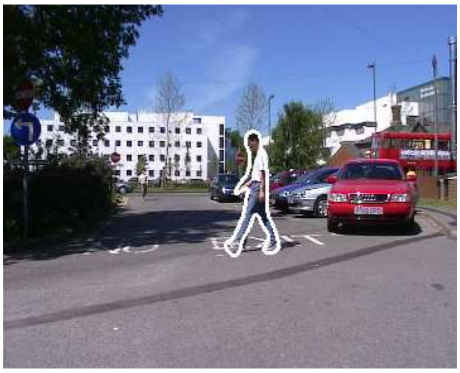
Subject 24: Last frame



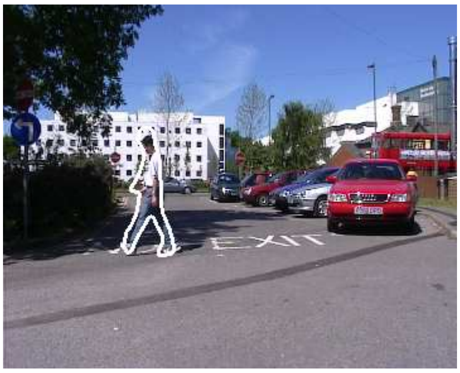
Subject 25: First frame



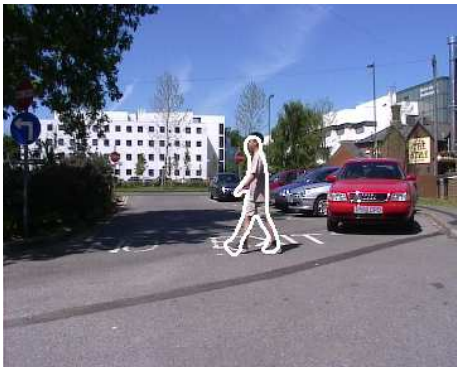
Subject 25: Last frame



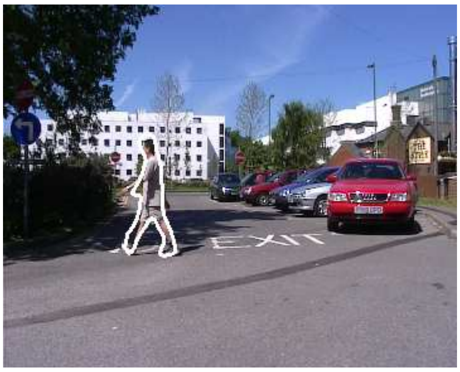
Subject 26: First frame



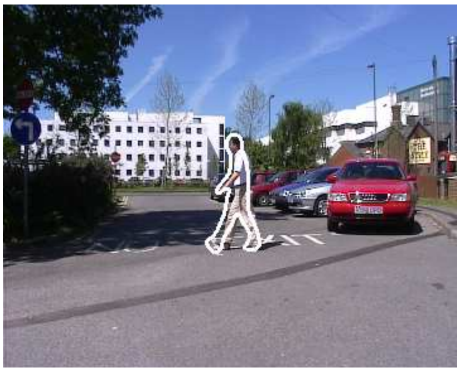
Subject 26: Last frame



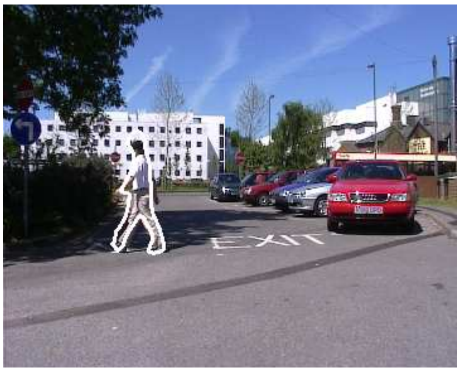
Subject 27: First frame



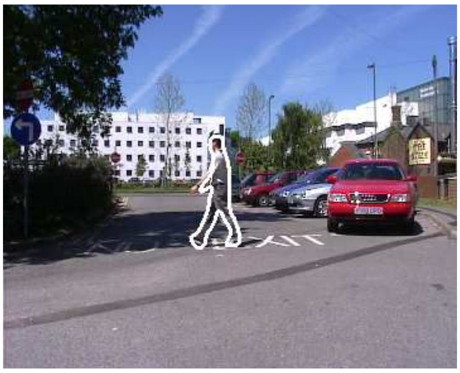
Subject 27: Last frame



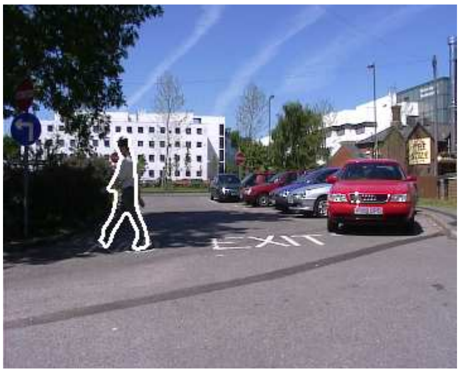
Subject 28: First frame



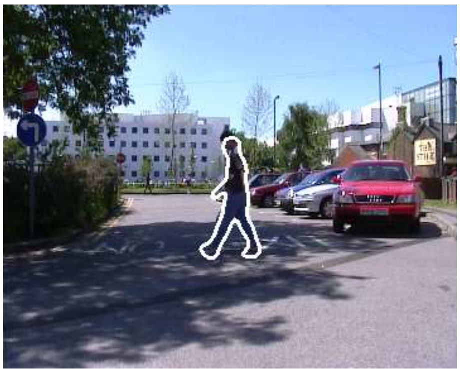
Subject 28: Last frame



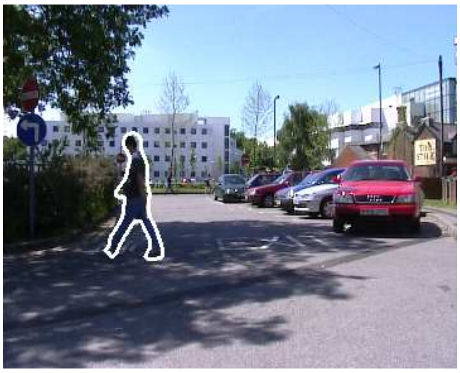
Subject 29: First frame



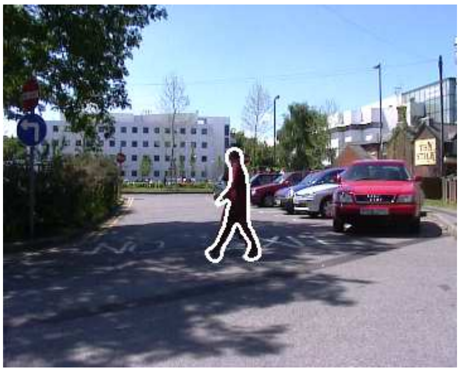
Subject 29: Last frame



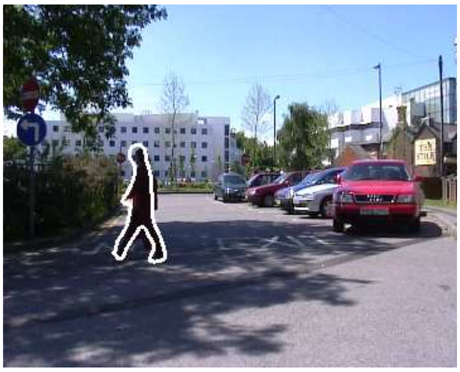
Subject 30: First frame



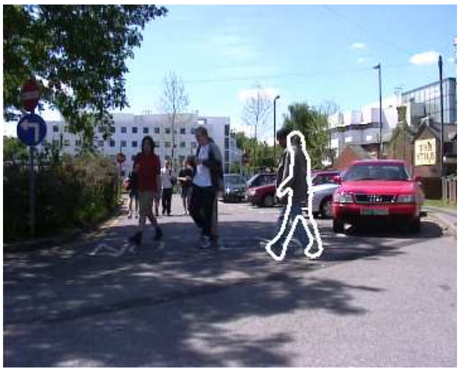
Subject 30: Last frame



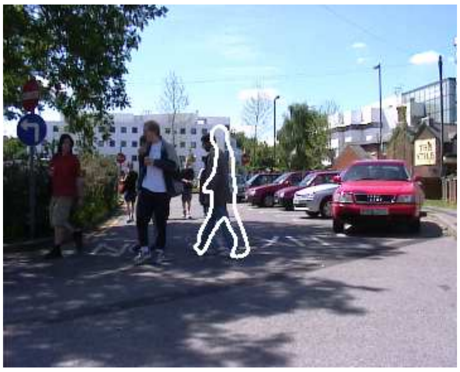
Subject 31: First frame



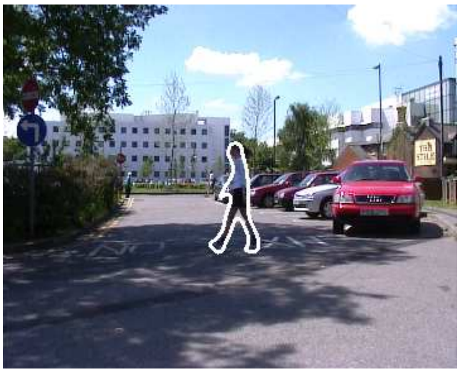
Subject 31: Last frame



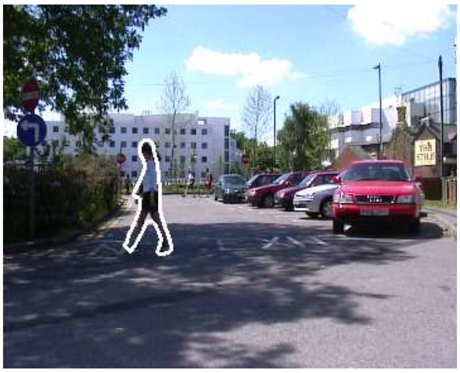
Subject 32: First frame



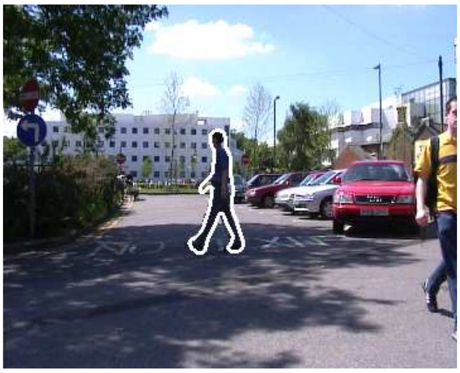
Subject 32: Last frame



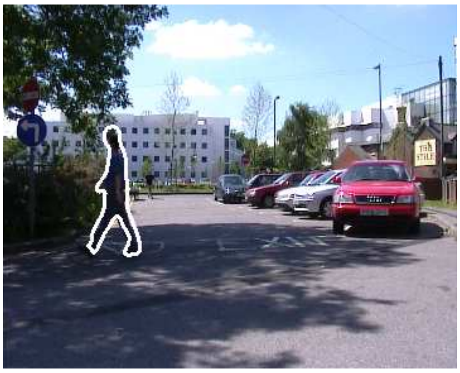
Subject 33: First frame



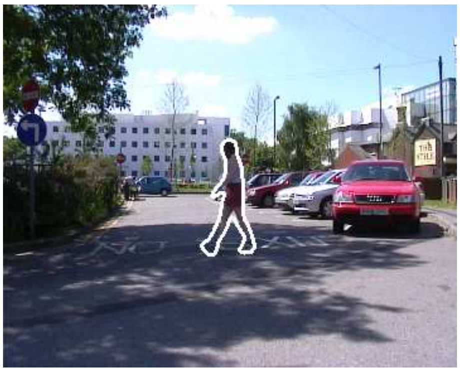
Subject 33: Last frame



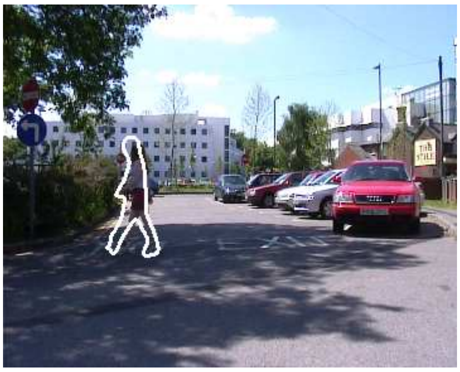
Subject 34: First frame



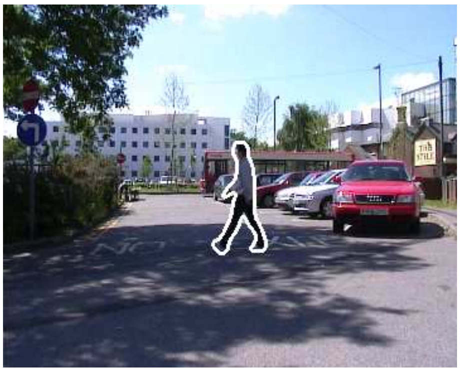
Subject 34: Last frame



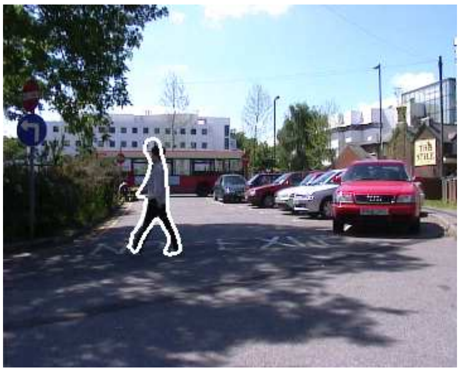
Subject 35: First frame



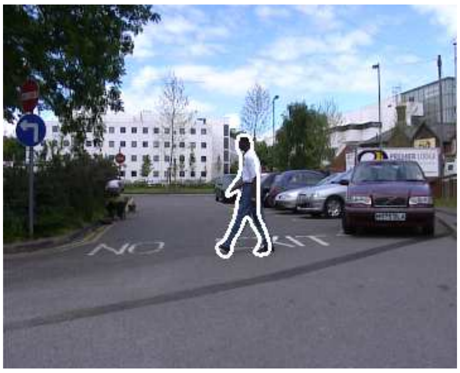
Subject 35: Last frame



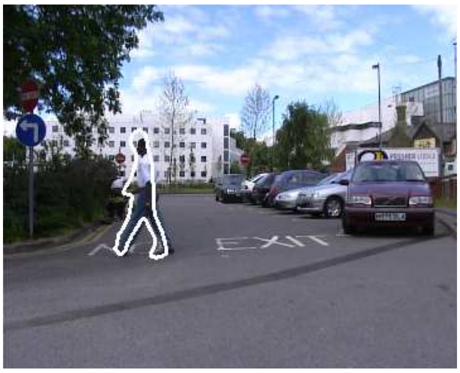
Subject 36: First frame



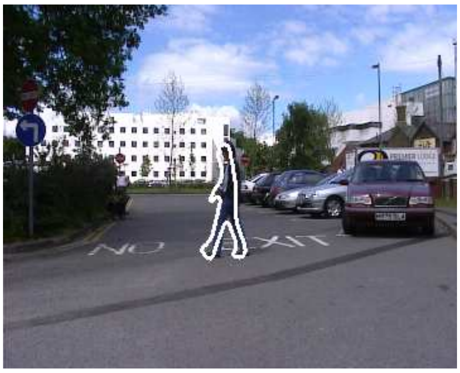
Subject 36: Last frame



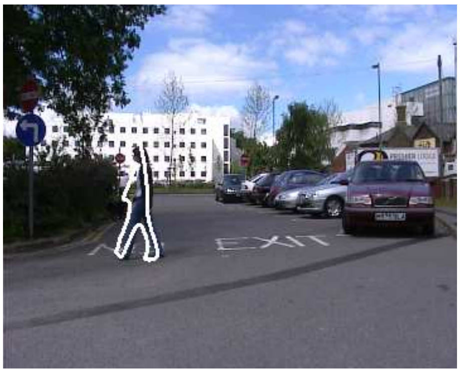
Subject 38: First frame



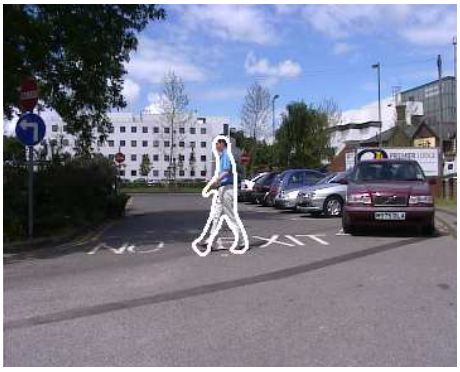
Subject 38: Last frame



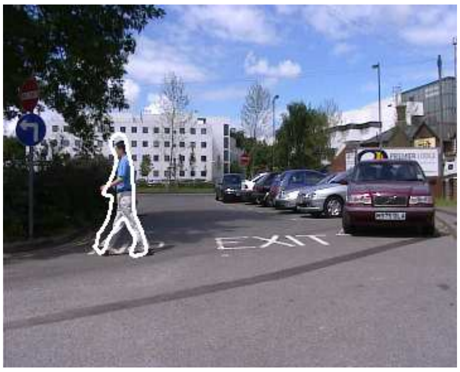
Subject 39: First frame



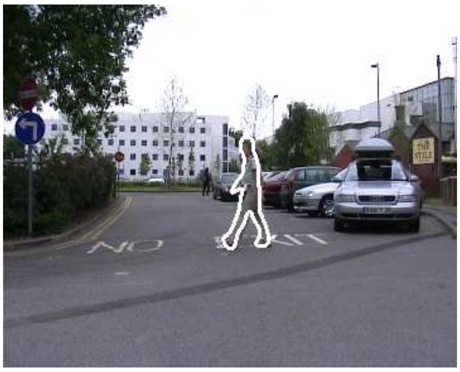
Subject 39: Last frame



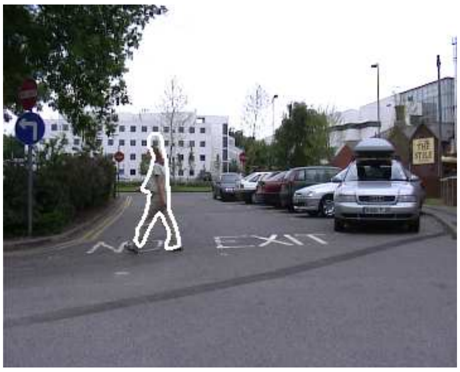
Subject 40: First frame



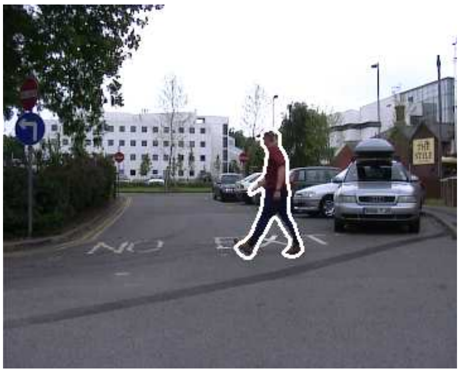
Subject 40: Last frame



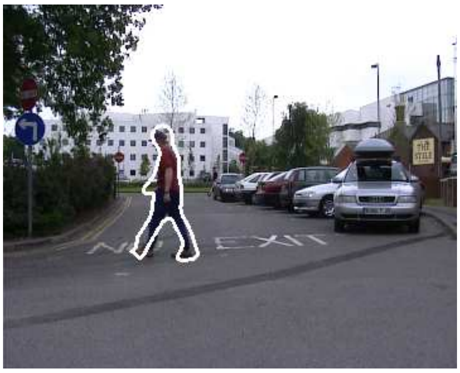
Subject 41: First frame



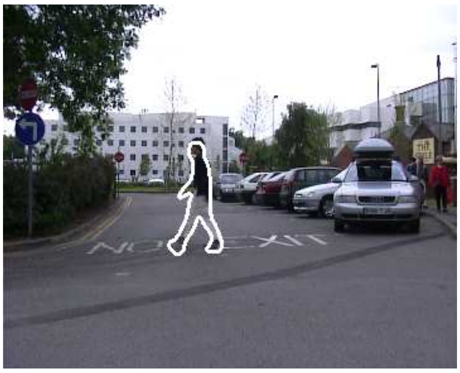
Subject 41: Last frame



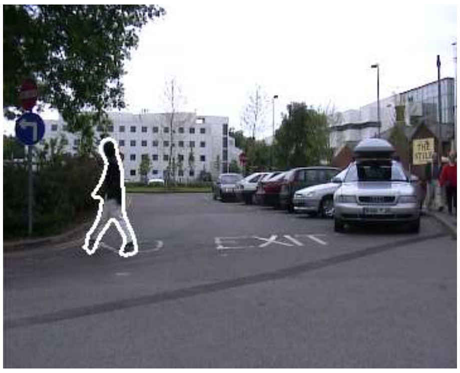
Subject 42: First frame



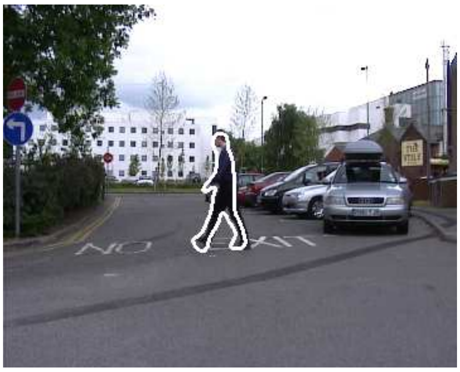
Subject 42: Last frame



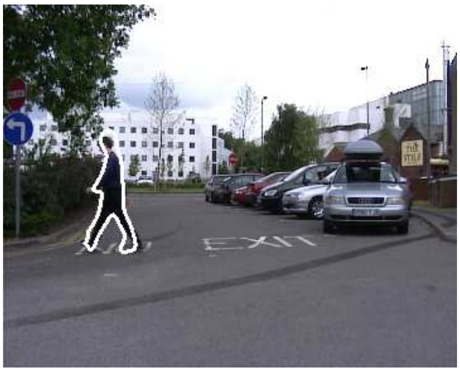
Subject 43: First frame



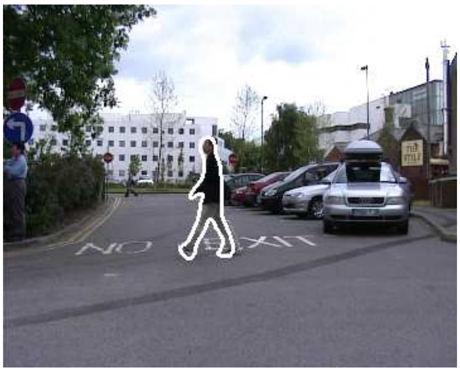
Subject 43: Last frame



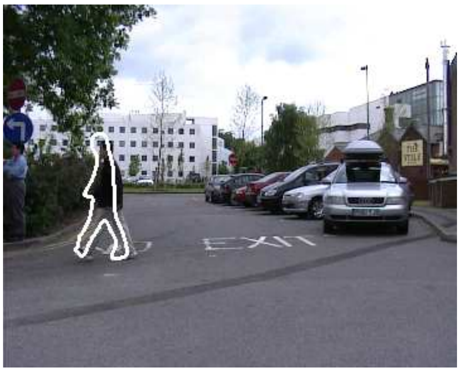
Subject 44: First frame



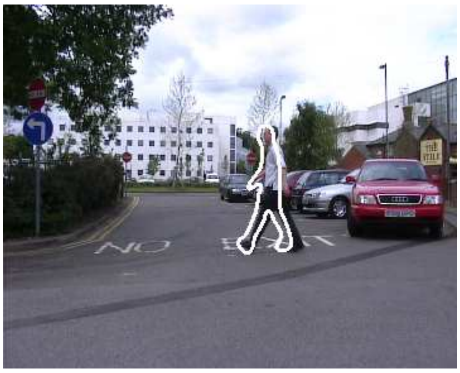
Subject 44: Last frame



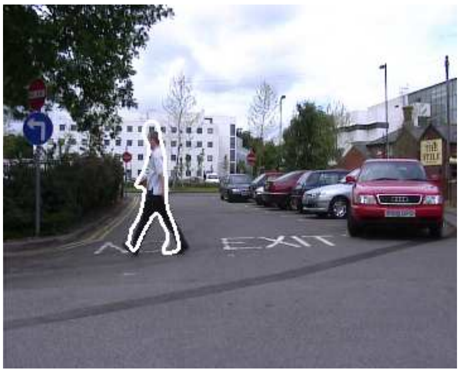
Subject 45: First frame



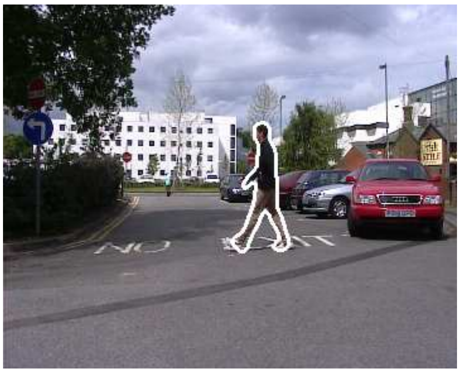
Subject 45: Last frame



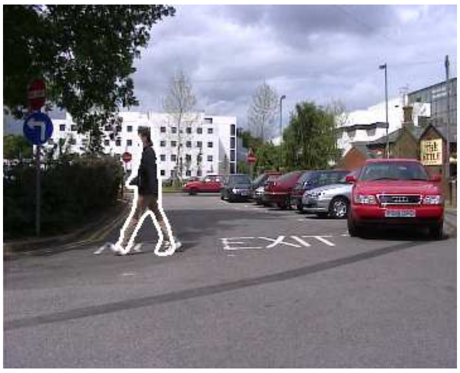
Subject 46: First frame



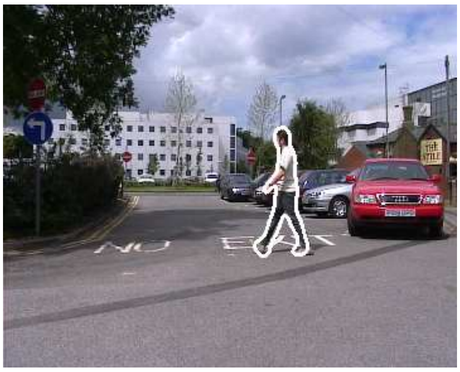
Subject 46: Last frame



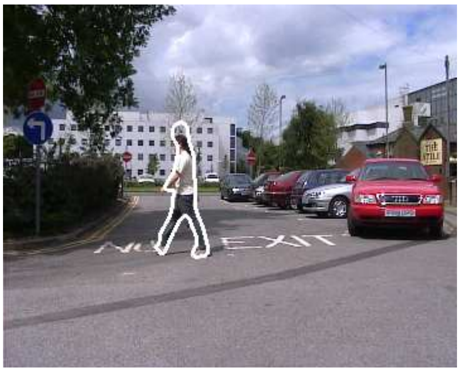
Subject 47: First frame



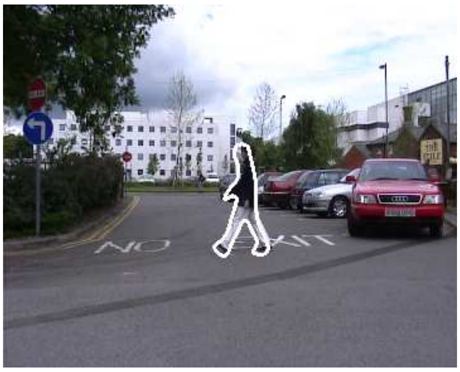
Subject 47: Last frame



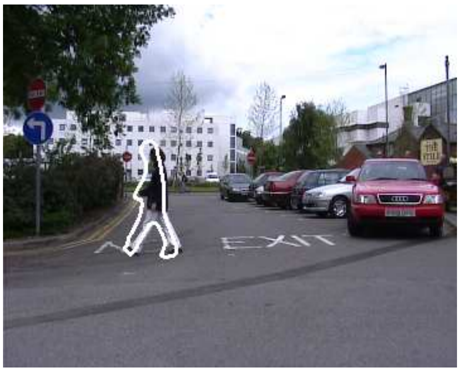
Subject 48: First frame



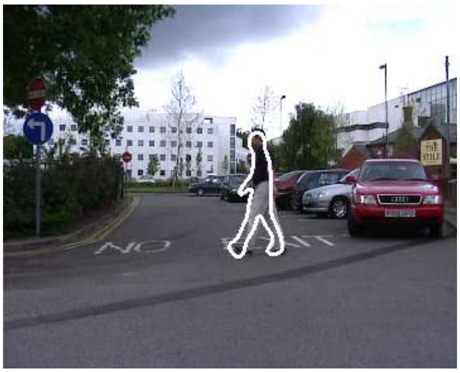
Subject 48: Last frame



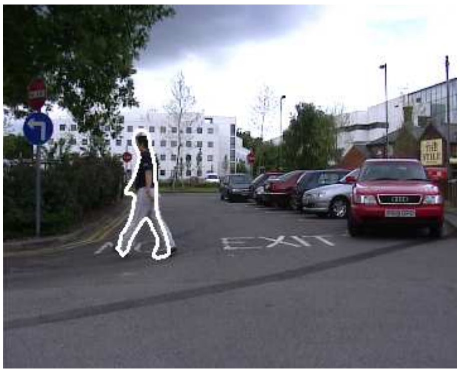
Subject 49: First frame



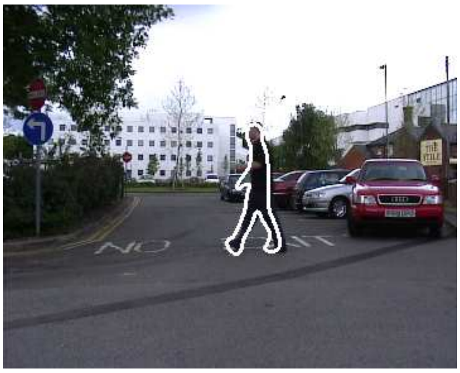
Subject 49: Last frame



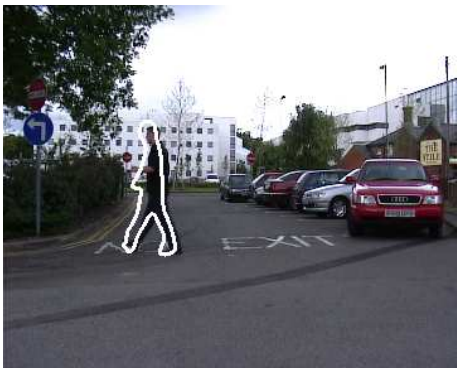
Subject 50: First frame



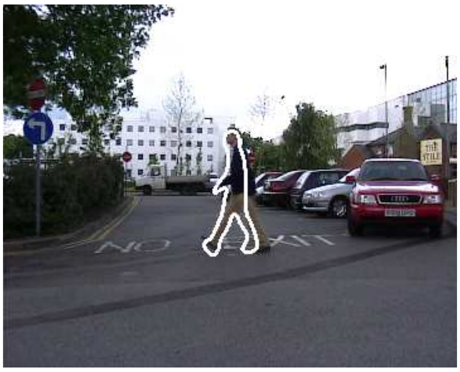
Subject 50: Last frame



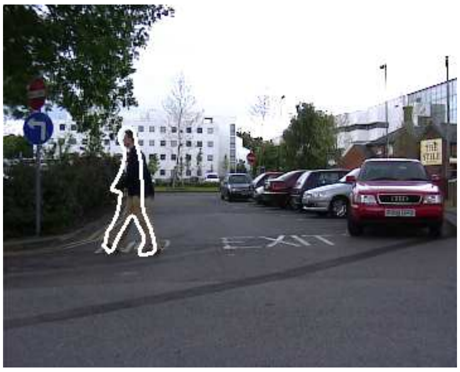
Subject 51: First frame



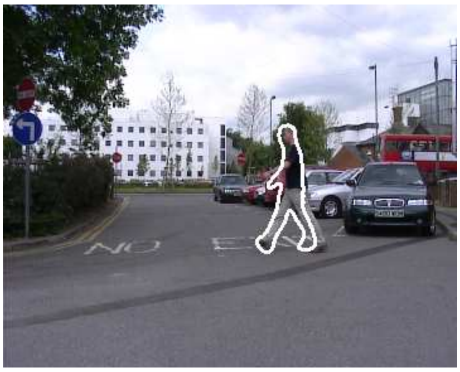
Subject 51: Last frame



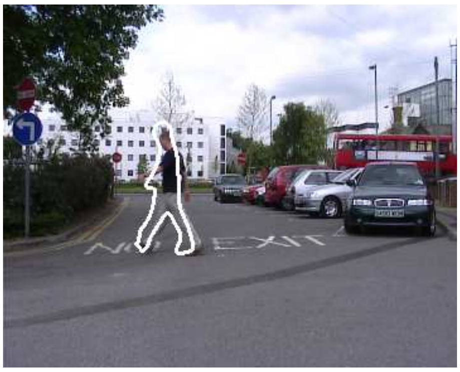
Subject 52: First frame



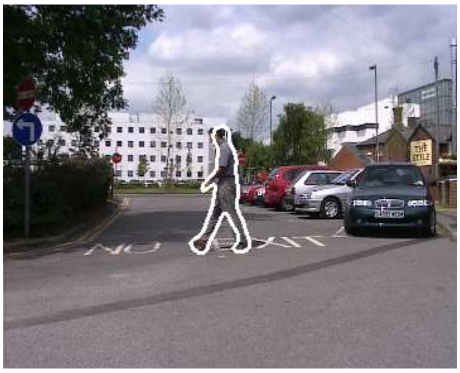
Subject 52: Last frame



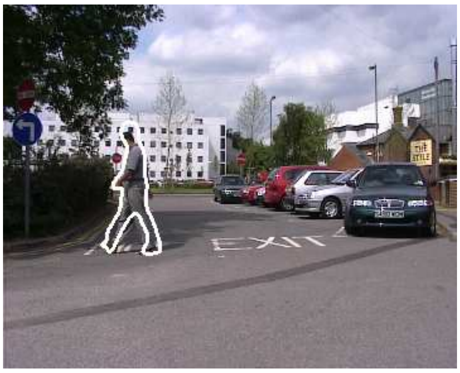
Subject 53: First frame



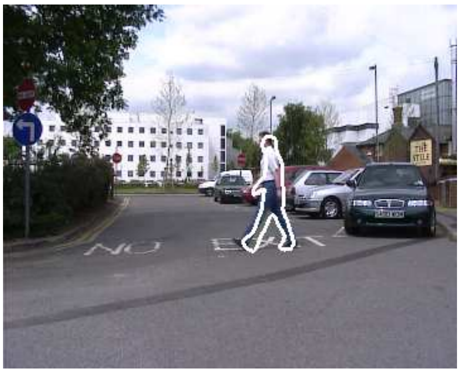
Subject 53: Last frame



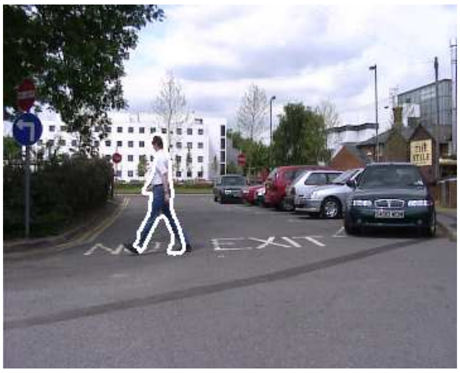
Subject 54: First frame



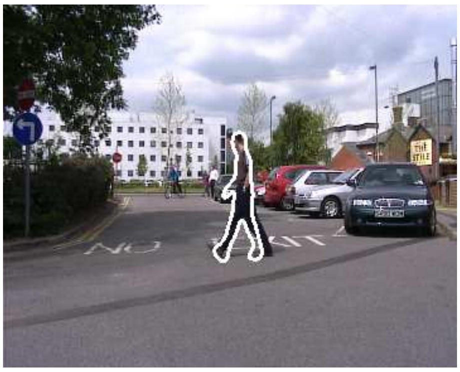
Subject 54: Last frame



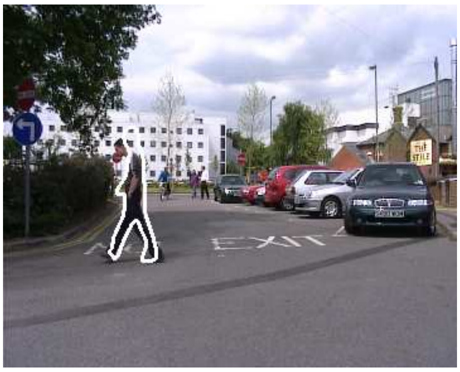
Subject 55: First frame



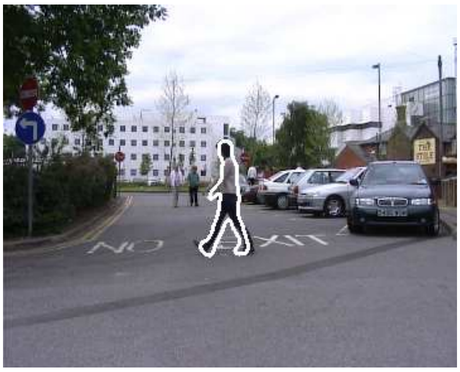
Subject 55: Last frame



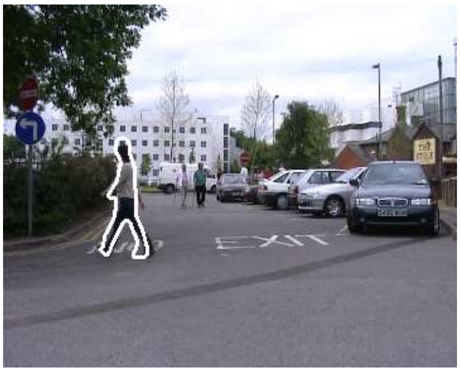
Subject 56: First frame



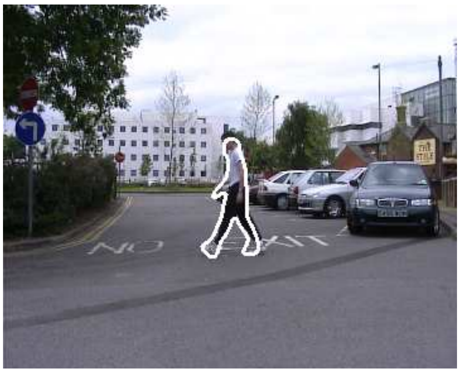
Subject 56: Last frame



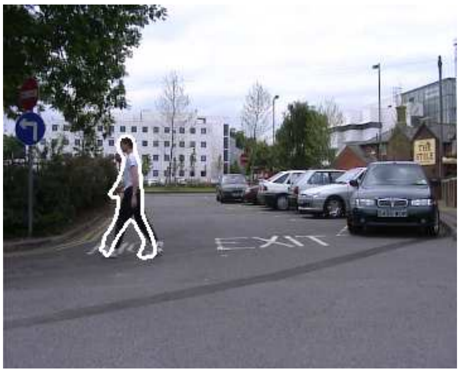
Subject 57: First frame



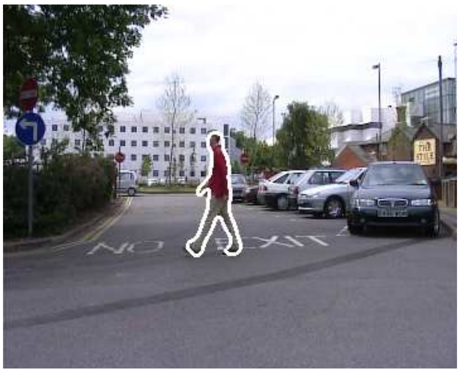
Subject 57: Last frame



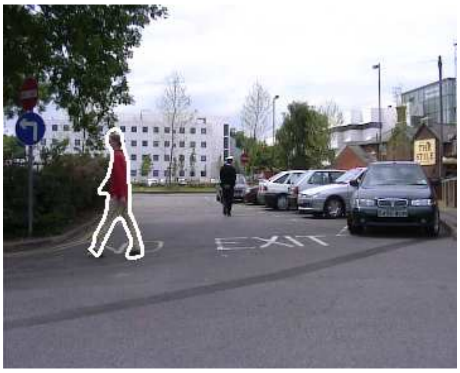
Subject 58: First frame



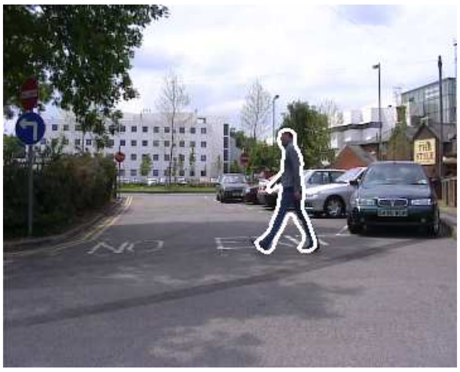
Subject 58: Last frame



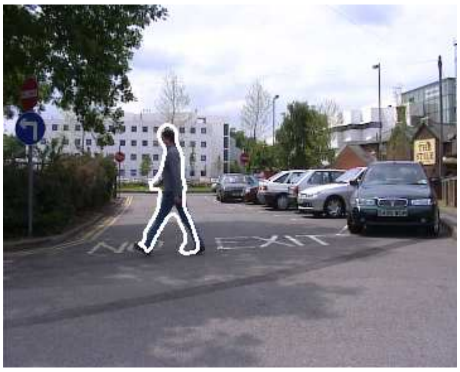
Subject 59: First frame



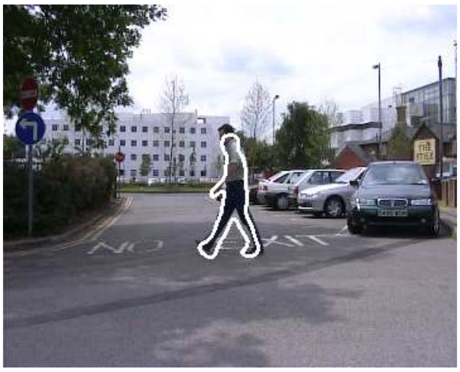
Subject 59: Last frame



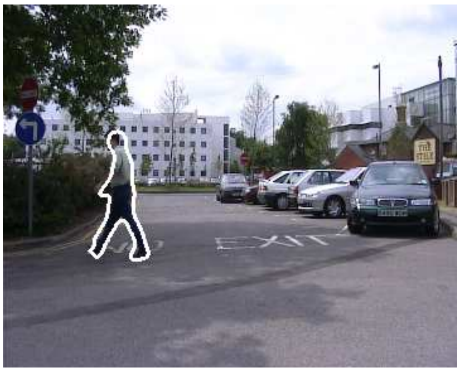
Subject 60: First frame



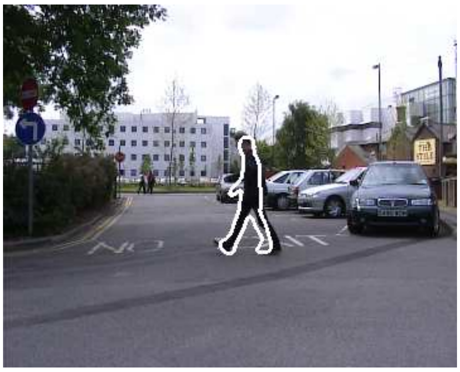
Subject 60: Last frame



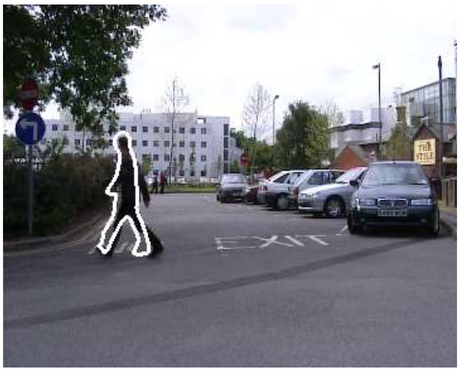
Subject 61: First frame



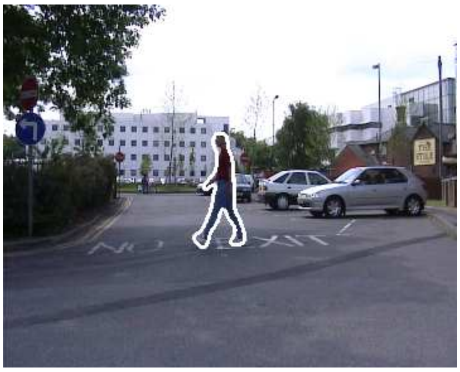
Subject 61: Last frame



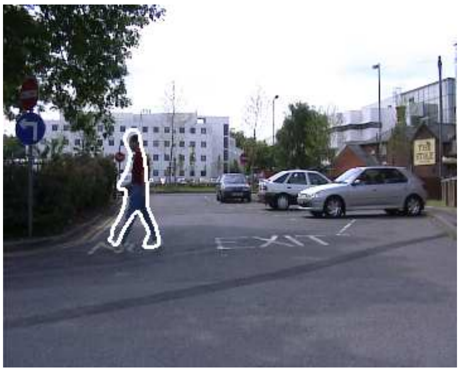
Subject 62: First frame



Subject 62: Last frame



Subject 63: First frame



Subject 63: Last frame

References

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

[2] A. S. Aguado, M. S. Nixon, and M. E. Montiel. Parameterizing Arbitrary Shapes via Fourier Descriptors for Evidence-Gathering Extraction. *Computer Vision and Image Understanding*, 69(2):202–221, 1998.

[3] C. Angeloni, P. O. Riley, and D. E. Krebs. Frequency content of whole body gait kinematic data. *IEEE Transactions on Rehabilitation Engineering*, 2(1):40–46, 1994.

[4] J. Atkinson. *The Developing Visual Brain. Series 32 Oxford Medical Publication.* Oxford University Press, 2000.

[5] D. H. Ballard. Generalizing the Hough Transform to Detect Arbitrary Shapes. *Pattern Recognition*, 13(2):111–122, 1981.

[6] S. Ben Yacoub and J. M. Jolion. Hierarchical line extraction. *IEE Proceedings on Vision, Image and Signal Processing*, 142(1):7–14, 1995.

[7] C. BenAbdelkader, R. Cutler, and L. Davis. Motion-based recognition of people in eigengait space. In *5th International Conference on Automatic Face and Gesture Recognition*, 2002.

[8] E. O. Brigham. *Fast Fourier Transform and Its Applications.* Prentice-Hall, 2 edition, 1988.

[9] D. Cunado, M. S. Nixon, and J. N. Carter. Automatic Gait Recognition via Model-Based Evidence Gathering. In *Proc. of AutoID '99: IEEE Workshop on Identification Advanced Technologies*, pages 27–30, 1999.

[10] J. Foster. Automatic Gait Recognition via Area Based Metrics. *PhD, Electronics and Computer Science, University of Southampton.*, 2003.

[11] J. P. Foster, M. S. Nixon, and A. Prugel-Bennett. New area based measures for gait recognition. In *Proceedings Audio- and Video-Based Biometric Person Authentication*, pages 312–317, 2001.

[12] H. Freeman. On the Encoding of Arbitrary Geometric Configurations. *IRE Transactions on Electronic Computers*, 10(2):260–268, 1961.

[13] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Morgan Kaufmann, 2 edition, 1990.

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

[15] G. H. Granlund. Fourier Preprocessing for Hand Print Character Recognition. *IEEE Transactions on Computers*, 21(2):195–201, 1972.

[16] M. G. Grant. Extraction of arbitrarily moving arbitrary shapes by evidence gathering. *PhD, Electronics and Computer Science, University of Southampton.*, 2002.

[17] J. B. Hayfron-Acquah. Automatic Gait Recognition by Symmetry Analysis. *PhD, Electronics and Computer Science, University of Southampton.*, 2002.

[18] J. B. Hayfron-Acquah, M. S. Nixon, and J. N. Carter. Automatic Gait Recognition by Symmetry Analysis. In *Proc. of Audio-and-Video-Based Biometric Person Authentication*, pages 272–277, 2001.

[19] P. V. C. Hough. *Method and Means for Recognizing Complex Patterns*. US Patent 3069654, 1962.

[20] P. S. Huang, C. J. Harris, and M. S. Nixon. Recognising humans by gait via parametric canonical space. *Journal of Artificial Intelligence in Engineering*, 13(4):359–366, 1999.

[21] J Illingworth and J Kittler. An adaptive hough transform. *IEEE Trans Pattern Analysis and Machine Intelligence*, 9:690–697, 1987.

[22] J. Illingworth J. Princen and J. Kittler. A hiererchical approach to line extraction based on the hough transform. *Computer Vision and Image Understanding*, 52:57–77, 1990.

[23] A. Johnson and A. Bobick. A multi-view method for gait recognition using static body parameters. In *3rd International Conference on Audio- and Video Based Biometric Person Authentication*, pages 301–311, June 2001.

[24] A. Kale, A. N. Rajagopalan, N. Cuntoor, and V. Kruger. Gait based recognition of humans using continuous HMMs. In *Proc. of the International Conference on Face and Gesture Recognition 2002*, 2002.

[25] F. P. Kuhl and C. R. Giardina. Elliptic Fourier Features of a Closed Contour. *Computer Graphics and Image Processing*, 18:236–258, 1982.

[26] V. F. Leavers. Which hough transform? *CVGIP: Image Understanding*, 58(2):250–264, 1993.

[27] Y. Liu, R. Collins, and Y. Tsin. Gait sequence analysis using frieze patterns. Technical report, Robotics Institute, Carnegie Mellon University, December 2001.

[28] R. G. Lyons. *Understanding Digital Signal Processing.* Addison Wesley, 1997.

[29] M. E. Montiel, A. S. Aguado, and E. Zaluska. Fourier Series Expansion of Irregular Curves. *Fractals*, 5(1):105–199, 1997.

[30] M. P. Murray, S. B. Sepic, G. M. Gardner, and W. J. Downs. Walking Patterns of Men with Parkinsonism. *American Journal of Physical Medicine*, 57:278–294, 1978.

[31] J. M. Nash, J. N. Carter, and M. S. Nixon. Extracting moving articulated objects by evidence gathering. In *British Machine Vision Conference, BMVC98*, volume 2, pages 609–618, 1998.

[32] M. S. Nixon and A. Aguado. *Feature Extraction & Image Processing.* Newnes, 2002.

[33] Mark S Nixon and John N Carter. Automatic recognition by gait. *Proceedings of the IEEE*, 94(11):2013–2024, November 2006.

[34] Mark S. Nixon, Tieniu N. Tan, and Rama Chellappa. *Human Identification based on Gait.* International Series on Biometrics. Springer, 2005.

[35] E. Persoon and K. Fu. Shape Discrimination Using Fourier Descriptors. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-7(3):170–179, 1977.

[36] S. Pheasant. *Bodyspace: anthropometry, ergonomics, and design.* Taylor & Francis, 1986.

[37] J. Shutler. Velocity Moments for Holistic Shape Description of Temporal Features. *PhD, Electronics and Computer Science, University of Southampton.*, 2002.

[38] 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 (RASC02)*, pages pp. 66–71, 2002.

[39] J. D. Shutler and M. S. Nixon. Zernike Velocity Moments for Description and Recognition of Moving Shapes. In *Proc. of BMVC 2001*, 2001.

[40] M. W. Whittle. *Gait Analysis.* Butterworth-Heinemann, 2003.

[41] C. Y. Yam. Model-based Approaches for Recognising People By The Way They Walk or Run. *PhD, Electronics and Computer Science, University of Southampton.*, 2002.

[42] C. Y. Yam, M. S. Nixon, and J. N. Carter. Extended Model-Based Automatic Gait Recognition of Walking and Running. In *Proc. of 3rd Int. Conf. on Audio and Video-Based Biometric Person Authentication, AVBPA 2001*, 2001.