ABSTRACT

Title of Dissertation: ALGORITHMS FOR
GAIT-BASED HUMAN IDENTIFICATION
FROM A MONOCULAR VIDEO SEQUENCE

Amit Kale, Doctor of Philosophy, 2003

Dissertation directed by: Professor Rama Chellappa
Department of Electrical and Computer Engineering

Human gait is a spatio-temporal phenomenon that characterizes the motion characteristics of an individual. It is possible to detect and measure gait even in low-resolution video. This makes it an attractive modality in surveillance applications, where it is often difficult to get face or iris information at high enough resolution for recognition applications. Psychophysical studies indicate that humans have the capability for recognizing people from even impoverished displays of gait, indicating the presence of identity information. From early medical studies it appears that there are twenty four different components to human gait, and that if all the measurements are considered, gait is unique. It is interesting, therefore, to study the utility of gait as a biometric. The goal of this thesis is to investigate the information contained in the video sequences of human gait and how to extract and exploit that information in ways that facilitate human identification.

In our work, we present both deterministic and stochastic approaches for gait recognition. Human identification using gait, similar to text-based speaker identification, involves different individuals performing the same task and a template-matching approach is suitable for such problems. In situations where the amount of training data is limited, we show the utility of a simple feature viz. the width of the outer contour of the binarized silhouette of the subject and its derivatives for gait recognition in a dynamic time warping framework. By virtue of their deterministic nature, template matching methods have limited noise resilience. A careful analysis of gait would reveal that it has two important components. The first is a structural component that captures the physical build of a person while the second is the motion kinematics of the body during a gait cycle. We propose a systematic approach to gait recognition by building representations for the structural and dynamic components of gait using exemplars and hidden Markov models (HMMs). The stochastic nature of the HMM yields better noise resilience than the template matching technique. To recognize a person walking at a large distance, humans try to combine information such as posture, arm/leg swing, hip/upper body sway or some unique movements that are characteristic of that person. We demonstrate the same effect through fusion of different dynamic and static gait features in both determinisitic and stochastic frameworks. Most gait recognition algorithms rely on the availability of an exact side view in the probe. However, it is not realistic to expect that this assumption will be valid in most real-life scenarios. We present a view invariant gait recognition algorithm which is based on synthesizing a side view of a person from an arbitrary monocular view. The method is based on the planar approximation of a person that is valid when human identification at a distance is desired.

ALGORITHMS FOR
GAIT-BASED HUMAN IDENTIFICATION
FROM A MONOCULAR VIDEO SEQUENCE

by

Amit Kale

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2003

Advisory Committee:

Professor Rama Chellappa, Chairman/Advisor
Professor Larry Davis, Deans Representative,
Professor David Jacobs,
Professor William S. Levine,
Professor Min Wu

© Copyright by

Amit Kale

2003

DEDICATION

To my parents

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to my advisor, Dr. Rama Chellappa. He introduced me to this problem that has absorbed most my professional life over the past few years, guided me through the process that reached fruition in this thesis, and held out a beacon of hope when the path seemed like a long dark tunnel. Moreover, through his daily interactions he has given me something to live up to by his ever-friendly attitude and his emphasis on honesty and integrity.

I am also grateful to my committee members, including Dr. Larry Davis, Dr. William Levine, Dr. David Jacobs and Dr. Min Wu. I would like to thank them for the time they spent discussing different technical issues that arose in the course of this work and for reading and commenting on the thesis. I am very grateful to Dr Navdeep Singh, who instilled in me the quest for knowledge and the courage to probe the unknown. I would like to take this opportunity to thank many of my colleagues for their help and advice, without which this thesis would not have seen the light of day. It is impossible to acknowledge all of them individually, but I would like to particularly mention Dr Amit RoyChowdhury, Dr. A. N. Rajagopalan, Dr. B. Yegnanarayana, Dr. Haiying Liu, Dr Gang Qian, Mr Naresh Cuntoor

and Mr Aravind Sundaresan. I would like to thank Mr Ashok Veeraraghavan and Mr Narayanan Ramanathan for their help in running experiments. I would also like to thank Mr Shaohua Zhou for providing data from his face recognition work. I have very much enjoyed my stay at Maryland for the last five years and this is in so small part due to the excellent friends I have had: Chirag Kathrani, Arunesh Mishra, Vivek Gautam, Sandeep Nayak, Nikhil Vichare, Himaanshu Gupta, Namrata Vaswani, Danitza Radichevich, Jan Neumann, Kaushik Chakraborty, Savinder Dhaliwal, Saurabh Dadu, Sanjay Gayakwad to name but a few.

Words cannot express my gratitude and indebtedness to my parents, who have given up so much in life in order that I could reach this stage today. Thinking about the sacrifices they made humbles me. We usually take their unconditional love for granted, but I would like to take this opportunity to say "Thank you". I would also like to thank my younger sister for her love and understanding. Last but not the least I would like to thank Barbara Slone, for being a constant source of encouragement and inspiration.

TABLE OF CONTENTS

List of Tables viii

List of Figures x

1 Introduction 1

1.1 Prior Work . . . 2

1.1.1 Studies in Psychophysics . . . 2

1.1.2 Medical Studies . . . 5

1.1.3 Computational approaches . . . 6

1.2 Thesis Contributions . . . 7

1.3 Thesis Organization . . . 10

2 Gait-Based Human Identification Using Appearance Matching 11

2.1 Introduction . . . 11

2.2 The Width Vector . . . 14

2.3 Gait Representation . . . 19

2.3.1 Direct Features . . . 20

2.3.2 Eigenfeatures . . . 22

2.4 Matching Gait Sequences using DTW . . . 23

2.4.1 Implementation details . . . 28

2.5 Data . . . 29

2.6 Experimental Results . . . 31

2.6.1 Results for the UMD Database . . . 32

2.6.2 Results for the CMU database . . . 35

2.6.3 Results for the MIT Database . . . 38

2.6.4 Results for the UMD3 database . . . 39

2.6.5 Results for the USF Database . . . 42

2.7 Conclusions . . . 42

3 A Framework for gait-based person identification using continuous HMMs 45

3.1 Feature Selection . . . 46

3.2 Proposed Algorithms 47
3.2.1 Overview 47
3.2.2 Approach 1: Indirect Approach 50
3.2.3 Approach 2: Direct Approach 56
3.3 Experimental Results 60
3.3.1 The CMU Database 61
3.3.2 The UMD Database 64
3.3.3 The USF Database 65
3.4 Conclusions 66

4 Fusion of multiple evidences for gait recognition 69
4.1 Methodology 70
4.1.1 Holistic fusion 71
4.1.2 Parts-based fusion 73
4.1.3 Multicamera fusion 76
4.2 Experiments 78
4.3 Conclusions 81

5 View Invariant Gait Recognition 83
5.1 Introduction 83
5.2 Framework for view-invariant gait recognition 86
5.2.1 Tracking 86
5.2.2 Estimation of 3-D Azimuth Angle 87
5.2.3 Statistical Error Analysis of Azimuth Estimation 88
5.2.4 View Synthesis 90
5.2.5 Alternative Approach for View Synthesis 91
5.2.6 Justification for generating the synthetic view 92
5.3 Matching Gait Sequences using DTW 93
5.4 Experimental Results 94
5.4.1 Gait-based Recognition 94
5.4.2 Video-based rendering of planar dynamic scenes 104
5.4.3 Fusion of Face and Gait 108
5.5 Conclusion 112

6 Conclusions and Future Work 115
6.1 Applications and Future Directions 118

Appendix 120

A Appendix **120**

Bibliography **124**

LIST OF TABLES

2.1 UMD database: Cumulative match scores for the first five ranks for the case of 2 full-cycles having a relative phase shift of 1 half-cycle. 32

2.2 UMD database: Cumulative match scores for the first five ranks for different directly-derived feature vectors. 33

2.3 Cumulative match scores for the UMD database using different eigen-features. 34

2.4 Effect of length of gait sequence for the UMD database: Cumulative match scores for the first five ranks. 34

2.5 UMD database: Cumulative match scores for the velocity profile. 35

2.6 CMU database: Recognition results for different experiments. . . 36

2.7 Cumulative match scoresfor the ball experiment in the CMU database using the eigen-smoothed feature. 38

2.8 UMD3 database: Cumulative match scores for the first three ranks for different azimuth angles using the direct and the eigen-feature vectors. 41

2.9 Probe Sets and match scores for the USF database using the baseline algorithm and our approach using width feature and entire binary silhouette. 43

3.1 Probe Sets and match scores for the USF database using the baseline algorithm and our indirect and direct approaches. 66

3.2 Recognition Scores for Baseline and HMM (122 people). 67

4.1 CMS scores: Combining left and right projection vectors. UMD Dataset. 79

4.2 Probe Sets and match scores for the USF database using the baseline algorithm and our approach using width feature and entire binary silhouette. 79

4.3 UMD database: CMS for the first five ranks for the case of 2 full-cycles having a relative phase shift of 1 half-cycle. 80

4.4 Cumulative match scores at rank 1 and rank 5 for CMU dataset: Combining leg dynamics and height using Sum rule 80

4.5 Cumulative match scores at rank 1 and rank 5 for UMD dataset: Combining leg dynamics and height using Sum rule. 80
4.6 Cumulative match scores at rank 1 and rank 3 for MIT dataset: Combining leg dynamics and height by adding the similarity scores. 81
4.7 Cumulative match scores at rank 1 and rank 5 for the UMD dataset: effect of frontal and side gait fusion 81

LIST OF FIGURES

2.1 Silhouettes for a full gait-cycle corresponding to (a) Person 1 (b) Person 2 (c) Person 3 (d) Person 4. 16

2.2 Width overlay plots for (a) Person 1 (b) Person 2 (c) Person 3 (d) Person 4. 17

2.3 Temporal plot of the width vector for several gait cycles of (a) Person 1 (b) Person 2 (c) Person 3 (d) Person 4. 18

2.4 Width plots for an arbitrary frame. (a) Unsmoothed raw feature. (b) 3-point smoothed feature. (c) 5-point smoothed feature. (d) 11-point smoothed feature. (e) 21-point smoothed feature. (f) Reconstructed width vector using the first two eigenvectors. . . . 21

2.5 A typical dynamic time-warping path. 24

2.6 Sum of widths as a function of time. 28

2.7 Sample images of the same subject corresponding to (a) slow-walk and (b) fast-walk (Notice the change in posture and body dynamics) (c) Warping path for person with largest reference sequence-length to probe sequence-length ratio (d) Warping path for the person in (a) and (b). 37

2.8 Width profile of an individual for different azimuth angles (a) 0 deg (b) 15 deg (c) 30 deg (d) 45 deg and (e) 60 deg. 40

3.1 Stances corresponding to the gait cycle of two individuals. 48

3.2 Rate-distortion curve for number of exemplars vs distortion. . . . 52

3.3 FED vector components plotted as a function of time. 53

3.4 Exemplars estimated using IP distance measure 59

3.5 Cumulative match characteristic for normal walk and walk when carrying an object for the CMU database. 62

3.6 Cumulative match characteristic for across speed testing for the CMU database. 63

3.7 Results for the UMD database using Algorithm 1 (44 people). . . 64

3.8 Cumulative match characteristic for USF database using our approach. 67

4.1 Illustrating the generation of (a) left projection vector, (b) right projection vector and (b) width vector. 71
4.2 Effect of eigen decomposition and reconstruction on the width vectors. (a) Overlapped raw width vectors (b) Smoothed width vectors. Notice that the leg region (the bottom half of the figures) contain a significant portion of the dynamics. 74
4.3 Outer contour plot for a frontal gait sequence in the UMD datasbase. (a) Subject walking towards the camera (b) 4 half-cycles chosen for recognition (c) The positions of head and foot in each frame is located to compute the normalizing factor. Two straight lines are then fit to the trajectories of the head and foot. (d) The 4 half-cycles normalized. 77

5.1 Framework for View Invariant Gait Recognition 85
5.2 Imaging Geometry 86
5.3 Tracked position of the head for 2 walk cycles 89
5.4 (a) represents different stances of a person walking parallel to the camera; (b) (d) and (f) represent different stances of a person walking at angles 15, 30 and 45 degrees to the camera; (c) (e) and (g) represent side-views synthesized from original videos where the person walks at angles of 15, 30 and 45 degrees to the camera. 96
5.5 Width profile as a function of time for (a) Canonical View ($\theta = 0$); (b)Unnormalized sequence for $\theta = 45$; synthesized views for: (c) $\theta = 30$ (d) $\theta = 45$. 97
5.6 Quality of reconstruction as a function of the angle of walk. . . . 98
5.7 Cumulative Match Characteristics for Original and Synthesized images for (a) $\theta = 15$ (b) $\theta = 30$ and (c) $\theta = 45$ for UMD3 database. 100
5.8 ROC curves for $\theta = 15, 30$ and 45 degrees for the case of leg and height fusion. 101
5.9 Distributions of $Tan(\alpha)$ and its transformation to θ at one time instant 101
5.10 Σ shaped walking pattern in the NIST database. 102
5.11 Examples for the NIST database (a) Gallery images of person walking parallel to the camera (b) Unnormalized images of person walking at 33^0 to the camera(c) Synthesized images for (b). 102
5.12 Gait recognition performance on the NIST database. 103
5.13 Computation of Homography for the CMU database. 104

5.14 Cumulative Match Characteristics for Original and Synthesized images for (a) Full body (b) Leg only for the CMU database. 105
5.15 Rendering of Non-planar objects 107
5.16 Width Vectors projected on to the principal eigen vector for azimuth angle $\theta = 0$ 108
5.17 The first column in (a) shows a frame from the canonical view; the second and third columns respectively show an image synthesized from a similar frame at an azimuth $\theta = 15$ before and after the correction for non-planarity is applied. (b) and (c) represent the same cases for azimuths $\theta = 30$ and 45 respectively. 109
5.18 Similarity matrices for (a) Gait Recognition ; (b)Face Recognition; CMC characteristics for: (c) Gait (d) Face 113
5.19 Histogram of the true matches (yellow) and false matches (black) 114
5.20 CMC curves for (a) Hierarchical and (b) Holistic Fusion 114

Chapter 1

Introduction

Automated person identification is an important component of surveillance. An effective approach to person identification is to reduce it to the problem of identifying physical characteristics of the person. This method of identification of persons based on his/her physiological/behavioral characteristics is called biometrics. The primary advantage of biometric identification over the methods of identification utilizing "something you possess" or "something that you know" approach is that biometrics cannot be misplaced or forgotten; it represents a tangible component of "something that you are". Established biometric methods range from fingerprint and hand-geometry techniques to more sophisticated methods based on face recognition and iris identification. Unfortunately, no single biometric is perfect or complete. Fingerprints and hand-geometry are reliable but require physical contact. Although, signatures based on face and iris are non-intrusive in nature, the applicability of all these methodologies is restricted to very controlled environments. In fact, current technology is capable of recognizing mostly frontal faces. At the time of writing, iris recognition is being attempted at distances of not more than five meters.

When person identification is attempted in natural settings such as those arising in surveillance applications, it takes on a new dimension. Biometrics such as fingerprint or iris are no longer applicable. Furthermore, night vision capability (an important component in surveillance) is usually not possible with these biometrics. Even though an IR camera would reveal the presence of people, the facial features are far from discernible in an IR image at large distances. A biometric that can address some of these shortcomings is human 'gait' or the walking style of an individual. The attractiveness of gait as a biometric arises from the fact that it is non-intrusive and can be detected and measured even in low resolution video. Furthermore, it is harder to disguise than static appearance features such as face and it does not require a cooperating subject.

1.1 Prior Work

Early research on gait can be primarily grouped into two categories. The first category involves psychophysical studies of gait viz. studying the ability of human observers to recognize gait. The second category includes biomechanical studies viz. studying kinetics and kinematics associated with gait. These studies in part led to the development of computational approaches to gait-based human identification. In this section we describe some of the gait research done in the areas of psychophysics, biomechanics and computer vision.

1.1.1 Studies in Psychophysics

The belief that humans can distinguish between gait patterns of different individuals is widely held. Intuitively, it is possible to think of the qualities of walk that help a perceiver identify an approaching figure even before the face becomes discernible. These gait related quantities include stride length, bounce, rythm, speed and perhaps even attributes such as swagger or body swing. The suggestion that humans can identify people by their gait was investigated in a series of early studies by Johansson [2]. He presented participants with images that had been reduced to point-light displays. Small light bulbs were attached to the body joints of a darkly dressed walker and the brightness of the display was reduced so that only these light points were visible. In this way only gait related cues were available and thus the perception of pure biological motion could be examined. When these point-light displays were static, the random collection of dots were variously interpreted as star constellations. However, as soon as the figures moved, the points of light were immediately perceived to be a human in motion. This suggests that we have some implicit notion of human movement, and can recognize temporal data within this context. Later work using point-light displays went further, demonstrating that not only could a walking figure rapidly be extracted from the moving lights, but also a perceiver could distinguish between different sorts of biological motion including walking, climbing stairs, jumping etc [3]. Additionally observers were also shown to be able to recognize facial expressions through reflecting patches placed on certain parts of the face [4]. These studies suggests that humans are adept at biological human perception.

Attempts to address the question of identification from gait have proceeded in small steps. Kozlowski and Cutting [5] first investigated whether perceivers could identify the gender of a point-light walker. Their results indicated an accuracy rate of 65% and 70% when the walker was viewed from the side. A minimum exposure time of two seconds was required for gender identification. In [6], it was suggested that gender may be identified indirectly through a determination of the “center of moment” of a walker.

The demonstration that gender could be extracted from gait provided insight into how perceivers might discriminate between gait patterns of different individuals. The prospect for perceivers being able to identify individuals from their gaits was thus encouraging. Cutting and Kozlowski [7] demonstrated that perceivers could reliably recognize themselves and their friends from dynamic point-light displays. Barclay et al. [8] suggested that individual walking styles might be captured by differences in a basic series of pendular limb motions. Interestingly Beardsworth and Buckner [9] have shown that the ability to recognize onself from a point-light display is greater than the ability to recognize one's friends, despite the fact that we rarely see our own gait from a third-person perspective. This suggests that, as well as the apparent sensitivity of the human visual system to biological motion, it is likely that there is some transfer of information from the kinaestetic modality to the visual modality.

Stevenage et al. [10] also explored the ability of people to identify people using gait information alone. In one experiment, people were given video footage of six walking subjects to study. After the studying phase was over, one group viewed the same walkers under simulated daylight in which the silhouette and motion of the walkers was clearly visible. A second group viewed the subjects under simulated dusk in which the outline and motion were difficult to see. A third group was shown only the point light displays of the walkers. It was found that all the three groups were able to learn the gait of the six walkers and label their identity regardless of the nature of data. In a second experiment, a different set of viewers were given only two seconds to view a target, and they had to pick the target from a parade of walkers under similar viewing conditions as the first experiment. It was found that even with such a brief exposure time and their unfamiliarity with the walking subjects, the perceivers could identify the target correctly at greater than chance rate.

1.1.2 Medical Studies

Although walking is one of the most universal of all human activities, it was only in the early part of the twentieth century that the detailed components of this act started being systematically examined. The studies from the Grecian era until the mid-nineteenth centuray were, for the most part, observational. The 1950s saw many diversified kinematic gait studies, most prominently at the University of California. These studies were spurred by a need for an improved understanding of locomotion for the treatment of World War II veterans. One of the goals of gait researchers was to build a model for normal gait, deviations from which could be used to study gait abnormalities. Such studies involved collecting the gait of a large number of subjects with no gait abnormalities. The presence of identity information in gait was a by-product of these studies. In [11], subjects with reflective targets attached to specific anatomical landmarks walk before a camera in the illumination of a strobe light flashing twenty times per second. A mirror was mounted over the walking area so that the target images projected in the overhead view are captured as well. Twenty different components of gait which include motions of the hip, knee, ankle etc. were studied and it was found that when considered together, the gait of the subjects was unique. Certain components such as pelvic and thoracic rotation had more pronounced inter-person difference. The ability of recognizing and or modifying specific movement patterns is of particular interest for successful and effective therapeutic interventions. A recent study by Schollhorn et al. [12] studied the gait of fifteen subjects to study the presence of identity information in gait. It was found in this study that kinetic variables (captured using a force platform) as well as kinematic variables (captured by reflective markers on the thigh, shank and hip) were both necessary for gait identification. Furthermore simply using the leg portion of the body was adequate for getting good identification performance.

1.1.3 Computational approaches

Medical and psychophysical studies, as discussed in Sections 1.1.1 and 1.1.2 indicate that there exists identity information in gait. It is therefore interesting to study the utility of gait as a biometric. In recent years, there has been an increase in research related to gait-based human recognition. We attempt to give a summary of some of examples below, but the listing is by no means complete.

As noted before, joint angles may be sufficient for recognizing people by their gait. However reliably recovering joint angles from a monocular video is a hard problem. Approaches to gait recognition problem can be broadly classified as being either model-based or model-free. Both methodologies follow the general framework of feature extraction, feature correspondence and high-level processing. The major difference is with regard to establishing feature correspondence between two consecutive frames. Methods which assume *a priori* models match the 2-D image sequences to the model data. Feature correspondence is automatically achieved once matching between the images and the model data is established. Examples of this approach include the work of Lee et al. [13], where several ellipses are fitted to different parts of the binarized silhouette of the person and the parameters of these ellipses such as location of its centroid, eccentricity etc. are used as a feature to represent the gait of a person. Recognition is achieved by template matching. In [14], Cunado et al. extract a gait signature by fitting the movement of the thighs to an articulated pendulum-like motion model. The idea is somewhat similar to an early work by Murray [11] who modeled the hip rotation angle as a simple pendulum, the motion of which was approximately described by simple harmonic motion. Model-free methods establish correspondence between successive frames based upon the prediction or estimation of features related to position, velocity, shape, texture and color. Alternatively, they assume some implicit notion of what is being observed. Examples of this approach include the work of Huang et al. [15], where optical flow is to derive a motion image sequence for a walk cycle. Principal components analysis is then applied to the binarized silhouette to derive what are called eigen gaits. Benabdelkader et al. [16] use image self-similarity plots as a gait feature. Little and Boyd [17] extract frequency and phase features from moments of the motion image derived from optical flow and use template matching to recognize different people by their gait. Appearance based methods work reasonably well in the face of inaccurate background segmentation, changes in speed etc. However, such methods cannot tolerate drastic changes in clothing.

1.2 Thesis Contributions

In this thesis we present appearance based approaches for gait-based human identification using monocular video.

The first approach is a template matching approach to the problem. Given the gait video of an individual, the images are binarized and the width of the outer contour of the silhouette of that individual is obtained for each image frame. Several gait features are derived from this basic width vector. Temporally-ordered sequences of the feature vectors are then used for representing the gait of a person. While matching the feature templates for recognition, dynamic time-warping (DTW), which is a non-linear time-normalization technique, is used to deal with naturally occurring changes in the walking speeds of individuals. The performance of the proposed method is tested on indoor as well as outdoor gait databases, and the efficacy of different gait features and their noise resilience is studied. The experiments also demonstrate the effect of change in the viewing angle, characteristics of the floor surface, and frame-rate of data capture on the accuracy of gait recognition. The approach is found to work well in cases when relatively clean binarized silhouettes can be obtained through pre-processing and the training data is limited.

The second approach takes a more structured approach to represent the structure and kinematics of gait. Two different image features are considered: the width of the outer contour of the binarized silhouette of the walking person and the entire binary silhouette itself. To obtain the observation vector from the image features we employ two different methods. In the first method referred to as the indirect approach, the high-dimensional image feature is transformed to a lower-dimensional space by generating what we call the Frame to Exemplar (FED) distance. The FED vector captures both structural and kinematic traits of each individual. For compact and effective gait representation and recognition, the gait information in the FED vector sequences is captured in a hidden Markov model (HMM). In the second method referred to as the direct approach, we work with the feature vector directly (as opposed to computing the FED) and train an HMM. We estimate the HMM parameters (specifically the observation probability B) based on the distance between the exemplars and the image features. In this way we avoid learning high-dimensional probability density functions. The statistical nature of the HMM provides overall robustness to representation and recognition. The performance of the methods is illustrated using several databases.

To recognize a person walking at a large distance, humans try to combine information such as posture, arm/leg swing, hip/upper body sway or some quixotic characteristic of that person. The same effect can be achieved through fusion of different gait features. Dynamic features such as the swing of the hands/legs, the sway of the upper body and static features like height, in both the frontal and side views are considered. Established techniques, both probabilistic and non-probabilistic in nature, are used for matching the features. Various combination strategies may be used depending upon the features being combined. Three simple rules: the Sum, Product and MIN rules that are relevant to our feature sets are used. Experiments on different datasets demonstrate that fusion can be used as an effective strategy in recognition.

The gait of a person is easily recognizable when extracted from a side-view of the person. Accordingly, gait-recognition algorithms work best when presented with images where the person walks parallel to the camera (i.e. the image plane). However, it is not realistic to expect that this assumption will be valid in most real-life scenarios. Hence it is important to develop view-invariant gait recognition algorithms. We adopt a strategy whereby we synthesize the side-view of the person from an arbitrary view using a single camera and without explicit 3-D reconstruction. The method is based on the planar approximation of a person which is valid when human identification at a distance is desired. The differential equations of structure from motion (SfM) are used to estimate the 3-D azimuth angle from the video followed by view-synthesis. Since an explicit 3D model is not constructed we call this an "implicit SfM" approach. Statistical measures of quality of the synthesis are incorporated to increase the recognition performance. Examples of synthesized views are presented and gait recognition performance on three databases is presented. An application to video-based rendering of planar dynamic scenes is also presented.

1.3 Thesis Organization

The thesis is organized as follows. In Chapter 2, we present the appearance based template matching approach to gait recognition. The framework for gait-based recognition using continuous HMMs is presented in Chapter 3. Chapter 4 describes the fusion of multiple evidences to boost gait recognition performance. In Chapter 5 we present a view-invariant gait recognition algorithm based on the planar approximation of the person when human identification at a distance is desired. We conclude and outline future extensions and applications of the thesis in Chapter 6.

Chapter 2

Gait-Based Human Identification Using Appearance Matching

2.1 Introduction

A gait cycle corresponds to one complete cycle from rest (standing) position to-right-foot-forward-to-rest-to-left-foot-forward-to-rest position. The movements within a cycle consist of the motion of the different parts of the body such as head, hands, legs etc. The characteristics of an individual are reflected not only in the dynamics and periodicity of a gait cycle but also in the size and shape of that individual. Our aim is to build a model for representation and recognition of individual gait.

In a pattern classification problem, choice of the feature as well as the classifier is important. We choose the width of the outer contour of the silhouette a person as the basic image feature. As will be shown, the outer contour contains sufficient information for recognizing gait. From the raw width vector, different low-dimensional features are derived. These include the smoothed and down-sampled width vector, the eigensmoothed width vector, and the velocity feature vector. Temporally-ordered sequences of these feature vectors are used for compactly representing the person's gait. For a normal walk, gait sequences are repetitive and exhibit nearly periodic behavior. Typically, $5-10$ contiguous half-cycles of gait data per subject may be available and the number of frames per cycle ranges from 8 to 20. The amount of training data is inadequate for adopting statistical model-based approaches such as the Markov model, as it may not be possible to reliably estimate the parameters of the model. Hence a template-matching is adopted for comparing the probe and the reference sequences of the temporally-ordered feature vectors. Typically, gait cycles when taken at different times tend to be unequal in length due to changes in walking speeds of the individuals. Hence, a classifier based on direct template-matching is not appropriate. To deal with this issue, dynamic time-warping (DTW) is employed for matching gait sequences. The DTW method was originally developed for isolated word recognition [18], and later adapted for text-dependent speaker verification [19]. The gait problem is analogous to text-based speaker identification/verification wherein different individuals utter the same text but differ only in the characteristics of their utterance [19]. DTW uses an optimum time expansion/compression function for producing non-linear time normalization so that it is able to deal with misalignments and unequal sequence lengths of the probe and the reference gait sequences. Importantly, DTW can be used even with limited training data.

The performance of our approach is tested on five standard gait databases; namely, the University of Maryland (UMD) database, the MIT database, the Carnegie Mellon University (CMU) database, the University of Maryland angular-walk database (UMD3) and the USF gait challenge database. A description of these databases is detailed in Section 2.5 and will be refered to in later chapters as well. These databases contain video sequences of individuals walking in a normal manner along certain pre-defined paths. The UMD and CMU databases have both frontal and side views, the MIT database has only side-view sequences. The UMD3 database contains videos of people walking at angles 0, 15, 30, 45 and 60 degrees. The USF database consists of 71 people walking along an elliptical path on grass and concrete. A recently released version has 122 subjects. The proposed DTW-based approach is equally applicable to side as well as frontal views and the results are analyzed to determine the factors that contribute to accuracy of gait recognition. The idea is to study the efficacy of the features derived from the basic width vector and their resilience to noise in the estimate of the width vector. Our experiments also reveal the effect of differences in walking speeds, and the relevance of different body parts for gait recognition. Low frame-rate during data capture and methods to mitigate its effect are discussed. The effect of change in viewing angle on gait recognition is examined. It will be shown that the accuracy also depends on the surface characteristics of the floor.

In Section 2.2, the basic width vector, its extraction, and its relevance for the gait problem are explained. Section 2.3 discusses different low-dimensional features derivable from the basic width vector for gait representation. In Section 2.4, we describe how gait sequences can be matched using dynamic time-warping. Section 2.5 briefly reviews the gait databases used in our studies. Experimental results are discussed in Section 2.6. Conclusions and suggestions for further studies are given in Section 2.7.

2.2 The Width Vector

An important issue in gait-based human identification is the extraction of salient features that will effectively capture gait characteristics. The feature must be reasonably robust to operating conditions and should yield good discriminability across individuals. As mentioned earlier, we assume that the side view of each individual is available. Intuitively, the silhouette appears to be a good feature as it captures the motion of most of the body parts. It also supports night vision capability as it can be derived from IR imagery also. While extracting this feature we are faced with two options:

1. Use the entire silhouette.

2. Use only the outer contour of the silhouette.

We propose to use only the outer contour as we believe that it contains adequate information for recognizing gait. We choose the width of the outer contour of the silhouette as our feature vector. Given the image sequence of a subject, the width vectors are generated as follows:

1. Background subtraction as discussed in [20] is first applied to the image sequence. The resulting motion image is then binarized into foreground and background pixels (see Figure 2.1).

2. A bounding box is then placed around the part of the motion image that contains the moving person. The size of the box is chosen to accomodate all the individuals in the database. These boxed binarized silhouettes can be used directly as an image features or further processed to derive the width vector as described below.

3. Given the binarized silhouettes, the left and right boundaries of the body are traced. The width of the silhouette along each row of the image is then stored. The width along a given row is simply the differences in the locations of the right-most and the left-most boundary pixels in that row.

The physical structure of the subject as well as the swing of the limbs and other details of the body are retained in the width vector thus derived but the pose information is lost. For the four individuals shown in Figure 2.1, an overlay of the width vectors derived from the silhouettes is given in Figure 2.2.

The width-vector plots clearly bring out the fact that there is relatively more swing in the middle region (corresponding to hands) and in the end region (corresponding to feet), as compared to other parts of the body. However, these plots do not depict the temporal information. To bring out the temporal effects explicitly, the width vectors are plotted as a sequence of gray-level patterns in Figure 2.3 for the same four individuals. In this figure, about 5 full gait cycles are shown. The vertical axis corresponds to the index of the width vector, while the horizontal axis corresponds to the frame number. The total number of frames for each plot was fixed at 110 for uniformity. We call this the *temporal* plot of the width vector. It is clear that the width vector is approximately periodic and gives the extent of movement of different parts of the body. It varies with time in a quasi-periodic fashion, with the time taken for the completion of a half-cycle treated as the period.

Every component of the width vector contributes to the gait signature of a subject. The brighter a pixel, the larger is the value of the width vector in that position. The top part in each of the plots corresponds to the head-region, the middle part corresponds to the torso while the bottom part corresponds to the

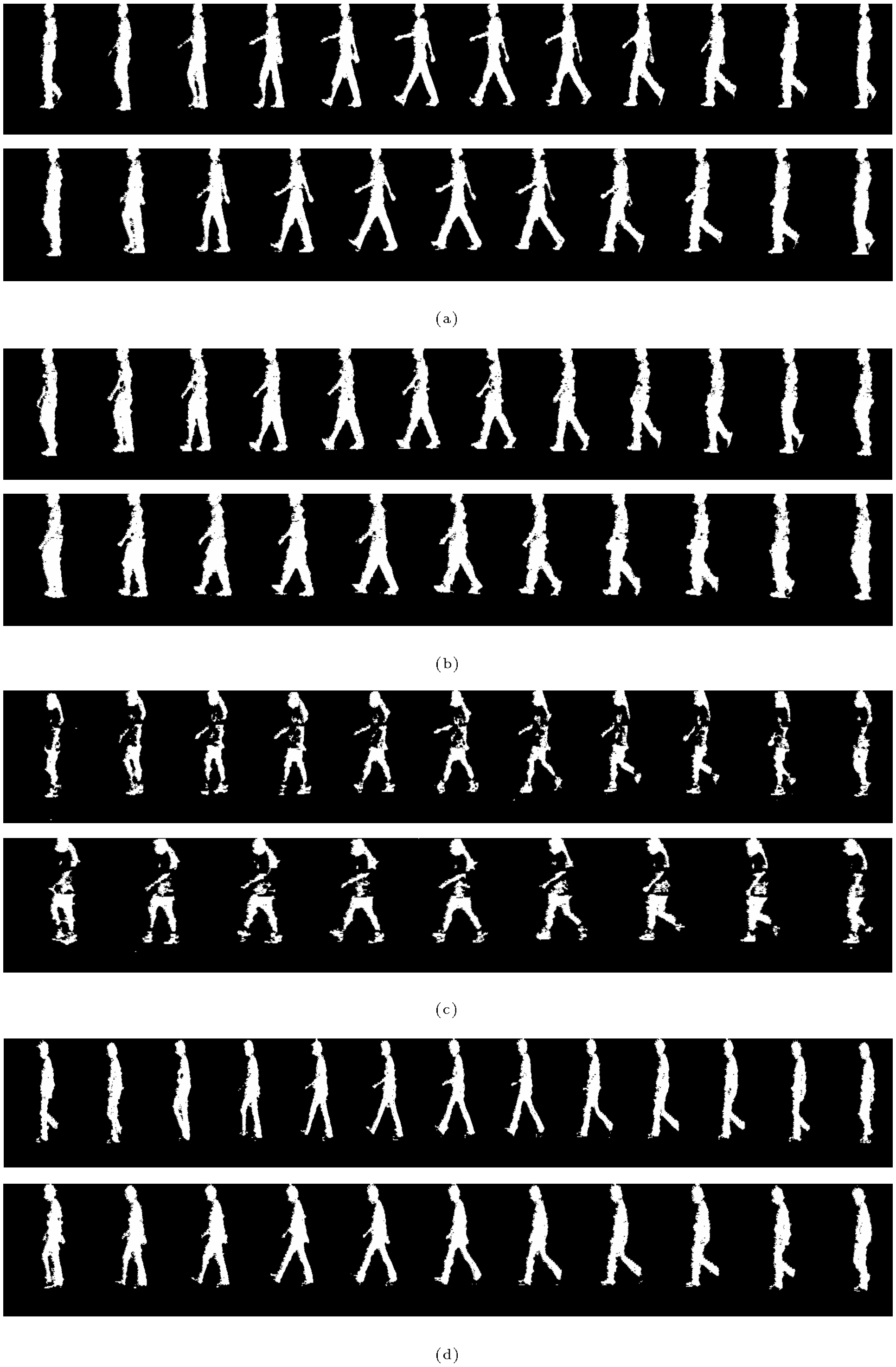
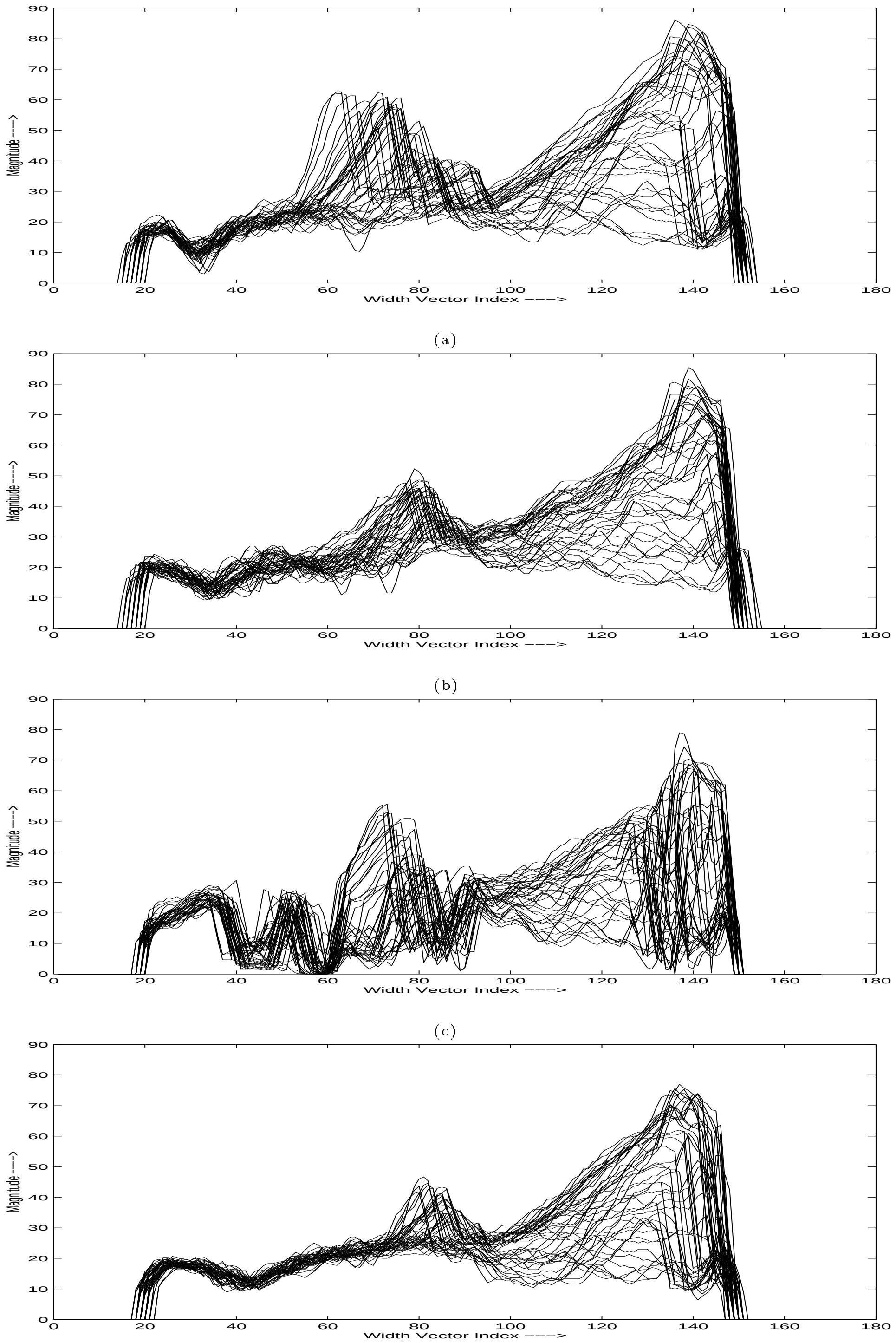


Figure 2.1: Silhouettes for a full gait-cycle corresponding to (a) Person 1 (b) Person 2 (c) Person 3 (d) Person 4.



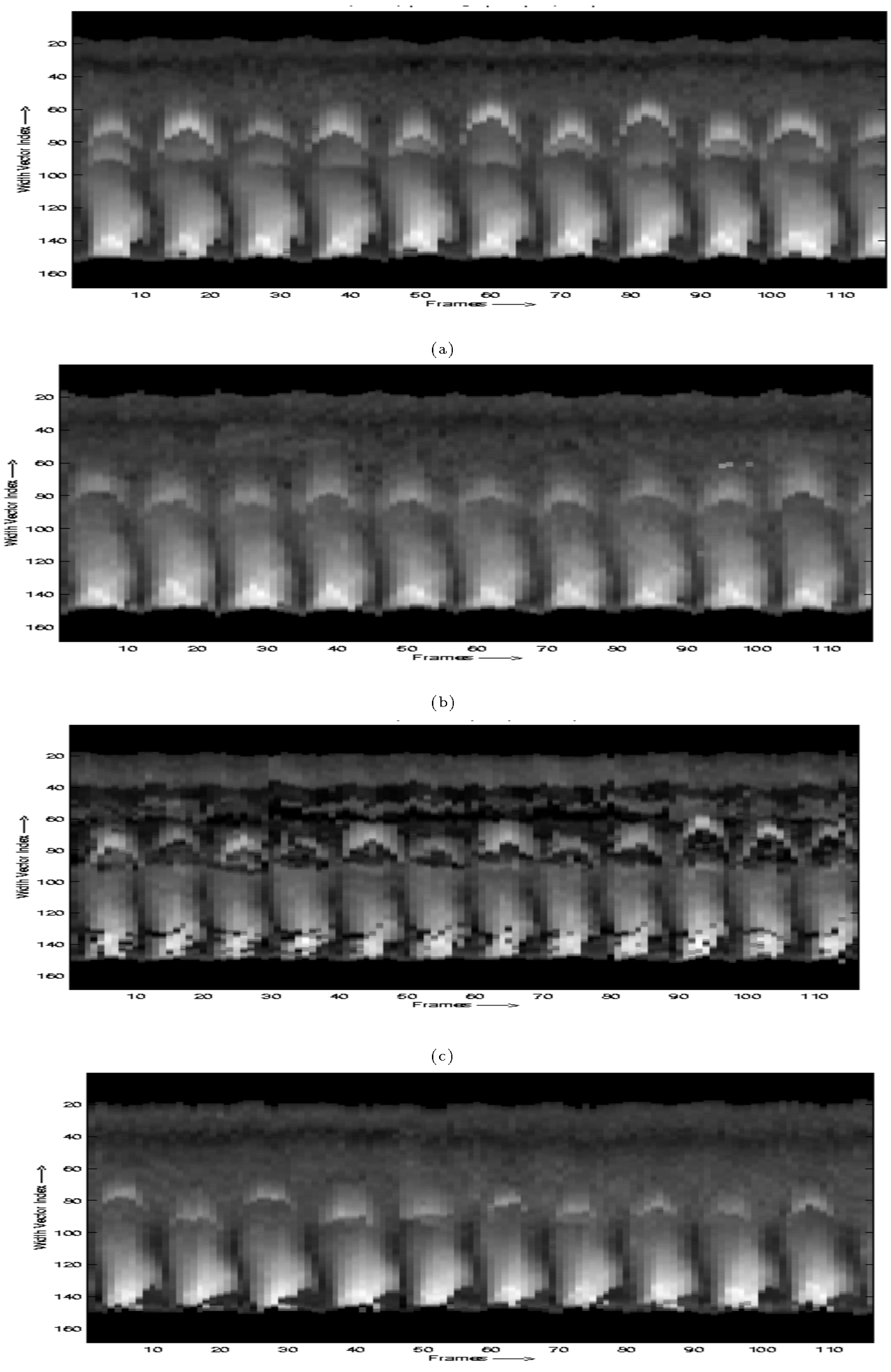
(a)

(b)

(c)

(d)

Figure 2.2: Width overlay plots for (a) Person 1 (b) Person 2 (c) Person 3 (d) Person 4.



(a)

(b)

(c)

(d)

Figure 2.3: Temporal plot of the width vector for several gait cycles of (a) Person 1 (b) Person 2 (c) Person 3 (d) Person 4.

foot-region of the individual. Note that the intensity variations in the torso and leg region, (corresponding to the hand and leg-swing) are larger than in the head region. The extent of the dark region near the top and bottom of the temporal width plots reflects the height differences between the individuals.

A comparison of the temporal width plots across individuals reveals interesting gait information. The structural or static differences among people is clearly revealed by the extent of the non-zero portions in these plots. One can also decipher kinematic information. For instance in Figure 3, the hand motion in the case of Person 1 is more pronounced as compared to Person 2. Note that the intensities are brighter for Person 1 in the middle region compared to Person 2. There also exist visible differences in the brightness gradients or velocities for different people. i.e., the velocity profile for each individual is different. There are differences in the repetition frequency which may be useful to distinguish people at a gross level. For instance, Person 3 has 12 half-gait cycles for the same number of frames compared to 10 $\frac{1}{2}$ cycles for Person 1 and 10 cycles for Person 2. Thus, considering a gait (half)-cycle as a unit of observation, we can capture both spatial and temporal characteristics of an individual's gait signature over several cycles.

2.3 Gait Representation

As discussed, the width vector captures important characteristics about an individual's gait. However, the dimension of the raw width vector can be quite high (168 in our experiments). In this section, we enumerate different low-dimensional but effective features that can be derived from the basic width vector. The idea is to arrive at a compact representation that exploits redundancy in the gait data for dimensionality reduction. Features that are directly obtained from the basic width vector are called direct features and these include the smoothed and down-sampled versions of the width vector. On the other hand, features derived using an eigen-analysis of the width vector across frames are called eigenfeatures. Corresponding to every frame of the gait cycle of an individual, one can extract any of the feature vectors discussed here. The gait of the individual is then represented by temporally ordering the feature vectors in accordance with the image frames in the gait cycle of that individual. We now describe each of the features in detail.

2.3.1 Direct Features

Examples of direct features are shown in Figure 2.4.

1. *The raw width vector*: This is the basic width vector discussed in Section 2.2, and is derived directly from the video sequence.

2. *3-point smoothed width vector*: This is derived from the raw width vector by smoothing it with a 3-point mean filter.

3. *5-point, 11-point, and 21-point smoothing and down-sampling by a factor of 4, 8 and 16, respectively*: The motivation behind smoothing and down-sampling stems from the fact that the original width vector has redundancies. Hence, it should be possible to discriminate reasonably well using lower dimensional features. Also, the original raw width vector is derived by simple and quick pre-processing. Hence, it is usually noisy and smoothing helps to mitigate the effect of noise.

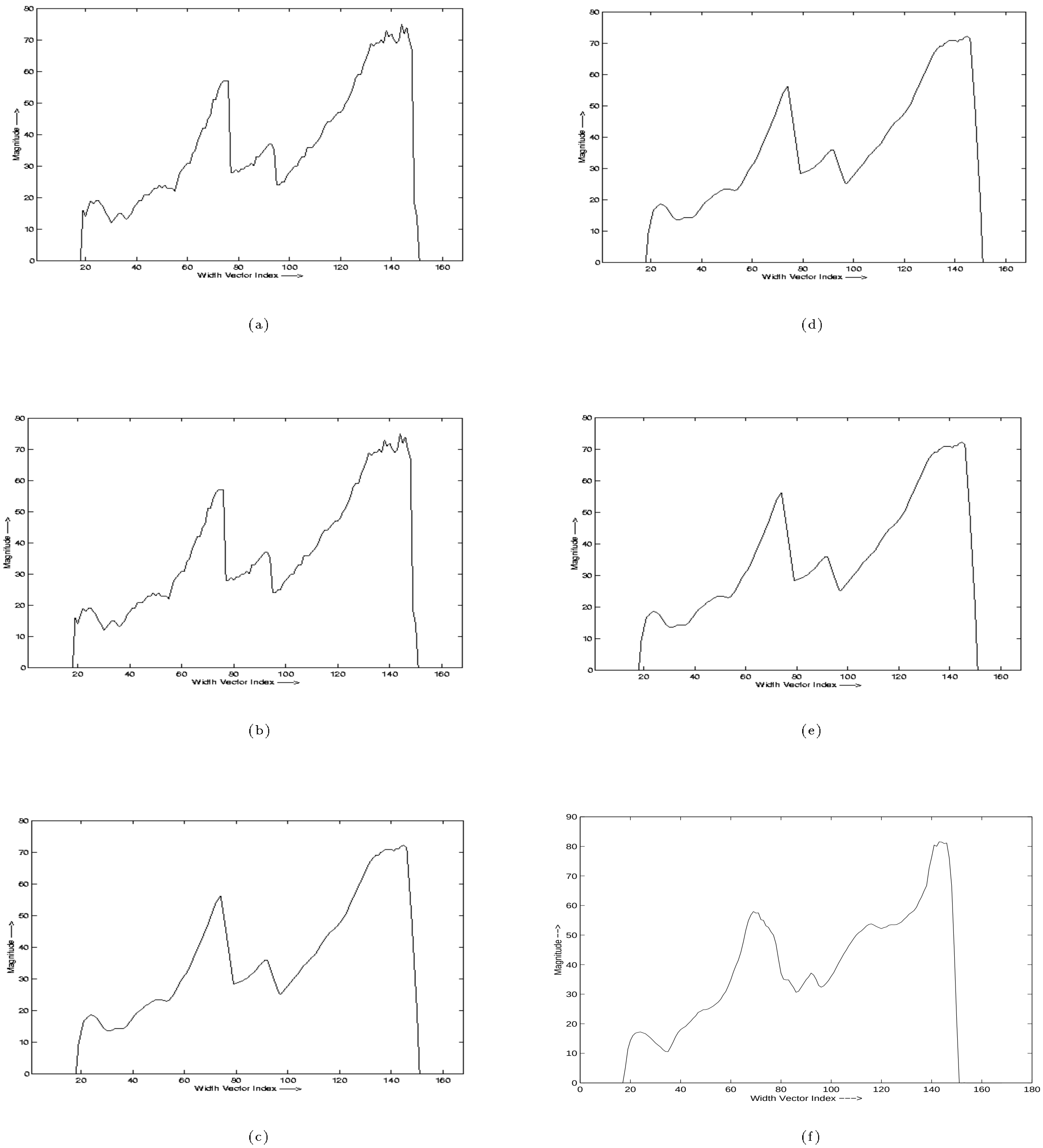


Figure 2.4: Width plots for an arbitrary frame. (a) Unsmoothed raw feature. (b) 3-point smoothed feature. (c) 5-point smoothed feature. (d) 11-point smoothed feature. (e) 21-point smoothed feature. (f) Reconstructed width vector using the first two eigenvectors.

4. *Difference feature vector*: This is useful to study the effect of kinematics for gait-based human identification. There are many different ways to extract kinematic information from the width vector. One would be to simply take the difference of successive raw width vectors across frames thus preserving only the changes that occur between frames, during a gait cycle followed by smoothing and downsampling. There is a trade-off between smoothing and extracting dynamic information. Although some degree of smoothing is required to counter noise, too much smoothing can alter the dynamic information. Obviously, most of the structural information, like girth of the person, is lost when we go to the velocity domain. It is to be expected that neither dynamic nor structural information, in isolation, will be sufficient to capture gait. As will be shown later, both are necessary and cannot be decoupled.

2.3.2 Eigenfeatures

From the temporal width plots, we note that although the width vector changes with time within a gait cycle, there is a high degree of correlation among the width vectors across frames. Most changes occur in the hand and in the leg regions. Hence, it is reasonable to expect that gait information in the width vector can be derived with much fewer coefficients. This is what we attempt to do in the eigenanalysis of gait.

All the width vectors corresponding to several training gait cycles of an individual are used to construct the covariance or scatter matrix for that individual. Principal components analysis is then performed to derive the eigenvectors for this data.

Given the width vectors $\{W(1),\cdots,W(N)\}$,for N frames $W(.) \in R^M$, we compute the eigen vectors $\{V(1,)\cdots,V(M)\}$ corresponding to the eigen values of the scatter matrix arranged in the descending order and reconstruct the corresponding width vectors using $m(< M)$ most significant eigen vectors as

$$W_r(i) = (\sum_{j=1}^{m} w_j V(j)) + \bar{W}$$

where $w_j =< W(i), V(j) >$ and $\bar{W} = \frac{W(1)+\cdots+W(N)}{N}$.

Note that every individual has his/her own eigengait space. Reconstruction of the width vector using only the first and the second eigenvectors is shown in Figure 2.4(f). The consequence of approximating the width vector using only the first two eigenvectors is that the effect of noise is suppressed and the width vector is smooth. The first two eigenvectors capture the physical structure of the person as well as typical arm and leg swings, and serve as a freeze-frame representation of the gait sequence.

2.4 Matching Gait Sequences using DTW

Since only limited training data is usually available for gait analysis, a template-matching approach is adopted for recognition. The input to the gait recognition algorithm is a sequence of image frames. Each image frame can be compactly represented by any of the feature vectors discussed in Section 2.2 to derive a template which consists of a temporally ordered sequence of feature vectors for representing gait. The number of frames in the reference gait sequence and the probe gait sequence depend on the number of gait cycles available. Larger the number of gait cycles used for matching, the better we can expect the performance to be. Typically, the number of frames in the reference and probe data

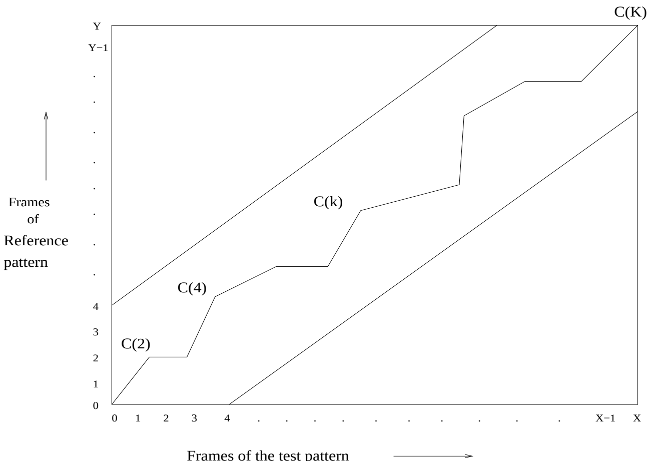

C(K)
Y
Y−1
Frames of Reference pattern
C(k)
C(4)
C(2)
4
3
2
1
0
0 1 2 3 4 . . . . X−1 X
Frames of the test pattern

Figure 2.5: A typical dynamic time-warping path.

will differ. Moreover, the reference and probe gait data are seldom synchronized. Therefore, direct matching of a probe gait sequence with a reference gait sequence should be avoided. In this work, a pattern-matching method based on dynamic programming paradigm is used for dealing with this situation. Dynamic Time Warping (DTW) [19] can be used to compensate for the variability in the speed of walking which in turn reflects in the number of frames for each gait cycle. A distance metric (usually the Euclidean distance) defined as a function of time is computed between the two feature sets representing the gait data. A decision function is arrived at by integrating the metric over time. The DTW method was originally developed for isolated word recognition [18], and later adapted for text-dependent speaker verification [19]. The gait problem is analogous to text-based speaker identification/verification wherein different individuals utter the same text but differ only in the characteristics of their utterance [19].

Consider the $x - y$ plane shown in Figure 2.5 where the x and the y axes represent the frame numbers of the probe and the reference patterns, respectively. Assume that the first frame of the reference and probe sequence are both indexed as 1. Let the last frames of the reference and probe sequences be indexed as X and Y, respectively. The match between the two sets can be represented by a sequence of K points $C(1), C(2), ..., C(k), ..., C(K)$, where $C(k) = (x(k), y(k))$, and $x(k)$ is a frame index of the probe sequence and $y(k)$ is a frame index of the reference sequence. Here, $C(k)$ represents mapping of the time axis of probe sequence onto that of the reference sequence. The sequence

$$F = C(1), C(2), ...C(k), ..., C(K)$$

is called the warping path.

The process of time normalization uses certain constraints depending on the problem at hand. Some of the constraints are the following [21].

(a) Endpoint Constraints: The fixed endpoints of the patterns lead to a set of constraints on the warping function. They are of the form

$$x(1) = 1,\ y(1) = 1,\ x(k) = X,\ \text{and}\ y(k) = Y \tag{2.1}$$

i.e., the first and the last frames of the probe sequence should be matched with the first and the last frames of the reference, respectively. Under these constraints, a typical warping would look like the one shown in Figure 2.5. The motivation behind the end point constraint arises from the fact that DTW involves matching two entities that are similar and where the start and and end point for the sequences correspond e.g. comparing the speech signals for the same word pronounced by two different speakers. The imposition of the end-point constraint ensures that extremal feature vectors are not lost. When matching gait sequences, the use of the width vector provides a simple way of segmenting the gait sequence into cycles (see Section 2.4.1), before DTW may be performed. In the presence of noise the exact start and end frames may be hard to determine. However the flexibility provided by the montonicity and local continuity constraints ensure that the errors arising from this imprecise localization are not large.

(b) Monotonicity Constraint: This constraint requires feature vectors belonging to a probe sequence (or the reference sequence) to be matched in monotonically increasing order. This maintains the temporal order of the sequence during time normalization. This is accomplished by ensuring that

$$x(k-1) \leq x(k), \text{ and } y(k-1) \leq y(k). \tag{2.2}$$

(c) Local Continuity Constraints: This constraint can be used to ensure use of each frame probe frame during normalization, and also to ensure that no more than one frame is skipped in the reference. This is achieved by setting the constraints

$$x(k) - x(k-1) = 1, \text{ and } y(k) - y(k-1) \leq 2 \tag{2.3}$$

(d) Global Path Constraints: This constraint restrains the extent of compression or expansion. The warping path can be constrained by defining a region around which the warping path is allowed to traverse. The region between the two parallel lines marked in the $x-y$ plane in Figure 6 defines the allowable grid or the region through which the warping path can traverse. The warping path can be limited to a band around the diagonal so that the search time for optimal path is significantly reduced. This is reasonable as it is unlikely that the probe sequence of the genuine case would deviate significantly from its reference.

After deciding the local and global constraints, the DTW algorithm is applied as follows.

1. Local Distance Computation: This involves computing the distance between the feature vectors representing the probe and reference frames. The distance between the probe and reference feature vectors $\mathbf{t_x}$ and $\mathbf{r_y}$ can be computed using

   $$d(C(k)) = |\mathbf{t_x} - \mathbf{r_y}|$$

   so that the total distance along the path F is given by

   $$E(F) = \sum_{k=1}^{K} d(C(k)) \tag{2.4}$$

   The distance $E(F)$ is the similarity score which can be used at the decision making stage.

2. Cumulative Distance Computation: The cumulative distance $D(x(k), y(k))$ is the minimum distance to reach $(x(k), y(k))$ at the k^{th} stage starting from $(1,1)$. This distance is the sum of all the local distances of points through which the warping path passes to reach $(x(k), y(k))$ from $(1,1)$. Initialize the cumulative distance $D(1,1)$ to $L(1,1)$, where $L(1,1) = d(C(1))$. For all the other points lying within the global region of search, compute the cumulative distance using the local path constraints, i.e., the point $(x(k), y(k))$ can only be reached from the points $((x(k)-1), y(k))$ or $((x(k)-1), (y(k)-1))$ or $((x(k)-1), (y(k)-2))$. If the distance at stage k is obtained as

   $$D(x(k), y(k)) = \min \begin{cases} D((x(k)-1), y(k)) + L((x(k), y(k)) \\ D((x(k)-1), (y(k)-1)) + L(x(k), y(k)) \\ D((x(k)-1), (y(k)-2)) + L(x(k), y(k)) \end{cases} \tag{2.5}$$

   then $D(X, Y)$ gives the distance between the probe and the reference sequences.

3. Backtracking: Using the cumulative distance matrix and the local path constraints, the warping path F can be obtained by backtracking.

2.4.1 Implementation details

Use of dynamic time warping relies on similar start and end points for the probe and gallery. Hence prior to applying the DTW algorithm it is necessary to parse the video sequence into cycles. A simple way to achieve this is using the width feature explained in Section 2. From Figure 2.3 it is easy to see that $N(t) = (\sum_{j=1}^{M} I(j))(t)$ where $I(j)$ denotes the width at row j of the image, viz. the sum of widths will show a periodic variation (see Figure 2.6).

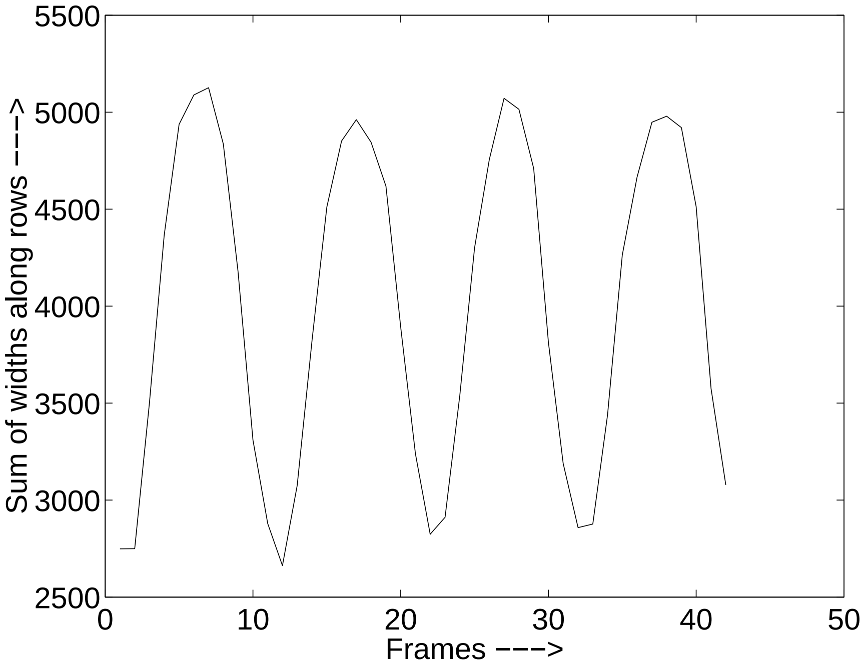


Figure 2.6: Sum of widths as a function of time.

The troughs of the resulting waveform correspond to the rest positions during the walk cycle while the crests correspond to the part of the cycle where the hands and legs are maximallly displaced. A half-cycle consists of frames between two successive troughs of sum of intensities plot. In general, the exact trough may be hard to determine in the presence of noise and picking a frame in its neighborhood is usually adequate. Given a video sequence, the width feature thus provides a natural parsing in terms of half-cycles. In our experiments, four half-cycles from the probe sequence are matched with four half-cycles from the gallery sequence. Euclidean distance was chosen as the local distance. A global constraint of $max(0.1N_{gallery}, 0.1N_{probe})$ was chosen. Using the similarity matrix resulting from the matching algorithm, the cumulative match characteristic as explained in [22]. Essentially the cumulative match characteristic is a plot of the number of times the right person occurs in the top n matches where $n < G$ where G denotes the number of people in the gallery, as a function of n.

2.5 Data

The UMD Database: Gait data was captured with the out-door video research (OVR) facility at the University of Maryland [1]. Cameras located at right angles and at a height of 4.5 meters above the ground were used for data capture. The UMD dataset consists of 44 individuals. There are 38 male and 6 female subjects with different ethnicity, physical build etc. The individuals were asked to walk normally along a T-shaped path. Because the cameras are orthogonally placed and the path is T-shaped, a side-view and a frontal view could be captured simultaneously as the person walked. The videos for each individual were recorded with intervals ranging from about half-a-day to a maximum of one month. Each image frame was of size 170 × 138 pixels. Labeling was done

[1] http://degas.umiacs.umd.edu/hid

to mark the gender, test/training sequences, direction of walk, number of gait cycles, and left/right foot forward.

The CMU Database: The CMU database [2] consists of different views of indoor gait sequences of 25 individuals walking on a treadmill. Different situations such as a person walking at a slow pace, fast pace, and walking while carrying a ball in his/her hands are considered. Even though it is an indoor database, it serves as a good testbed for evaluating performance under variations in walking speed and also to assess the relative importance of body parts. For example, in the sequence where the person is carrying the ball, there is very little upper-body movement. The size of each image frame in this database is 640 × 486 pixels.

The MIT Database: The MIT database [3] consists of side-views of gait sequences of 25 subjects captured indoors, and collected on four different days. There are 4-8 segments per subject per day with 2-5 contiguous half cycles per segment. The number of subjects on a particular day varies from 4 to 16. As a result, the degree of overlap across days is less. The frame rate of the camera is only 15 frames per second (as compared to 25 for the UMD database), and the image size is 128 × 104 pixels. A lower frame rate leads to sparse temporal sampling of the gait information and must be tackled appropriately. This database provides a good testbed for evaluating training and test databases captured on different days.

The UMD3 Database: The UMD3 database consists of outdoor gait sequences of 12 people walking along straight-line paths at azimuth angles of $0, 15, 30, 45$ and 60 degrees from the side view. The data was captured at the rate of 25

[2] http://hid.ri.cmu.edu

[3] http://www.ai.mit.edu/people/llee/HID/intro.htm/

frames per second by a tripod-mounted camera. Two sequences of each person were captured with a time interval of at least 1 hr between them. One of the sequences was used as reference and the other was used as the probe. This database provides a testbed for assessing gait recognition performance as a function of the viewing angle. The image size for zero azimuth is 210 × 105 pixels.

The USF Database: In this database [23], each subject was required to walk in a counter-clockwise manner along two elliptical courses. The elliptical courses are located outdoors and approximately 15 meters along the major-axis and 5 meters along the minor-axis. To enable testing on different floor surfaces, the first course was laid out on flat concrete, while the second was laid out on a typical grass-lawn. Each course was viewed by the two cameras. Their lines of sight were not parallel but verged so that the whole ellipse was just visible from the two cameras. When the persons walked along the rear portion of the ellipse, their view was approximately sideways. Subjects were also asked to bring a second pair of shoes, so that they could walk a second time in a different pair of shoes. A little over half of the subjects walked in two different shoe-types. The image size of the subjects is 100 × 50 pixels. The database has 71 individuals of which 75% are males. The USF database also has people walking with a briefcase in their hand and data collected six months apart.

2.6 Experimental Results

In this section, we present the results of our approach to gait recognition on all the databases described in Section 2.5. Each of the databases has its own interesting features, and our exercise reveals several interesting facts about the gait-based human identification problem. The performance of all the features

$Feature\backslash Rank$	1	2	3	4	5
Set of 4 half cycles 1(a)	68.1	77.2	84.0	84.0	84.0
Set of 4 half cycles shifted by one 1(b)	70.4	79.5	81.8	86.3	86.3
Minimum of 1(a) and 1(b)	79.0	81.4	83.2	86.0	86.0

Table 2.1: UMD database: Cumulative match scores for the first five ranks for the case of 2 full-cycles having a relative phase shift of 1 half-cycle.

discussed in Section 2.3 were first studied for the UMD database to get a feel for what might be a good feature vector for gait representation and recognition. The method described in Section 2.4 was used for recognition.

2.6.1 Results for the UMD Database

For this database there were approximately 40 frames corresponding to 4 half-cycles in the gallery and probe. It must be mentioned here that similar to hand-dominance (right/left handedness), foot-dominance (right/left leggedness) also exists in individuals. In practice, it is difficult to align the reference and the probe sequences accurately with respect to heel-strike. The result of ignoring heel-strike information (left-foot forward or right-foot forward first) in gait analysis is reflected in Table 2.1.

Suppose there are five half-cycles in both reference and probe sequences for a particular subject. The first four half-cycles of the two sequences are matched to generate a matrix of similarity scores. Then, the reference sequence is matched with a probe sequence shifted by half-cycle to generate another matrix of similarity scores. Of the two shifted probe sequences, only one can provide a better match if the subject exhibits foot-dominance. The two similarity scores are combined using the minimum-error criterion. The improvement in recognition when

$Feature\backslash Rank$	1	2	3	4	5
Unsmoothed 168-dim feature	79.0	81.4	83.2	86.0	86.0
3-point smoothed 168-dim feature	79.0	81.4	83.7	86.0	86.0
5-point smoothed 42-dim feature	76.7	83.7	83.7	88.3	88.7
11-point-smoothed 21-dim feature	76.7	83.7	83.7	83.7	90.0
21-point-smoothed 11-dim feature	79.0	86.0	86.5	88.3	90.0

Table 2.2: UMD database: Cumulative match scores for the first five ranks for different directly-derived feature vectors.

heel-strike is taken into account can be noted from Table 2.1. In the results that follow, we assume that fusion of the similarity scores from two phase-shifted cycles has been carried out.

The recognition results for the direct features are given in Table 2.2. From the table, we note that all the features have done well despite the presence of noise in the estimate of the basic width vector. The experiment also shows that dimensionality reduction is possible to a great extent by exploiting the redundancy in the gait data. In the experiments to come, we choose the 5-point smoothed 42-dimensional width vector, which is between the two extremities of no-smoothing and very heavy-smoothing, as the direct feature for performance analysis.

For the eigenfeatures, the performance results obtained by combining different numbers of eigenvectors is given in Table 2.3. Again, four half-cycles of each subject were used for matching. Note that by using just the first two eigenvectors an accuracy of 80% is achievable. Other eigenvectors are noisy and, in fact, tend to lower the accuracy. Hence, we have used only the first two eigenvectors for computing the eigenfeatures in our experiments. It is also of interest to study the effect of the number of half cycles on matching. Intuitively, one would expect

$Feature\backslash Rank$	1	2	3	4	5
Eigenvector 1	73	75	80	80	84
Eigenvectors 1,2	80	87	90	90	91
Eigenvectors 1,2,3	68	80	84	84	84
Eigenvectors 1,2,3,4	73	77	84	84	84
Eigenvectors 1,2,3,4,5	70	73	79	82	84
Eigenvector 2	58	63	67	72	74
Eigenvectors 2,3	61	64	67	82	76
Eigenvectors 2,3,4	61	65	68	68	72
Eigenvectors 2,3,4,5	68	68	72	74	74
Eigenvector 3	51	54	60	62	62
Eigenvectors 3,4	55	60	68	68	72
Eigenvector 4	22	28	35	42	42
Eigenvector 5	20	24	36	40	44

Table 2.3: Cumulative match scores for the UMD database using different eigenfeatures.

$Feature\backslash Rank$	1	2	3	4	5
Four half-cycles	80.0	87.0	90.0	90.0	91.0
Two half-cycles	71.0	80.0	82.0	84.0	87.0
One half-cycle	65.0	67.0	77.0	83.0	87.0

Table 2.4: Effect of length of gait sequence for the UMD database: Cumulative match scores for the first five ranks.

to obtain a better match with more number of half cycles, i.e., with a longer gait sequence. When the number of half-cycles is reduced, systematically from four to one, the corresponding results are given in Table 2.4.

While reducing the number of half-cycles, the number of frames was kept approximately constant by linearly interpolating the components of the width vector. As expected, the performance degrades as the length of the sequence is reduced.

To assess the relative discriminability of the structural component of the

$Feature \backslash Rank$	1	2	3	4	5
Smoothed and differenced 168-dim feature	41.9	51.6	61.2	70.9	74.1
Eigen decomp. of velocity profile	56	75	76	80	83
Eigen decomp. of width vector and successive difference	32.0	40.0	51.0	61.0	65
of projected weights					
Eigen decomp. of width vector and using 2nd eigenvector	58.0	63.0	67.0	72.0	74.0

Table 2.5: UMD database: Cumulative match scores for the velocity profile.

width vector versus its dynamic component, we computed the difference of the width vectors corresponding to successive frames in the gait sequence. Eigendecomposition of the velocity vector is carried out and reconstructed using the first two eigenvectors. Table 2.5 shows the results obtained by considering only the velocity profile as the feature of interest.

Alternatively, the width vector could have been first subjected to eigendecomposition, and the successive differences of the projected values used for matching. Note that the accuracy drops significantly if only the velocity information is used. Clearly, for gait-based human identification both structural and kinematic information are important.

2.6.2 Results for the CMU database

The following experiments were performed on this database: (i) Train on slow-walk and test on slow-walk, (ii) train on fast-walk and test on fast-walk, (iii) train on slow-walk and test on fast-walk and (iv) train on walk carrying a ball and test on walk carrying a ball.For this database, the number of frames corresponding to 4 half-cycles varied between 55 for slow walk to about 75 for fast walk. The results are given in Table 2.6. We note that the eigensmoothed feature performs better than the direct-smoothed feature. This can be attributed to

		Rank				
Experiment	Feature	1	2	3	4	5
Slow vs Slow	Direct-smoothed feature	70.8	83.3	87.5	95.8	95.8
	Eigen-smoothed feature	95.8	95.8	95.8	95.8	100.0
Fast vs Fast	Direct-smoothed feature	83.3	83.3	83.3	83.3	87.5
	Eigen-smoothed feature	95.8	95.8	95.8	95.8	100.0
Fast vs. Slow	Direct-smoothed feature	54.1	75.0	87.5	87.5	87.5
	Eigen-smoothed feature	75.0	83.3	83.3	83.3	87.5

Table 2.6: CMU database: Recognition results for different experiments.

the fact that eigensmoothing exploits spatio-temporal redundancy unlike direct-smoothing which uses only spatial smoothing. When the reference is the slow-walk sequence and the probe is the fast-walk sequence, the performance is found to be inferior than for the case when the reference and probe are both slow-walk sequences. DTW is known to perform poorly [24] when the ratio of reference sequence-length to probe sequence-length is less than 0.5 or more than 2. In the CMU dataset, this ratio for the worst case was 1.36. From Table 7, we note that the DTW-based method is reasonably robust to changes in walking speed. In fact, in our experiments, the value of the ratio for one of the mismatched cases was 1.15. A few frames in the gait cycles of this incorrectly recognized person under slow and fast-walk modes are shown in Figures 2.7 (a) and (b).

As is apparent from the figure, the posture as well as hand-swings for the person are quite different in cases of fast-walk and slow-walk. Figures 2.7 (c) and (d) show the warping paths for the person with the highest ratio and the incorrectly recognized person, respectively. Note that the warping path for the correctly recognized person is much more regular as compared to that of the incorrectly recognized individual. Hence, it is the change in the posture and body dynamics of the person rather than the mismatch in the length of the

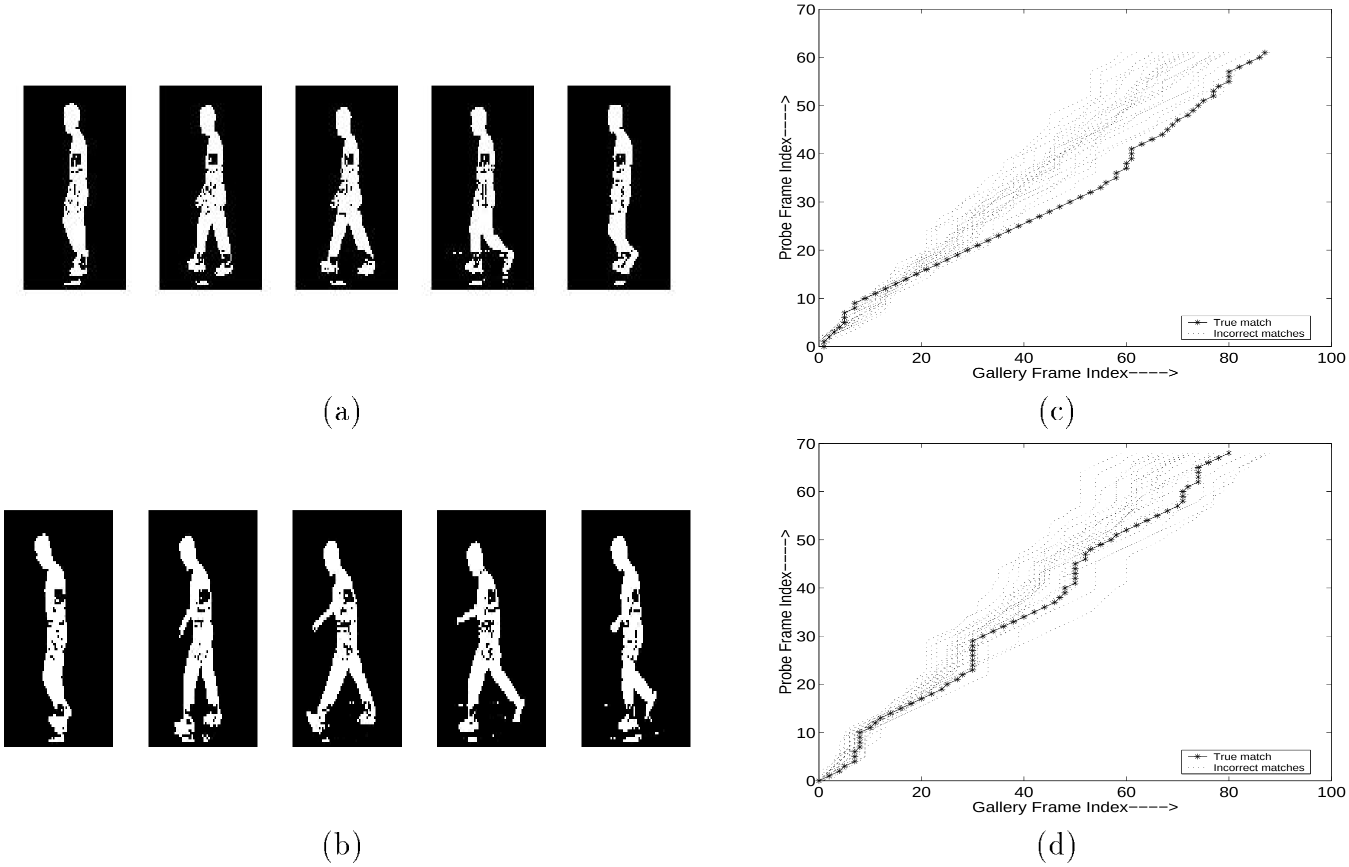


Figure 2.7: Sample images of the same subject corresponding to (a) slow-walk and (b) fast-walk (Notice the change in posture and body dynamics) (c) Warping path for person with largest reference sequence-length to probe sequence-length ratio (d) Warping path for the person in (a) and (b).

$Experiment\backslash Rank$	1	2	3	4	5
Ball vs Ball	95.45	100.00	100.00	100.00	100.00

Table 2.7: Cumulative match scoresfor the ball experiment in the CMU database using the eigen-smoothed feature.

reference and probe sequences that was responsible for the mismatch.

Finally, we consider the case when the person is walking with a ball in his hand. In this situation, most of the gait dynamics is confined to the leg region. For this experiment, we observe from Table 2.7 that the recognition performance is very good (person identified correctly within the top 2 matches). This experiment suggests that for the purpose of recognition certain parts of the body may contain more information than the others. In particular, using the leg motion alone provides more discriminating evidence as compared to what might be obtained by weighting the evidences from the hand and leg motion coequally. This fact can be used for building more efficient gait recognition algorithms in some cases. Incidentally, this has been noted in kinesiology research as well [12].

2.6.3 Results for the MIT Database

The evaluation scheme for the MIT database is as follows. For training, data collected on days 2, 3 and 4 were used. Data collected on day 1 was used for testing. Overall, there are 8 subjects. We extract 2 half-cycles (which is the maximum number of half-cycles common to all the subjects in the database) and build the direct-smoothed width vector. Since the frame rate is 15 frames/second, on the average, there are only 7-8 frames per half-cycle so that there were about 28 frames in four half-cycles. We recognize that the length of the gait sequence is not adequate. As discussed earlier, the more the number of half-cycles, the better would be the possibility of obtaining a correct match. Nevertheless, 4 out of the 8 people were correctly recognized. Since the frame-rate is low for this database, it results in coarse-sampling in the temporal domain and presents a problem while matching. When we used linear interpolation between frames to partially compensate for the coarse sampling, the recognition rate increased to 5 matches out of 8. When the eigenfeatures were tried on this database, the results were not very different.

2.6.4 Results for the UMD3 database

The UMD3 database contains gait data corresponding to different azimuth angles. The gait sequences were captured at different times at angles $\theta = 0$, 15, 30, 45 and 60 deg with respect to the side-view. To account for the foreshortening effect as a person walks at non-zero azimuths, the height of each individual was computed from the zero-azimuth width plots, and this was used to normalize the height corresponding to non-zero azimuth angles. Figure 2.8 shows the width plots for one of the individuals in the database for each of the above angles.

There were approximately 60 frames corresponding to 4 half-cycles for the UMD3 database. It can be seen that for large azimuth angles the hand swings appear only in alternate half-cycles. This is because the body occludes the presence of hands in the alternate half-cycles. Another observation is that the range of variations in the width amplitude is small for large azimuth angles. Therefore, recognition performance can be expected to be more sensitive to the presence of noise in the gait sequence for higher values of θ.

Two full-cycles having a relative shift of one half-cycle were used as gallery sequence (similar to the first UMD database) and two full-cycles captured at a

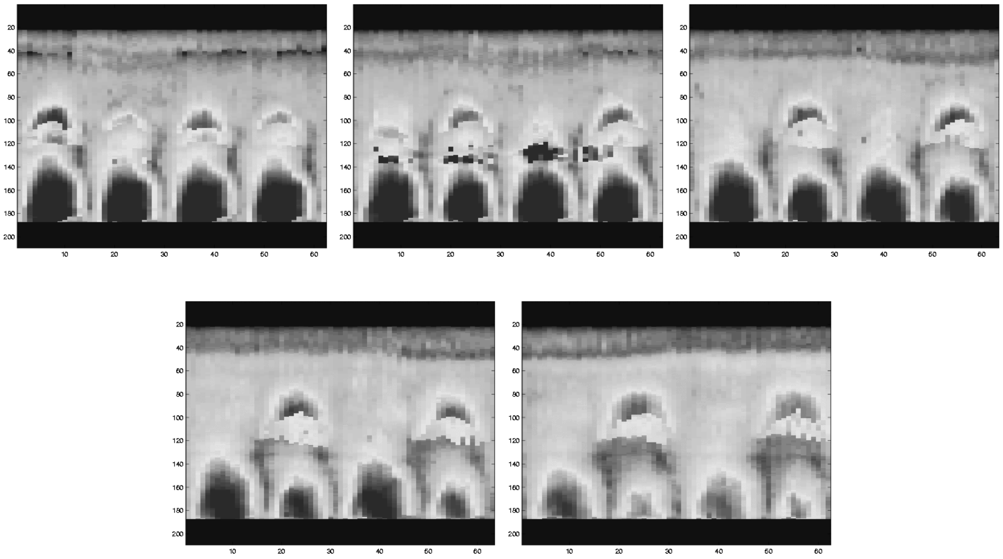


Figure 2.8: Width profile of an individual for different azimuth angles (a) 0 deg (b) 15 deg (c) 30 deg (d) 45 deg and (e) 60 deg.

later time were chosen as the probe sequence. Both the 5-point direct-smoothed feature and the eigensmoothed feature were tested for recognition and the results are given in Table 2.8.

In the table, L and R refer to two full gait-cycles with a relative shift of one half-cycle, while F refers to the case when we accept smaller of the two entries corresponding to L and R. It is clear that $\theta = 0$, which is the side-view, provides the best viewing direction for gait recognition. Since the variation in the amplitude of the width vector was small for large θs, eigensmoothing results in better performance. For smaller values of θs when the width vector components are relatively large, eigensmoothing does not seem to yield any significant improvement over direct smoothing.

		Rank (direct)			Rank (eigen)		
Angle from side-view	Sequence chosen	1	2	3	1	2	3
0 deg	L	75.00	91.67	100.00	91.67	91.67	100.00
	R	75.00	83.33	91.67	83.33	83.33	91.67
	F	91.67	91.67	100.00	91.67	91.67	100.00
15 deg	L	66.67	83.33	100.00	91.67	91.67	100.00
	R	58.33	75.00	83.33	91.67	91.67	100.00
	F	91.67	91.67	100.00	91.67	91.67	100.00
30 deg	L	66.67	83.33	91.67	83.33	83.33	91.67
	R	25.00	50.00	50.00	58.33	91.67	100.00
	F	83.33	83.33	91.67	83.33	83.33	100.00
45 deg	L	33.33	41.67	50.00	41.67	66.67	91.67
	R	66.67	75.00	83.33	83.33	91.67	100.0
	F	50.00	83.33	100.00	58.33	83.33	100.00
60 deg	L	33.33	50.00	58.33	58.33	75.00	91.67
	R	41.67	58.33	58.33	75.00	83.33	91.67
	F	58.33	83.33	91.67	75.00	83.33	91.67

Table 2.8: UMD3 database: Cumulative match scores for the first three ranks for different azimuth angles using the direct and the eigen-feature vectors.

2.6.5 Results for the USF Database

The USF database has the largest number of individuals among all the databases. It has variations with respect to floor surface (grass (G) or concrete(C)), shoe type (A or B), and camera viewing direction (left (L) or right (R)). The reference for all the experiments was chosen to be (G, A, R). The number of frames corresponding to four half cycles varied from 65 to 90. Different probe sequences for the experiments along with the cumulative match scores are given in Table 2.9 for the baseline algorithm [23] as well as our method using the eigensmoothed width feature. Note that recognition performance suffers most due to difference in surface characteristics, and least due to difference in viewing angle. However, unlike the UMD3 experiment, both the cameras see an 'almost' side-view of the subjects. An examination of the USF database revealed that the silhouettes provided were noisier compared to the previous datasets. We wanted to see what the performance would be by using the binary silhouettes directly as the feature. In this case we used the binary correlation distance in the local distance computation. As can be seen from the last two columns of Table 2.9 usage of the binarized silhouettes yields better performance numbers compared to the width vector in this case.

2.7 Conclusions

In this chapter, we have proposed a new approach to gait recognition using the DTW technique. The method was tested on five different databases using several features. The width of the outer contour of the binarized silhouette of a person was used as the gait feature. Different feature vectors were derived from

Experiment (Probe)	Baseline		Width Vector		Binary Silhouette	
	Rank 1	Rank 5	Rank 1	Rank 5	Rank 1	Rank 5
A (G, A, L)	79	96	79	91	84	97
B (G, B, R)	66	81	67	79	83	91
C (G, B, L)	56	76	30	55	59	79
D (C, A, R)	29	61	17	42	41	64
E (C, B, R)	24	55	15	39	24	53
F (C, A, L)	30	46	16	30	27	51
G (C, B, L)	10	33	9	31	24	38

Table 2.9: Probe Sets and match scores for the USF database using the baseline algorithm and our approach using width feature and entire binary silhouette.

the basic width vector using either direct smoothing and down-sampling or by way of eigenanalysis for gait representation and matching. During the matching process, the DTW algorithm was used to effectively handle unequal lengths of the reference and the probe gait sequences. Various situations such as changes in walking speed, sensitivity to viewing angle, different floor characteristics, low frame-rate etc were considered. The performance of our method was found to be quite satisfactory on all the databases. The eigensmoothed feature performed the best followed by the direct-smoothed feature vector. The velocity profile of the width vector is not sufficient to capture gait characteristics when used in isolation. Our experiments confirm that the side-view is the optimal one for capturing the gait characteristics. As the azimuth angle increases, the recognition accuracy falls. The method is capable of recognizing even frontal gait but with lower accuracy as compared to that of the side-view.

It would be interesting to analyze if the warping path characteristics of the DTW algorithm can be used to determine the direction of motion. Also, a study that combines evidences from different component level features of the human body for the purpose of recognition deserves investigation.

Chapter 3

A Framework for gait-based person identification using continuous HMMs

A careful analysis of gait would reveal that it has two important components. The first is a structural component that captures the physical build of a person e.g. body dimensions, length of limbs etc. The second component is the motion kinematics of the body during a gait cycle. Our effort in this chapter is directed towards deriving and fusing information from these two components. We propose a systematic approach to gait recognition by building representations for the structural and kinematic components of gait.

We considered two image features, one being the width of the outer contour of the binarized silhouette, and the other being the binary silhouette itself. A set of exemplars that occur during a gait cycle is derived for each individual. To obtain the observation vector from the image features we employ two different methods. In the *indirect approach* the high-dimensional image feature is transformed to a lower-dimensional space by generating the FED. The FED vector captures both structural and kinematic traits of each individual. For compact and effective gait representation and recognition, the gait information in the FED vector sequences is captured using a HMM for each individual. In the *direct approach*, we work with the feature vector directly and train a HMM for gait representation. The difference between the direct and indirect methods is that in the former the feature vector is directly used as the observation vector for the HMM whereas in the latter, the FED is used as the observation vector. In the direct method, we estimate the observation probability by an alternative approach based on the distance between the exemplars and the image features. In this way we avoid learning high-dimensional probability density functions. The performance of the methods is tested on different databases.

Section 3.1 explores the issue of feature selection. Section 3.2 describes the two algorithms. In Section 3.3 , we present experimental results and Section 3.4 concludes the chapter.

3.1 Feature Selection

An important issue in gait is the extraction of appropriate salient features that will effectively capture the gait characteristics. The features must be reasonably robust to operating conditions and should yield good discriminability across individuals. We assume that the side view of each individual is available. As discussed before, the silhouette appears to be a good feature to look at as it captures the motion of most of the body parts. While extracting this feature we are faced with two options:

1. Use the entire silhouette.

2. Use only the outer contour of the silhouette.

The choice of using either of the above features depends upon the quality of the binarized silhouettes. If the silhouettes are of good quality, the outer contour retains all the information of the silhouette and allows a representation, the dimension of which is an order of magnitude lower than that of the binarized silhouette. However for low quality, low resolution data, the extraction of the outer contour from the binarized silhouette may not be reliable. In such situations, as seen from Section 2.6.5 direct use of the binarized silhouette may be more appropriate The utility of the width feature has been discussed before in Chapter 2.

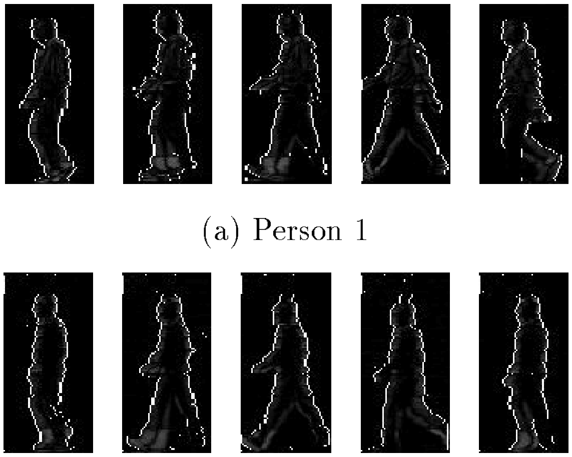
3.2 Proposed Algorithms

Given a sequence of image features for person j, $\mathcal{X}^j = \{\mathbf{x}^j(1), \mathbf{x}^j(2), \cdots, \mathbf{x}^j(T)\}$, we wish to build a model for the gait of person j and use it to recognize this person from M different subjects in the database.

3.2.1 Overview

A closer examination of the physical process behind the generation of gait signature reveals that, during a gait cycle, it is possible to identify certain distinct phases or stances. In Figure 3.1, we show five frames that we have picked from a gait cycle for two individuals.

In the first stance, the person holds the two feet together. In the second, he is just about to start and his hand is slightly raised. In the third stance, the hands and the feet are separated, while in the fourth, the hands and feet are displaced to a maximum. Finally, in the fifth stance, the person is returning to the rest state. Clearly, every person transits among these successive stances as he/she walks.



(a) Person 1

(b) Person 2

Figure 3.1: Stances corresponding to the gait cycle of two individuals.

Although, these stances are generic, there exist differences not only in their image appearance based on the physical build of an individual but also in the way an individual transits across these stances as he/she walks which represents the gait kinematics of the individual. A reasonable way to build a structural representation for a person is to pick N exemplars (or stances) $\mathcal{E} = \{\mathbf{e}_1, \cdots, \mathbf{e}_N\}$ from the pool of images that will minimize the error in representation of all the images of that person. The specifics of choice of exemplars may differ for different approaches. Given the image sequence for an unknown person $\mathcal{Y} = \{\mathbf{y}(1), \cdots, \mathbf{y}(T)\}$, these exemplars can be directly used for recognition as

$$ID = \text{arg min}_j \sum_{t=1}^{T} \text{min}_{n \in \{1, \cdots, N\}} d(\mathbf{y}(t), \mathbf{e}_n^j),$$

where $\mathbf{y}(t)$ represents the image of an unknown person at the tth time instant, while $\mathbf{e}_n^j$ represents the nth exemplar of the jth individual. Note, however, that such a simple discrimination criterion is susceptible to failures not only due to noise but more importantly due to the presence of structural similarities among people in the database. To improve discriminability, the kinematics of the data must be exploited. A closer look at the gait cycle reveals that there is a temporal progression in the proximity of the observed silhouette to the different exemplars. Note that at the start of the gait cycle, a frame is closer to the first exemplar as compared to the other four. As time progresses, the frame will be closer to the second exemplar as compared to the others and so on. A similar behavior is reflected with regard to the remaining exemplars as well. Underlying the proximity of the silhouette to the exemplars is a probabilistic dependence across the exemplars. This encompasses information about how long a gait cycle persists in a particular exemplar as well as the way in which the gait cycle transits from one exemplar to the other. For two people who are similar in physical build, this kinematic knowledge can be used to improve the recognition performance. Because the transitions are systematic, it is possible to model this probabilistic dependence by a Markov matrix as shown below.

$$A = [P(\mathbf{e}_i(t)|\mathbf{e}_j(t-1))] \tag{3.1}$$

for $i, j \in \{1, \cdots, N\}$. The matrix A encodes the kinematics in terms of state duration densities and transition probabilities. Often, in a practical situation, only a finite amount of training data is available and modeling can be difficult if the feature dimensionality is high. The dimension of the feature vectors described in the previous section is at least 100. Directly using the feature vectors to estimate the structure of the person and the kinematics of gait is clearly not advisable. We propose two different approaches to model the structure and kinematics of gait.

3.2.2 Approach 1: Indirect Approach

Gait representation

In this approach we pick N exemplars (or stances) $\mathcal{E} = \{\mathbf{e}_1, \cdots, \mathbf{e}_N\}$ from the pool of images that will minimize the error in representation of all the images of that person. If the overall average distortion is used as a criterion for codebook design, the selection of the N exemplars is said to be optimal if the overall average distortion is minimized for that choice. There are two conditions for ensuring optimality. The first condition is that the optimal quantizer is realized by using a nearest neighbor selection rule.

$$q(\mathbf{x}) = \mathbf{e}_i, \Longleftrightarrow d(\mathbf{x}, \mathbf{e}_i) \leq d(\mathbf{x}, \mathbf{e}_j), j \neq i, 1 \leq i, j \leq N,$$

where $\mathbf{x}$ represents an image in the training set, $d(\mathbf{x}, \mathbf{e}_i)$ is the distance between $\mathbf{x}$ and $\mathbf{e}_i$ while N is the number of exemplars. The second condition for optimality is that each codeword/exemplar $\mathbf{e}_i$ is chosen to minimize the average distortion in the cell C_i, i.e.,

$$\mathbf{e}_i = \text{argmin}_{\mathbf{e}} E(d(\mathbf{x}, \mathbf{e}) | \mathbf{x} \in C_i),$$

where the C_is represent the Voronoi partitions [25] across the set of training images. To iteratively minimize the average distortion measure, the most widely used method is the k-means algorithm [25, 26]. However, implementing the k-means algorithm raises a number of issues. It is difficult to maintain a temporal order of the centroids (i.e. exemplars) automatically. Even if the order is maintained, there could be a cyclical shift in the centroids due to phase shifts in the gait cycle (i.e. different starting positions). In order to alleviate these problems, we divide each gait cycle into N equal segments. We pool the image features corresponding to the ith segment for all the cycles. The centroids (essentially the mean) of the features of each part were computed and denoted as the exemplar for that part. Doing this for all the N segments gives the optimal exemplar set $\mathcal{E} = \{\mathbf{e}_1^*, \cdots, \mathbf{e}_N^*\}$.

Of course, there is the issue of picking N. This is the classical problem of choosing the appropriate dimensionality of a model that will fit a given set of observations e.g., choice of degree for a polynomial regression. The notion of 'best fit' can be precisely defined by an objective function involving a penalty for the model complexity. Examples include minimum Bayes information criterion [27], minimum description length [28] etc. Similar problems exist for the case where there exists no parametric model for the data set e.g. vector quantization. In problems like image compression, it is a common practice to look at the rate-distortion curves to examine the marginal reduction in the distortion as the bits per pixel are increased. We take a similar approach here. In Figure 3.2, we plot the average distortion as a function of the number of centroids for the UMD database. It can be observed that beyond five centroids, the average distortion does not decrease as rapidly with the increase in the number of centroids.

In order to reliably estimate the gait kinematics we propose a novel way of compactly encoding the observations observations, while retaining all the relevant information. Let $\mathbf{x}(t)$ denote the feature extracted from the image at time t. The distance of $\mathbf{x}(t)$ from the corresponding exemplars $\mathbf{e}_n \in \mathcal{E}$ can be computed to build an FED, $\mathbf{f}(t)$, which serves as a lower (N-)dimensional representation of the image at time t. For instance, for the jth individual we compute the nth element of the FED vector as

$$[\mathbf{f}_j^{\mathcal{X}^j}(t)]_n = d(\mathbf{x}^j(t), \mathbf{e}_n^j), \tag{3.2}$$

where, $t \in \{1, \cdots, T\}$, $\mathbf{e}_n^j$ denotes the nth exemplar of the jth person and

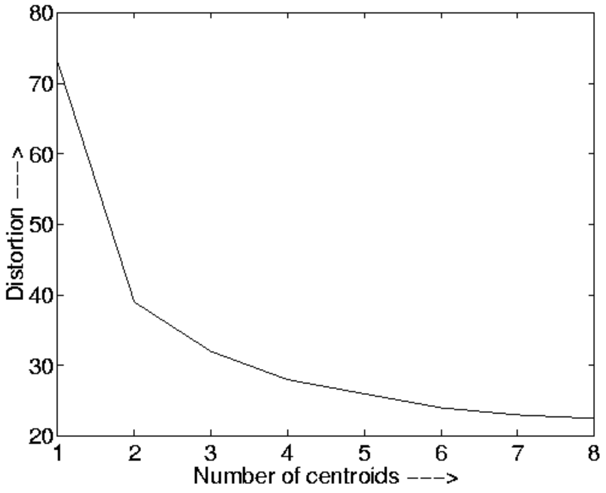


Figure 3.2: Rate-distortion curve for number of exemplars vs distortion.

$n \in \{1, \cdots, N\}$. Thus, $\mathbf{f}_j^{\mathcal{X}^j}(t)$ constitutes an observation vector for person j. Similarly, $\mathbf{f}_j^{\mathcal{X}^i}(t)$ represents the observation sequence of the person i encoded in terms of the exemplars of person j. Note that the dimension of $\mathbf{f}_j^{\mathcal{X}^j}(t)$ is only N. These observations can be derived for several such gait cycles in the database.

It is clear that as we examine a gait cycle, the proximity of a frame from each of the stances changes with time. Correspondingly, the elements of the vector $\mathbf{f}_j^{\mathcal{X}^j}(t)$ would reflect this feature. To elaborate this further, note that for a frame at the start of the gait cycle, the first element of the observation vector will be smaller in magnitude as compared to the remaining four elements. As time progresses, the first element will increase in magnitude because the frame moves closer to the second stance. The magnitude of the second element will decrease as long as the frame is close to the second stance and then it will start to increase as well. A similar behavior is observed in the rest of the elements of the vector. The duration for which an element of this vector stays low encodes the stance duration density as also the probability of transition to another stance. Figure

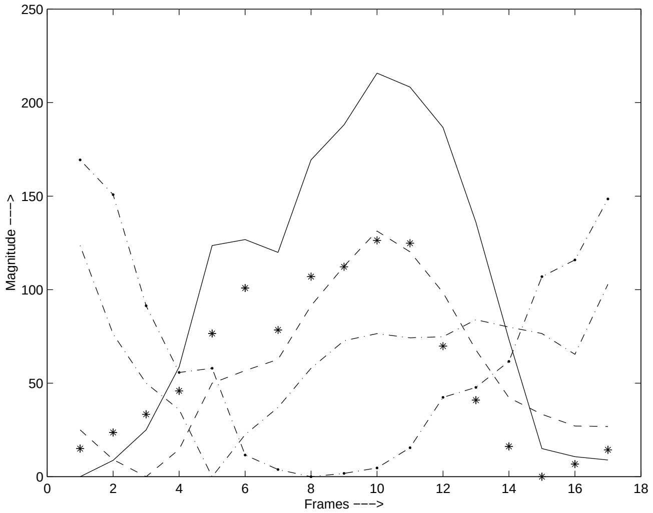


Figure 3.3: FED vector components plotted as a function of time.

3.3 shows the evolution of the different components of the FED vector $\mathbf{f}_j^{\mathcal{X}^j}(t)$ for a half-gait cycle.

As can be seen, there is a systematic succession of troughs for the different FED vector components across time. The FED vector representation is independent of the choice of features. The distance in (3.2) will change depending upon the feature. For example, for the case of the width feature, d corresponds to the Euclidean distance, whereas for the binary silhouette d corresponds to the binary correlation.

As described before, it is possible to model transition across exemplars by a Markov matrix. For the person j, it is possible to look upon the FED vector sequence $\mathbf{f}_j^{\mathcal{X}^j}(t)$ as the observed manifestation of the transition across exemplars (a hidden process). An HMM is appropriate for such a signal. HMMs use a Markov process to model the changing statistical characteristics that are manifested in the actual observations. The state sequence is *hidden*, and can only be observed through another set of observable stochastic processes. Each hidden state of the model is associated with a set of output probability distributions which can be either discrete probability mass functions or continuous probability density functions. Details on HMMs can be found in [29]. For the gait problem, the exemplars can be considered as analogues of states of the HMM while the FED vector sequence can be considered as the observed process. Since the feature vectors are transformed to the intermediate FED vector representation, we refer to this approach as an indirect approach. In the proposed model for gait, the primary HMM parameters of interest are the number of states, the initial probability vector (π), the transition probability matrix (A) and the output probability distribution B which we model as a continuous probability distribution. A brief explanation of each of these terms follows.

- Initial probability (π): The initial probability vector is given by $\pi = \{\pi_i\}$ where π_i represents the probability of being in state i at the beginning of the experiment. For the gait problem, π_i can be thought of as the probability of starting in a particular stance.

- Transition probability matrix (A): The entries of this matrix are given by a_{ij} where $a_{ij} = P(i_{t+1} = j | i_t = i)$. This represents the probability of being in state j at time $t+1$ given that the previous state was i. An HMM in which every state of the HMM can be reached from any other state, viz. every coefficient a_{ij} of A is positive, is refered to as ergodic HMM. When the coefficients have the property $a_{ij} = 0$ for $j < i$, i.e., if no transitions are allowed to states whose indices are lower than the current state, the HMM is refered to as a left-to-right model.

- Probability of observation(B): The probability of observing symbol $\mathbf{x}$ while in state j is given by $b_j(\mathbf{x})$. Since the observations in our experiment are continuous-valued, finding $b_j(\mathbf{x})$ turns out to be a problem of estimating the underlying probability density function of the observations. In the literature on HMMs, a Baum-Welch type of re-estimation procedure has been formulated [29] for a mixture of any log concave or elliptically symmetric density function (such as the Gaussian).

In this paper, $\lambda = (A, B, \pi)$ will be used to compactly represent an HMM.

Gait Recognition

The HMM model parameters $\lambda = (A, B, \pi)$ serve as a means to represent the gait of different people. For robust recognition, it is reasonable that one must examine several gait cycles before taking a decision i.e., instead of looking at a single walking cycle, it would be prudent to examine multiple cycles of a person to derive any conclusion about his/her gait. We assume that several gait cycles of an individual are given. The problem is to recognize this individual from a database of people whose models for gait are known a priori.

To begin with, the image sequence of the unknown person $\mathcal{Y} = \{\mathbf{y}(1), \cdots, \mathbf{y}(T)\}$ is subjected to the same image processing operations as the training image sequence i.e., the image feature $\mathbf{y}(t)$ of this person is generated for each frame and the FED vector $\mathbf{f}_j^{\mathcal{Y}}(t)$ is computed for all $j \in \{1, \cdots, M\}$ using (2). We wish to compute the likelihood that the observation sequence $\mathbf{f}_j^{\mathcal{Y}}$ was generated by the HMM corresponding to the jth person. This can be deciphered by using the forward algorithm [29] which computes this log probability as

$$P_j = \log(P(\mathbf{f}_j^{\mathcal{Y}}|\lambda_j)) \tag{3.3}$$

Here, λ_j is the HMM model corresponding to the person j. We repeat the above procedure for every person in the database thereby producing $P_j, j \in \{1, \cdots, M\}$. Suppose that the unknown person was actually person m. We would then expect P_m to be the largest among all P_j's. A larger value of P_m will be the result of two factors.

1. The distance between $\mathcal{Y}$ and the stances of person m will be smaller than that between $\mathcal{Y}$ and any other person.

2. The pattern of transitions between stances/states for $\mathcal{Y}$ will be closest to that for person m.

Note that the observed image sequence must be in accordance with the transition probability matrix A as well as the observation probability B in order to yield a larger value for the log-probability. If the values of $P_1, \cdots, P_M$ are observed for a sufficient number of gait cycles of the unknown person, one would expect that in a majority of cases P_m would be lower as compared to the rest of the P_is. For smaller databases, the performance can be easily examined in terms of a confusion matrix. For larger databases, a more convenient way of reporting recognition performance is to report the number of times the right person occurs in the top n matches where $n < M$ i.e. by way of cumulative match scores (CMS).

3.2.3 Approach 2: Direct Approach

Gait representation

In this approach we use the feature vector in its entirety to estimate the HMM $\lambda = (A, B, \pi)$ for each person. Hence we refer to this approach as the direct approach. One of the important issues in training is learning the observation probability B. In general, if the underlying distribution of the data is non-Gaussian, it can be characterized by a mixture of Gaussians. As discussed before, the reliability of the estimated B depends on the number of training samples available and the dimension of the feature vector. In order to deal with the high dimensionality of the feature vector, we propose an alternative representation for B.

As discussed in the previous section it is possible, during a gait cycle, to identify certain distinct phases or stances. We build a structural representation for a person by picking N exemplars (or stances) from the training sequence, $\mathcal{X} = \{\mathbf{x}(1), \cdots, \mathbf{x}(T)\}$. We now define B in terms of the distance of this vector from the exemplars as follows.

$$b_n(\mathbf{x}(t)) = P(\mathbf{x}(t)|\mathbf{e}_n) = \beta e^{-\alpha D(\mathbf{x}(t), \mathbf{e}_n)} \tag{3.4}$$

The probability, $P(\mathbf{x}(t)|\mathbf{e}_n)$ is defined as a function of $D(\mathbf{x}(t), \mathbf{e}_n)$, the distance of the feature vector $\mathbf{x}(t)$ from the n^{th} exemplar, $\mathbf{e}_n$. The motivation behind using an exemplar-based model in the above manner is that the recognition can be based on the *distance measure* between the observed feature vector and the exemplars. During the training phase, a model is built for all the subjects in the gallery. Note that B is completely defined by $\mathcal{E}$ if α and β are fixed beforehand. An initial estimate of $\mathcal{E}$ and λ is formed from $\mathcal{X}$, and these estimates are refined iteratively using Expectation-Maximization [30]. We can iteratively refine the estimates of A and π by using the Baum-Welch algorithm [29] with $\mathcal{E}$ fixed. The algorithm to refine estimates of $\mathcal{E}$, while keeping A and π fixed, is determined by the choice of the distance metric. We describe in the following sections the methods used to obtain initial estimates of the HMM parameters, the training algorithm, and, finally, identification from a probe sequence.

Initial Estimate of HMM Parameters

In order to obtain a good estimate of the exemplars and the transition matrix, we first obtain an initial estimate of an ordered set of exemplars from the sequence and the transition matrix and then iteratively refine the estimate. We observe that the gait sequence is quasi-periodic and we use this fact to obtain the initial estimate $\mathcal{E}^{(0)}$. We first divide the sequence into cycles. We can further divide each cycle into N temporally adjacent clusters of approximately equal size. We visualize the frames of the n^{th} cluster of all cycles to be generated from the n^{th} state. Thus we can get an initial estimate of $\mathbf{e}_n$ from the feature vectors belonging to the n^{th} cluster of all cycles. In order to get reliable initial estimates of the exemplars, we need to robustly estimate the cycle boundaries (see [31]). A corresponding initial estimate of the transition matrix, $A^{(0)}$ (with $A^{(0)}_{j,j} = A^{(0)}_{j,j \bmod N+1} = 0.5$, and all other $A^{(0)}_{j,k} = 0$) is also obtained. The initial probabilites $\pi^{(0)}_n$ are set to be equal to $1/N$.

Training the HMM Parameters

The iterative refinement of the estimates is performed in two steps. In the first step, a Viterbi evaluation [29] of the sequence is performed using the current values for the exemplars and the transition matrix. We can thus cluster feature vectors according to the most likely state they originated from. Using the current values of the exemplars, $\mathcal{E}^{(i)}$ and the transition matrix, $A^{(i)}$, Viterbi decoding on the sequence $\mathcal{X}$ yields the most probable path $\mathcal{Q} = \{q^{(i)}(1), q^{(i)}(2), \ldots, q^{(i)}(T)\}$, where $q^{(i)}(t)$ is the estimated state at time t and iteration i. Thus the set of observation indices, whose corresponding observation is estimated to have been generated from state n is given by $\mathcal{T}_n^{(i)} = \{t : q^{(i)}(t) = n\}$. We now have a set of frames for each state and we would like to select the exemplars so as to maximise the probability in (3.5). If we use the definition in (3.4), (3.6) follows.

$$\begin{aligned} \mathbf{e}_n^{(i+1)} &= \arg_{\mathbf{e}} \max \textstyle\prod_{t \in \mathcal{T}_n^{(i)}} P(\mathbf{x}(t)|\mathbf{e}) & (3.5) \\ \mathbf{e}_n^{(i+1)} &= \arg_{\mathbf{e}} \min \textstyle\sum_{t \in \mathcal{T}_n^{(i)}} D(\mathbf{x}(t), \mathbf{e}) & (3.6) \end{aligned}$$

The actual method for minimizing the distance in (3.6) however depends on the distance metric used. We use the inner product (IP) distance (3.7). We have experimented with other distance measures, namely the Euclidean (EUCLID) distance and the sum of absolute difference (SAD) distance [31].

$$D_{IP}(\mathbf{x}, \mathbf{e}) = 1 - \frac{\mathbf{x}^{\mathsf{T}}\mathbf{e}}{\sqrt{\mathbf{x}^{\mathsf{T}}\mathbf{x}\mathbf{e}^{\mathsf{T}}\mathbf{e}}} \qquad (3.7)$$

Note that though $\mathbf{x}$ and $\mathbf{e}$ are 2-dimensional images, they are represented as vectors of dimension $D \times 1$ for ease of notation. $\mathbf{1}_{D\times 1}$ is a vector of D ones. The equations for updating the exemplars is given by (3.8). $\tilde{\mathbf{x}}$ denotes the normalized vector $\mathbf{x}$.

$$\mathbf{e}_n^{(i+1)} = \textstyle\sum_{t \in \mathcal{T}_n^{(i)}} \tilde{\mathbf{x}}(t) \qquad (3.8)$$

The exemplars estimated for one observation sequence are displayed in Figure 3.4.

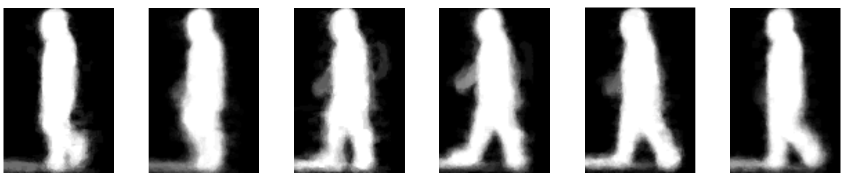


Figure 3.4: Exemplars estimated using IP distance measure

Given $\mathcal{E}^{(i+1)}$ and $A^{(i)}$, we can calculate $A^{(i+1)}$ using the Baum-Welch algorithm [29]. We set $\pi_n^{(i+1)} = \frac{1}{N}$ at each iteration. Thus we can iteratively refine our estimates of the HMM parameters. It usually takes only a few iterations to obtain an acceptable estimate.

Identifying from a Test Sequence

Given the sequence of the unknown person, $\mathcal{Y}$, and the exemplars and HMM model parameters for the different people in the database, we wish to recognize the unknown person. As before, the given image sequence of the unknown person is subjected to the same image processing operations as the training image sequence to extract the relevant image features. As explained before the likelihood that the observation sequence was produced by the jth individual in the database is computed using the forward algorithm as

$$P_j = \log(P(\mathcal{Y}|\lambda_j)). \tag{3.9}$$

Note that λ_j implicitly includes the exemplar set corresponding to person j.

3.3 Experimental Results

In this section, we demonstrate the performance of the proposed algorithms on different databases. Our experiments are aimed at finding how well the two methods perform with respect to several different variations such as size of database, speed of walking, clothing, illumination etc. We have considered normal walk as well as the treadmill data for analysis. The video sequences were taken from the CMU UMD and USF databases. For the sake of brevity, we present detailed results of the indirect approach using width vectors on the

CMU and UMD databases while the results of the direct and indirect approaches with the binarized silhouette feature will be presented for the USF database.

Silhouettes and the feature vectors for the person are extracted using the procedure described in Section 3.1. In general the choice of N depends on the frame rate. For the UMD and CMU databases, we chose N to be five. However for the USF database which has a higher frame rate, we found that $N = 6$ is a better choice. The Viterbi algorithm was used to identify the probe sequence, since it is efficient and can operate in the logarithmic domain using only additions. For every gait cycle, we rank order the probabilities and the corresponding person indices in descending order. We then evaluate performance by letting each person in the database be the unknown u and plot the fraction of times that the right person is within the top n matches as a function of n. This curve known as the cumulative match score characteristic (CMS) was first used in the context of face recognition by Philips et al. [22]. Details of these databases are given in Section 2.5.

3.3.1 The CMU Database

We did the following experiments on this database: (i) train on slow-walk and test on slow-walk, (ii) train on fast-walk and test on fast-walk, (iii) train on slow-walk and test on fast-walk, (iv) train on fast-walk and test on slow-walk (v) train on walk carrying a ball and test on walk carrying a ball. In cases (i), (ii) and (v), for each person, the sequences were divided into two halves, one half used for training and the other for testing, while in the cases (iii) and (iv), the entire slow/fast sequence was used for training and the other fast/slow sequence was used for evaluation.

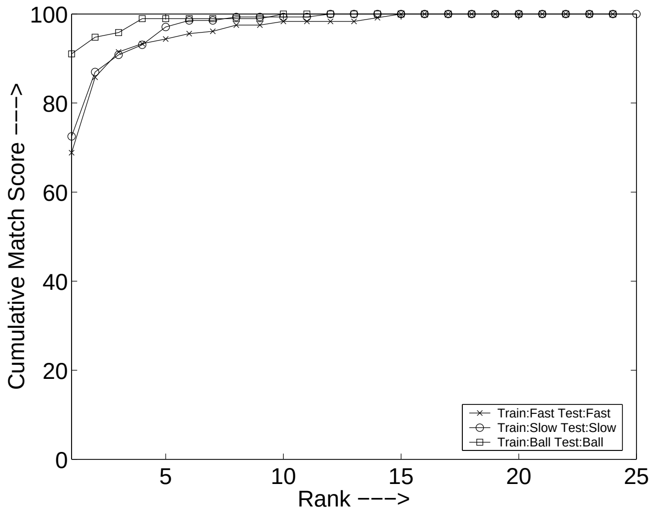


Figure 3.5: Cumulative match characteristic for normal walk and walk when carrying an object for the CMU database.

The results obtained using the proposed method are given in Figures 3.5 and 3.6.

It can be seen that the right person in the top 3 matches 90% of the times for the cases where the training and testing sets correspond to the same walking styles. Observe that the results on CMU database when the HMM is trained using cycles from slow walk and tested using cycles from fast walk, the result is poor compared to the situation when the training and testing scenarios are reversed. In an effort to understand this, we ran an experiment whereby we artificially increased the number of frames per activity cycle using interpolation and observed the resulting HMM. It was seen that the A matrix tends towards diagonal dominance. This occurs on account of the fact that the HMM does not provide adequate representation of extreme temporal durations of activity. The probability of t consecutive observations in state i can be written as

$$d_i(t) = a_{ii}^t(1 - a_{ii})$$

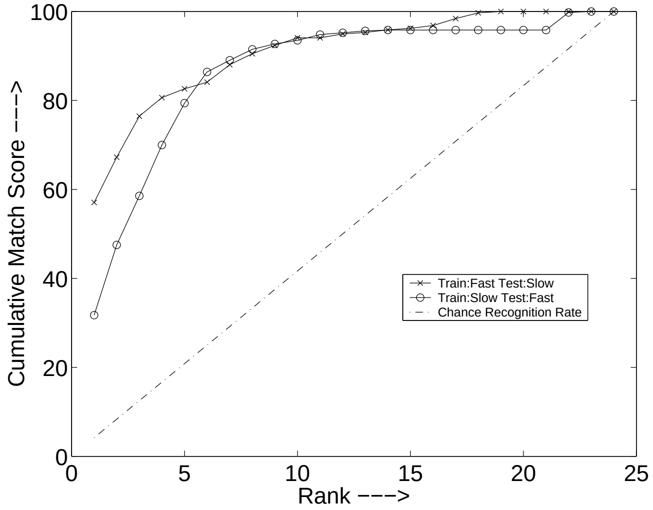
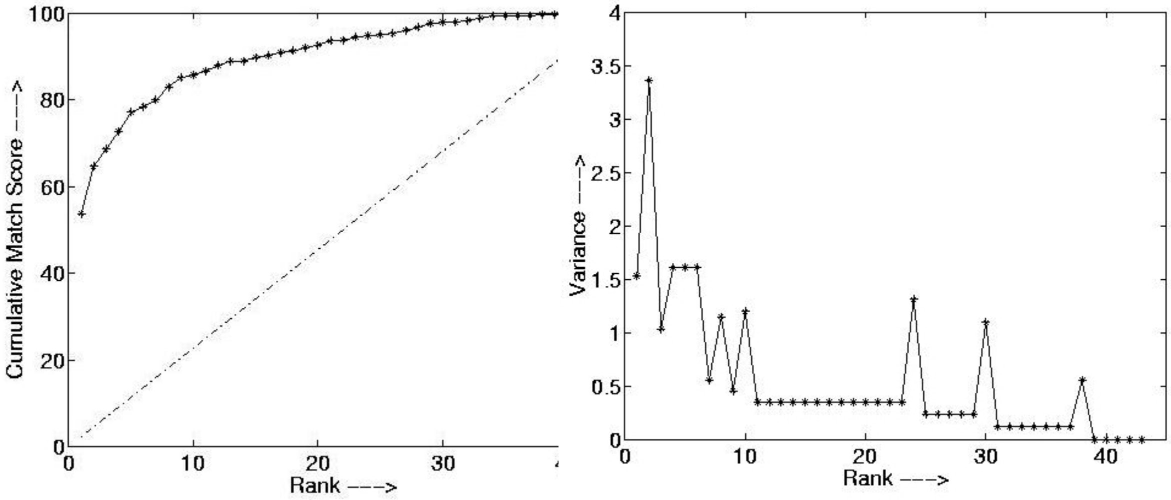


Figure 3.6: Cumulative match characteristic for across speed testing for the CMU database.

where $d_i(t)$ is the probability of taking a self-loop at state i for t times viz. a geometric distribution. As the duration of the activity $T \rightarrow \infty$ for a fixed N, this causes $a_{ii} \rightarrow 1$ and $a_{ij} \rightarrow 0$. Clearly, the geometric distribution does not represent a realistic description of the state duration density in the gait-modeling problem. Similar issues have been raised in the context of speech recognition and a solution is to explicitly model the distribution of state duration as has been done by Russell [32]. For the case of training with fast-walk and testing on slow-walk, the dip in performance, similar to the appearance matching case (see Figure 2.7(a) and (b)) is caused due to the fact that for some individuals as biomechanics suggests, there is a considerable change in body kinematics and stride length as a person changes his speed.

When the subjects are walking with a ball in their hands, most of the gait dynamics are confined to the leg region. For this experiment i.e. case (v), we observe from Figure 3.5, that the top match is the correct match 90% of the time which is higher than the top match score (around 70%) in the normal walk



(a) Cumulative Match Characteristic

(b) Recognition Confidence

Figure 3.7: Results for the UMD database using Algorithm 1 (44 people).

cases. This observation is in line with the DTW method as well (Section 2.6.2).

3.3.2 The UMD Database

The result using the proposed method for the UMD database is shown in Figure 3.7(a).

In Figure 3.7(a), the dashed-dotted line represents the chance recognition rate which corresponds to the line connecting $\left(1, \frac{100}{P}\right)$ and $(P, 100)$, where P denotes the number of people in the gallery. In order to assess the confidence in our recognition capability, we computed the recognition performance by dropping one subject from the gallery and the probe when we compute the cumulative match characteristic. Essentially this amounts to computing the CMS characteristics by eliminating row i and column i from the similarity matrix. By this procedure, we ensure that no individual in the gallery leads to a bias in the performance computation. Having computed the CMS characteristics thus, we find the variance in the cumulative match score at every rank. Smaller the value of this variance, the more reliable is the gait recognition performance. This is shown for the UMD database in Figure 3.7(b).

3.3.3 The USF Database

Finally, we consider the USF database[1] which has been identified as the gait challenge database [1]. The database has variations as regards viewing direction, shoe type, surface type. Also the subjects were asked to carry a briefcase for one testing condition. We present the results of both our methods and a comparative analysis on this dataset. Different probe sequences for the experiments along with the cumulative match scores are given in Table 3.1 for the baseline algorithm [23], our direct and indirect approaches The image quality for the USF database is worse than the previous two databases in terms of resolution and amount of noise. We experimented with both the width feature as well as the binarized silhouette for the USF dataset. However, the extraction of the outer contour in this case is not reliable and the width vectors were found to be noisy. In Table 3.1, we report only the results of our methods using the silhouettes as the image feature. G and C indicate grass and concrete surfaces, A and B indicate shoe types and L and R indicate left and right cameras, respectively. From Table 3.1 we observe that the direct method is more robust to the presence of noise than the indirect method. We also note that the recognition performance suffers most

[1] More details about this database can be found at http://figment.csee.usf.edu/GaitBaseline/

Experiment (Probe)	Baseline		Indirect Approach		Direct Approach	
	Rank 1	Rank 5	Rank 1	Rank 5	Rank 1	Rank 5
A (G, A, L)	79	96	91	100	99	100
B (G, B, R)	66	81	76	81	89	90
C (G, B, L)	56	76	65	76	78	90
D (C, A, R)	29	61	25	61	35	65
E (C, B, R)	24	55	29	39	29	65
F (C, A, L)	30	46	24	46	18	60
G (C, B, L)	10	33	15	33	24	50

Table 3.1: Probe Sets and match scores for the USF database using the baseline algorithm and our indirect and direct approaches.

due to differences in surface and background characteristics, and least due to difference in viewing angle. Results from other research groups using this data can be found in [33] and websites (http://degas.umiacs.umd.edu/links.html). From Tables 2.9 and 3.1, we note that the HMM approach indeed surpasses the performance using the appearance matching method. Recently, the gallery in the USF database was extended by adding subjects who walked with only one shoe type on grass, which happened to be labelled as Shoe B. Since the shoe type labeling is arbitrary, they were put in the gallery to increase the gallery size to 122. In Table 3.2 we present results using the new gallery. In this table $t2$ refers to the probe sequences captured in November and BF represents the people walking with a briefcase. The corresponding CMS curves are shown in Figure 3.8.

3.4 Conclusions

In this chapter, we have presented two approaches to represent and recognize people by their gait. The width of the outer contour of the binarized silhouette

Experiment (Probe)	Baseline	HMM
A (G, A, L)	73	89
B (G, B, R)	78	88
C (G, B, L)	48	68
D (C, A, R)	32	35
E (C, B, R)	22	28
F (C, A, L)	17	15
G (C, B, L)	17	21
H (G, A, R, BF)	61	85
I (G, B, R, BF)	57	80
J (G, A, L, BF)	36	58
K $(G, A, R, t2)$	3	17
L $(C, A, R, t2)$	3	15

Table 3.2: Recognition Scores for Baseline and HMM (122 people).

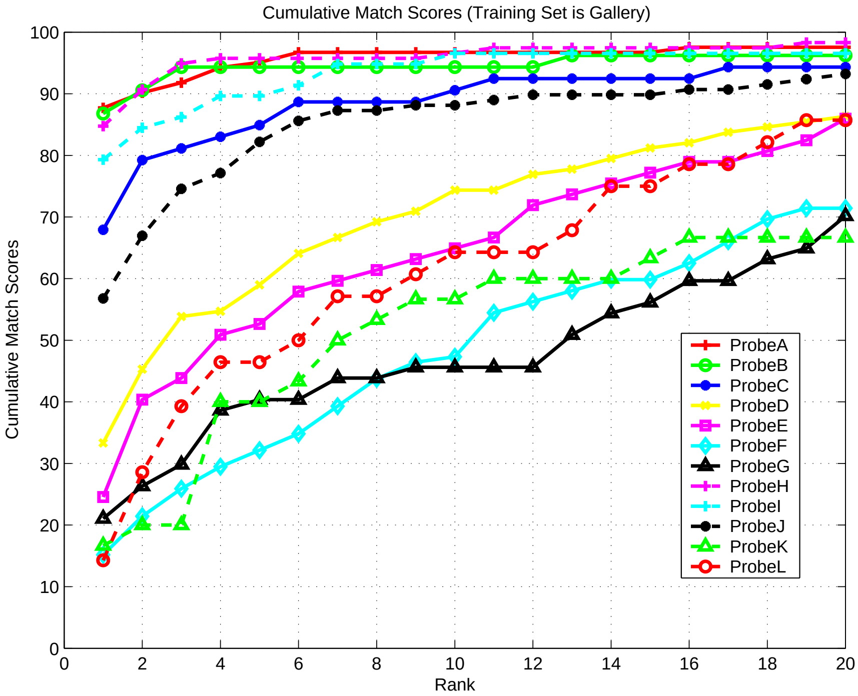


Figure 3.8: Cumulative match characteristic for USF database using our approach.

as well as the silhouette itself were used as features to represent gait. In one approach, a low-dimensional observation sequence is derived from the silhouettes during a gait cycle and an HMM is trained for each person. Gait identification is performed by evaluating the probability that a given observation sequence was generated by a particular HMM model. In the second approach the distances between an image feature and exemplars were used to estimate the observation probability B. The performance of the methods was illustrated using different gait databases.

Chapter 4

Fusion of multiple evidences for gait recognition

Consider a person walking at a sufficiently large distance so that the face is not clearly visible to the naked eye. To recognize the person, we may try to combine information such as posture, arm/leg swing, hip/upper body sway or some unique characteristic of that person. Generally speaking, information may be fused in two ways. The data available may be fused and a decision can be made based on the fused data or each signal/feature can be matched separately, using possibly different techniques and the decisions made may be fused. The former is called data fusion while the latter is referred to as decision fusion. Kokar et al. [34] have shown that decision fusion is a special case of data fusion.

In this chapter, we investigate different techniques to combine classification results of multiple evidences extracted from the gait sequences and demonstrate the improvement in recognition performance. For gait signatures, we use established techniques such as DTW and HMMs discussed in Chapters 2 and 3. The fusion methodologies can be grouped into three categories

1. Holistic fusion: This involves fusing evidence extracted from the entire silhouette. In this strategy we investigate the swing in the hands and legs. Since gait is not completely symmetric in that the extent of forward swing

of hands and legs is not equal to the extent of the backward swing, we build the left and right projection vectors. We also consider the issue of foot dominance.

2. Parts-based fusion: This involves fusing evidence from parts of the silhouette. Specifically we consider evidence obtained from the leg portion of the silhouette alone and fuse it with the height information.

3. Multicamera fusion: This involves fusing information from different cameras. Specifically we consider fusing information from the front and side views of the subject.

We characterize the performance of the recognition system based on CMS [22] computed using the aforementioned matrix of similarity scores. As in any recognition system, we would like to obtain the best possible performance in terms of recognition rates. Combination of evidences obtained is not only logical but also statistically meaningful. We show that combining evidence using simple strategies such as SUM, PRODUCT and MIN rules improves the overall performance.

4.1 Methodology

As before, background subtraction is used to extract the moving person and the resulting motion image is binarized. All the features of interest are extracted from the aforementioned sequences of binarized images. In what follows, unless stated otherwise, gait sequence refers to the sequence of binarized images of the subject. Different strategies such as Sum, Product and MIN rules [35], as applicable in each of the cases are used. Let $s^c_{i,j}$ represent the similarity score between the i^{th} subject in the probe set and the j^{th} subject in the gallery set for

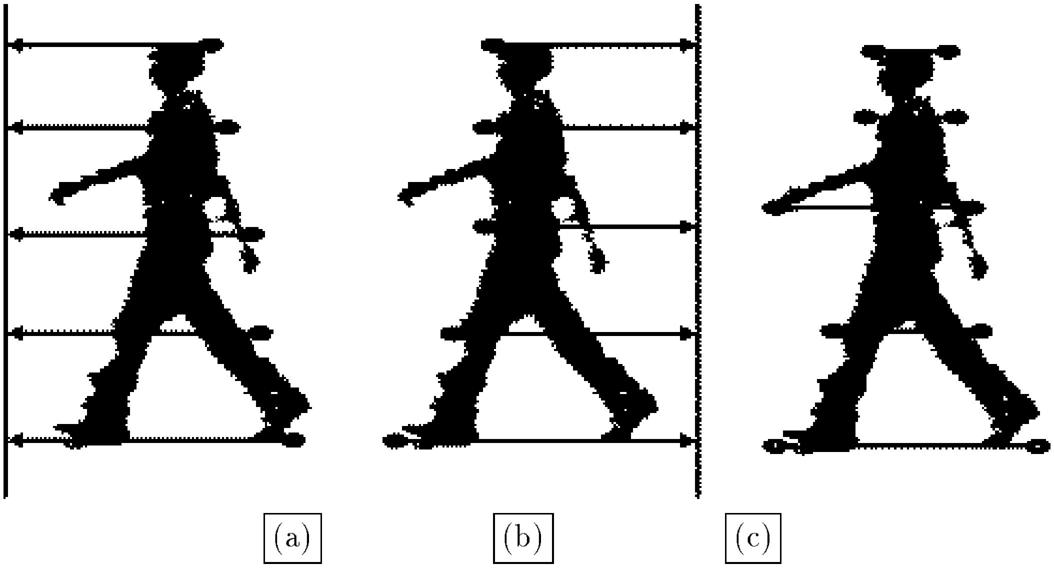


Figure 4.1: Illustrating the generation of (a) left projection vector, (b) right projection vector and (b) width vector.

$c = 1, \ldots, C$ and C be the number of classifiers. Let $\hat{s_{i,j}}$ be the overall similarity score.

- SUM rule: $\hat{s_{i,j}} = \sum_{c=1}^{C} s_{i,j}^{c}$
- PRODUCT rule: $\hat{s_{i,j}} = \prod_{c=1}^{C} s_{i,j}^{c}$
- MIN rule: $\hat{s_{i,j}} = min_{c=1}^{C} s_{i,j}^{c}$

As noted in [35] if the estimation errors of the different classifiers are assumed to be uncorrelated and unbiased, then variance reduces to $\hat{\sigma}^2 = \sigma_e^2/C$.

4.1.1 Holistic fusion

Previously we considered the width of the outer contour of the silhouette as a feature. The temporal plot of width captures the total motion of the body parts within a gait cycle. Note however that the this total motion has two components to it, namely the forward and backward swings of the body. It would be interesting to study if there is something to be gained by independently studying the forward and backward swings. In order to study this we built the right and left projection vectors. The method for constructing the projection vectors is described next:

Given a binarized image, we first align the box so that the subject is in the center of the bounding box. Traversing along each row, starting from the rightmost column, we identify the first foreground pixel (allowing for adequate resilience to noise). This constitutes the elements of 'left projection vector' (see Figure 4.1(a)). Similarly, a 'right projection vector' (see Figure 4.1(b)) is built by starting the traversal from the left-most column. Each frame is thus represented by means of a left and a right projection vector.

The pair of left and right projection vectors is extracted for each of the gallery subjects. When an unknown test sequence comes in, we extract its left and right projection vectors as outlined above. In order to identify the unknown subject, his/her projection vectors are matched with those of the gallery subjects. In order to deal with speed changes during normal walk the DTW framework discussed in Chapter 2 is used. Dynamic Time Warping provides for such a mathematical framework [36] in that it allows for non-linear time normalization. We form two matrices of similarity scores by matching the left and right projection vectors in the gallery (reference/training) with those in the probe (testing) set, separately. The SUM rule is used for combining the similarity scores.

The second aspect of the total body motion is concerned with foot dominance. Like hand dominance (right/left handedness), foot dominance (right/left leggedness) also exists. While matching therefore, we may assume that improperly aligned (i.e. right/left leg forward) reference and test sequences affects the performance. This is an issue because it is not possible to distinguish between the left/right limbs from the 2-D binarized silhouettes. Suppose there are five (half-) cycles in both the gallery and probe sequences for a particular subject. To account for foot-dominance, we match the first four half-cycles of the two sequences and generate a matrix of similarity scores. Then, we match the gallery sequence with a phase-shifted probe sequence to generate another matrix of similarity scores. Of the two phase-shifted test sequences, only one can provide a match that is in-phase unless the subject does not exhibit foot dominance. Without loss of generality, we may assume that foot dominance exists in all subjects. Then one of the two test sequences is a better match unless corrupted by noise. Therefore, the two similarity scores are combined using the MIN rule.

4.1.2 Parts-based fusion

Previously, both the hands and legs were considered while selecting the features. If the movement of the hands is restricted (if the subject is carrying an object in his/her hands) or if the torso portion of the sequence is unreliable either due to a systematic failure in background subtraction or due to image synthesis (to be discussed in Chapter 5), then leg dynamics carries information about the subject's gait. We construct the width vector (width of the outer contour of the binarized silhouette) of size $N \times 1$ from each of the images of size $N \times M$ in the sequence, as illustrated in Figure 4.1(c). All the width vectors corresponding to several training gait cycles of an individual are used to construct the covariance or scatter matrix for that individual. Principal components analysis is then performed to derive the eigenvectors for this data. Given the width vectors

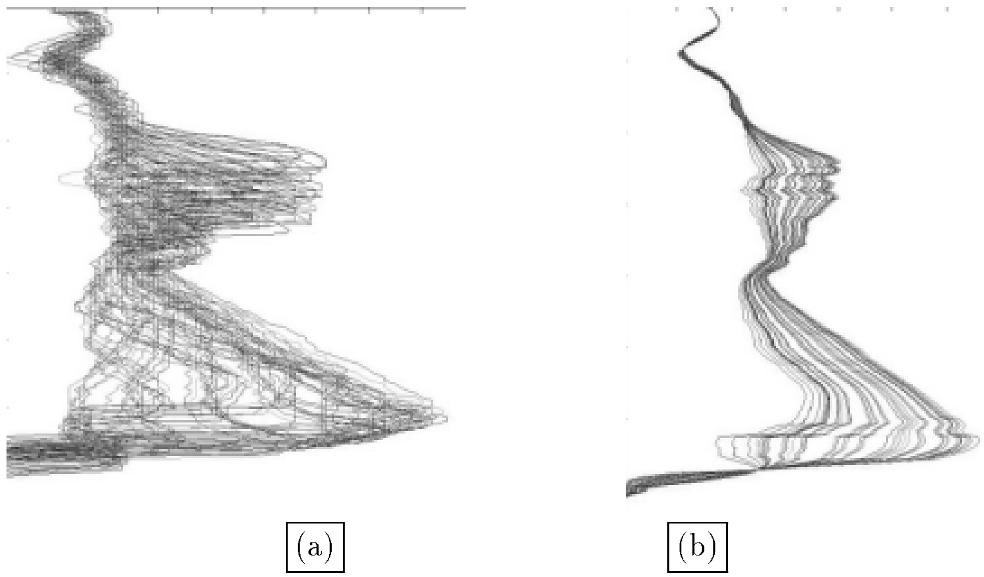


Figure 4.2: Effect of eigen decomposition and reconstruction on the width vectors. (a) Overlapped raw width vectors (b) Smoothed width vectors. Notice that the leg region (the bottom half of the figures) contain a significant portion of the dynamics.

$\{W(1), \cdots, W(N)\}$,for the N frames $W(.) \in R^M$, we compute the eigen vectors $\{V(1,)\cdots, V(M)\}$ corresponding to the eigen values of the scatter matrix arranged in the descending order and reconstruct the corresponding width vectors using $m(< M)$ most significant eigen vectors as

$$W_r(i) = (\sum_{j=1}^{m} w_j V(j)) + \bar{W}$$

where $w_j =< W(i), V(j) >$ and $\bar{W} = \frac{W(1)+\cdots+W(N)}{N}$. As seen before $m = 2$ yields good recognition performance. Figure 4.2 illustrates the effect of eigensmoothing on the gait sequence using two eigenvectors.

A cursory examination of the width vectors suggests that the leg region may exhibit a more consistent pattern compared to other parts of the body such as the arms. At the same time, the gross structure of the body, as contained in the say, the height is also useful in discrimintating between subjects. While leg dynamics concentrates on the variation of the width vector in the horizontal direction in the leg region alone, the height of the subject varies in an orthogonal direction. The width vector is truncated so that only the information about the leg is retained. This sequence of truncated width vectors is the first feature set, say set $\mathcal{A}$. We estimate the height of the subject from the image sequence using robust statistics. The estimated height of the individuals forms the second feature set, say set $\mathcal{B}$. The Euclidean distance measure is used to compare the feature set $\mathcal{B}$ of estimated height of the subjects in the probe and gallery sets.

To compare the truncated width vectors that contain information about leg dynamics, we use the indirect HMM approach described in Chapter 3. The sequence of images is projected on to the exemplar set creating the FED representation for each frame and the sequence of resulting FED vectors is used to train an HMM model using the Baum-Welch algorithm. model $\lambda = (\mathcal{A}, \mathcal{B}, \Pi)$, in a qualitative sense, describe respectively the time of persistence in a stance, the description of the stances and the initial probability distribution of the stances. Specifically, $\mathcal{A}$ denotes the transition probability matrix, $\mathcal{B}$, the observation probability of each of the states and Π denotes the initial probability distribution of the states. The Viterbi algorithm is used in the evaluation phase to compute the forward probabilities. The absolute values of the log probability values are recorded as the similarity scores.

If the decisions made are statistically independent, conditioned on the feature sets then we may write the final error probability $\mathcal{P}_\rceil = \prod_{c=1}^{C} \mathcal{P}_\rceil^\rfloor$. In practice, however it is difficult to validate this assumption. Instead, we use the low correlation of decisions across feature sets as corroboration to the hypothesis that the errors in the two feature sets, the leg dynamics and the height, are uncorrelated. We use the PRODUCT rule to combine the scores to compute the overall similarity scores.

4.1.3 Multicamera fusion

Hithero we have studied gait in its canonical view so that the apparent motion of the walking subject is maximal. This does not preclude the possibility of using other views ranging from the frontal view to any arbitrary angle of viewing. Even in the frontal view where the apparent leg/arm swing is the least, there may be several cues that can be used toward human recognition. More specifically, the head posture, hip sway, oscillating motion of the upper body among other features may be useful for recognition. As before, to focus our attention on gait, we extract the outer contour of the subject from the binarized gait sequence in the form of the width vector

Figure 4.3(a) shows the variation of the width vector (that reflects in the outer contour) as a function of time. The size of the width vector grows indicating that the subject is approaching the camera. We retain the last four half cycles of each sequence as shown in Figure 4.3(b), and the raw width vector itself is used for matching. To account for the apparent change in the size of the subject, we normalize the width vectors by computing an appropriate scale factor. The positions of the head and the feet (top and bottom pixels) are identified in each frame and smoothed using a median filter. To the resulting sinusoidal patterns shown in Figure 4.3(c), two straight lines, one to the top and one to the bottom, are fit. The distance between the two lines in each frame is used to compute the normalizing factor. The normalized plot is shown in Figure 4.3(d).

For matching these sequences, we use the DTW technique. When both the

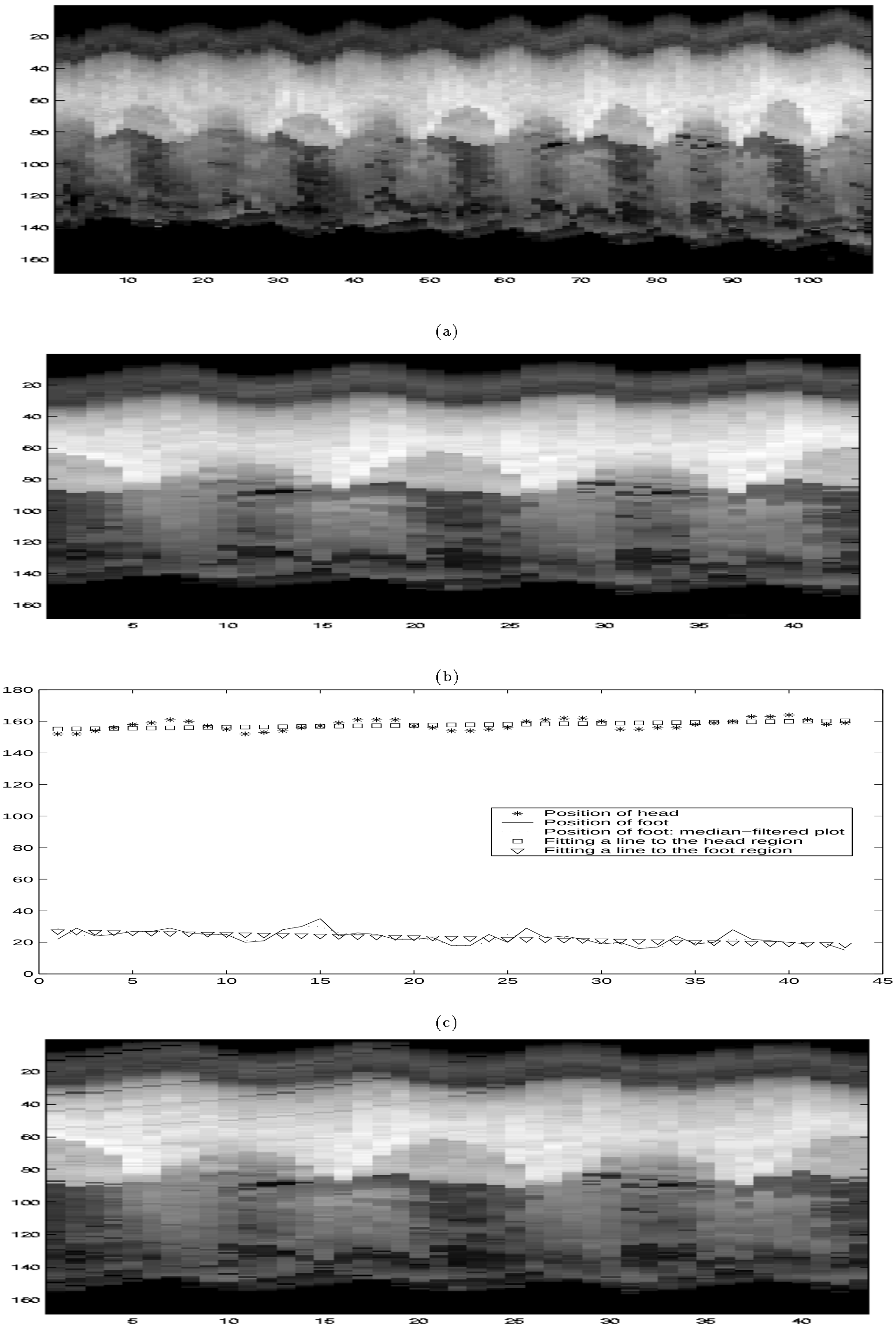


Figure 4.3: Outer contour plot for a frontal gait sequence in the UMD datasbase. (a) Subject walking towards the camera (b) 4 half-cycles chosen for recognition (c) The positions of head and foot in each frame is located to compute the normalizing factor. Two straight lines are then fit to the trajectories of the head and foot. (d) The 4 half-cycles normalized.

frontal and fronto-parallel (side) gait sequences are available, it is natural to combine these two orthogonal views before making the final decision about the identity of the subject. One way to combine multiple views is through the use of 3-D models. Currently, 3-D models have been built using sequences captured inside the lab under controlled conditions. [37] takes the visual hull approach while Bobick et al. extract parameters insensitive to the angle of viewing [38]. We adopt the decision fusion approach and combine the matching scores obtained from the frontal and side gait sequences using the SUM rule.

4.2 Experiments

We performed experiments on the UMD, CMU, MIT and USF databases (Section 2.5). We consider the use of left and right projection vectors first. As discussed before, the similarity scores obtained by matching the left and right projection vectors (separately) are combined to produce the overall matching score. Table 4.2 demonstrates the increase in recognition rate when the results are combined using the Sum rule. Data fusion is used to form the width vector by combining the two projection vectors and four half cycles are used for matching. Using the width vector for matching produces recognition rates that are in general, slightly lesser than those obtained by combining the similarity scores of the two projection vectors. However, comparison of the width vectors requires fewer number of computations since it has N components whereas the two projection vectors together have $2N$ components per frame.

We observe in Table 4.2 that the right projection vector performs better than its left counterpart suggesting that, at least in this dataset, the forward swing carries a more consistent gait signature. We also considered the effect of

Feature	CMS at rank 1	CMS at rank 5
Left projection vector	64	81
Right projection vector	65	81
Fusion	86	89

Table 4.1: CMS scores: Combining left and right projection vectors. UMD Dataset.

Experiment (Probe)	Baseline	Width Vector		Fusion
		Left	Right	
A (G, A, L)	79	19	39	53
B (G, B, R)	66	25	35	42
C (G, B, L)	56	10	15	20
D (C, A, R)	29	7	10	21
E (C, B, R)	24	7	12	19
F (C, A, L)	30	5	10	17
G (C, B, L)	10	4	6	9

Table 4.2: Probe Sets and match scores for the USF database using the baseline algorithm and our approach using width feature and entire binary silhouette.

$Feature\backslash Rank$	1	2	3	4	5
Set of 4 half cycles 1(a)	68.1	77.2	84.0	84.0	84.0
Set of 4 half cycles shifted by one 1(b)	70.4	79.5	81.8	86.3	86.3
Minimum of 1(a) and 1(b)	79.0	81.4	83.2	86.0	86.0

Table 4.3: UMD database: CMS for the first five ranks for the case of 2 full-cycles having a relative phase shift of 1 half-cycle.

Feature	CMS at rank 1	CMS at rank 5
Leg dynamics	91	100
Fusion: leg dynamics and height	96	100

Table 4.4: Cumulative match scores at rank 1 and rank 5 for CMU dataset: Combining leg dynamics and height using Sum rule

foot-dominance on the UMD database. As can be seen from Table 4.3, fusing evidences from two phase shifted gait sequences does give an improvement in the recognition performance.

Table 4.2 shows the results of combining recognition scores of matching leg dynamics and the overall physical feature, height. Clearly, the leg region retains information that is useful for the purpose of recognition. Tables 4.2 and 4.2 present the recognition scores for the UMD and MIT datasets respectively.

Finally we analyze the effect of multi-camera fusion. We considered the UMD

Feature	CMS at rank 1	CMS at rank 5
Leg dynamics	31	65
Fusion: leg dynamics and height	49	72

Table 4.5: Cumulative match scores at rank 1 and rank 5 for UMD dataset: Combining leg dynamics and height using Sum rule.

Evaluation Scheme	CMS at rank 1	CMS at rank 3
Day 1 vs. Days 2,3,4	29	50
Day 2 vs. Days 1,3,4	50	100
Day 3 vs. Days 1,2,4	20	54
Day 4 vs. Days 1,2,3	30	52

Table 4.6: Cumulative match scores at rank 1 and rank 3 for MIT dataset: Combining leg dynamics and height by adding the similarity scores.

Feature	CMS at rank 1	CMS at rank 5
Frontal Gait	52	74
Side gait	66	86
Frontal and side	85	95

Table 4.7: Cumulative match scores at rank 1 and rank 5 for the UMD dataset: effect of frontal and side gait fusion

database for this experiment. As described in Section 4.1.3, the width vectors for the frontal gait are normalized appropriately. The recognition rates for the frontal gait are shown in Table 4.2. Note that the accuracy for the frontal gait sequences is lower as compared to the side-view case. This is because the swing is less pronounced now, and the dynamics are not as effectively captured as in the case of the side-view. Note that the numbers quoted for the side-view are prior to normalization for foot-dominance. Thus fusion again helps to improve the recognition scores.

4.3 Conclusions

Different features that affect gait such as the swing of the hands and legs, the sway in the body as observed in frontal gait, static features like height were systematically analyzed. Dynamic time warping and HMMs were chosen for matching. The matrices of similarity scores between the gait sequences in the gallery and probe sets were computed. Sum, Product and MIN rules were used to combine the decisions made using the separate features. As expected, the overall recognition performance improved due to fusion. Experimental results using four different datasets, with each dataset presenting different types of challenges, have been presented.

Chapter 5

View Invariant Gait Recognition

5.1 Introduction

The gait of a person is best reflected when he/she presents a side view (referred to in this chapter as a canonical view) to the camera. Hence, most of the above gait recognition algorithms rely on the availability of the side views of the subject. The situation is analogous to face recognition where it is desirable to have frontal views of the person's face. In realistic scenarios, however, gait recognition algorithms need to work in a situation where the person walks at an arbitrary angle to the camera. The most general solution to this problem is to estimate the 3-D model for the person. Features extracted from the 3-D model can then be used to provide the gait model for the person. This problem requires the solution of the structure from motion (SfM) or stereo reconstruction problems [39, 40], which are known to be hard for articulating objects. In the absense of methods for recovering accurate 3-D models, a simple way to exploit existing appearance based methods is to synthesize the canonical views of a walking person. In [41], Shakhnarovich et al. compute an image based visual hull from a set of monocular views which is then used to render virtual canonical views for tracking and recognition. Gait recognition is achieved by matching a set of image features based on moments extracted from the silhouettes of the synthesized probe video to the gallery. An alternative to synthesizing canonical views is the work of Bobick and Johnson [42]. In this work, two sets of activity-specific static and stride parameters are extracted for different individuals. The expected confusion for each set is computed to guide the choice of parameters under different imaging conditions (viz. indoor vs outdoor, side-view vs angular-view etc). A cross-view mapping function is used to account for changes in viewing direction. The set of stride parameters (which is smaller than the set of static parameters) is found to exhibit greater resilience to viewing direction. Representation using such a small set of parameters may not give good recognition rates on large databases.

In this chapter we present a view-invariant gait recognition algorithm for the single camera case. We show that it is possible to synthesize a canonical view from an arbitrary one without explicitly computing the 3-D depth. Consider a person walking along a straight line which subtends an angle θ with the image plane (AC in Figure 5.2). If the distance, Z_0, of the person from the camera is much larger than the width, ΔZ, of the person, then it is reasonable to replace the scaling factor $\frac{f}{Z_0+\Delta Z}$ for perspective projection by an average scaling factor $\frac{f}{Z_0}$. In other words, for human identification at a distance, we can approximate the actual 3-D human as a planar object. Assume that we are given a video of a person walking at a fixed angle θ (Figure 5.2). We show that by tracking the direction of motion, α, in the video sequence, we can estimate the 3-D angle θ. This can be done by using the optical flow based SfM equations. Using the planarity assumption, knowing angle θ and the calibration parameters, we can synthesize side-views of the sequence of images of an unknown walking person

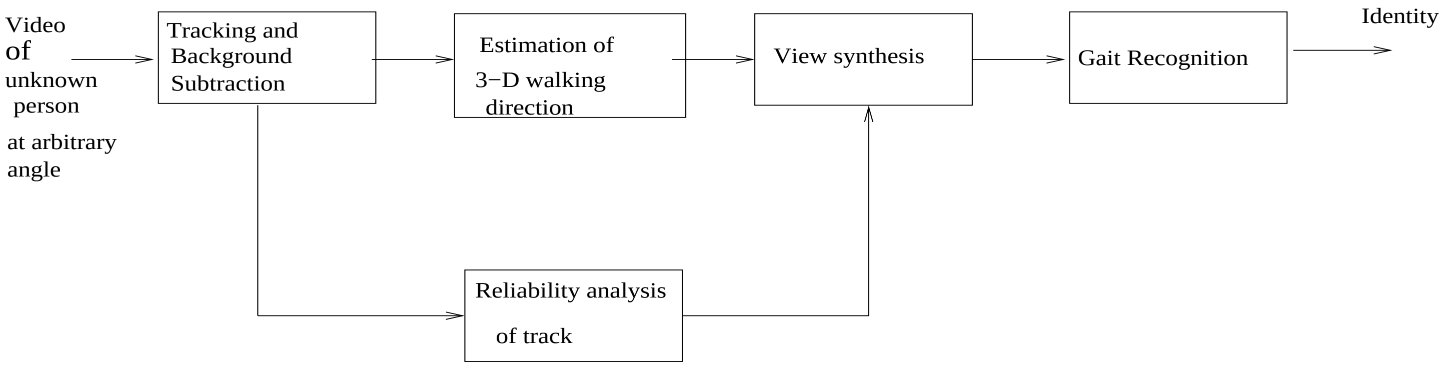


Figure 5.1: Framework for View Invariant Gait Recognition

without explicitly computing the 3D model of the person. We refer to this approach as the “implicit SfM” approach. In the case where there is no real translation of the person e.g. person walking on a treadmill, an alternative approach is employed to obtain the synthesized views of the person. Given a set of point correspondences for a planar surface between the canonical and non-canonical views in a set of training images, we compute a homography. This homography is then applied to the binary silhouette of the person to obtain the synthesized views. We refer to this approach as the “homography approach”.

An overview of our gait recognition framework is given in Figure 5.1. We present recognition performance using three gait databases (UMD3, NIST and CMU). The implicit SfM approach is used for the UMD3 and NIST databases while the homography approach is used for the CMU database. Keeping in view the limited quantity of training data, the DTW algorithm [19] is used for gait recognition. A by-product of the above method is a simple algorithm to synthesize novel views of a planar scene.

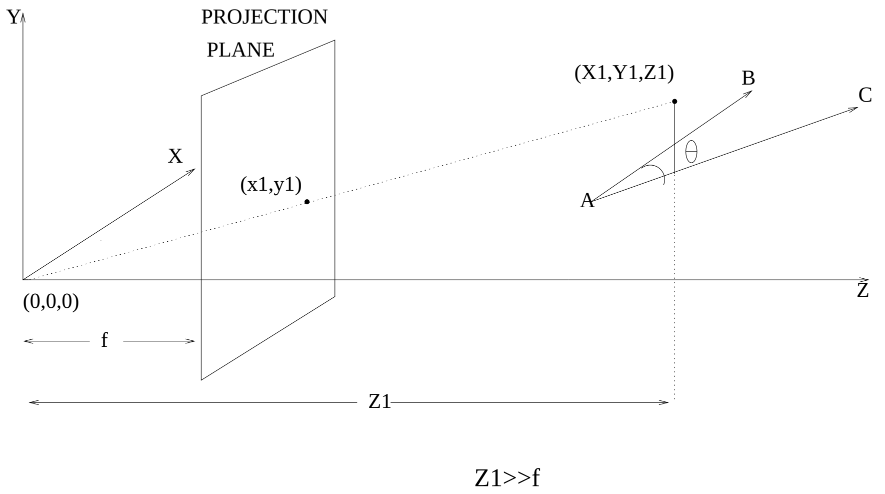


Figure 5.2: Imaging Geometry

5.2 Framework for view-invariant gait recognition

The imaging setup is shown in Figure 5.2. The coordinate frame is attached rigidly to a camera with the origin at the center of perspective projection and the Z-axis perpendicular to the image plane. Assume that the person walks with a translational velocity $\mathbf{V} = [v_X, 0, v_Z]^T$ along the line AC. The line AB is parallel to the image plane XY and this is the direction of the canonical view which needs to be synthesized. The angle between the straight line AB and AC, i.e. θ, represents a rotation about the vertical axis referred to as the azimuth angle. In what follows, we will use the notation that $[X, Y, Z]$ denotes the coordinates of a point in 3D and $[x, y]$ its projection on the image plane.

5.2.1 Tracking

We assume that only one subject is present in the field of view and the availability of a high level motion detection module which identifies abrupt changes in direction of motion and segments of *approximately* constant heading directions.

Assuming that we can find the location (x_{ref}, y_{ref}) of the persons head at the start of such a segment, we use a sequential Monte Carlo particle filter [43] to track the head of the person to get $\{(x^i(t), y^i(t)), w^i(t)\}$ where the superscript denotes the index of the particle and $w^i(t)$ denotes the probability weight for the estimate $((x^i(t), y^i(t))$.

5.2.2 Estimation of 3-D Azimuth Angle

Assume that the motion between two consecutive frames in the video sequence is small. Using the optical flow based SfM equations, let $p(x(t), y(t))$ and $q(x(t), y(t))$ represent the horizontal and vertical velocity fields of a point $(x(t), y(t))$ (e.g. centroid of the head) in the image plane where t denotes time. Since we consider a straight line motion along AC, p and q are related to the 3D object motion and scene depth by [44]

$$\begin{aligned} p(x(t), y(t)) &= (x(t) - f x_f(t))h(x(t), y(t)) \quad (5.1) \\ q(x(t), y(t)) &= y(t)h(x(t), y(t)), \quad (5.2) \end{aligned}$$

where f denotes the focal length, $h(x(t), y(t)) = v_Z(t)/Z(x(t), y(t))$ is the scaled inverse scene depth and $x_f(t) = cot(\theta(t)) = \frac{v_X}{v_Z}(t), y_f(t) = \frac{v_Y}{v_Z}(t)$ is the focus of expansion (FOE). When $v_Z = 0$ but $v_X \neq 0$, we see that $\theta = 0$, i.e. the canonical direction of walk, AB. Also, in this case $q(x, y) = 0$.

For the constant velocity models, $v_Z(t) = v_Z(\neq 0)$ and $v_X(t) = v_X(\neq 0)$, $cot(\theta(t)) = \frac{v_X}{v_Z}$. In this case, given the initial position of the tracked point (x_{ref}, y_{ref}), it can be shown that

$$cot(\alpha)(x_{ref}, y_{ref}) = \frac{p(x(t), y(t))}{q(x(t), y(t))} = \frac{x_{ref} - f cot(\theta)}{y_{ref}}. \quad (5.3)$$

For a derivation of (5.3) see the Appendix. Thus, given f and (x_0, y_0), the azimuth angle θ can be computed as

$$cot(\theta) = \frac{x_{ref} - y_{ref} cot(\alpha(x_{ref}, y_{ref}))}{f}, \tag{5.4}$$

Knowing (x_0, y_0),$cot(\alpha)$ and θ, f can be computed as part of a calibration procedure.

5.2.3 Statistical Error Analysis of Azimuth Estimation

Let $f_X(x)$ denote the distribution of a random variable X. Defining $r = cot(\alpha)$, from (5.3) we have $r = \frac{p}{q}$. In order to obtain the distribution of r we define an auxiliary random variable $s = q$. Then it can be shown using the properties of a bijective transformation of a pair of random variables [45], that $f_{RS}(r, s) = f_{PQ}(rs, s)|s|$. The pdf of r follows by marginalization as

$$f_R(r) = \int_{-\infty}^{\infty} f_{RS}(r, s) ds = \int_{-\infty}^{\infty} |s| f_{PQ}(rs, s) ds. \tag{5.5}$$

In general, computing the above integral is non-trivial. We derive expressions for the pdf of r for the following special cases.

1. Uniform additive noise for p and q:

   Given $p = \bar{p} + n_p$ and $q = \bar{q} + n_q$ where $n_p \sim U(-\Delta_1, \Delta_1)$, $n_q \sim U(-\Delta_2, \Delta_2)$ and $\bar{p}$ and $\bar{q}$ denote the true image-plane velocities, it can be shown that

   $$f_R(r) = \frac{1}{4\Delta_1 \Delta_2} \int_{I_s} |s| ds \tag{5.6}$$

   where $I_s = \{v : \bar{p} - \Delta_1 \leq s \leq \bar{p} + \Delta_1 \bigcap (\bar{q} - \Delta_2) r \leq s \leq (\bar{q} + \Delta_2) r\}$

2. Zero mean multiplicative Gaussian noise for p and q:

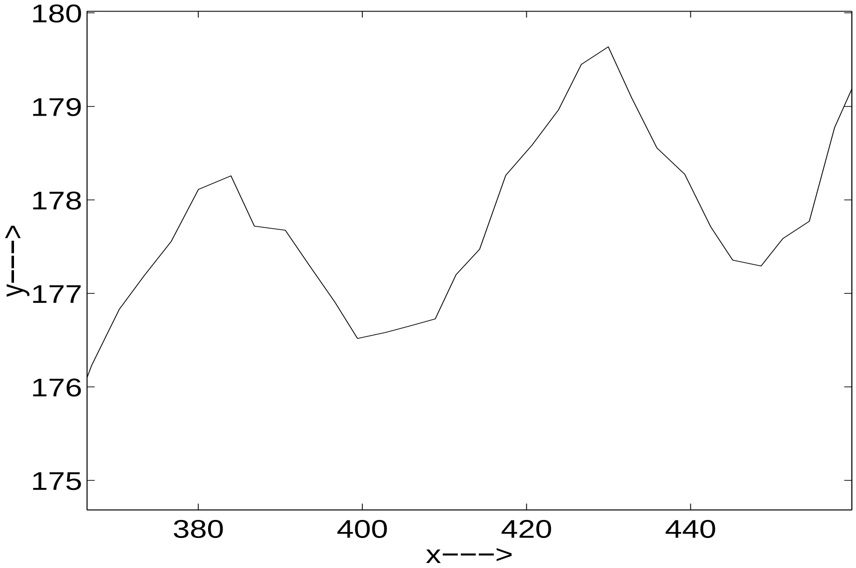


Figure 5.3: Tracked position of the head for 2 walk cycles

Given $p = \bar{p}n_p$ and $q = \bar{q}n_q$ where $n_p \sim \mathcal{N}(0, \sigma_1^2$ and $n_q \sim \mathcal{N}(0, \sigma_2^2)$ where $\bar{p}$ and $\bar{q}$ denote the true image-plane velocities, we can explicitly solve (5.5) to get $f_R(r) = \frac{\sigma_1 \sigma_2}{\pi} \frac{1}{\sigma_2^2 r^2 + \sigma_1^2}$ viz. Cauchy with scale parameter $\frac{\sigma_1}{\sigma_2}$.

From the standpoint of implementation, under arbitrary noise distributions, the sequential Monte Carlo filter can be used to get the distribution of $cot(\alpha)$. As a person walks, his/her head bobs periodically as shown in Figure 5.3 for 2 cycles. Hence in order to estimate the distribution of $cot(\alpha)$ it is reasonable to consider time-instants separated by the approximate period (T) of the walk cycle. Thus

$$cot(\alpha(t+T)) \sim \left(\frac{x^i(t+T) - \tilde{x}(t)}{y^i(t+T) - \tilde{y}(t)}, w^i(t+T) \right), \tag{5.7}$$

where $(\tilde{x}(t), \tilde{y}(t)) = \arg\max_w^i (x^i(t), y^i(t), w^i(t))$ The distribution for $cot(\theta)$ follows from (5.4) using the theory of propagation of random variables [45].

5.2.4 View Synthesis

Having obtained the angle θ, we need to synthesize the canonical view. Let Z denote the distance of the object from the image plane. If the dimensions of the object are small compared to Z, then the variation in θ, $d\theta \approx 0$. This essentially corresponds to assuming a planar approximation to the object. Let $[X_\theta, Y_\theta, Z_\theta]'$ denote the coordinates of any point on the person (as shown in the Figure 5.2) who is walking at an angle $\theta \geq 0$ to the plane passing through the starting point $[X_{ref} Y_{ref} Z_{ref}]'$ and parallel to the image plane which we shall refer to, hereafter, as the canonical plane. Computing the 3-D coordinates of the synthesized point involve a rotation about the line passing through the starting point.

Then

$$\begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \end{bmatrix} = R(\theta) \cdot \begin{bmatrix} X_\theta - X_{ref} \\ Y_\theta - Y_{ref} \\ Z_\theta - Z_{ref} \end{bmatrix} + \begin{bmatrix} X_{ref} \\ Y_{ref} \\ Z_{ref} \end{bmatrix} \tag{5.8}$$

where

$$R(\theta) = \begin{bmatrix} cos(\theta) & 0 & sin(\theta) \\ 0 & 1 & 0 \\ -sin(\theta) & 0 & cos(\theta) \end{bmatrix}. \tag{5.9}$$

Denoting the corresponding image plane coordinates as $[x_\theta, y_\theta]'$ and $[x_0, y_0]'$ (for $\theta = 0$) and using the perspective transformation, we can obtain the equations for $[x_0, y_0]'$ as (see Appendix)

$$\begin{aligned} x_0 &= f\frac{x_\theta cos(\theta) + x_{ref}(1 - cos(\theta))}{-sin(\theta)(x_\theta + x_{ref}) + f} \\ y_0 &= f\frac{y_\theta}{-sin(\theta)(x_\theta + x_{ref}) + f}, \end{aligned} \tag{5.10}$$

where

$$x = f\frac{X}{z} \text{and } y = f\frac{Y}{z}.$$

(5.10) is attractive since it does not involve the 3D depth; rather it is a direct transformation of the 2D image plane coordinates in the non-canonical view to get the image plane coordinates in the canonical one. Thus using the estimated azimuth angle θ we can obtain a synthetic canonical view using (5.10). A brief derivation of (5.10) is given in the Appendix. In summary obtaining the synthesized views from the non-canonical views involves a three step procedure:

1. Estimation of the image plane angle α from the video.

2. Estimation of the azimuth angle θ via Equation (5.4).

3. View synthesis using Equation (5.10).

5.2.5 Alternative Approach for View Synthesis

The above procedure relies on the true translation of the walking person in the video. For the case when there is no translation (e.g. person walking on a treadmill), neither α nor f can be estimated from the video. An alternative approach to view synthesis can be used in this case. We observe that Equation(5.10) is a homography of the form

$$H(\theta) = \begin{bmatrix} a & 0 & b \\ 0 & c & 0 \\ d & 0 & e \end{bmatrix}. \tag{5.11}$$

where $a = fcos(\theta)$, $b = fx_{ref}(1 - cos(\theta)$, $c = f$, $d = -sin(\theta)$ and $e = -x_{ref}sin(\theta) + f$. Hence if we have point correspondences for any planar surface between the canonical and non-canonical views in a set of training images, we can estimate a homography of the type in (5.11). This homography can then be applied to the binary silhouette of the person in the non-canonical view to get the synthesized views.

5.2.6 Justification for generating the synthetic view

Although it is possible to recognize a person purely from joint-angles, considering the difficulty in estimating them from video, we choose to take an appearance based approach for gait recognition. For a person walking along a non-canonical direction, appearance based features used for recognition get distorted. To explain this better we consider a width-vector based image feature. For a given image the width-vector is computed as the distance between the extremities of the silhouette along each row. The top part of the width vector corresponds to the head-region, the middle part corresponds to the torso while the bottom part corresponds to the foot-region of the individual. When the person is at rest, the values of the width vector in the hand and leg regions will be smaller as compared to when the legs are apart. Temporal plots of the width-vector for the same person walking in the canonical and non-canonical($\theta = 45$) direction are shown in Figures 5.5 (a) and (b) respectively. In these plots, the vertical axis corresponds to the row index of the width vector, while the horizontal axis corresponds to the frame number. The dark parts in the leg region correspond to the rest phases while the bright parts correspond to when the legs are apart. A simple gait feature, viz. the stride length or the maximal separation of the feet, can be derived from the width plots by measuring the highest intensity in the leg regions. Observation of the width plots for the canonical and non-canonical views reveals that the apparent stride-length is smaller for the non-canonical view. The second effect that is obvious from the plots is a foreshortening effect as the person walks away from the camera. In order to obtain good gait recognition performance, it is necessary to correct for both of these effects through view synthesis. We will show later in our experiments how view synthesis provides for a correction of both of these effects (see Figures 5.5 (c) and (d)) and improves the gait recognition performance appreciably.

5.3 Matching Gait Sequences using DTW

Since only limited training data is usually available for gait analysis, we adopt a template-matching approach for recognizing gait. The number of frames in the reference gait sequence and the probe gait sequence depend on the number of gait cycles available. Larger the number of gait cycles used for matching, the better will be the performance. Typically, the number of frames in the reference and probe data will differ. Moreover, the reference and probe gait data are seldom synchronized. Therefore, direct matching of a probe sequence with a reference sequence is not possible. A pattern-matching method based on dynamic programming paradigm is useful to deal with this situation. The DTW technique [19] is used to compensate for the variability in the speed of walking which in turn reflects in the number of frames for each gait cycle. Different appearance based features may be extracted from the binarized silhouettes e.g. width of the outer contour of the silhouette [46], smoothed versions of moment features in image regions [13] etc with corresponding distance metrics for matching. In order to assess the utility of our method without being affected by the choice of a particular image feature, we choose to use the entire image as the feature with binary correlation as a distance measure.

Given $\{\theta^i(t), w^i(t)\}$ using (5.4) and (5.7) we choose the MAP estimate of θ to synthesize the corresponding probe sequence image which we denote as $P(t)$. The gallery sequence is obtained from the video of the person walking in the canonical plane. Use of DTW relies on similar start and end points for the probe and gallery. Hence prior to applying the DTW algorithm it is necessary to segment the video sequence into cycles. A simple way to achieve this is using the width feature explained in Section 2.6. Figure 5.5(a) shows the width vector plot as a function of time for a given video sequence. It is easy to see that the sum of intensities along successive columns will show a periodic variation. The troughs of the resulting waveform correspond to the rest positions during the walk cycle while the crests correspond to the part of the cycle where the hands and legs are maximallly displaced. A half-cycle consists of frames between two successive troughs of sum of intensities plot. We take gallery images corresponding to 4 successive half cycles as the gallery and similarly for the probe. Details of the DTW are given in Section 2.4.

5.4 Experimental Results

5.4.1 Gait-based Recognition

In this section, we present gait recognition results using three databases. We would like to mention here that at present the USF gait challenge database is the most widely used database in the community. In the USF database the subject walks on an elliptical path and the back portion of the ellipse is used for gait recognition experiments. However, for most of the path travelled there is very little angle variation and hence this database is not a suitable testbed for our approach.

As explained in Sections 2.2 and 2.3, the person's head is tracked using a sequential Monte Carlo filter. Using (5.4), $\theta \sim (\theta^i(t), w^i(t))$ is obtained. The distributions corresponding to two different time instants are shown in Figure

5.9. Using $\tilde{\theta}(t) = \arg\max_{w^i}(\theta^i(t), w^i(t))$ the image of the unknown person $X(t)$ is synthesized using (5.10). A bounding box is then placed around the part of the motion image that maximally contains the moving person. In order to depict the performance of the gait recognition algorithms, we use the CMC plots which show the percentage of times the right person occurs in the top n matches where $n < P$ where P denotes the number of subjects in the gallery.

UMD3 database This consists of 12 people, who walk along straight lines at different values of azimuth angle $\theta = 0, 15, 30$ and 45. The image sequences corresponding to $\theta = 0$ were used as the gallery while the other sequences were used as a probe.

The synthesized images for Database 1 are shown in Figure 5.4, along with the images from the original video sequences. The width profile plot for the canonical view and the view synthesized from $\theta = 45$ are shown in Figure 5.5. As can be seen from this plot, our method has compensated for both the foreshortening effect as well as restored the true leg-swing.

From the synthesized images we note that in the torso region, the reconstructed silhouette is broader than the original. The reason for this is the limitation of the planarity assumption for the torso region viz. parts of the torso unseen in the canonical view, appear in non-canonical views. The synthesis algorithm, which interprets this as a plane, renders a broader reproduction of the torso part. Notice however that this effect is somewhat lesser in the leg portions of the silhouette. This fact, as will be explained later, can be exploited to improve gait recognition performance. Quality of the synthesized images can be evaluated by computing the binary correlation of the synthesized images at $\theta = 15, 30, 45$ with images in the canonical view. This is plotted as a function of

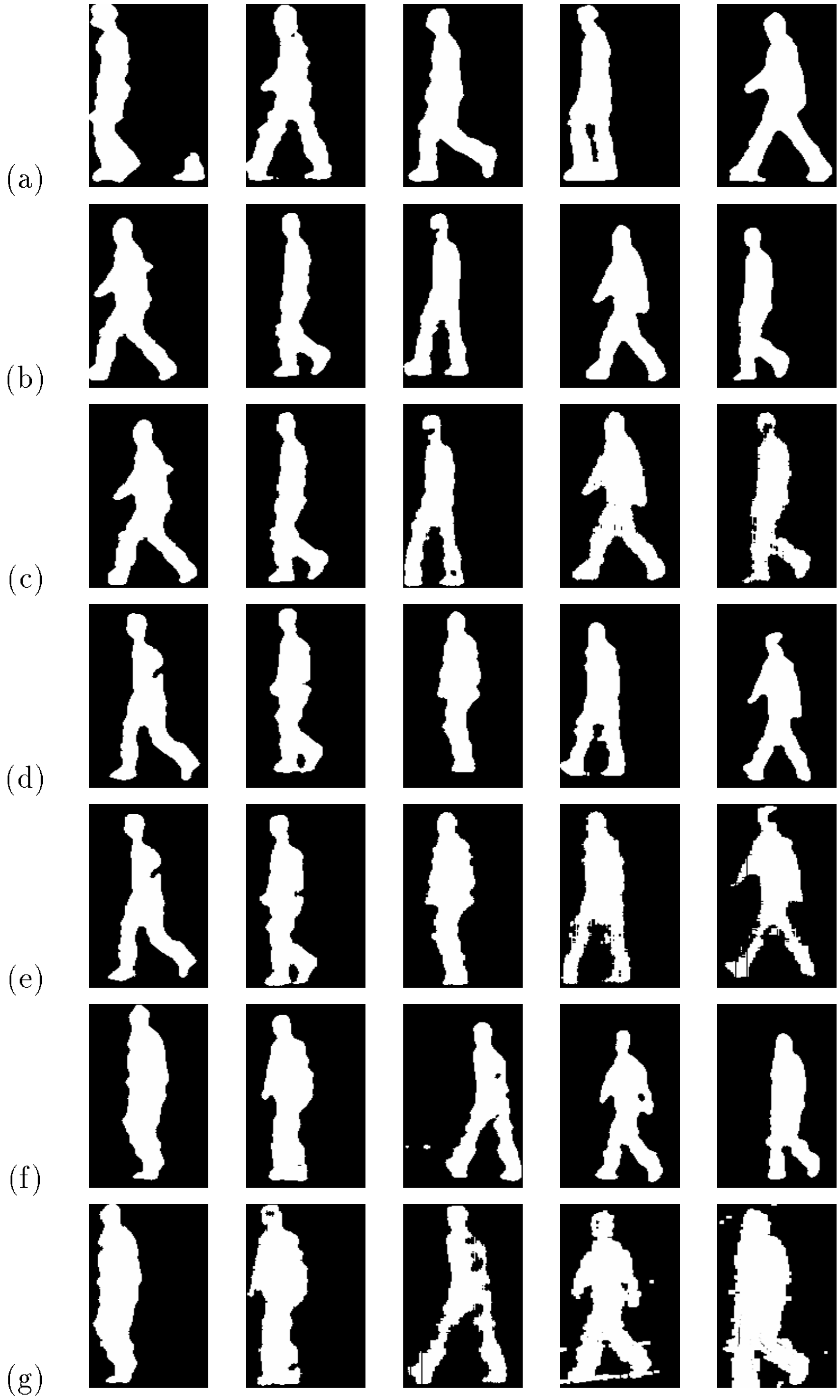


Figure 5.4: (a) represents different stances of a person walking parallel to the camera; (b) (d) and (f) represent different stances of a person walking at angles 15, 30 and 45 degrees to the camera; (c) (e) and (g) represent side-views synthesized from original videos where the person walks at angles of 15, 30 and 45 degrees to the camera.

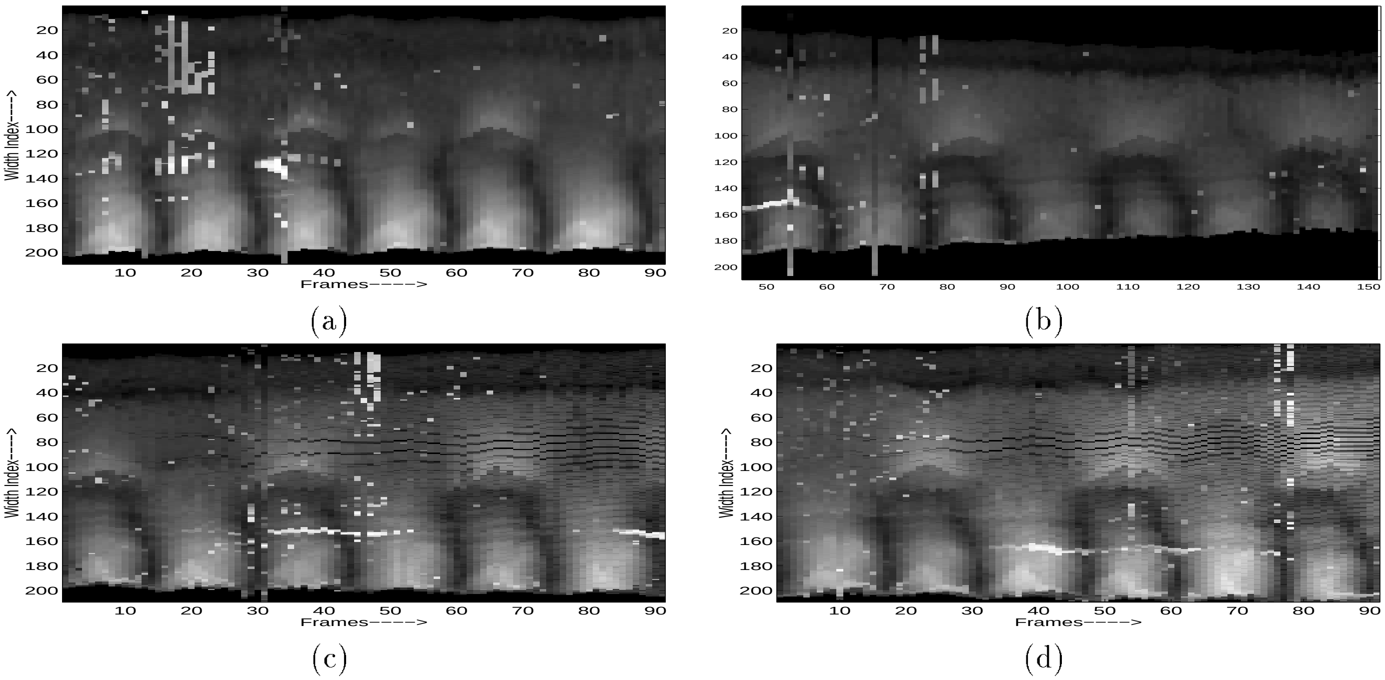


Figure 5.5: Width profile as a function of time for (a) Canonical View ($\theta = 0$); (b)Unnormalized sequence for $\theta = 45$; synthesized views for: (c) $\theta = 30$ (d) $\theta = 45$.

θ in Figure 5.6. Two consecutive cycles in the canonical view are chosen as the gallery to be compared with two consecutive gait cycles in the probe sequence. The DTW technique is used to match a given probe sequence to the different gallery sequences using binary correlation as a local distance measure and a similarity matrix $S = s(i,j)$ is obtained, where $s(i,j)$ refers to the similarity between the probe i and the gallery j.

Gait recognition performance for $\theta = 15, 30$ and 45^0 is shown in Figures 5.7(a), (b) and (c) using the synthesized and raw images in terms of a cumulative match characteristic. Clearly, our method has improved the recognition performance . As noted before, the algorithm results in a broader reproduction of the torso region. The situation can be remedied by assigning a lower weight to the torso region when computing the binary correlation or simply ignoring it.

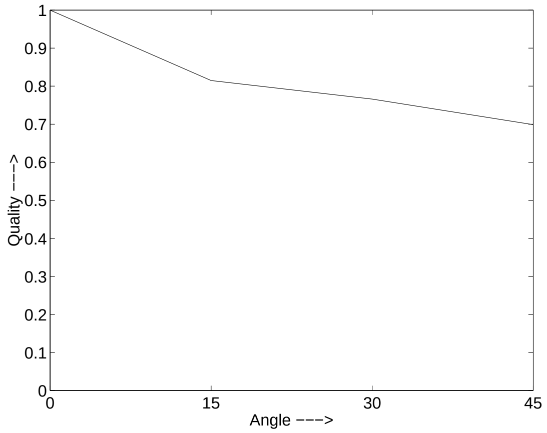


Figure 5.6: Quality of reconstruction as a function of the angle of walk.

We take the latter approach by computing the binary correlation only over the lower half of the boxed image. The result using only the leg region is shown as the dashed lines in the Figure 5.7. It can be seen that the gait recognition result is better than what is obtained by using the entire body. Interestingly, [23] notes that the lower 20 % of the silhouette accounts for roughly 90% of the recognition. Similarly [46] showed that the gait recognition in the case when the subject was carrying a ball, viz. no upper body dynamics, the recognition rates were better. To boost the gait recognition performance further, certain structural characteristics of the individual that are extracted subsequent to view synthesis e.g. height can be fused with the leg dynamics. The height of the probe sequence is estimated robustly from the synthesized video as $h(i) = \text{median} h^j(i), j = 1 \cdots M$ M, being the length of the probe sequence. We fuse height information together with the leg dynamics by scaling each entry $s(i,j)$ of the similarity matrix by the corresponding height ratio, $\max(\frac{h(i)}{h(j)}, 2 - \frac{h(i)}{h(j)})$. The results for this case are shown as the solid line with circles in Figure 5.7. This strategy clearly gives the best gait recognition performance.

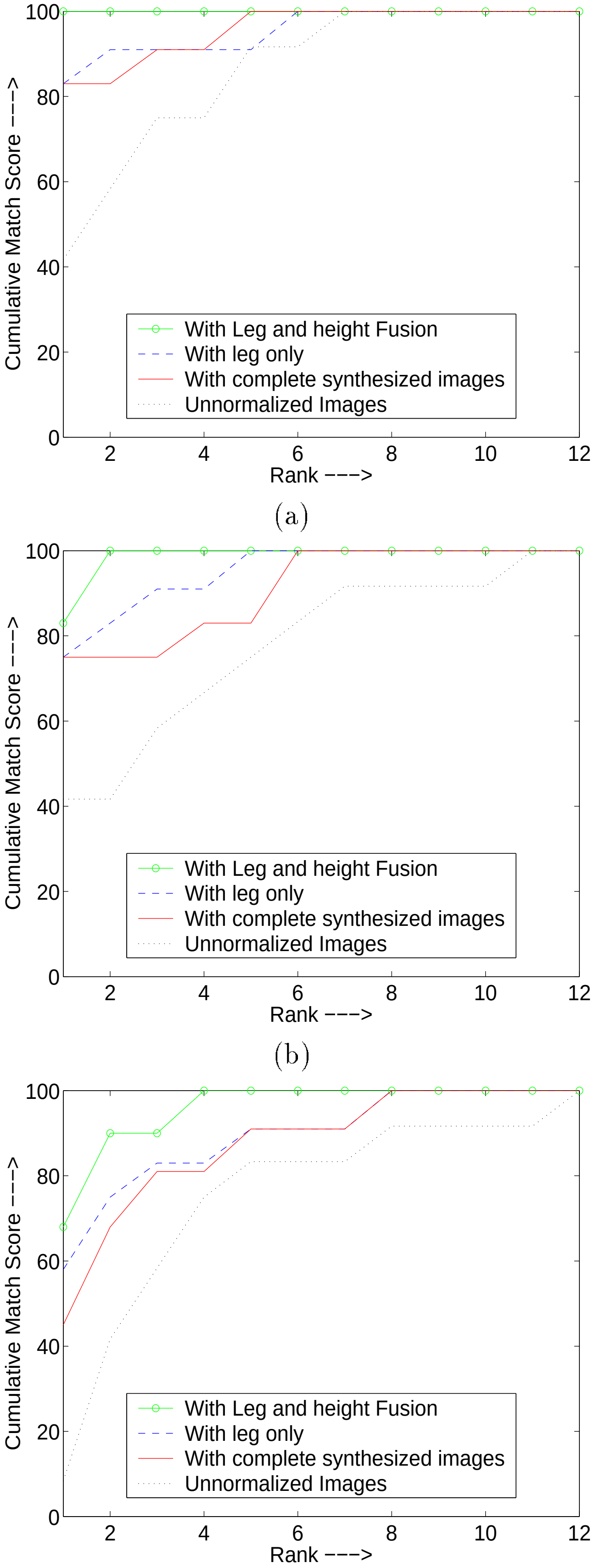
The fact that the gait recognition results are encouraging upto angles of 45 degrees allows us to hypothesize that it is possible to do reasonable human identification using gait with only two cameras (installed perpendicular to each other). In order to study the efficiency of gait recognition with synthesized views, we compute the Receiver Operating Characteristic (ROC), which is a plot of the probability of detection (i.e. correct recognition), p_D, vs. the probability of a false alarm (i.e. false acceptance), p_F, for azimuth angles $\theta = 15, 30$ and 45 degrees. For a given threshold, we define the notion of detection to correspond to $s(i, i) >$ threshold and that of a false alarm to correspond to $s(i, j) >$ threshold. The plots are shown in Figure 5.8. The performance degradation with increasing θ can be understood from these plots. The ROC curves indicate that the proper detection threshold should vary with θ, so as to obtain a performance characteristic with small p_F and large p_D.

NIST database

This consists of 30 people walking along a Σ -shaped walking pattern as shown in Figure 5.10.

There are two cameras looking at the top horizontal part of the sigma. The camera that is located further away is used in our experiments since the planar approximation we make is more valid in that case. The segment of the sigma next to the top horizontal part is used as a probe. This segment is at an angle 33^0 to the horizontal part. A few images from the NIST database are shown in Figure 5.11. The method described before was used to estimate the walking direction and it turned out to be very close to the true value (Figure 5.9)

To do gait recognition we employed the fusion of the leg-dynamics with the height since it gave the best performance for Database 1. The gait recognition result is shown in Figure 5.12. As can be seen the recognition rate is about 60%.



(a)

(b)

(c)

Figure 5.7: Cumulative Match Characteristics for Original and Synthesized images for (a) $\theta = 15$ (b) $\theta = 30$ and (c) $\theta = 45$ for UMD3 database.

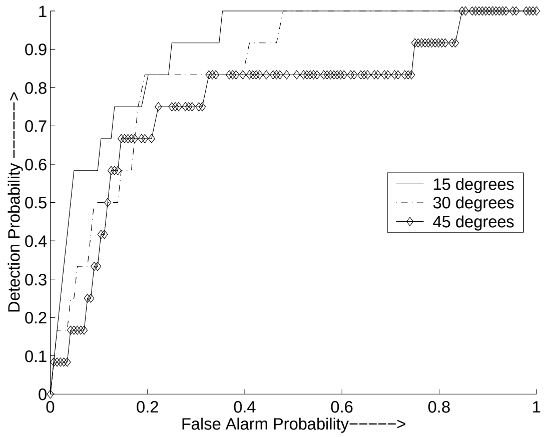


Figure 5.8: ROC curves for $\theta = 15, 30$ and 45 degrees for the case of leg and height fusion.

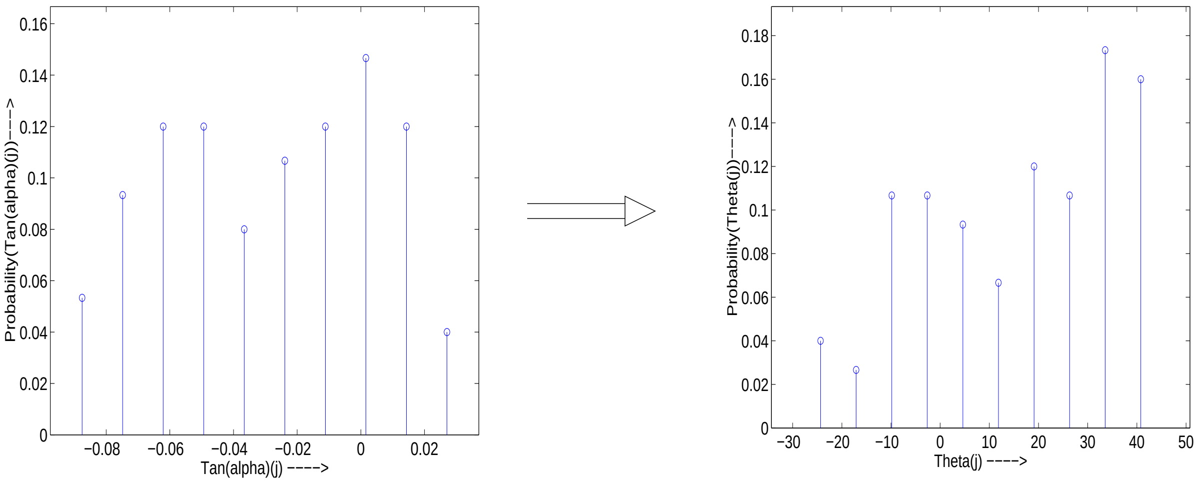


Figure 5.9: Distributions of $Tan(\alpha)$ and its transformation to θ at one time instant

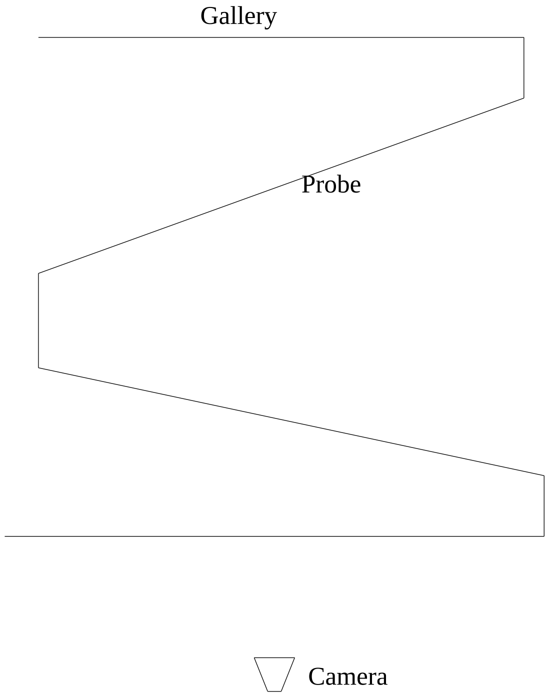


Figure 5.10: Σ shaped walking pattern in the NIST database.

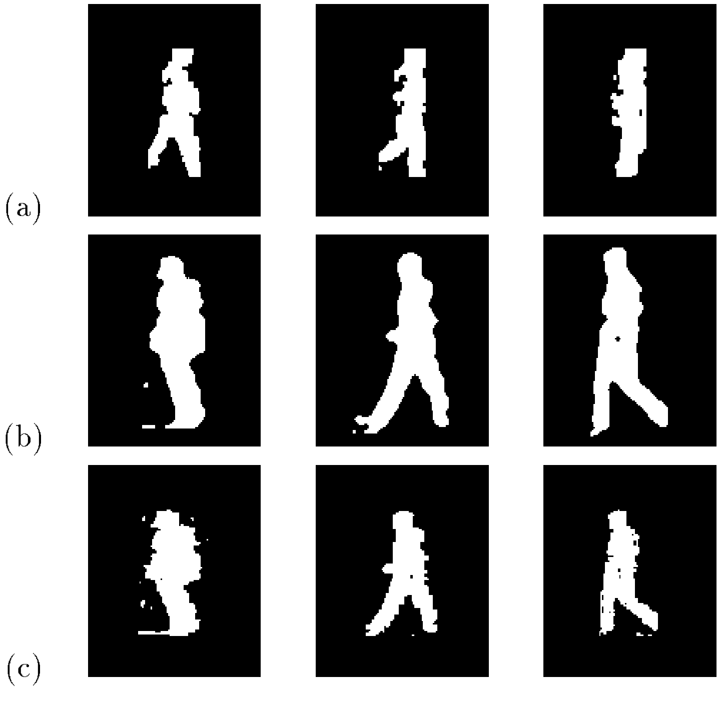


Figure 5.11: Examples for the NIST database (a) Gallery images of person walking parallel to the camera (b) Unnormalized images of person walking at 33^0 to the camera(c) Synthesized images for (b).

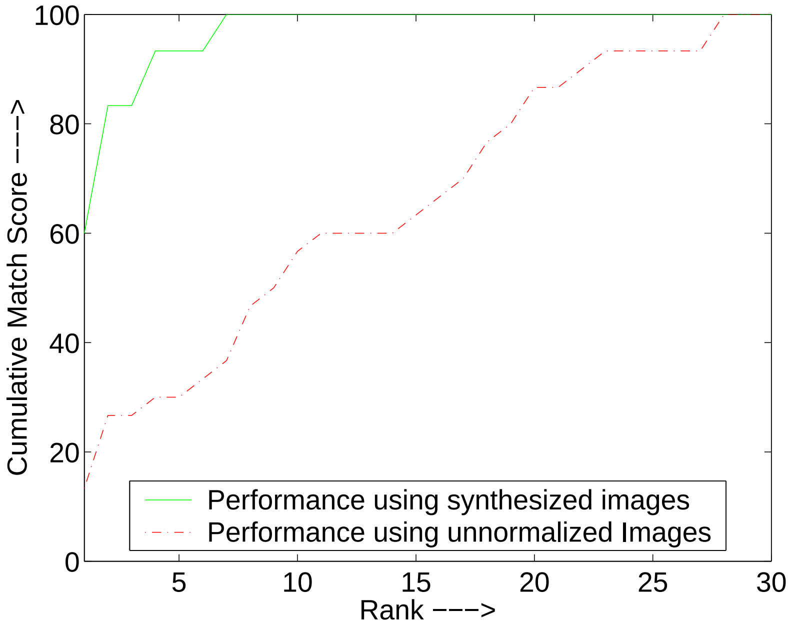


Figure 5.12: Gait recognition performance on the NIST database.

One of the reasons for the lower recognition performance in this case is that the image size is rather small. Note however that the recognition goes to 100% within 6 ranks.

CMU MoBo database

The CMU MoBo database [47] consists of 25 people walking on a treadmill in the CMU 3D room. There are six synchronized cameras evenly distributed around the subject walking on the treadmill. For testing our algorithm we considered views of the slow walk sequences from cameras 3 and 13. The camera 3 captures the exact side view while the camera 13 captures a non-canonical view of the subjects. The sequence from camera 3 was used as the gallery while the sequence from camera 13 was used as the probe. Since there is no actual translation involved in this case, it is not possible to estimate the image plane angle α or the focal length f. Hence in order to obtain the synthesized

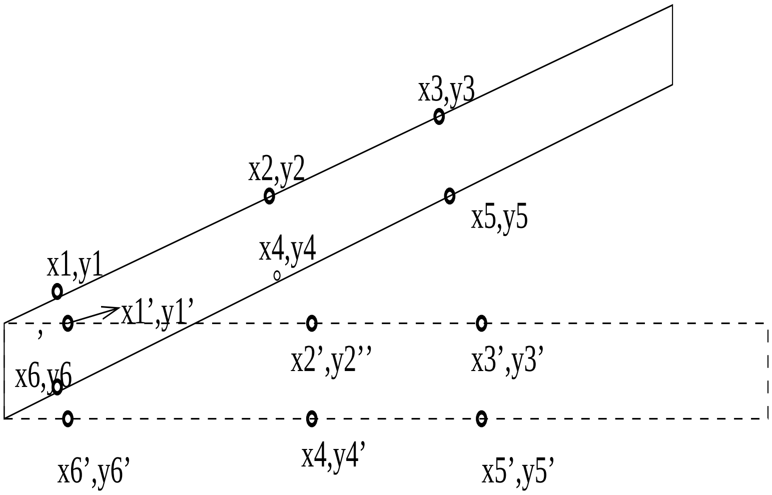


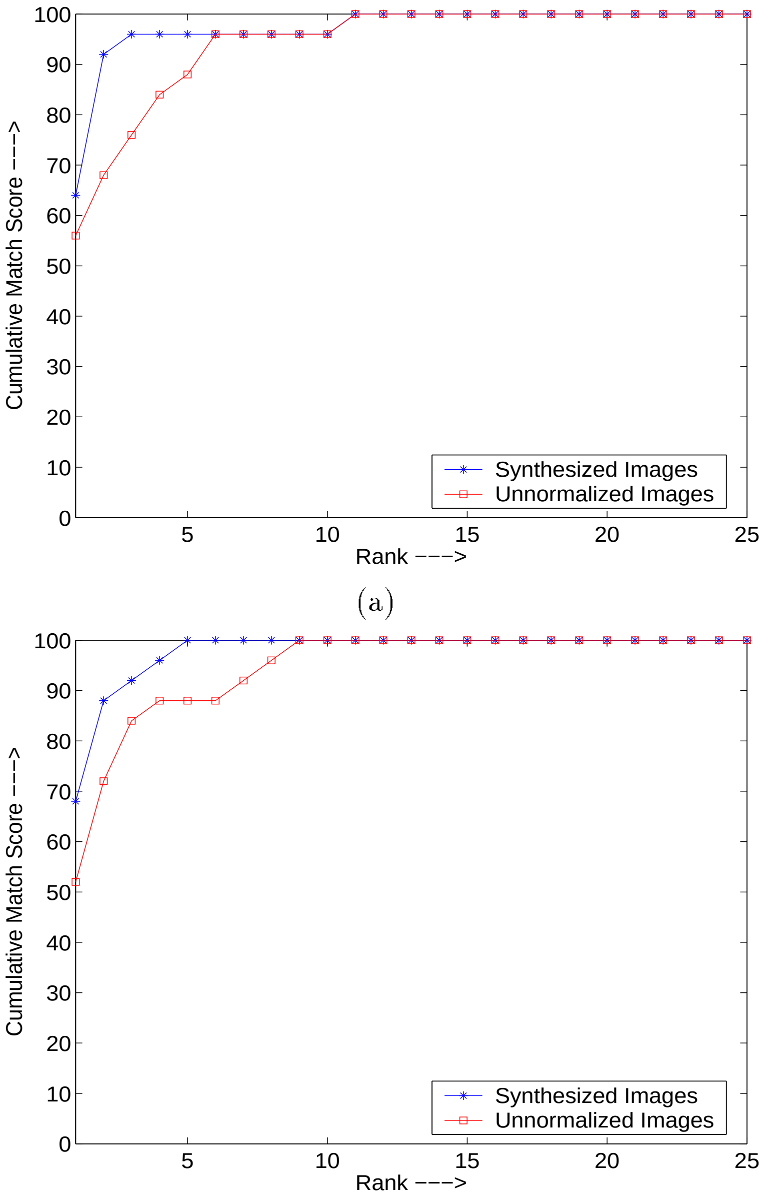
Figure 5.13: Computation of Homography for the CMU database.

images for this database, we use an alternate approach. We considered several points $S = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ on the side of the treadmill which is a planar rectangular surface as shown in Figure 5.13. We then constructed the view of this rectangular patch as it would appear in the canonical view. A set of points $S' = \{(x'_1, y'_1), \cdots, (x'_n, y'_n)\}$ is considered on this hypothesized patch. A homography of the nature of (5.11) is estimated using the sets S and S' [40]. This homography is then used to obtain the synthesized images for the person.

Gait recognition performance using the unsynthesized and synthesized images is shown in Figure 5.14. Again we see that synthesis results in better gait recognition performance and that using the leg region alone achieves better performance as compared to using the entire body. In the latter case, the true person is identified within the top 6 matches.

5.4.2 Video-based rendering of planar dynamic scenes

The view synthesis algorithm can also be used for rendering planar dynamic scenes with interesting applications in video compression, video indexing and



(a)

(b)

Figure 5.14: Cumulative Match Characteristics for Original and Synthesized images for (a) Full body (b) Leg only for the CMU database.

retrieval. It would be interesting to study how the method performs when the object being viewed is non-planar. When the imaged object is non-planar, portions unseen in the canonical view appear. An example is shown in Figure 5.15 which shows the top views of a rectangular block. The dashed rectangle represents the canonical case while the solid one represents the case when the block is at an angle θ. Under perspective projection, the length of projection of the tilted rectangular block on the image plane for $\theta \neq 0$ can be shown to be a function of the spatial coordinates of the block(see Appendix), while under a weak perspective projection it is a function of θ alone. To simplify analysis we used the weak perspective approximation. The error introduced by this approximation is small provided $Z >> a, b, x$. The dotted lines represent the true perspective projections of the edge of the block while the solid lines represent the weak perspective approximation. Under weak perspective projection, the apparent width of the block visible in the image plane is

$$W_a = \frac{f}{Z}(acos(\theta) + bsin(\theta)). \tag{5.12}$$

Rendering this image of the block directly using (5.10) will lead to an incorrect synthesized view. In general this problem can be avoided only if multiple cameras are available. For the present single camera case, a simple way to circumvent this problem is to synthesize only the portion of the image corresponding to $acos(\theta)$.

From Figure 5.17, we see that the torso appears wider than in the canonical view. This is due to the fact that the torso is really non-planar. The effects of non-planarity become more severe as the azimuth angle increases. To deal with this we need to ignore parts of the torso which are not seen in the canonical view. We assume the torso to be a rectangular block. Given the extent of the torso (U, L) the widths at different row positions $a(x) : U < x < L$ can be learned

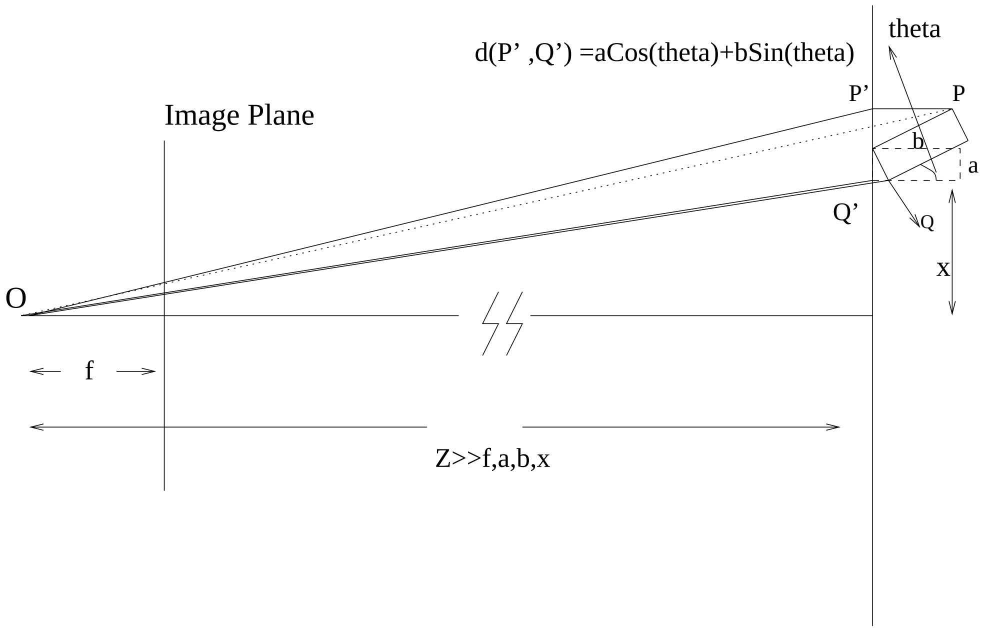


Figure 5.15: Rendering of Non-planar objects

from different images of the person in the canonical view. Given θ, and an image in the non canonical view at azimuth θ only pixels upto a distance of $a(x)Cos(\theta)$ (see (5.12)) from the left extremity of the silhouette are synthesized using (5.10). The next issue is how to obtain U and L. To this end, the width of the outer contour of the silhouette is computed for different images in the canonical view. Principal components analysis is performed on the width vectors, which are then projected on to the principal eigen vector. This has the effect of removing noise and retaining only the true variations which occur in the torso(hand)and leg regions. The result is shown in the Figure 5.16. From this overlay the extent of the torso region can be determined.

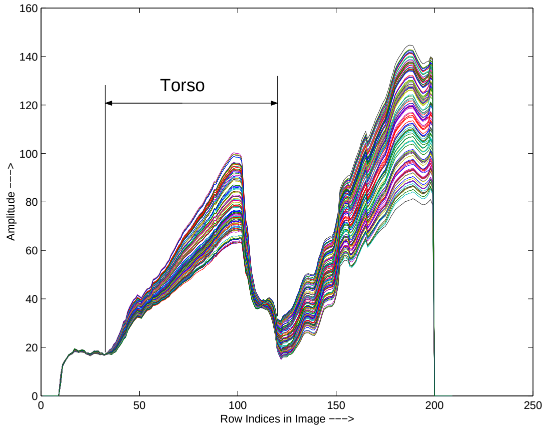
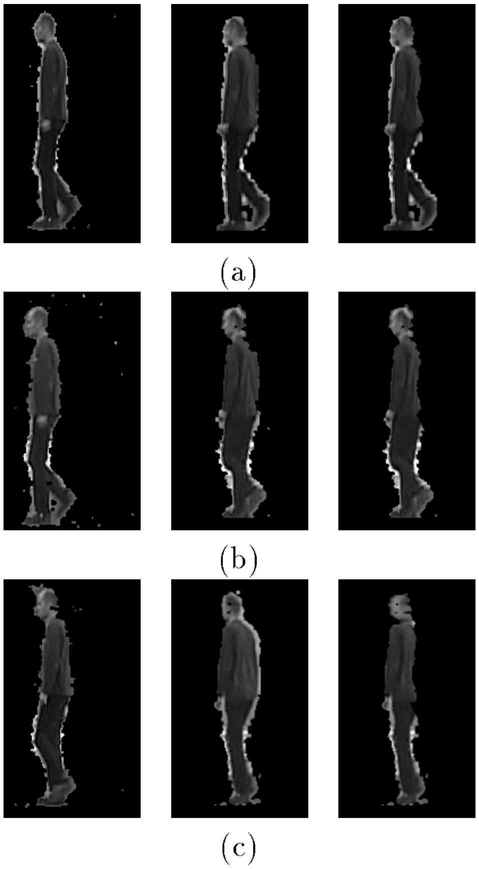


Figure 5.16: Width Vectors projected on to the principal eigen vector for azimuth angle $\theta = 0$

5.4.3 Fusion of Face and Gait

For optimal performance, the system must use as many cues as possible and combine them in meaningful ways. Different modalities can be used for identification based on the distance of the individual from the camera. If the person is far away from the camera, it is hard to get face information at a high enough resolution for recognition tasks. However when available, it yields a very powerful cue for recognition. A modality which can be detected and measured when the subject is far away from the camera is gait. One of the features of the NIST database is that it provides for recognition of multiple modalities. The far portions of the walking path are suitable for gait recognition, while at the end of the walking path the person walks providing a frontal view of his face to the camera. This final segment can be used for face-recognition. In such situations it is possible to fuse both face and gait cues as part of a higher level recognition system to get better recognition performance than from any of the single recognition units.

In [48], an algorithm for probabilistic recognition of human faces was proposed. In still-to-video recognition, where the gallery consists of still images, a



(a)

(b)

(c)

Figure 5.17: The first column in (a) shows a frame from the canonical view; the second and third columns respectively show an image synthesized from a similar frame at an azimuth $\theta = 15$ before and after the correction for non-planarity is applied. (b) and (c) represent the same cases for azimuths $\theta = 30$ and 45 respectively.

time series model is used to fuse temporal information in a probe video, which simultaneously characterizes the kinematics and identity using a motion vector and an identity variable. The joint posterior density of the motion vector and the identity variable is estimated at every time instant and then propagated to the next time instant. Marginalization over the motion vector yields a robust estimate of the posterior distribution of the identity variable. A computationally efficient SIS algorithm is used to estimate the posterior distribution. It was shown that a degeneracy in the posterior probability of the identity variable occurs, leading to improved recognition. The computational load involves a term that is linearly proportional to the size of the database N. We employ decision fusion which is a special case of data fusion (see Kokar et al. [34]) to combine the results of our gait recognition algorithm and the face recognition algorithm in [48]. We consider two fusion scenarios: hierarchical and holistic. The first involves using the gait recognition algorithm as a filter to pass on a smaller set of candidates to the face recognition algorithm. The second involves combining the similarity scores obtained individually from the face and gait recognition algorithms.

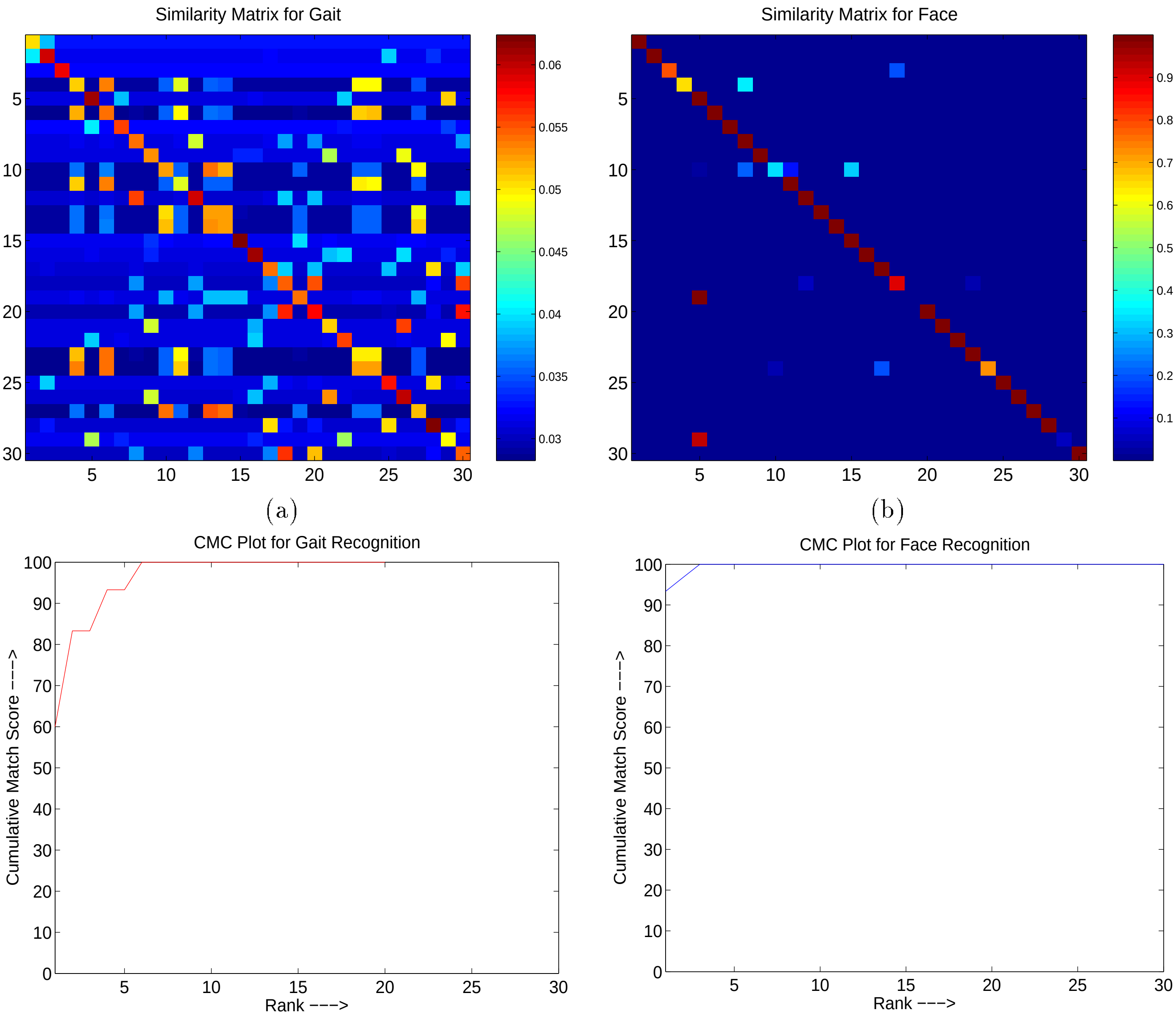
The face recognition algorithm yields a match score which is a probability while the gait recognition algorithm yields a distance measurement. In order to make the scores comparable before fusing them, it is necessary to apply a score transformation to the classifiers. The transformation should be such that the relative ordering of the scores is not altered. In other words the transformation function should be monotone. Some of the commonly used transformations include linear, logarithmic, exponential and logistic. The purpose of these transformations is, first, to map the scores to the same range of values and second, to change the distribution of the scores. For example, the logarithmic transformation puts strong emphasis on the top ranks, whereas the lower ranked scores which are transformed to very high values, have a quickly decreasing influence. A detailed discussion of score transformation is given in [49] in the context of combining classifiers for face recognition. It was shown that all the above mentioned transformation gave comparable results. For the problem of face and gait fusion we chose the exponential transformation for score normalization. The similarity matrices and CMC characteristics for the face and gait recognition algorithms are shown in Figure 5.18.

Hierarchical Fusion: Given the similarity matrix for the gait recognition algorithm, we plot the histograms of the diagonal and non diagonal terms of the normalized similarity matrix. From Figure 5.19 we note that the distributions of the true matches and false matches have limited overlap. This suggests that a threshold can be determined from the histogram and only individuals whose score is higher than this threshold need be passed to the face recognition algorithm. Although it is tempting to choose this threshold as high as possible, it should be noted that due to overlap in the two histograms, choosing a very high value may lead to the true person not being in the set of individuals passed to the face recognition algorithm . For the NIST database we chose a threshold of 0.035. This results in passing approximately the top six matches from the gait recognition unit to the face recognition algorithm. The CMC plot for the resulting hierarchical fusion is shown in Figure 5.20 (a). Note that the top match performance has gone up to 97% from 93 % for this case. The more important gain however is in terms of the number of computations required. This number drops to one-fifth of its previous value. This demonstrates the value of gait as a filter.

Holistic Fusion If the main requirement is that of accuracy as against computational speed, alternate fusion strategies can be employed. Assuming that gait and face can be considered to be independent cues, a simple way of combining the scores is to use the SUM or PRODUCT rule [35]. Both the strategies were tried. The CMC curve for either case is as shown in Figure 5.20(b). In both cases the recognition rate is 100 %.

5.5 Conclusion

In this chapter, we have proposed a method for synthesizing arbitrary views of planar objects, and applying the synthesized views for gait recognition when humans are walking at an arbitrary angle to the camera. Our method used the differential form model of structure from motion for estimating the azimuth angle of the original view from monocular video data. Thereafter, a video sequence at the new view was synthesized. The entire process was done in 2D, though 3D structure of the scene played an implicit role. Examples of synthesized views were presented. Gait recognition performance was reported for three databases and the benefit of using our method was demonstrated. One of the future areas of our work is to achieve reliable gait recognition over the entire field of view by using only two cameras. Though the method has been explained from the motivation of the gait recognition problem, it has important applications in other areas too, like multimedia and video processing. That forms a part of our future research into this problem.



(a)

(b)

(c)

(d)

Figure 5.18: Similarity matrices for (a) Gait Recognition ; (b)Face Recognition; CMC characteristics for: (c) Gait (d) Face

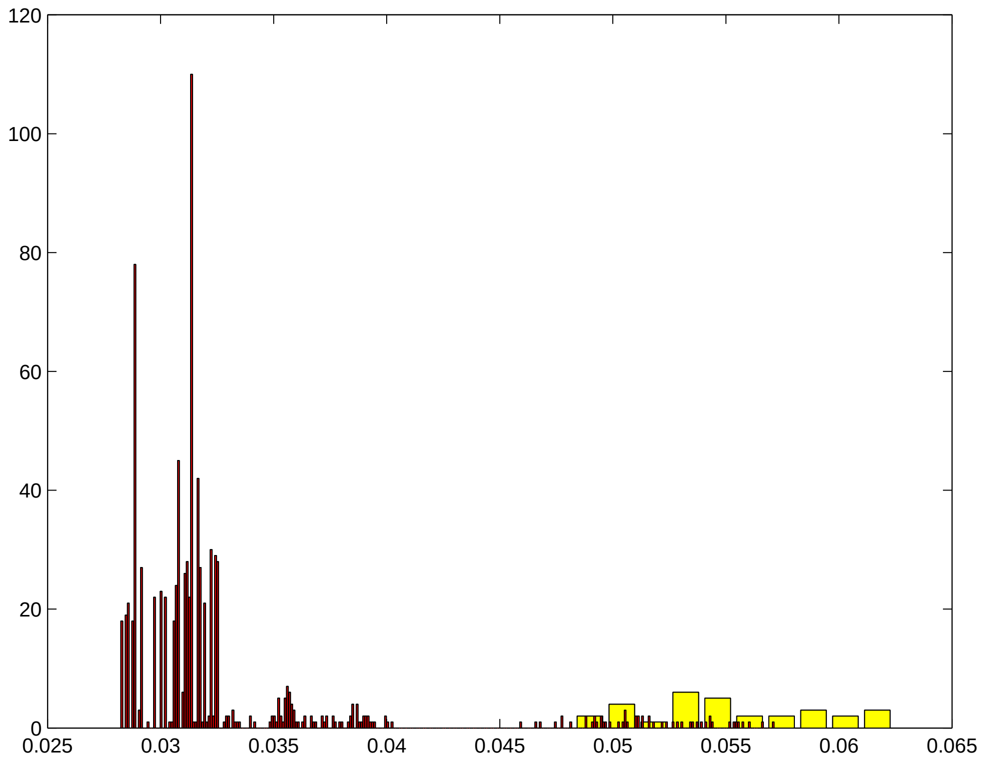
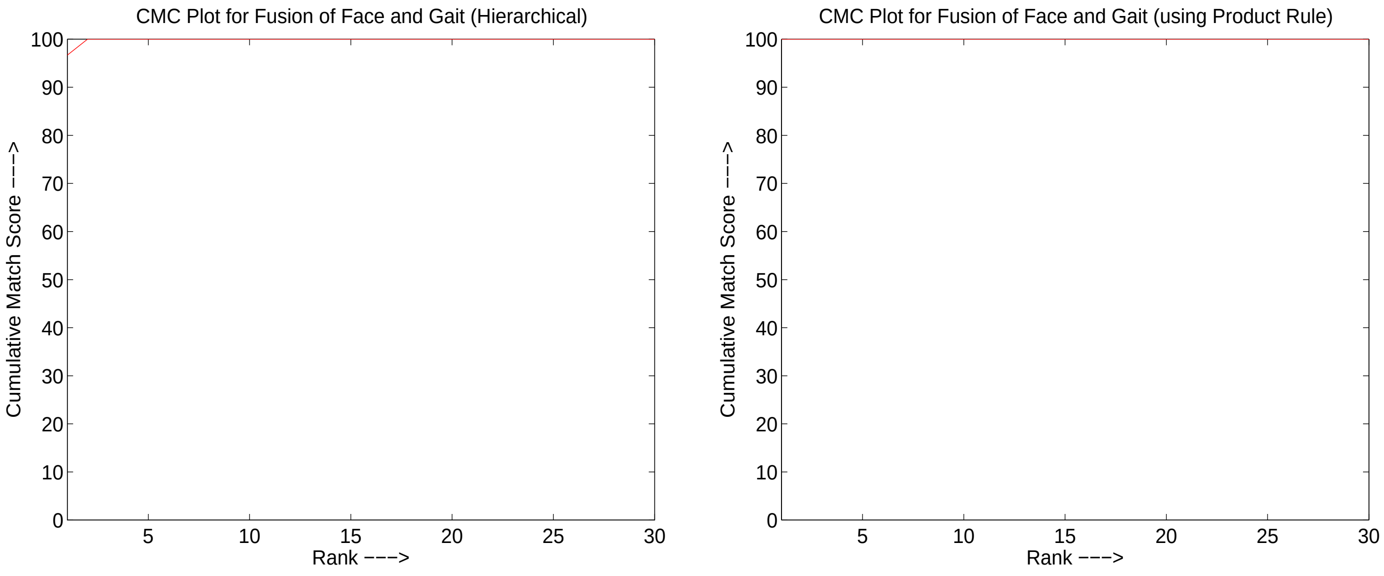


Figure 5.19: Histogram of the true matches (yellow) and false matches (black)



(a)

(b)

Figure 5.20: CMC curves for (a) Hierarchical and (b) Holistic Fusion

Chapter 6

Conclusions and Future Work

The goal of this thesis is to investigate the information contained in the video sequences of human gait and how to extract and represent that information in ways that facilitate human identification. The attractiveness of gait as a biometric arises from the fact that it is non-intrusive and can be detected and measured even in low resolution video. Furthermore, it is harder to disguise than static appearance features such as face and it does not require a cooperating subject. Unlike traditional biometrics like face, fingerprints, iris etc., gait is a spatio-temporal signature viz. it contains a structural component and a dynamic component. In this thesis, we presented algorithms for implicit and explicit representation of these components. The following were the main contributions of the thesis:

In Chapter 2, an appearance based method for gait recognition based on template matching was presented. The width of the outer contour of the binarized silhouette of a person was used as the gait feature. Different feature vectors were derived from the basic width vector using either direct smoothing and down-sampling or by way of eigenanalysis for gait representation and matching. The results showed that since the gait measurements were highly structured, the two most significant eigenvectors retain most of the identity information. Also, eigen-analysis furnished an elegant way of smoothing the gait data. During the matching process, the DTW algorithm was used to effectively handle unequal lengths of the reference and probe gait sequences. Experimental results were reported on several standard gait databases. The method was shown to be robust to changes in walking speed. Other situations such as sensitivity to viewing angle, different floor characteristics, low frame-rate etc were considered. The performance of our method was found to be quite satisfactory on all the databases. The contribution of dynamics via the velocity profile of the width vector was studied and it was shown that, by itself, it is not sufficient to capture the gait characteristics when used in isolation. Our experiments also confirm that the side-view is the optimal one for capturing gait characteristics.

In Chapter 3, we proposed a systematic approach to gait recognition by building representations for the structural and dynamic components of gait in an HMM framework. A set of exemplars that occur during a gait cycle was derived for each individual. To obtain the observation vector from the image features two different methods were used. In the *indirect approach* the high-dimensional image feature was transformed to a lower-dimensional space by generating the Frame to Exemplar (FED) distance vector sequence. For compact and effective gait representation and recognition, the gait information in the FED vector sequences was captured using a HMM for each individual. In the *direct approach*, the feature vector is directly used to train an HMM for gait representation. An alternative approach based on the distance between the exemplars and the image features is used for estimating the observation probability. The performance of the methods was tested on different gait databases. In situations where the binary silhouettes had significant noise e.g. the USF gait challenge database, the statistical nature of the hidden Markov model was shown to provide more robust performance compared to appearance matching. On the other hand for the CMU and UMD database where the silhouettes of good quality but the amount of training data was limited, the appearance matching method gave a better performance. This observation can guide the choice of the gait recognition algorithm to be used in a particular situation.

In Chapter 4, combination of multiple instances of different gait features was studied. Experiments were designed to analyze the significance of different body parts such as the hands and legs for recognition. It was shown that there is an asymmetry in gait as far as forward and back swings go and a simple way to boost gait recognition via right and left projection vectors was demonstrated. We found that the leg region exhibits a more consistent pattern and when combined with simple physical features such as the apparent height, higher recognition rates may be obtained. Frontal gait was also studied for the purpose of identification and the width vector was computed as the frontal gait feature. Sum, product and MIN rules were used to combine the decisions made using the separate features. As expected, the overall recognition performance improved due to fusion.

In Chapter 5, we proposed a method for synthesizing arbitrary views of planar objects, and applying the synthesized views for gait recognition when humans are walking at an arbitrary angle to the camera. Our method used the differential form model of structure from motion for estimating the azimuth angle of the original view from monocular video data. Thereafter, a video sequence at the new view was synthesized. The entire process was done in 2D, though 3D structure of the scene played an implicit role. Examples of synthesized views were presented.

Gait recognition performance was reported for three databases and the benefit of using our method was demonstrated. An application to video-based rendering of planar dynamic scenes was also presented.

6.1 Applications and Future Directions

This thesis provided a suite of algorithms for the problem of human identification using gait. The ideas presented here can be applied to a variety of gait and non-gait applications. The techniques developed apply to general activity-specific person identification. One of the uses of the temporal width plot for example, can be in the qualitative and/or quantitative assessment of gait abnormalities. It would also be interesting to analyze if the warping path characteristics of the DTW algorithm can be used to study subtle changes in individual gait patterns. The gait recognition algorithms can also be used in conjunction with other cues such as the color of clothing etc. for short time verification problems viz "was this the same person who walked in front of this camera t minutes ago?". We also presented a view invariant gait recognition algorithm. The fact that the gait recognition results are encouraging upto angles of 45 degrees allows us to hypothesize that it is possible to do reasonable human identification using gait with only two cameras (installed perpendicular to each other). This could prove to be less restrictive than the visual hull approach that needs at least 4 cameras. A simple approach for correcting the view synthesized from a non-canonical view, independent of the spatial location of the object was discussed. This approach can be extended to more general 3-D models of objects. One of the uses of the approach can be for example, for synthesizing canonical views of distant faces. The approach also has potential for use in video communications. For example, if the azimuth angle of a dynamic planar patch can be estimated by the method discussed in Chapter 5, then in order to reproduce the sequence at the receiver end all we would need to send is a canonical image of the object, the azimuth angle along with a few motion parameters. We also presented preliminary results of a face and gait fusion algorithm. It would be interesting to study the fusion at the data level of face and gait cues via 3-D models for view invariant recognition.

Appendix A

Appendix

Proof of Equation (5.3):

Assuming a constant velocity model

$$X(t) = X_{ref} + v_X t \tag{A.1}$$

$$Y(t) = Y_{ref} \tag{A.2}$$

$$Z(t) = Z_{ref} + v_Z t \tag{A.3}$$

Under perspective projection,

$$x(t) = f\frac{X(t)}{Z(t)} = f\frac{X_{ref} + v_X t}{Z_{ref} + v_Z t} \tag{A.4}$$

$$y(t) = f\frac{Y(t)}{Z(t)} = f\frac{Y_{ref}}{Z_{ref} + v_Z t} \tag{A.5}$$

Taking the time derivatives and simplifying

$$\dot{x}(t) = f\frac{Z_{ref}v_X - X_{ref}v_Z}{(Z_{ref} + v_Z t)^2} \tag{A.6}$$

$$\dot{y}(t) = f\frac{Y_{ref}v_Z}{(Z_{ref} + v_Z t)^2} \tag{A.7}$$

Dividing Equation A.6 by Equation A.7, we get

$$cot(\alpha) = \frac{\dot{x}(t)}{\dot{y}(t)} \tag{A.8}$$

$$= \frac{Z_{ref}v_X - X_{ref}v_Z}{-Y_{ref}v_Z} \tag{A.9}$$

Noting that $cot(\theta) = \frac{v_X}{v_Z}$ and multiplying and dividing Equation A.9 by $\frac{f}{Z_{ref}v_Z}$, we get

$$cot(\alpha) = \frac{fcot(\theta) - f\frac{X_{ref}}{Z_{ref}}}{-f\frac{Y_{ref}}{Z_{ref}}} \tag{A.10}$$

$$= \frac{x_{ref} - fcot(\theta)}{y_{ref}} \tag{A.11}$$

where $x_{ref} = f\frac{X_{ref}}{Z_{ref}}$ and $y_{ref} = f\frac{Y_{ref}}{Z_{ref}}$

Proof of Equation (5.10):

Let $[X_\theta, Y_\theta, Z_\theta]'$ denote the coordinates of any point on the person who is walking at an angle $\theta \geq 0$ to the plane passing through the starting point $[X_{ref}Y_{ref}Z_{ref}]'$ and parallel to the image plane which we shall refer to, hereafter, as the canonical plane. Rotation of this point on to the canonical plane by θ yields

$$\begin{bmatrix} X_{int} \\ Y_{int} \\ Z_{int} \end{bmatrix} = \begin{bmatrix} Cos(\theta) & 0 & Sin(\theta) \\ 0 & 1 & 0 \\ -Sin(\theta) & 0 & Cos(\theta) \end{bmatrix} \cdot \begin{bmatrix} X_\theta - Xref \\ Y_\theta - Y_{ref} \\ Z_\theta - Z_{ref} \end{bmatrix} = \begin{bmatrix} (X_\theta - Xref)Cos(\theta) + (Z_\theta - Z_{ref})Sin(\theta) \\ (Y_\theta - Y_{ref}) \\ -(X_\theta - X_{ref})Sin(\theta) + (Z_\theta - Z_{ref})Cos(\theta) \end{bmatrix}, \tag{A.12}$$

The coordinates of the synthesized point with respect to the coordinate frame attached to the camera are given as

$$\begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \end{bmatrix} = \begin{bmatrix} X_{int} \\ Y_{int} \\ Z_{int} \end{bmatrix} + \begin{bmatrix} X_{ref} \\ Y_{ref} \\ Z_{ref} \end{bmatrix} \tag{A.13}$$

which simplifies to

$$\begin{bmatrix} X_0 \\ Y_0 \\ Z_0 \end{bmatrix} = \begin{bmatrix} X_\theta Cos(\theta) + Z_\theta Sin(\theta) - X_{ref} Cos(\theta) - Z_{ref} Sin(\theta) + X_{ref} \\ Y_\theta \\ -X_\theta Sin(\theta) + Z_\theta Cos(\theta) - X_{ref} Sin(\theta) - Z_{ref} Cos(\theta) + Z_{ref} \end{bmatrix} \tag{A.14}$$

Under perspective projection,

$$x_0 = f\frac{X_0}{Z_0}, y_0 = f\frac{Y_0}{Z_0} \tag{A.15}$$

$$x_\theta = f\frac{X_\theta}{Z_\theta}, y_\theta = f\frac{Y_\theta}{Z_\theta.} \tag{A.16}$$

Substituting from Equation (A.14) in (A.15), we get

$$x_0 = f\frac{X_\theta Cos(\theta) + Z_\theta Sin(\theta) - X_{ref} Cos(\theta) - Z_{ref} Sin(\theta) + X_{ref}}{-X_\theta Sin(\theta) + Z_\theta Cos(\theta) - X_{ref} Sin(\theta) - Z_{ref} Cos(\theta) + Z_{ref}}, \tag{A.17}$$

$$y_0 = f\frac{Y_\theta}{-X_\theta Sin(\theta) + Z_\theta Cos(\theta) - X_{ref} Sin(\theta) - Z_{ref} Cos(\theta) + Z_{ref}}. \tag{A.18}$$

Multiplying and dividing the numerator and denominator of Equations A.17 and A.18 by $\frac{f}{Z_\theta}$ we get

$$x_0 = f\frac{x_\theta Cos(\theta) + fSin(\theta) + f\frac{X_{ref}}{Z_\theta}(1 - Cos(\theta)) - f\frac{Z_{ref}}{Z_\theta} Sin(\theta)}{-x_\theta Sin(\theta) + fCos(\theta) - f\frac{X_{ref}}{Z_\theta} Sin(\theta) + f\frac{Z_{ref}}{Z_\theta}(1 - Cos(\theta))}, \tag{A.19}$$

$$y_0 = f\frac{y_\theta}{-x_\theta Sin(\theta) + fCos(\theta) - f\frac{X_{ref}}{Z_\theta} Sin(\theta) + f\frac{Z_{ref}}{Z_\theta}(1 - Cos(\theta))}. \tag{A.20}$$

Now for $Z_{ref}, Z_\theta >> 0$, $f\frac{X_{ref}}{Z_\theta} \approx f\frac{X_{ref}}{Z_{ref}} = x_{ref}$ and $\frac{Z_{ref}}{Z_\theta} \approx 1$ This means that we consider the person far from the camera and for a few cycles only when he does not move too far from the starting point.

Thus Equations A.19 and A.20 simplify to

$$x_0 = f\frac{x_\theta Cos(\theta) + x_{ref}(1 - Cos(\theta))}{-Sin(\theta)(x_\theta + x_{ref}) + f}, \tag{A.21}$$

$$y_0 = f\frac{y_\theta}{-Sin(\theta)(x_\theta + x_{ref}) + f}. \tag{A.22}$$

Proof of Equation (5.12):

Considering only the front face of the rectangular block in Figure 5.15 the projected length under true and weak perspective perspective projection are given by (A.23)and (A.24)as

$$W_{persp} = \frac{aCos(\theta) - \frac{xaSin(\theta)}{Z}}{1 + \frac{aSin(\theta)}{Z}} \tag{A.23}$$

$$W_{weakpersp} = aCos(\theta) \tag{A.24}$$

BIBLIOGRAPHY

[1] P. J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. W. Bowyer, "The gait identification challenge problem: Data sets and baseline algorithm," *Proc of the International Conference on Pattern Recognition*, 2002.

[2] G Johansson, "Visual motion perception," *Scientific American*, vol. 232, pp. 76–88, 1975.

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

[4] J. N. Bassili, "Facial motion in the perception of faces and emotional expression," *Journal of Experimental Psychology:Human Perception and performance*, vol. 4, pp. 373–379, 1978.

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

[6] J. E. Cutting and D.R. Proffitt, *Gait perception as an example of how we perceive events*, Plenum Press, London, 1981.

[7] J. Cutting and L. Kozlowski, "Recognizing friends by their walk:gait perception without familiarity cues," *Bulletin of the Psychonomic Society*, vol. 9, pp. 353–356, 1977.

[8] C.D. Barclay, J. E. Cutting, and L.T. Kozlowski, “Temporal and spatial factors in gait perception that influence gender recognition,” *Perception and Psychophysics*, vol. 23, pp. 145–152, 1978.

[9] T. Beardsworth and T. Buckner, “The ability to recognize oneself from a video recording of ones movements without seeing ones body,” *Bulletin of the Psychonomic Society*, vol. 18, no. 1, pp. 19–22, 1981.

[10] S. V. Stevenage, M. S. Nixon, and K. Vince, “Visual analysis of gait as a cue to identity,” *Applied Cognitive Psychology*, , no. 13, pp. 513–526, March 1999.

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

[12] W. I Scholhorn, . Nigg B.M, D.J.Stephanshyn, and W. Liu, “Identification of individual walking patterns using time discrete and time continuous data sets,” *Gait and Posture*, vol. 15, pp. 180–186, 2002.

[13] L. Lee and W.E.L. Grimson, “Gait analysis for recognition and classification,” *Proceedings of the IEEE Conference on Face and Gesture Recognition*, pp. 155–161, 2002.

[14] D. Cunado, J.M. Nash, M.S. Nixon, and J. N. Carter, “Gait extraction and description by evidence-gathering,” *Proc. of the International Conference on Audio and Video Based Biometric Person Authentication*, pp. 43–48, 1995.

[15] P.S. Huang, C.J. Harris, and M.S. Nixon, “Recognizing humans by gait via parametric canonical space,” *Artificial Intelligence in Engineering*, vol. 13, no. 4, pp. 359–366, October 1999.

[16] R. Cutler C. Benabdelkader and L.S. Davis, “Motion based recognition of people in eigengait space,” *Proceedings of the IEEE Conference on Face and Gesture Recognition*, pp. 267–272, 2002.

[17] J. Little and J. Boyd, “Recognizing people by theirgait: the shape of motion,” *Videre*, vol. 1, no. 2, pp. 1–32, 1998.

[18] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-26 no. 1, pp. 43–49, 1978.

[19] Sadaoki Furui, “Cepstral analysis technique for automatic speaker verification,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, no. 2, pp. 254–272, April 1981.

[20] A. Elgammal, D. Harwood, and L. Davis, “Non-parametric model for background subtraction,” *FRAME-RATE Workshop, IEEE*, 1999.

[21] Jinu Mariam Zacharia, “Text-independent speaker verification using segmental, suprasegmental and source features,” M.S. thesis, IIT Madras Chennai India, March 2002.

[22] P. J. Philips, H. Moon, and S. A. Rizvi, “The feret evaluation methodology for face-recognition algorithms,” *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 22, no. 10, pp. 1090–1100, October 2000.

[23] P. J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. W. Bowyer, "Baseline results for the challenge problem of human id using gait analysis," *Proc. of the 5th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2002.

[24] L. Rabiner and H. Juang, *Fundamentals of speech recognition*, Prentice Hall, 1993.

[25] M. Anderberg, "Cluster analysis for applications," 1973.

[26] R. Duda and P. Hart, "Pattern classification and scene analysis," 1973.

[27] G. Schwartz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 497–511, 1978.

[28] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Transactions on Information Theory*, vol. IT-30, no. 4, pp. 629–636, 1984.

[29] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–285, February 1989.

[30] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[31] A. Sundaresan, A. RoyChowdhury, and R. Chellappa, "A hidden markov model based framework for recognition of humans from gait sequences," *Proc. of the International Conference on Image Processing*, 2003.

[32] M. Russell and R. K. Moore, "Explicit modelling of state occupancy in hidden markov models for automatic speech recognition," *Proceedings of IEEE Conference on Acoustics Speech and Signal Processing*, June 1985.

[33] D. Tolliver and R. Collins, "Gait shape estimation for identification," *Proceedings of AVBPA*, pp. 734–742, 2003.

[34] M.M. Kokar and J.A. Tomasik, "Data vs. decision fusion in the category theory framework," *FUSION 2001*, 2001.

[35] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp. 226–239, March 1998.

[36] B. H. Juang, "On the hidden markov model and dynamic time warping for speech recognition - a unified view," *Technical Journal*, vol. 63, pp. 1213–1243, 1984.

[37] G. Shakhnarovich and T. Darrell, "On probabilistic combination of face and gait cues for identification," *Proc. of the Int. Conf. on Automatic Face and Gesture Recognition*, 2002.

[38] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 23, no. 3, pp. 257–267, March 2001.

[39] O.D. Faugeras, *Three-Dimensional Computer Vision: A Geometric Viewpoint*, MIT Press, 1993.

[40] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.

[41] G.Shakhnarovich, L.Lee, and T.Darrell, "Integrated face and gait recognition from multiple views," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, December 2001.

[42] A.F. Bobick and A. Johnson, "Gait recognition using static activity-specific parameters," *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[43] Michael Isard and Andrew Blake, "Contour tracking by stochastic propagation of conditional density," *Proceedings of ECCV*, , no. 1, pp. 343–356, 1996.

[44] Vishwjit Nalwa, *A Guided Tour of Computer Vision*, Addison-Wesley, 1993.

[45] Athanasios Papoulis, *Probability, Random Variables and Stochastic Processes*, McGraw Hill, 1991.

[46] A. Kale, N. Cuntoor, and R. Chellappa, "A framework for activity-specific human recognition," *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (Orlando, FL)*, May 2002.

[47] Ralph Gross and Jianbo Shi, "The cmu motion of body (mobo) database," Tech. Rep. CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, June 2001.

[48] Shaohua Zhou and R. Chellappa, "Probabilistic human recognition from video," *Proceedings of ECCV*, 2002.

[49] B. Achermann and H. Bunke, "Combination of classifiers on the decision level for face recognition," Tech. Rep., Institut fur Informatik und angewandte, Mathematik,, Universitat Bern, 1996.