University of South Florida

DIGITAL COMMONS
@ UNIVERSITY OF SOUTH FLORIDA

Digital Commons @ University of South Florida

USF Tampa Graduate Theses and Dissertations

USF Graduate Theses and Dissertations

11-10-2004

Gait-Based Recognition at a Distance: Performance, Covariate Impact and Solutions

Zongyi Liu
University of South Florida

Follow this and additional works at: https://digitalcommons.usf.edu/etd

 Part of the American Studies Commons

Scholar Commons Citation

Liu, Zongyi, "Gait-Based Recognition at a Distance: Performance, Covariate Impact and Solutions" (2004). *USF Tampa Graduate Theses and Dissertations.*
https://digitalcommons.usf.edu/etd/1134

This Dissertation is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact digitalcommons@usf.edu.

Gait-Based Human Recognition at a Distance: Performance, Covariate Impact and Solutions

by

Zongyi Liu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Major Professor: Sudeep Sarkar, Ph.D.
Dmitry Goldgof, Ph.D.
Kevin Bowyer, Ph.D.
Tapas Das, Ph.D.
Rangachar Kasturi, Ph.D.

Date of Approval:
November 10, 2004

Keywords: Gait Biometrics, Baseline, Population HMM, Eigen-Stance, LDA

© Copyright 2004, Zongyi Liu

DEDICATION

To my fiancee and my family members!

ACKNOWLEDGEMENTS

This research was supported by funds from the DARPA Human ID program (F49620-00-1-00388). Here I sincerely thank my major professor, Dr. Sudeep Sarkar, for his kind and patient help and guidance in the past 3 years, which greatly helped my academic background and research abilities. Ning Yang from Electrical Engineering of USF helped groundtruthing the face images. Laura Malave, Adebola Osuntugun, Preksha Sudhakar, and Christine Bexley helped to meticulously put together the manual silhouette database. The code developed by Ross Beveridge *et al.* at CSU is used to perform face recognition. I also thank Dr. Kevin Bowyer and P. Jonathon Phillips for their comments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

GAIT-BASED HUMAN RECOGNITION AT A DISTANCE: PERFORMANCE, COVARIATE IMPACT AND SOLUTIONS

Zongyi Liu

ABSTRACT

It has been noticed for a long time that humans can identify others based on their biological movement from a distance. However, it is only recently that computer vision based gait biometrics has received much attention. In this dissertation, we perform a thorough study of gait recognition from a computer vision perspective. We first present a parameterless baseline recognition algorithm, which bases similarity on spatio-temporal correlation that emphasizes gait dynamics as well as gait shapes. Our experiments are performed with three popular gait databases: the USF/NIST HumanID Gait Challenge outdoor database with 122 subjects, the UMD outdoor database with 55 subjects, and the CMU Mobo indoor database with 25 subjects. Despite its simplicity, the baseline algorithm shows strong recognition power. On the other hand, the outcome suggests that changes in surface and time have strong impact on recognition with significant drop in performance. To gain insight into the effects of image segmentation on recognition – a possible cause for performance degradation, we propose a silhouette reconstruction method based on a Population Hidden Markov Model (pHMM), which models gait over one cycle, coupled with an Eigen-stance model utilizing the Principle Component Analysis (PCA) of the silhouette shapes. Both models are built from a set of manually created silhouettes of 71 subjects. Given a sequence of machine segmented silhouettes, each frame is matched into a stance by pHMM using the Viterbi algorithm, and then is projected into and reconstructed by the Eigen-stance model. We demonstrate that the system dramatically improves the

silhouette quality. Nonetheless, it does little help for recognition, indicating that segmentation is not the key factor of the covariate impacts. To improve performance, we look into other aspects. Toward this end, we propose three recognition algorithms: (i) an averaged silhouette based algorithm that deemphasizes gait dynamics, which substantially reduces computation time but achieves similar recognition power with the baseline algorithm; (ii) an algorithm that normalizes gait dynamics using pHMM and then uses Euclidean distance between corresponding selected stances – this improves recognition over surface and time; and (iii) an algorithm that also performs gait dynamics normalization using pHMM, but instead of Euclidean distances, we consider distances in shape space based on the Linear Discriminant Analysis (LDA) and consider measures that are invariant to morphological deformation of silhouettes. This algorithm statistically improves the recognition over all covariates. Compared with the best reported algorithm to date, it improves the top-rank identification rate (gallery size: 122 subjects) for comparison across hard covariates: briefcase, surface type and time, by 22%, 14%, and 12% respectively. In addition to better gait algorithms, we also study multi-biometrics combination to improve outdoor biometric performance, specifically, fusing with face data. We choose outdoor face recognition, a "known" hard problem in face biometrics, and test four combination schemes: score sum, Bayesian rule, confidence score sum, and rank sum. We find that the recognition power after combination is significantly stronger although individual biometrics are weak, suggesting another effective approach to improve biometric recognition. The fundamental contributions of this work include (i) establishing the "hard" problems for gait recognition involving comparison across time, surface, and briefcase carrying conditions, (ii) revealing that their impacts cannot be explained by silhouette segmentation, (iii) demonstrating that gait shape is more important than gait dynamics in recognition, and (iv) proposing a novel gait algorithm that outperforms other gait algorithms to date.

CHAPTER 1

INTRODUCTION

1.1 Overview of Gait Study History

The observation that people are able to identify a friend from a distance is a commonly reported experience. The speculation is that it is the pattern of walking, i.e., gait, that provides the information for recognition. The major early study was done by Johansson [38], who used point light displays to demonstrate the ability of humans to rapidly distinguish human locomotion from other motion patterns. Cutting and Kozlowski [17] showed that this ability also extends to recognition of friends. Since then, there has been various experiments to show that humans can recognize gender [74, 4], direction of motion [38, 19, 32], and weight-carrying conditions [73]. Perhaps the most recent evidence comes from the experiments by Stevenage *et al.* [82] who show that humans can identify individuals on the basis of their gait signature, without reliance on shape, in the presence of lighting variations and under brief exposures. In addition, people also investigated movement perception of partial body, usually the limbs. For example, Pollick *et al.* [68] examined visual perception of affect from point-light displays of arm movement. Their results showed that the first dimension of psychological space was highly correlated to the kinematics, for both natural and scrambled arm movements, so that the perceived affect should be a formless cue directly related to the kinematics.

Human motion recognition has also been studied by neuroscientists. Grossman *et al.* [27, 21] studied the activities of different brain regions during the viewing of point-light figures. The comparison of areas involved in coherent-motion perception and kinetic-boundary perception indicated the existence of neural mechanisms in brain specialized

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

Figure 1.1. Examples of Important Gait Phases in One Cycle, (a) Right Heel Strike, (b) Left Toe-Off, (c) Left Heel-Strike, (d) Right Toe-Off, and (e) Right Heel-Strike.

for biological motion (kinematics) analysis. Grezes *et al.* [28] explored the hemodynamic responses of 10 healthy people to seven types of visual motion displays. Their results showed that non-rigid biological motions are perceived by both the posterior portion of superior temporal sulcus and the left intraparietal cortex.

Due to the periodic nature of human walking, a gait cycle is usually defined as the basic unit of gait. According to Murray *et al.* [64], a gait cycle is the time interval starting from the right heel striking the floor, followed by the swing of left leg advancing forward, the the left heel striking the floor, the right leg swinging to advance forward, and ending at the right heel striking the floor again, as illustrated in Fig. 1.1. There are 2 important phases in a gait cycle: the left/right heel-strike (see Fig. 1.1 (a), (c) and (e)) where the two legs are fully apart, and the left/right toe-off (see Fig. 1.1 (b) and (d)) where one leg just leaves the ground. A gait cycle can also be partitioned into four periods: (i) right stance period when the right foot is in contact with the floor, beginning from "right heel-strike" and ending at "right toe-off"; (ii) left swing period when the left foot is not in contact with the floor, beginning from "left toe-off" and ending at "left heel-strike"; (iii) left stance period when the left foot is in contact with the floor, beginning from "left heel-strike" and ending at "left toe-off"; and (iv) right swing period when the right foot is not in contact with the floor, beginning from "right toe-off" and ending at "right heel-strike". Moreover, the time

between these periods, i.e., when both feet are in contact with the floor, is called "double limb support".

In computer vision, much progress has been made in studying human motion since the early days of analyzing human motion in terms of groups of rigidly moving points [22, 96]. An excellent snapshot into current work on human movement modeling is available in a recent special issue [31]. Work in computer vision based human motion modeling can be classified according to the model employed: articulated vs. elastic non-rigid, with and without prior shape modeled [2]; or in terms of whether 2D or 3D models are implicitly or explicitly employed [24]. A more recent, extensive survey [63] looks at over 130 publications in computer vision-based human motion analysis and classifies them based on the issues addressed: initialization (8 publications), tracking (48 publications), pose estimation (64 publications), and recognition (16 publication). The review also finds that the three most common assumptions used effectively constrain the scene to be (i) indoors, (ii) with static background, and (iii) with uniform background color. These assumptions makes it difficult to judge the autonomous operation of the developed ideas in real life outdoor situations.

In the specific area of gait recognition, most works have focused on discriminating between different human motion types, such as running, walking, jogging, or climbing stairs [77]. It is only recently that human identification (HumanID) from gait has received attention and become an active area of computer vision [65, 54, 81, 7, 85, 79, 30, 6, 53, 46, 15, 93, 89, 45, 84, 16, 86, 39]. Compared with traditional biometrics, such as face, iris and fingerprint, gait has the advantages that it can be acquired from a distance, while face or fingerprint data collection usually requires subjects be close to sensors. Of course, gait recognition also suffers several drawbacks, for example, data are easily affected by various sources of noise, such as shadows, moving background objects, and low image quality, that cause problems in segmentation. In addition, gait can also be impacted by surface, shoe-type and weight carried.

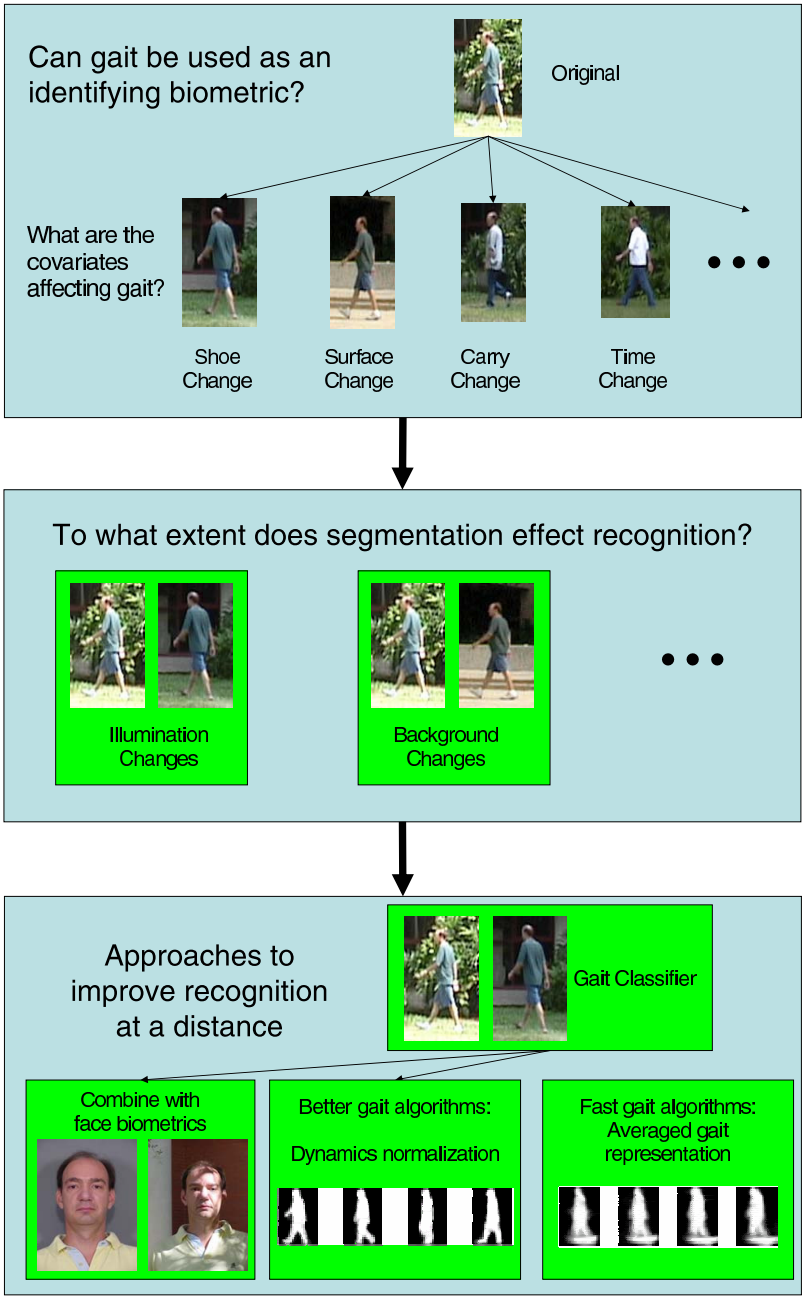


Figure 1.2. Scientific Issues Addressed in this Work.

1.2 Scientific Issues Addressed in this Work

Scientific issues considered in this dissertation, as Fig. 1.2 shows, include (i) potential of gait biometrics, (ii) impact of silhouette quality on recognition, and (iii) investigation of methods to improve gait recognition performance.

In Chapter 2 we begin our study by reviewing the state of art algorithms, which can be grouped into 3 types: temporal alignment based, shape based, and static parameters based algorithm. We introduce several large gait databases, including the USF/NIST HumanID database with 122 subjects, UMD database with 55 subjects, CMU Mobo database with 25 subjects, and U. of Southampton (SOTON) database with about 100 subjects.

In Chapter 3 we study the impact of different covariates on recognition. We present a parameterless baseline algorithm based on a simple spatio-temporal correlation technique. Despite its simplicity, the algorithm shows strong recognition power. However, it also suggests that surface and time changes are the "hard" problems where significant performance degradation are observed. This drop in recognition performance with surface and time changes has also been reported by gait research groups for the HumanID gait challenge dataset [16, 44, 94, 89].

To discover underlying factors that affect performance, we study the impact of silhouette quality on recognition in Chapter 4. It is reasonable to speculate that since silhouette quality varies with background and environmental changes, the drop in performance can be attributed to silhouette quality. Instead of proposing a new segmentation algorithm, our approach is to refine silhouettes produced by other algorithms. We build a Population Hidden Markov Model (pHMM) to model a full gait cycle, starting from the right leg stride and ending at the same stance in the next cycle. The pHMM is trained on a set of manual silhouettes comprised of 71 subjects using Baum-Welch algorithm [70]; and unlike other sequence-specific HMMs, our pHMM does not perform classification, instead, it matches frames of a given sequence into a set of predefined states following the Viterbi algorithm [70], by picking up the most likely transition sequence. Then, each frame is projected into a corresponding eigen-stance space, also built from manual silhouettes,

and reconstructed. We show that the silhouette quality after reconstruction improves, as measured by the traditional detection rate (P_D) and false positive prediction rate ($\overline{PPV}$). However, the improved silhouette does not help recognition. On the contrary, the performance even dropped slightly, due to the removal of error correlations for some of the experiences that involved matching sequences taken roughly around the same time of day and same background. So that we assert that the impact of the "hard" covariates can not be explained by the silhouette quality. To improve the recognition, one should look at other aspects of the problem.

In Chapter 5 we investigate methods to improve recognition performance. We propose three algorithms. First, we propose an averaged silhouette based algorithm. Compared with the baseline algorithm, it deemphasizes gait dynamics. It substantially reduces computation time but achieves similar recognition power. Our second algorithm normalizes gait dynamics using a Population Hidden Markov Model (pHMM), and bases similarity measurements on Euclidean distances between gait shapes of selected stances. This algorithm shows stronger recognition power for the "hard" problems involving surface and time changes. The third algorithm also performs gait dynamics using pHMM, but differs with respect to the similarity computation stage. Instead of shape Euclidean distances, it uses distances in the Linear Discriminant Analysis (LDA) shape space, which suppresses within-subject shape variations but emphasizes differences across subjects. The distance measure is also structured to be invariant to morphological deformation so as to be robust with respect to within-subject body width differences that might arise due to segmentation. We show that this algorithm significantly improves the performance with variation of covariates, especially for surface, briefcase, and time changes. In addition to the USF/NIST HumanID database, we also demonstrate its generalizability, without re-training, to other databases, such as the UMD outdoor database (involving time differences) and the CMU Mobo database (involving speed differences).

In Chapter 6 we study another approach to improve performance: multi-biometrics combination, specifically, fusing with face data. For gait recognition, we use dynamics

normalization with Euclidean distance based similarities. For face recognition, we use the Elastic Bunch Graph Match algorithm (EBGM) proposed by Wiskott [60]. Early studies found this algorithm to perform better than others [66]. The experiment is designed to evaluate the power of biometric fusion to overcome low performance on the "hard" problems of individual biometrics, specifically, involving outdoor data, time and surface changes. We experiment with four combination schemes: score sum, Bayesian rule, confidence score sum, and rank sum. The results show that although the accuracy of individual classifier is low, their combination has much stronger recognition power even with simple fusing schemes. It also reveals that the inter-modal (face+gait) is better than intra-modal (face+face or gait+gait), due to the low correlation between the intra-modal scores.

In summary, the specific scientific contributions of this dissertation are that

1. We establish that matching across surface type and time are "hard" problems for gait recognition. The effect of shoe-type changes is small. Surprisingly, carrying a briefcase does not impact recognition to as large an extent as surface. Viewpoint variation of 30° does not impact recognition, suggesting that 2D silhouettes might be insufficient as an input features for a wide range of viewing directions.

2. We demonstrate that silhouette quality is not the key factor affecting recognition. Quality of silhouettes produced by standard background subtraction based algorithms appear to be sufficient.

3. We present four new gait algorithms, including one that achieves the best recognition performance over all recognition algorithms to date. We also illustrate that just gait shape is sufficient for recognition – dynamics normalization helps.

4. We investigate the effects of biometrics combination, and show its power for improving performance over individual biometrics. Specifically, we show that in outdoor conditions fusing face with gait would significantly improve performance.

CHAPTER 2

RELATED WORK

Before we begin our study, we first review the state of the art in gait recognition. Table 2.1 summarizes the recent works in terms of the algorithmic approaches, datasets used, and identification performance for matching across different covariates, which we will describe in more details in the following sections.

A few words regarding biometrics nomenclature are in order before we describe recently proposed approaches to gait recognition. The term *gallery* is used to refer to the set of templates or sequences stored in the model base. *Probes* are the unknown templates to be identified or verified. In an identification scenario, one is interested in finding a match to a given probe from the whole gallery set, i.e. one-to-many match. In a verification scenario, one is interested in deciding whether a given probe matches a hypothesized or claimed gallery identity, i.e. one-to-one match. Performance for the identification scenario is captured by the Cumulative Match Characteristic (CMC) [41], which plots identification rates (P_I) within a given rank k. For the verification scenario the standard Receiver Operator Characteristic (ROC) is used. ROC curve plots the correct detection rate against the false alarm rate for various choices of the decision threshold.

2.1 Approaches

Gait recognition approaches, especially those that have been shown to work for more than 20 persons, are basically of three types: (i) temporal alignment based, (ii) silhouette shape based, and (iii) static parameter based approaches.

Table 2.1. Summary of Recent Gait Recognition Algorithms and their Performance.

Algorithm	Scene	Data Covariates	P_I at rank 1	Size (subjects)
Spatio-temporal correlation (USF, Baseline v1) [43, 42]	Outdoor	Viewpoint	79%	71
	Outdoor	Shoe	66%	71
	Outdoor	Surface	29%	71
Space of Probability Functions (USF) [91, 92]	Outdoor	Viewpoint	68%	71
	Outdoor	Shoe	61%	71
	Outdoor	Surface	12%	71
Temporal body length vector correlation (CAS) [93, 94]	Indoor	Time(minutes)	91%	28
	Outdoor	Viewpoint	71%	71
	Outdoor	Shoe	59%	71
	Outdoor	Surface	34%	71
Area Based Metrics (SOTON) [23]	Indoor	Sessions	75%	114
Fourier Descriptor (SOTON) [83, 101]	Indoor	Sessions	85%	115
		Speed	86%	116
Silhouette region moment and HMM (MIT) [53, 52]	Indoor	Temporal (days)	30-60%	24
	Outdoor	Viewpoint	88%	71
	Outdoor	Shoe	75%	71
	Outdoor	Surface	25%	71
Shape and kinematics (UMD) [46, 47, 90]	Indoor	Speed, viewpoint	58%	25
	Indoor	Time(3 months)	30%	24
	Outdoor	Session	55%	43
	Outdoor	Viewpoint	98%	122
	Outdoor	Shoe	85%	122
	Outdoor	Surface	36%	122
Body shape and template correlation (CMU) [15, 89]	Indoor	Speed, viewpoint	76%	25
	Indoor	Time (3 months)	45%	24
	Outdoor	Sessions	85%	55
	Outdoor	Viewpoint	87%	71
	Outdoor	Shoe	81%	71
	Outdoor	Surface	21%	71
Static body parameters (Georgia Tech.) [39, 85, 86]	Indoor	Viewpoint	$> 90\%$	18
	Indoor	Sessions	73%	18
	Magnetic	Speed	20-40%	15
Motion energy and history (UC Riverside) [29]	Outdoor	Viewpoint	91%	122
	Outdoor	Shoe	94%	122
	Outdoor	Surface	51%	122
	Outdoor	Briefcase	62%	122
	Outdoor	Time	18%	122

2.1.1 Temporal Alignment Based Approaches

The most common approach treats the sequence as a time series and involves three components. The first component is the extraction of features such as whole silhouettes, silhouette width vectors, silhouette boundary, silhouette moments, and so on. The second step involves the alignment of sequences of these features, corresponding to the given two sequences to be matched. The third aspect is the distance measure used. A simple version of this approach, Sarkar *et al.* (USF/NIST) proposed the spatio-temporal correlation idea as a baseline gait recognition algorithm v1, along with the HumanID Gait Challenge problem [43, 42]. The input feature is the whole silhouette. Given two sequences, one of them (the probe) is first partitioned into subsequences of approximately one gait cycle length. Each probe subsequence is temporally correlated with the other full sequence (the gallery) and the maximum correlation is noted. The similarity between two silhouette frames is chosen to be the Tanimoto distance, i.e. ratio of the number of pixels in the intersection of the silhouettes to the number in their union. The median value of these maximum correlations for the probe subsequences is chosen as the overall similarity value. The breaking up of the probe into subsequences helps in overcoming silhouette segmentation errors that may occur in bursts along a sequences due to background or illumination changes. Despite its simplicity, the baseline algorithm achieves very competitive performance.

Robledo and Sarkar [91, 92] (USF) proposed an approach of relational distributions and space of probability functions (SoPF). This method includes four stages: (i) segment person from a motion sequence, using the binary silhouette representation, (ii) extract low level features and build relational distributions – accumulated occurrences of each relationship between paired image features, (iii) build a space of probability functions (SoPF) from the relational distributions of a training dataset, and use PCA to reduce dimensions, and (iv) project relational distributions of test data into SoPF, and compute similarities based on their coordinates.

Tan *et al.* [93, 94] (CAS) considered the body length vector as the feature, which they compute from the vector of distances to the silhouette boundary from the silhouette center.

The distance vector is then normalized with respect to the magnitude and size. These one-dimensional vectors are then represented in a smaller dimensional space using PCA. Two sequences of body length vector are aligned by simple correlation and normalized Euclidean distances are computed.

The UMD group's approach uses the silhouette width vector as the feature [46, 47]. The width vector is defined to be the vector of silhouette widths at each row. The silhouettes are height normalized to arrive at vector of fixed lengths. They have exhaustively experimented with various modification of this basic idea as a feature. Sequence alignment is achieved based on *person specific* Hidden Markov models. In this approach, the gait of *each* person is represented as sequence of state transitions. The states correspond to the different gait stances and the observation model for each state is represented as distances from the average stance shape for that person. The HMM is built is using the Baum-Welch algorithm and recognition is performed by matching any given sequence to the HMMs using the Viterbi algorithm. The identity of the HMM that results in the maximum probability match is selected. In a more recent version of this basic approach [90], they use better shape representation, with much improved performance. First, they extract pre-shape vector by subtracting the centroid and normalizing for scale. Then, correlation with these shape vectors is performed using dynamic time-warping in shape space, or using HMM with shape cues based on Procrustes distance.

We also use HMMs in our work. However, there are a number of essential differences. First, we do not have *person specific* HMMs; we use a *population* HMM model, which can be looked upon as a *generic* walking gait model. Second, the HMM is *not* used for recognition; it is used just to stance align the frames of two sequences. Third, temporal dynamics play no role in the similarity computation.

Lee and Grimson [53] opt for more high level features. They partitioned a silhouette into 7 elongated regions, corresponding to the different body parts. Each region is represented using four features: centroid coordinate pair, aspect ratio, and orientation. Gait similarity is computed from either the average of the features in all the frames of a sequence, or the

magnitudes and phases of each region feature related to the dominant walking frequency. In a more recent work [52], they have experimented with HMM based alignment of two sequences, with improved results.

Lee and Elgammal [50] first interpolate gait cycles into N equally spaced time instances. To reduce the gait dimension, they use nonlinear dimensionality reduction frameworks, such as locally linear embedding and isometric feature mapping. Then, they learn a symmetric bilinear model from the synthesized gait frames. A support vector machine (SVM) is used for final classification.

2.1.2 Shape Based Approaches

This class of approaches emphasizes the silhouette shape similarity and disregards temporal information. One direction involves the transformation of the silhouette sequence into a single image representation. The simplest such transformation is the averaged silhouette, computed by simply summing the silhouettes over one gait cycle [58]. Similarity is based on just the Euclidean distance between the average silhouettes from two silhouette sequences. The performance is as good as the baseline algorithm, discussed earlier. A sophisticated version of this idea, with enhanced performance, was also proposed by Huan and Bhanu [29]. Instead of just over one gait cycle, they sum all the silhouettes in the sequence, followed by a reduced dimensional representation built using the PCA or Linear Discriminant Analysis (LDA). The training process is enhanced by synthesizing training data based on expected errors of the silhouettes. Similarity is computed in this linear subspace.

Nixon *et al.* (SOTON) have performed various analysis on indoor gait data. One of their approaches is based on body shape area [23]. It first masks selected body parts in a sequence of silhouettes, then measures area as a time varying signal. The dynamic temporal signal is used as a signature for automatic gait recognition. They tested the approach on an indoor gait database consisting of 114 subjects filmed under laboratory condition and achieved 75% recognition rate. They also proposed a Fourier descriptor

based algorithm [83, 101]. It models subject's boundary and spatio-temporal deformations with Fourier descriptors. They found that Fourier descriptors obtained from individuals appear to be unique and can be used for recognition. And they showed that over 85% of people are correctly identified in an indoor dataset of around 115 people.

Another way of using shape information preserves individual silhouettes but disregards the sequence ordering, and treats the sequences as just a collection of silhouette shapes [15, 89] (CMU). The silhouettes are first vertically normalized and horizontally centered, based on the first and second order moments. Then, silhouettes with similar shapes are clustered using the spectral partitioning framework, based on graph weights built out of the correlation of high variance areas in the shape, representing parts such as arms and legs. The power of this representation partly derives from the ability to identify and disregard low-quality silhouettes in a sequence; bad silhouettes form a separate cluster. A probe is identified by comparing the collection of its silhouette shapes to the gallery shape clusters.

Our best gait algorithm (dynamics normalization + LDA + morphological deformation) also falls in this category of gait algorithms that emphasize the silhouette shape over dynamics. However, unlike the approaches that arrive at one representations averaged over all the stances, we use stance specific representations. Like the CMU approach we do ignore the dynamics between the stances, but unlike the CMU approach, we do exploit the temporal ordering of the individual gait stances.

2.1.3 Static Parameters

The third class of approaches opts for parameters that can be used to characterize gait dynamics, such as stride length, cadence, and stride speed [39]. Sometimes static body parameters such as the ratio of sizes of various body parts are considered in conjunction with these parameters [39, 85, 86]. However, these approaches have not reported high performances on common databases, partly due to their need for 3D calibration information.

2.2 Gait Database Overview

Gait databases consist of a set of video sequences that require larger storage capacity than biometrics such as face database and fingerprint database. So usually the number of subjects in a gait database is relatively few. Currently there are around 10 gait databases with size varying from 6 subjects to 122 subjects. And in this section we describe the large databases that are available.

2.2.1 USF/NIST HumanID Gait Challenge Database

The USF/NIST HumanID Gait Challenge Database is currently the largest gait database consisting of 122 subjects and spanning up to 32 different conditions. These conditions are the result of all combinations of five covariates with two values each. It has been used by a number of research groups and has become a standard database for performance measurement. It is also the database used most in the dissertation. So here we describe it in detail.

2.2.1.1 Dataset

The gait video data was collected at the University of South Florida on May 20-21 and November 15-16, 2001. Participation in the collection process was voluntary. The collection process started with subjects being asked to read, understand, and sign an Institutional Review Board (IRB) approved consent form. The collection protocol had each subject walk multiple times counterclockwise around each of two similar sized and shaped elliptical courses. The basic setup is illustrated in Fig. 2.1. The elliptical courses were approximately 15 meters on the major axis and 5 meters on the minor axis. Both courses were outdoors. One course was laid out on a flat concrete walking surface. The other was laid out on a typical grass lawn surface. Each course was viewed by two cameras, whose lines of sight were not parallel, but verged at approximately 30°, so that the whole ellipse was just visible from each of the two cameras. When a person walked along the rear portion of the ellipse, their view was approximately fronto-parallel. Fig. 2.2 shows one

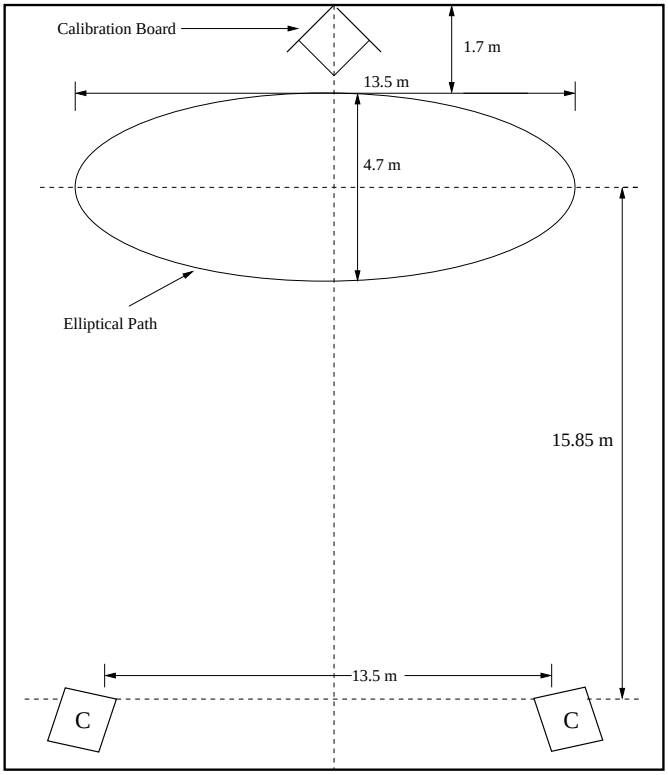


Figure 2.1. Camera Setup for the USF HumanID Data Acquisition.

sample frame from each of the four cameras on the two surfaces. The orange traffic cones marked the major axes of the ellipses. Although data from one full elliptical circuit for each condition is available, the challenge experiments are presented on the data from the rear portion of the ellipse.

Subjects were asked to bring a second pair of shoes, so that they could walk the two ellipses a second time in a different pair of shoes. A little over half of the subjects walked in two different shoe types. In addition, subjects were also asked to walk the ellipses carrying a briefcase of known weight (approximately 6 kilograms). Most subjects walked both carrying and not carrying the briefcase. In this dissertation we denote the values of each of the covariates by the following: 1) surface type by G for grass and C for concrete; 2) camera by R for right and L for left; 3) shoe type by A or B; 4) NB for not carrying a briefcase and BF for carrying a briefcase; and 5) the acquisition time, May and November, simply

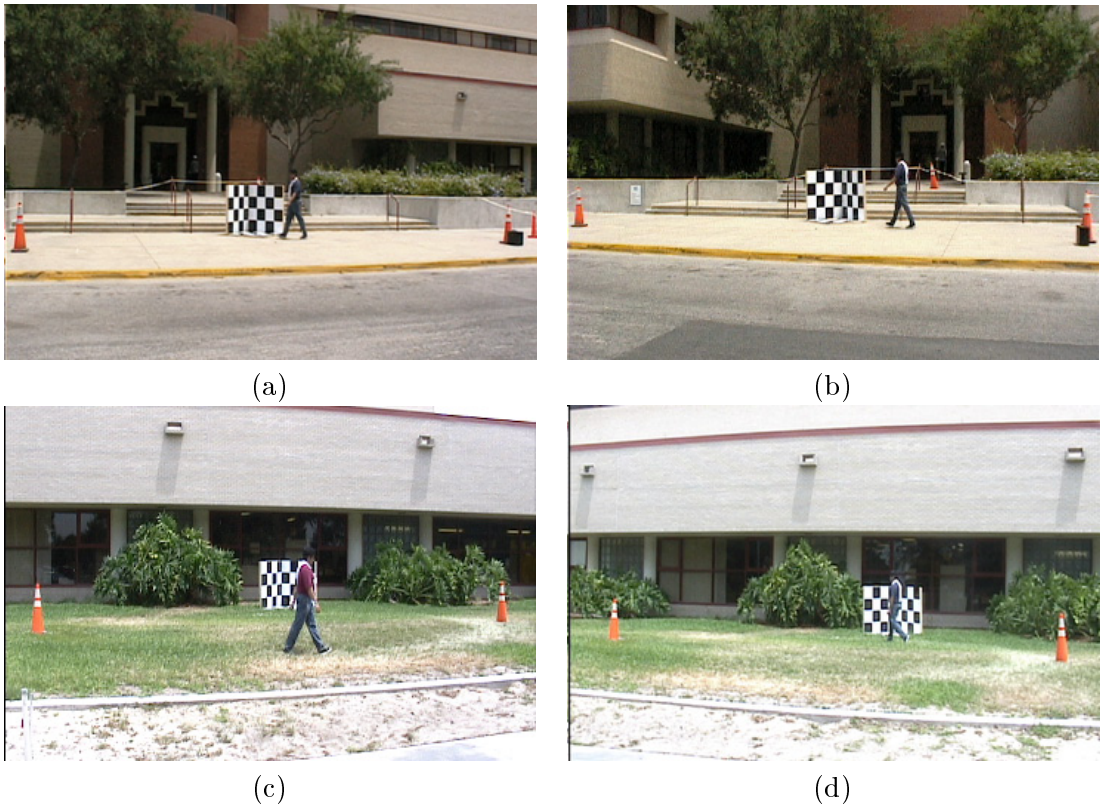


Figure 2.2. Frames from (a) the Left Camera for Concrete Surface, (b) the Right Camera for Concrete Surface, (c) the Left Camera for Grass Surface, (d) the Right Camera for Grass Surface.

by M and N. There are 33 subjects who were common between the May and November collections, so for them we also have data that exercise the time covariate. Table 2.2 shows the number of sequences for subjects that participated in the data collection for different covariate combinations.

2.2.1.2 Evaluation scheme

There are twelve challenge experiments defined for this dataset. The twelve experiments are designed to investigate the effect of five factors: *View, Shoe, Surface, Carry, Time*, on performance. The five factors are studied both individually and in combinations. The results of the baseline algorithm, described later, for the twelve experiments provide an ordering on the difficulty of the experiments.

Table 2.2. Number of Sequences for Each Combination of Possible Surface (G or C), Shoe (A or B), Camera View (L or R), Carry Condition (BF, NB) for People Who Participated in the Data Collection. The Last Row Lists Numbers of People Who were in Both Data Collections for Two Cases.

Surface	Carry	Shoe	Camera	Time	
				M or N	N
Concrete	NB	A	(L, R)	121	33
	NB	B	(L, R)	60	
	BF	A	(L, R)	121	
	BF	B	(L, R)	60	
Grass	NB	A	(L, R)	122	33
	NB	B	(L, R)	54	
	BF	A	(L, R)	120	
	BF	B	(L, R)	60	

We structured the challenge tasks in terms of gallery and probe sets, patterned on the FERET evaluations [41]. In biometrics nomenclature, the gallery is the set of people known to an algorithm or system, and probes are signatures given to an algorithm to be recognized. In this paper, signatures are video sequences of gait.

To allow for a comparison among a set of experiments and limit the total number of experiments, we fixed one gallery as the control. Then we created twelve probe sets to examine the effects of different covariates on performance. The gallery consists of sequences with the following covariates: Grass, Shoe Type A, Right Camera, No Briefcase, and collected in May along with those from the *new* subjects from November. This set was selected as the gallery because it was one of the largest for a given set of covariates. The structure of the twelve probe sets is listed in Table 2.3. The last two experiments study the impact of time. The time covariate implicitly includes a change of shoes and clothes because we did not require subjects to wear the same clothes or shoes in both data collections. We do have record of the shoe types that were used, but since subjects did not necessarily wear the same shoe six months later, the shoes did not match across time for all the subjects; for a subject, a "Shoe A" label in the May data does not necessarily refer to the same shoe as the "Shoe A" label in the November data. That is why in Table 2.3, we use A/B

Table 2.3. The Probe Set for Each of Challenge Experiments. The Gallery for All of the Experiments is (G, A, R, NB, M Or N) and Consists of 122 Individuals. Key experiments with only one covariate are in italics.

Exp.	Probe (Surface, Shoe, Camera, Carry, Time) (C/G, A/B, L/R, NB/BF, time)	Number of Subjects	Difference
A	*(G, A, L, NB, M or N)*	*122*	*View*
B	*(G, B, R, NB, M or N)*	*54*	*Shoe*
C	(G, B, L, NB, M or N)	54	Shoe, View
D	*(C, A, R, NB, M or N)*	*121*	*Surface*
E	(C, B, R, NB, M or N)	60	Surface, Shoe
F	(C, A, L, NB, M or N)	121	Surface, View
G	(C, B, L, NB, M or N)	60	Surface, Shoe, View
H	*(G, A, R, BF, M or N)*	*120*	*Briefcase*
I	(G, B, R, BF, M or N)	60	Shoe, Briefcase
J	(G, A, L, BF, M or N)	120	View, Briefcase
K	*(G, A/B, R, NB, N)*	*33*	*Time (Shoe, Clothing)*
L	(C, A/B, R, NB, N)	33	Surf., Time (Shoe, Clothing)

for shoe type in experiments K and L. However, the shoe labels within the May data and within the November data are consistent.

This database and the corresponding experiments are also known as the *gait challenge problem* [75]. Due to its large number of subjects, it is most often used for benchmarking algorithms in this dissertation.

2.3 UMD Database

There are two UMD gait datasets: dataset-1 consists of walking sequences of 25 subjects, and Dataset-2 contains walking sequences of 55 subjects walking along a T-shape pathway. In this paper, we use the larger one: dataset-2, taken outdoor by two surveillance cameras (Philips G3 EnviroDome camera system) at a height of 4.57 meters. Fig. 2.3 shows one sample frame. Each video sequence has approximately 10 gait cycles, viewed frontly and sideways. The database is diverse in terms of gender, age, and ethnicity. Moreover, like the gait challenge database, data collected on different days differ with respect to clothing

Figure 2.3. Sample of UMD Gait Database in which Subjects Walked Along a T-Shape Pathway in Outdoor, where the Side-View Portion for is Used for Recognition.

as well. There are two significant differences in imaging protocol with the gait challenge dataset: (i) the camera sample rate of the UMD data is 20 frames per second (f/s) but that of the gait challenge data is 30 f/s, and (ii) the camera was setup at 4.57 meters from the ground for the UMD data but it was 1.65 meters high for the gait challenge data.

The UMD dataset-2 offers us an opportunity to test gait recognition with short term (days) *time* differences for 55 subjects. Specifically, we use the UMD specifications of *experiment 1 for dataset-2*, which compares sequences taken on different days. For a more detailed description of the dataset and the experiment specification, please refer to the website *http://degas.umiacs.umd.edu/Hid/data.html*.

2.4 CMU Mobo Database

Unlike the HumanID database and UMD database, the CMU Mobo database is indoor data, which makes the recognition task relatively easier. It has also been used by several reported algorithms. Here we briefly describe it.

The CMU Mobo dataset [25] consists of sequences from 25 subjects walking on a treadmill, positioned in the middle of a room. Fig. 2.4 shows some sample frames. Each subject is recorded performing four different types of walking: Slow walk (2.06 miles/hr), fast walk

(a) (b) (c)

Figure 2.4. Samples of CMU Gait Database Walking on a Treadmill in the Middle of a Room under the Condition of (a) Slow Walk, (b) Fast Walk, and (c) Slow Walk Holding a Ball.

(2.82 miles/hr), slow walk holding a ball, and walk on an inclined plane. Each sequence is 11 seconds long, recorded at 30 frames per second. Six cameras were set up to take images from side view, diagonal view, frontal view, and back view, as Fig. 2.5 shows.

There are four studies defined for this database [26]:

1. How well does the gait recognition algorithm perform within each gait?
2. How well does the gait recognition algorithm perform within each view?
3. How well does the gait recognition algorithm generalize across different types of gaits for the same view?
4. How well does the gait recognition algorithm generalize across different types of gaits using two views?

For each study several experiments are specified. Table 2.4 lists the probe and gallery specifications for each experiment, which is named with the study number as the prefix.

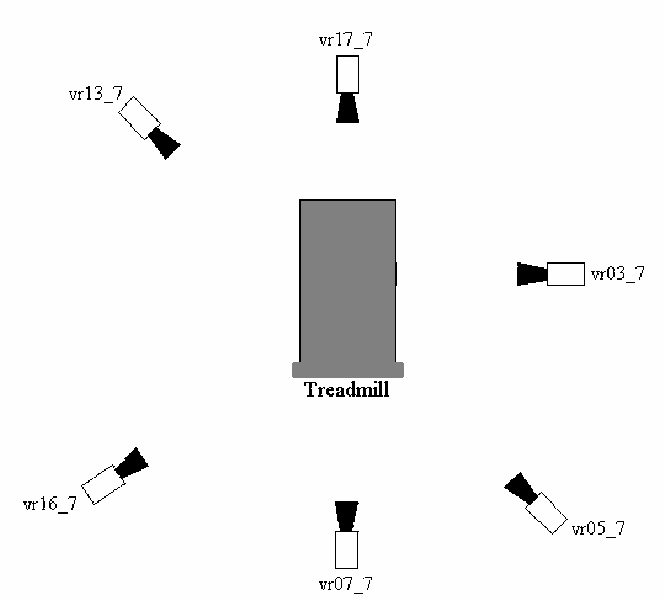


Figure 2.5. Camera Setup for the CMU Mobo Data Acquisition.

(a) (b) (c)

Figure 2.6. Samples of U. of Southampton Gait Database with Large Population (100) (a) Normal Track, (b) Normal Treadmill, and (c) Normal Outdoor.

Table 2.4. The Probe Set and Gallery Set for Each Experiment Defined for CMU Mobo Database with 25 Subjects. The Experiments are Named With Study Number as Prefix.

Exp.	Difference	Probe Walk Condition	Gallery Walk Condition	Probe View	Gallery View
1.1	Session[1]	Slow	Slow	Side	Side
1.2	Session[1]	Fast	Fast	Side	Side
1.3	Session[1]	Ball	Ball	Side	Side
2.1	Session[1]	Slow	Slow	Side	Side
2.2	Session[1]	Slow	Slow	Angle	Angle
2.3	Session[1]	Slow	Slow	Frontal	Frontal
3.1	Gait	Slow	Fast	Side	Side
3.2	Gait	Slow	Fast	Angle	Angle
3.3	Gait	Slow	Fast	Frontal	Frontal
3.4	Gait	Slow	Ball	Side	Side
3.5	Gait	Slow	Ball	Angle	Angle
3.6	Gait	Slow	Ball	Frontal	Frontal
4.1	Gait+View	Slow	Fast	Side+frontal	Side+frontal
4.2	Gait+View	Slow	Ball	Side+frontal	Side+frontal
4.3	Gait+View	Slow	Fast	Side+frontal	Angle
4.4	Gait+View	Slow	Ball	Side+frontal	Angle

[1] The gallery and probe are different frames in a same sequence.

2.5 University of Southampton (SOTON) Database

The Southampton database consists of two major segments: a large population and a small population database. Here we only briefly introduce the first one. For detailed description, please refer to *http://www.gait.ecs.soton.ac.uk/database/*.

The large database consists of around 100 subjects. It was collected over successive days in the summer of 2001. As Fig. 2.6 shows, subjects were collected under 3 scenarios: normal outdoor, indoor track and indoor treadmill. Two cameras were set up for each scenario with different views: fronto-parallel and oblique, so that there are totally 6 views for each subject.

The large database is intended to address two questions: (i) whether gait can be used as a biometrics with a significant number of people in normal conditions, and (ii) how much research effort needs to be directed towards biometric algorithms or towards subject segmentation algorithms in computer vision field.

CHAPTER 3

STUDY OF GAIT RECOGNITION FEASIBILITY: PARAMETERLESS BASELINE ALGORITHM (VERSION 2)

The feasibility of recognition using gait biometrics is the fundamental problem for our study. Can people be recognized by a computer from the way they are walking? And to what extent does gait offer potential as an identifying biometric? To answer these questions, we need to perform gait recognition over different condition variations, and identify the "difficulty" of each. Toward this end, in this chapter we first propose a parameterless baseline algorithm (v2). It uses a simple technique of background subtraction to segment a person from the image, and temporal-spatio correlation between two sequences to compute similarities. So it emphasizes gait dynamics as well as gait shapes. Using this baseline algorithm, we run the USF/NIST HumanID gait challenge experiments that include 5 covariates: viewpoint, shoe, surface, briefcase and time, and the CMU Mobo experiment that tests the effect of walking speed. To reduce the impact of the gallery, we evaluate the experiments under gallery variation. We also measure the effect of covariates by performing a statistical test on similarity score variations across condition changes. Our study also includes failure pattern in subjects.

3.1 Parameterless Baseline Algorithm (v2)

The baseline algorithm v2 [75] is an improvement over the first release, v1 [43, 42]. The first release requires three parameters: foreground/background threshold, frame number in one gait cycle and minimum size of foreground components. The new version presented in this dissertation does not need the specification of any parameter – it is parameter free. Fig. 3.1 demonstrates the flowchart showing the old and new version of the algorithm.

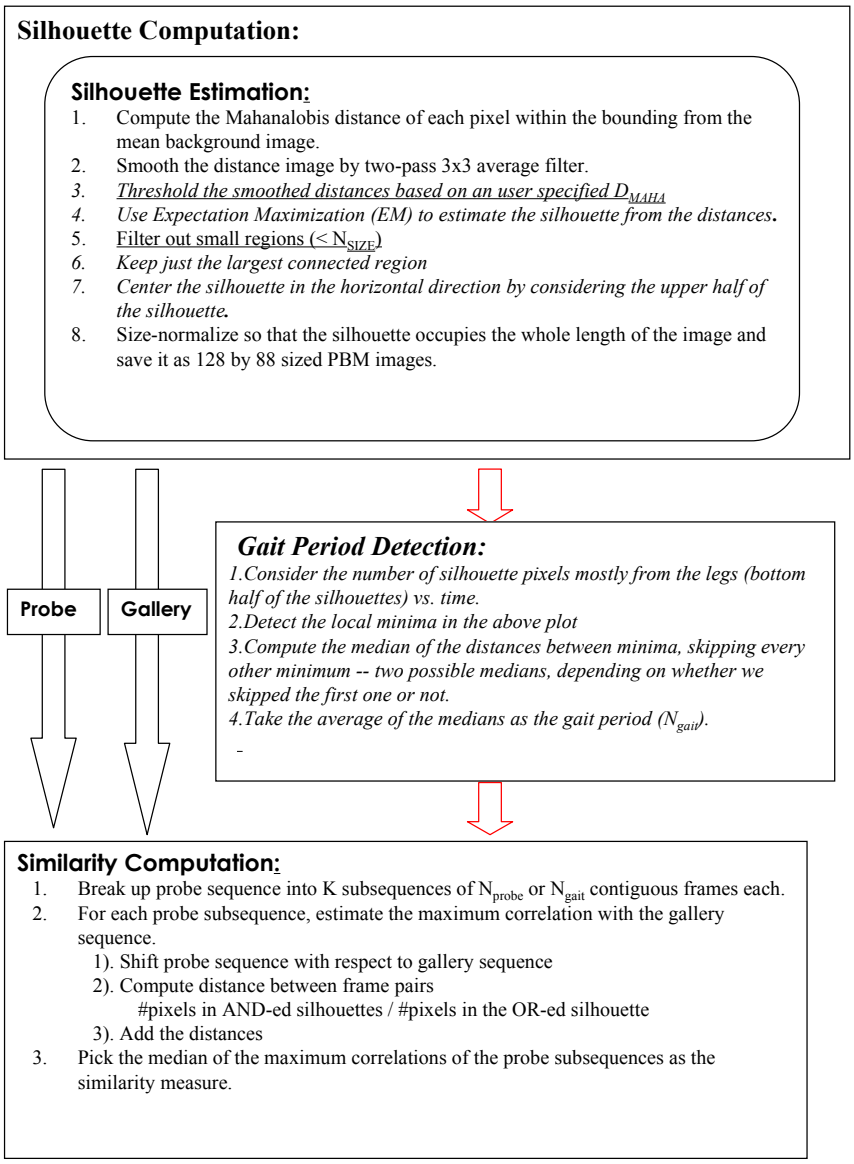


Figure 3.1. The Flowchart of the Baseline Algorithm of Both Versions (v1 and v2). The Parts Unique to the Parameterized Baseline Algorithm (v1) are Underlined, Parts Unique to the Parameter-free Baseline Algorithm (v2) are in *italic*, and the Remaining are Parts Common to Both.

(a) (b) (c) (d)

Figure 3.2. Sample Bounding Boxed Image Data as Viewed from (a) Left Camera on Concrete, (b) Right Camera on Concrete, (c) Left Camera on Grass, and (d) Right Camera on Grass.

Baseline v2 is a four-part algorithm that relies on silhouette template matching. The first part semi-automatically defines bounding boxes around the moving person in each frame of a sequence. The second part extracts silhouettes from the bounding boxes. The third part computes the gait period from the silhouettes. The gait period is used to partition the sequences for spatial-temporal correlation. The fourth part performs spatial-temporal correlation to compute the similarity between two gait sequences.

Locating the bounding boxes in each frame is a semi-automatic procedure. In the manual step, the bounding box is outlined in the starting, middle, and ending frames of a sequence. The bounding boxes for the intermediate frames are linearly interpolated from these manual ones, using the upper-left and the bottom-right corners of the boxes. This approximation strategy works well for cases where there is nearly fronto-parallel, constant velocity motion, which is true for the experiments reported here. Fig. 3.2 shows some examples of the image data inside the bounding box. The bounding boxes are conservatively specified, and results in background pixels around the person in each box. These bounding boxes are part of the information distributed with the dataset.

3.1.1 Silhouette Extraction

The second step in the baseline algorithm is to extract the silhouette in the bounding boxes. Following common practice in gait recognition work, we define the silhouette to be the *region* of pixels from a person. Prior to extracting the silhouette, a background model of the scene is built. In the first pass through a sequence, we compute the background statistics of the RGB values at each image location, (x, y), using pixel values *outside* the manually defined bounding boxes in each frame. We compute the mean $\mu_B(x, y)$ and the covariances $\Sigma_B(x, y)$ of the RGB values at each pixel location. For pixels within the bounding box of each frame, we compute the Mahalanobis distance in RGB-space for the pixel value from the estimated mean background value. Based on the Mahalanobis distance, pixels are classified into foreground or background. In our earlier version of the baseline algorithm [43], this decision used a fixed, user defined threshold. The present version adaptively decides on the foreground and background labels for each frame by estimating the foreground and background likelihood distributions using the iterative expectation maximization (EM) procedure. At each pixel, indexed by k, we have a two-class problem based on a scalar observation – the Mahalanobis distance, D_k. We model the observations as a two-class, $\{\text{Foreground} = \omega_1, \text{Background} = \omega_2\}$, Gaussian Mixture Model (GMM), $P(D_k) = \sum_{i=1}^{2} P(\omega_i) p(D_k|\omega_i, \mu_i, \sigma_i)$, where the class likelihood $p(D_k|\omega_i, \mu_i, \sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(D_k-\mu_i)^2}{2\sigma_i^2}}$. For each pixel, we would like to estimate the posterior $P(\omega_1|D_k)$. We iteratively estimate this using the standard EM update equations reproduced below [20]. The estimates from different iterations are distinguished using the superscript.

$$
\begin{aligned}
P^{(n+1)}(\omega_i) &= \tfrac{1}{N} \textstyle\sum_{k=1}^{N} P^{(n)}(\omega_i|D_k) \\
\mu_i^{(n+1)} &= \left(\textstyle\sum_{k=1}^{N} P^{(n)}(\omega_i|D_k) D_k\right) / \left(\textstyle\sum_{k=1}^{N} P^{(n)}(\omega_i|D_k)\right) \\
\sigma_i^{(n+1)} &= \left(\textstyle\sum_{k=1}^{N} P^{(n)}(\omega_i|D_k)(D_k - \mu_i)^2\right) / \left(\textstyle\sum_{k=1}^{N} P^{(n)}(\omega_i|D_k)\right) \\
P^{(n+1)}(\omega_i|D_k) &= \left(p(D_k|\omega_i, \mu_i, \sigma_i) P(\omega_i)\right) / \left(\textstyle\sum_{i=1}^{2} p(D_k|\omega_i, \mu_i, \sigma_i) P(\omega_i)\right)
\end{aligned}
\tag{3.1}
$$

The EM process is initialized by choosing class posterior labels based on the observed distance; the larger the Mahalanobis distance of a pixel, the greater is the initial posterior probability of being from the foreground.

$$P^{(0)}(\omega_1|D_k) = \min(1.0, D_k/255) \quad \text{and} \quad P^{(0)}(\omega_2|D_k) = 1 - P^{(0)}(\omega_1|D_k) \tag{3.2}$$

We found that with this initialization strategy the process stabilizes fairly quickly within 15 or so iterations.

It is worth mentioning a few words about pre- and post-processing steps that impact overall performance. We have found that if we smooth the computed Mahalanobis distance array (image) using a 9 by 9 pyramidal-shaped averaging filter, or equivalently, two passes of a 3 by 3 averaging filter, the quality of the silhouette and recognition performance improves. This smoothing compensates for DV compression artifacts. The convergence of the EM process is faster with these smoothed distances than without, possibly due to reduction in the noise of the computed Mahalanobis distances. There are two post-processing steps on the silhouette image computed by EM. First, we eliminate isolated, small, noisy regions by keeping only the foreground region with the largest area. Second, we scale this foreground region so that its height is 128 pixels and occupies the whole height of the 128 by 88 pixels sized output silhouette frame. The scaling of the silhouette offers some amount of scale invariance and facilitates the fast computation of a similarity measure. We also center the silhouette along the horizontal direction to compensate for errors in the placement of the bounding boxes. The silhouette is shifted in the horizontal direction so that the center column of the top portion of the silhouette is at column 44.

In most cases, the above strategy results in good quality silhouettes, but there are cases when it has problems. Fig. 3.3 shows some of these cases. Segmentation errors occur due to: (i) shadows, especially in the concrete sequences, (ii) inability to segment parts because they fall just below the threshold and are classified as background, (iii) moving objects in the background, such as the fluttering tape in the concrete sequences or moving leaves in

Figure 3.3. The Bottom Row shows Sample Silhouette Frames with a Variety of Segmentation Errors. The Raw Image Corresponding to each Silhouette is Shown on the Top Row.

the grass sequences, or other moving persons in the background, and (iv) lingering DV compression artifacts near the boundaries of the person.

There are many other possible scaling and centering options that might reduce the problems that we see in the current silhouettes. One option could be to take into account the entire sequence to decide upon the scaling parameters. However, such strategies would be dependent on the actual path taken by the subject. For instance, in our dataset, as the person moves along the elliptical path, the distance of the person from the camera changes, which changes the projected image size. The strategy we use does not use, assume, or estimate the shape of the path taken by the subject. Of course, the chosen frame by frame method might and does result in erroneous scaling when some part, such as the head, is not detected, but the employed matching strategy, which we shall see later, is resistant to some extent to such errors.

3.1.2 Gait Period Detection

The next step in the baseline algorithm is gait period detection. Gait periodicity, N_{gait}, is estimated by a simple strategy. We count the number of foreground pixels in the silhouette in each frame over time, $N_f(t)$. This number will reach a maximum when the two legs are farthest apart (full stride stance) and drop to a minimum when the legs overlap (heels together stance). To increase the sensitivity, we consider the number of foreground pixels mostly from the legs, which are selected simply by considering only the bottom half of the silhouette. Fig. 3.4 shows an instance of the variation of $N_f(t)$. Notice that two consecutive strides constitute a gait cycle. We compute the median of the distances between minima, skipping every other minimum. Using this strategy, we get two estimates of the gait cycle, depending on whether we skipped the first minimum or not. We estimate the gait period by the average of these two medians. Note that this strategy works for near fronto-parallel views, which is the view of choice for gait recognition, and would not work for frontal views. However, the failure with respect to viewpoint variation is not drastic. The views in the present dataset, on which we show the results, are not strictly fronto-parallel; it includes up to 30 degrees variation.

3.1.3 Similarity Computation

The output from the gait recognition algorithm is a complete set of similarity scores between all gallery and probe gait sequences. Similarity scores are computed by spatial-temporal correlation. Let a probe sequence of M frames be denoted by $\mathbf{I_P} = \{\mathbf{I_P}(1), \cdots, \mathbf{I_P}(M)\}$ and a gallery sequence of N frames be denoted by $\mathbf{I_G} = \{\mathbf{I_G}(1), \cdots, \mathbf{I_G}(N)\}$. The final similarity score is constructed out of matches of disjoint portions of the probe with the gallery sequence. Specifically, we partition the probe sequence into disjoint subsequences of N_{gait} contiguous frames, where N_{gait} is the estimated period of the probe sequence from the previous step. Note, we do not constrain the starting frame of each partition to be from a particular stance. Let the k-th probe subsequence be denoted by $\mathbf{I_{Pk}} = \{\mathbf{I_P}(kN_{gait}), \cdots, \mathbf{I_P}((k+1)N_{gait})\}$. The gallery gait sequence $\mathbf{I_G} = \{\mathbf{I_G}(1), \cdots, \mathbf{I_G}(N)\}$

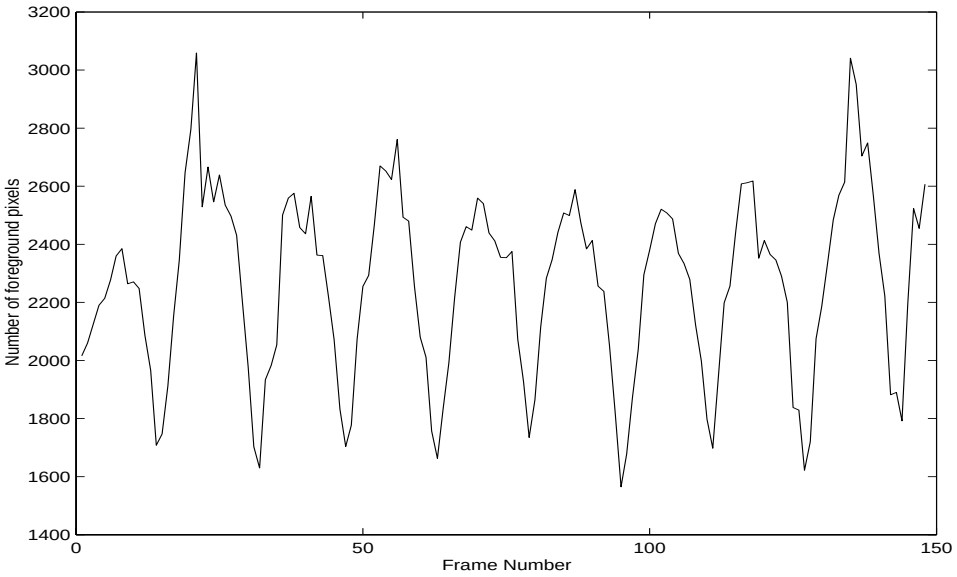


Figure 3.4. Cue for Gait Period – the Number of Foreground Pixels from the Bottom Half of the Silhouettes.

consists of all silhouettes extracted in the gallery sequence from the back portion of the elliptical path. Note, this gallery sequence is not partitioned. We then correlate each of the subsequences $\mathbf{I_{Pk}}$ with the entire gallery sequence $\mathbf{I_G}$.

There are three ingredients to the correlation computations: frame correlation, correlation between $\mathbf{I_{Pk}}$ and $\mathbf{I_G}$, and similarity between a probe sequence and a gallery sequence, comparing $\mathbf{I_P}$ and $\mathbf{I_G}$.

At the core of the above computation is, of course, the need to compute the similarity between two silhouette frames, $\text{FrameSim}\,(\mathbf{I_P}(i), \mathbf{I_G}(j))$, which we simply compute to be the ratio of the number of pixels in their intersection to their union. This measure is also called the Tanimoto similarity measure, defined between two binary feature vectors [20]. Thus, if we denote the number of foreground pixels in silhouette $\mathbf{I}$ by $\text{Num}(\mathbf{I})$ then we have,

$$\text{FrameSim}(\mathbf{I_P}(i), \mathbf{I_G}(j)) = \frac{\text{Num}(\mathbf{I_P}(i) \cap \mathbf{I_G}(j))}{\text{Num}(\mathbf{I_P}(i) \cup \mathbf{I_G}(j))} \tag{3.3}$$

Note that since the silhouettes have been pre-scaled and centered, we do not have to consider all possible translations and scales when computing the frame to frame similarity. The next step is to use frame similarities to compute the correlation between $\mathbf{I_{Pk}}$ and $\mathbf{I_G}$:

$$\mathrm{Corr}(\mathbf{I_{Pk}}, \mathbf{I_G})(l) = \sum_{j=0}^{N_{gait}-1} \mathrm{FrameSim}\left(\mathbf{I_P}(k+j), \mathbf{I_G}(l+j)\right) \tag{3.4}$$

For robustness, the similarity measure is chosen to be the median value of the maximum correlation of the gallery sequence with each of these probe subsequences. Other choices such as the average, minimum, or maximum did not result in better performance. The strategy for breaking up the probe sequence into subsequences allows us to address the case when we have segmentation errors in some contiguous sets of frames due to some background subtraction artifact or due to localized motion in the background.

$$\mathrm{S}(\mathbf{I_P}, \mathbf{I_G}) = \mathrm{Median}_k \left(\max_l \mathrm{Corr}(\mathbf{I_{Pk}}, \mathbf{I_G})(l) \right) \tag{3.5}$$

3.2 Performance of Parameterless Baseline Algorithm (v2)

The performance of the baseline algorithm on the challenge experiments establishes a "minimum" performance expected from any vision based gait recognition algorithm. We show that our baseline algorithm is a reasonable choice by reporting its performance on the CMU Mobo dataset described in Section 2.4. The heart of this section is the baseline performance on all twelve experiments of USF HumanID database, which is comprised of most subjects and condition variations, thus giving a better understanding toward the effectiveness of gait recognition. From the results on the twelve experiments, we are able to rank the difficulty of the experiments. We identify the error modes of the baseline algorithm so that better algorithms can be designed by concentrating on these subjects and investigating the causes of failure.

3.2.1 Base Results

The performance results for the twelve challenge experiments are reported as follows. We match each probe sequence to the gallery sequences, thus obtaining a similarity matrix with size that is the number of probe sequences by the gallery size. Following the pattern of the FERET evaluations [41], we measure performance for both identification and verification scenarios using cumulative match characteristics (CMCs) and receiver operating characteristics (ROCs), respectively. In the identification scenario, the task is to identify a given probe to be one of the given gallery images. To quantify performance, we sort the gallery images based on computed similarities with the given probe. In terms of the similarity matrix, this corresponds to sorting the rows of the similarity matrix. If the correct gallery image corresponding to the given probe occurs within rank k in this sorted set, then we have a successful identification at rank k. A cumulative match characteristic plots these identification rates (P_I) against the rank k.

In the verification scenario, a system either rejects or accepts if a person is who they claim to be. Operationally, a person presents a new signature, the probe, and an identity claim. The system then compares the probe with the stored gallery sequence that corresponds to the claimed identity. The claim is accepted if the match between the probe and gallery is above an operating threshold, otherwise it is rejected. This decision made solely on the similarity between a probe signature and the gallery signature that corresponds to the claimed identity, which is the usual practice, is optimal only if the underlying distributions are not dependent on the probe. However, recent experiments with face recognition methods (FRVT 2002 [40]), showed similarity score normalization can dramatically increase performance, possibly because it removes the dependencies of the non-match scores on the probe. This issue, however, needs a deeper theoretical look in future. Following FRVT 2002, instead of the raw similarity scores, we also report verification performance on gallery normalized similarity scores.

In *normalization* a similarity score, $\mathrm{S}(I_{P_i}, I_{G_j})$ between probe, I_{P_i}, and gallery signature, I_{G_j}, is adjusted by the statistics of the similarity scores between a probe and the full

gallery set, $\{I_{G_1}, \cdots, I_{G_N}\}$. We present results for two normalization functions. The first is z-norm [40], which is

$$\mathrm{S}_z(I_{P_i}, I_{G_j}) = \frac{\mathrm{S}(I_{P_i}, I_{G_j}) - \mathrm{Mean}_j \mathrm{S}(I_{P_i}, I_{G_j})}{\mathrm{s.d.}_j \mathrm{S}(I_{P_i}, I_{G_j})}, \tag{3.6}$$

where s.d. is standard deviation. For each probe, the normalized scores, most of which are non-match scores except for the one correct match one, will have zero mean and unit standard deviations. The second is MAD-norm, which is

$$\mathrm{S}_{\mathrm{MAD}}(I_{P_i}, I_{G_j}) = \frac{\mathrm{S}(I_{P_i}, I_{G_j}) - \mathrm{Median}_j \mathrm{S}(I_{P_i}, I_{G_j})}{\mathrm{Median}_j |\mathrm{S}(I_{P_i}, I_{G_j}) - \mathrm{Median}_j \mathrm{S}(I_{P_i}, I_{G_j})|}, \tag{3.7}$$

where the denominator is the median of the absolute deviations (MAD) around the median values. The MAD-norm is a robust version of z-norm. For each probe, the MAD normalized scores, will have zero first order and unit second order robust statistics. Given these normalized similarity scores, for a given operating threshold, there is a verification rate (or detection rate) and a false accept rate. Changing the operating threshold can change the verification and false accept rates. The complete set of verification and false accept rates is plotted on a receiver operating characteristic (ROC).

Table 3.1 summarizes the key performance indicators: the identification rate (P_I) at ranks 1 and 5, and the verification rate (P_V) for a false alarm rate of 1% and 10%. Verification rates are reported for un-normalized, z-normed, and MAD-normed similarity scores. Identification ranges from 3% to 78% at rank 1, and improves to a range from 12% to 93% at rank 5. The most striking feature of the verification results is the significant impact that normalization has on performance. At a false accept rate of 1%, the z-norm is superior to the MAD-norm, and at a false accept rate of 10%, both types of normalization are roughly equivalent. Because of the superiority of the z-norm at a false accept rate of 1%, all remaining verification results use the z-normalization procedure. With the z-norm, verification rates at a false accept rate of 1% range from 6% to 82%; at a false accept rate of 10%, verification rate ranges from 24% to 94%. These are very encouraging performances

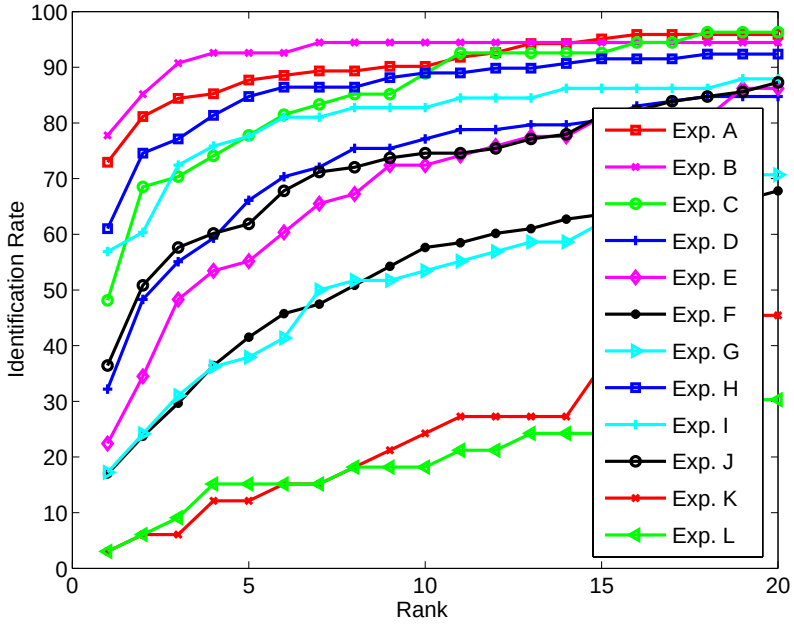


Figure 3.5. Baseline Performances for the 12 Experiments of USF HumanID Database in terms of the CMC Curves.

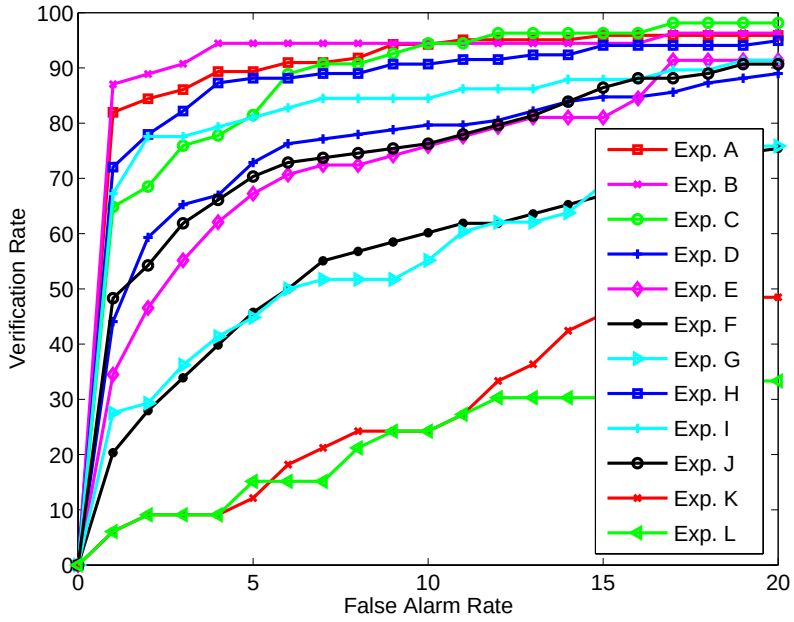


Figure 3.6. Baseline Performances for the 12 Experiments of USF HumanID Database in terms of the ROCs Plotted Upto a False Alarm Rate of 20%.

Table 3.1. Baseline Performances for the Experiments of USF HumanID Database: the Identification Rate P_I at Ranks 1 and 5, and the Verification Rate P_V at a False Alarm Rate of 1% and 10% of Unnormalized (UN), Z-Norm (ZN), and MAD-Norm (MAD). All Performance Scores are in Percent.

Exp.	Difference	P_I (%) at rank		P_V (%) at P_F=1%			P_V (%) P_F=10%		
		1	5	UN	MAD	ZN	UN	MAD	ZN
A	View	73	88	52	80	82	81	94	94
B	Shoe	78	93	48	80	87	82	94	94
C	Shoe, View	48	78	32	57	65	69	89	94
D	Surface	32	66	24	36	44	61	80	80
E	Surface, Shoe	22	55	16	33	35	52	76	76
F	Surface, View	17	42	10	22	20	45	59	60
G	Surface, Shoe, View	17	38	12	24	28	40	57	55
H	Briefcase	61	85	46	68	72	80	90	91
I	Briefcase, Shoe	57	78	48	60	67	76	85	85
J	Briefcase, View	36	62	22	45	48	64	75	76
K	Time, Shoe, Clothes	3	12	0	3	6	15	27	24
L	Surface, Time, Shoe, Clothes	3	15	0	3	6	18	27	24

given the straightforward nature of the baseline algorithm. The range of results for the twelve experiments allows for improvement by new algorithms. Fig. 3.5 and 3.6 plot the CMCs and ROCs of the twelve challenge experiments.

Table 3.2 lists the identification rates that have been reported by other algorithms on an earlier, smaller, release of the USF HumanID dataset. For comparison, we also list the reported performance of the baseline algorithm on the reduced dataset. We see that (i) the ranked order of performance on the different experiments follows that for the baseline algorithm, and (ii) the performance of the baseline algorithm is very competitive with respect to the other algorithms, especially on the hard problems.

We can rank the difficulty of the twelve experiments by their identification and verification rates, as reported by the baseline algorithm and corroborated by other algorithms. For instance, Experiment A, where the difference between probe and gallery is just the viewpoint change, is easier than Experiment G, where the difference between the gallery and probe is three covariates. The rank of experiments allows for a ranking of the difficulty

Table 3.2. Reported Top Rank Recognition for Earlier, Smaller, Release of the Gait Challenge Dataset. The Numbers for the First Two Columns are as Read from Graphs in the Cited Papers.

Exp.	Width Vectors (UMD) [16]	DTW (UMD) [44]	HMM (UMD) [84]	Body Shape (CMU) [89]	HMM (MIT) [52]	Body (CAS) [94]	Baseline [75]
A	52%	78%	99%	87%	88%	70%	87%
B	40%	65%	89%	81%	75%	59%	80%
C	20%	28%	78%	66%	70%	51%	53%
D	18%	10%	36%	21%	25%	34%	39%
E	20%	10%	29%	19%	15%	21%	33%
F	15%	10%	24%	27%	20%	27%	28%
G	15%	10%	18%	23%	10%	14%	26%
# subjects in gallery	71	71	71	71	71	71	71

of the five covariates. From early reported results, this ranking also appears to be somewhat independent of the choice of the gait recognition algorithm, as we see in Table 3.2. The baseline algorithm based rankings suggest that shoe type has the least impact, next is about 30° viewpoint change, the third is briefcase, then surface type, and time has the most impact, based on the drop in the identification rate due to each of these covariates. We quantify these effects next.

3.2.2 Impact of Variation in Gallery

The results presented so far are for one gallery set choice. It is well known that changing the gallery and corresponding probe set changes the recognition rate [41, 40]. In this section we examine the effect of changing the gallery and corresponding probe set and examine if the order of experiments, based on the baseline recognition rates, change.

The base challenge experiments presented so far use the set (G,A,R,NB,M or N) as the gallery. To examine the effect of gallery variation, we reran the twelve challenge experiments with different galleries and appropriately modified probe sets. In the challenge experiments, Experiment A examined the effect of change in view. To maintain consistency,

the corresponding probe set A for each gallery is a change in view. For example, if the gallery is (C,A,L,NB,M or N), then the probe set for experiment A should be (C,A,R,NB,M or N), and so on. We vary the gallery to be one of the following 8 cases: (G,A,R), (G,A,L), (G,B,R), (G,B,L), (C,A,R), (C,B,R), (C,A,L), and (C,B,L), with all the remaining two conditions, i.e., Carry and Time, fixed at NB, M or N. Table 3.3 summarizes the verification rates at a false alarm of 1% for the challenge experiments. The first column lists one the eight galleries, and remaining columns report recognition rates for changing different covariates. For example, the column labeled Surface + Shoe reports experimental results when the gallery and probe set have difference surface and shoe types. The remaining covariates are the same between the gallery and probe set. The performance scores establish bounds on the verification rates for each experiment. The mean and the median score for each experiment provide a proxy for the difficulty level for each experiment. The standard deviation (s.d.) provides a measure of the stability of a covariate. The camera angle or view covariate has the greatest variability in terms of performance.

It is interesting to note that the ordering of the experiments in terms of their difficulty level, as measured by the verification rates, is somewhat invariant to the choice of the gallery set. To quantify the statistical correlation among the ranking of the experiments for the different gallery variations, we use the Friedman test, which is a two-way analysis of performance scores of the n gallery variations for the k experiments. The null hypothesis is that the ratings for the gallery variations are not related. For the data in Table 3.3, the computed underlying test parameter, which is the Kendall's coefficient of concordance, is found to be 0.96; the maximum correlation being one. The P-value is found to be < 0.0001, which implies that the null hypothesis can be easily rejected. Rejection of the null hypothesis implies that the verification rates for the experiment are different *and* the rates for the different gallery variations are strongly correlated.

The Friedman test does not provide us with a statistical ranking between the experiments, it just tell us if there is one. To rank the experiments, particularly the ones where only one covariate is varied, we use a pairwise Wilcoxon signed rank test [97]. It com-

Table 3.3. Verification Performance Variation at $P_F = 1\%$ of Baseline Algorithm due to Variations in Gallery Type over 8 Possible Combinations; the Fixed Condition over the Being No-briefcase and the Non-repeat, i.e. NB, M or N.

Gallery	Experiments											
	View	Shoe	View + Shoe	Surface	Surface + Shoe	Surface + View	Surface + View + Shoe	Briefcase	Briefcase + Shoe	Briefcase + View	Time	Time + Surface
	A	B	C	D	E	F	G	H	I	J	K	L
(G,A,R)	82	87	65	44	35	20	28	72	67	48	6	6
(G,A,L)	76	82	59	44	35	25	10	75	62	40	3	6
(C,A,R)	54	86	44	32	16	20	14	75	57	24	6	3
(C,A,L)	63	88	49	37	28	17	20	72	59	29	3	6
(G,B,R)	91	82	61	34	24	18	12	69	56	48		
(G,B,L)	89	87	54	34	31	20	20	69	60	46		
(C,B,R)	68	92	41	28	28	26	22	78	67	29		
(C,B,L)	73	83	53	34	35	16	17	67	56	35		
Mean	75	85	53	36	29	20	18	72	61	37	5	5
Median	75	87	54	34	30	20	19	72	60	38	5	6
s.d.	12.8	3.4	8.3	5.6	6.6	3.6	5.9	3.7	4.5	9.5	1.6	1.4

putes the statistical significance of the null hypothesis that medians of two distributions are equal. Based on this test, along with modified Bonferroni corrections [35] to account for multiple comparisons, for an overall $\alpha = 0.05$ (95% significance), we arrive at the following difficulty ranking: (ExpB–Shoe, ExpA–View) $\geq$ (ExpA–View, ExpH–Briefcase) $>$ ExpD–Surface $>$ ExpK–Time.

3.2.3 Performance of the Baseline Algorithm on Mobo Dataset

So far, we have analyzed gait recognition based on the performance of the baseline algorithm. The simplicity of this algorithm might raise skepticism about its performance, hence call into question the conclusions based those performance numbers. Earlier, we had shown by listing the reported performance of other algorithms on an earlier, smaller, release of the gait challenge dataset that the performance of the baseline algorithm is at par with more sophisticated methods. Here we benchmark the performance of the baseline

Table 3.4. Top Rank Identification Rates for CMU Mobo Dataset Reported by Different Algorithms.

Gallery Probe	Slow Walk Slow Walk	Fast Walk Fast Walk	Ball Carry Ball Carry	Slow Walk Fast Walk
CMU(Body Shape) [15]	100%[1]	100%[1]	100%[1]	76%
UMD(HMM) [45, 46]	72%	71%	96%	31%
Georgia Tech.(Body Parameters)				50%[2]
MIT(Moment Based Features)[53]	100%	96%	96%	54%[3]
Baseline(Spa.-Temp. Correlation)	92%	96%	96%	72%

[1] As reported in http://www.hid.ri.cmu.edu/HidEval/evaluation.html
[2] As reported in http://www.cc.gatech.edu/cpl/projects/hid/CMUexpt.html
[3] As reported in http://www.ai.mit.edu/people/llee/HID/cmu_data_feat_sel.htm

algorithm on a different dataset, on which performance has also been reported by different algorithms.

The dataset chosen here is the CMU Mobo dataset, on which several papers have published results, hence it is a good external dataset to benchmark the performance of the baseline algorithm. As described in Section 2.4, this database consists of sequences from about 25 subjects walking on a treadmill positioned in the middle of the room, viewed from 6 different view points. Here we use the experiments defined on the side-view. Table 3.4 lists the reported identification rates for different algorithms on some of the most commonly reported experiments. The last row lists the performance of the baseline algorithm. We used the silhouettes that were provided with the dataset. We see that the performance of the baseline is not the lowest one and is near perfect for experiments comparing sequences under the same condition. For the experiment comparing sequences with walking speeds, the performance of the baseline algorithm is the second highest reported performance.

3.2.4 Covariate Effects

Which covariate has the most impact on recognition? From the baseline recognition results, it appears that time has the most impact as the recognition rates for Experiments K and L are the lowest. However, using recognition rates as indicators of covariate impact has problems and is at best a gross measure of impact. The recognition rate is a function

of both the match and the non-match score distributions. This rate can change due to change in either the match scores or non-match scores, or both. This is problematic since the non-match scores are a function of identity differences *and* any covariate difference that is present between the gallery and probes. The effect of a covariate is more cleanly captured by its impact on just the match scores.

We quantify the effect of a covariate on recognition by comparing the match scores for two probe sets, over the same set of individuals, that differ with respect to a specific covariate, but are similar in all other aspects. Therefore, for instance, if we want to study the effect of viewpoint on performance, then we could consider the probes in Experiments B and C, which differ with respect to just viewpoint. For shoe type we use the probes for Experiments A and C; for surface we use the probes for Experiments B and E; for briefcase we use the probes in Experiments B and I; and for time we use the probe in Experiment A and the probe specified by (G, A/B, L, NB, N).

Let a similarity score for the i-th subject in two choices of the probe sets, Probe 1 and Probe 2, be $\mathrm{S}_1(\mathbf{I}_{\mathbf{P}_i}, \mathbf{I}_{\mathbf{G}_i})$ and $\mathrm{S}_2(\mathbf{I}_{\mathbf{P}_i}, \mathbf{I}_{\mathbf{G}_i})$, respectively. The change in similarity for subject i, given by

$$\Delta \mathrm{S}_{12}(i) = \frac{\mathrm{S}_1(\mathbf{I}_{\mathbf{P}_i}, \mathbf{I}_{\mathbf{G}_i}) - \mathrm{S}_2(\mathbf{I}_{\mathbf{P}_i}, \mathbf{I}_{\mathbf{G}_i})}{\mathrm{S}_2(\mathbf{I}_{\mathbf{P}_i}, \mathbf{I}_{\mathbf{G}_i})},$$

quantifies the effect of a covariate on subject i. The distribution of these $\Delta \mathrm{S}_{12}(i)$ for all the subjects that are common between the probes and the gallery would provide an idea of the net effect of the covariate. If the distribution is centered around zero, this would signify no impact. If the drop is large then we can infer that the distribution of the match scores, upon changing that covariate, would overlap more with the non-match scores, with consequent drop in recognition performance.

Fig. 3.7 shows the distribution of the score changes between probes differing with respect to view point, shoe type, surface type, briefcase, and time. Notice how the distribution shifts as we go from shoe type to viewpoint to briefcase to time to surface type differences. The median percentage increase in similarity scores for shoe, viewpoint, briefcase, time, and surface are 0.84, 1.56, 2.73, 4.25, and 6.55, respectively. The Wilcoxon signed rank

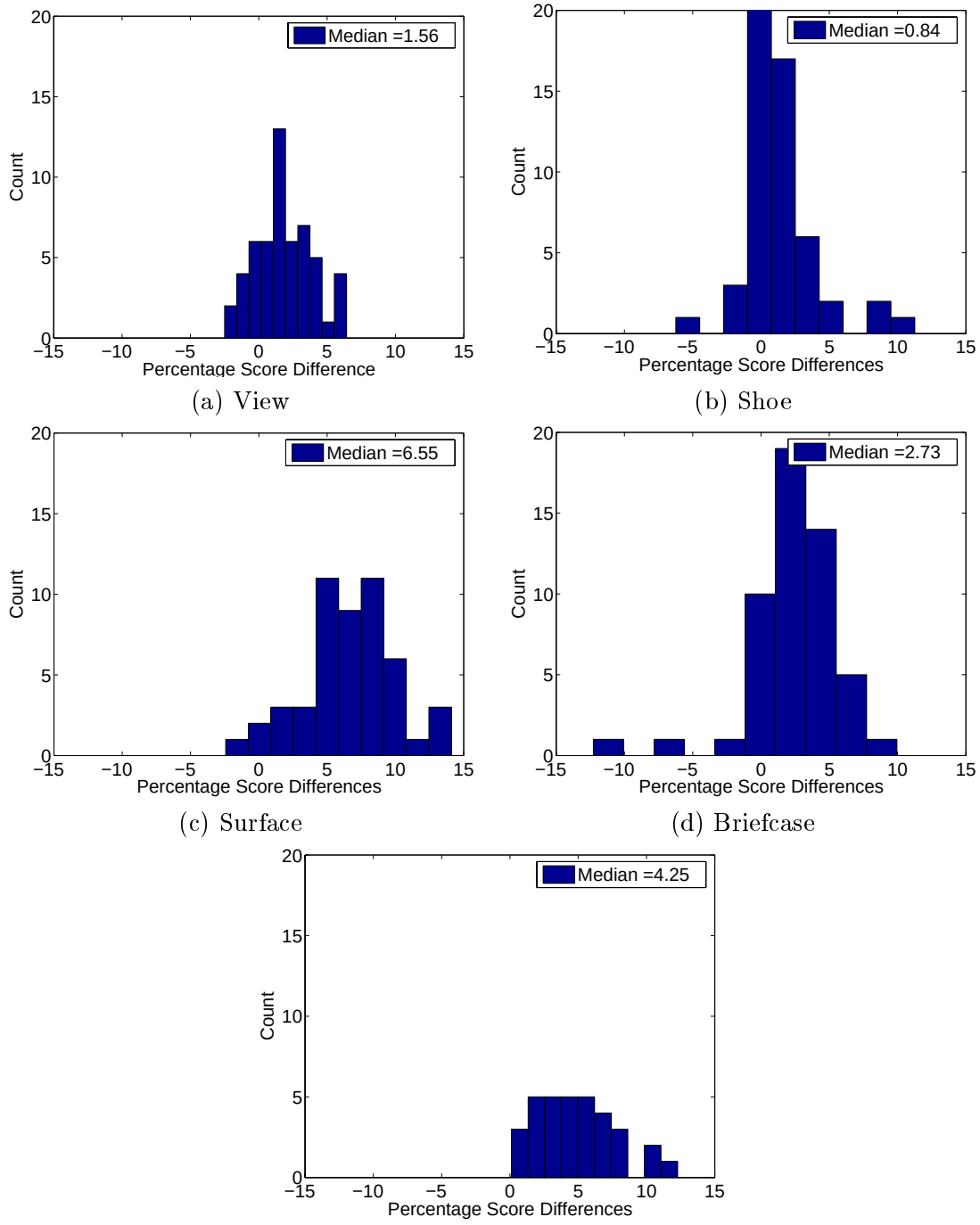


Figure 3.7. The Distribution of the Percentage Change in Similarity Values, $\Delta \mathrm{Sim}_{12}(i)$, Between two Probes Differing with Respect to (a) View Point, (b) Shoe Type, (c) Surface Type, (d) Briefcase, and (e) Time.

Table 3.5. Modified Bonferroni Test for 10 Pairwise Tests of the Impact of the Covariates to achieve an Overall Significance of 0.05.

Factor Pairs	Surf. Brief	Surf. Shoe	Surf. View	Shoe Brief	View Time	View Brief	Shoe Time	Brief Time	View Shoe	Surf. Time
Wilcoxon P-value	0	0	0	0.0038	0.0068	0.0582	0.0674	0.0674	0.0992	0.2783
Modified-α	0.0055	0.0055	0.0062	0.0071	0.0083	0.0100	0.0125	0.0166	0.0250	0.0500
Reject Null Hypo	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No

test [97] can be used to compute statistical significance of the null hypothesis that the population median of the score changes is 0. It is a nonparametric test that takes into account the magnitude as well as the rank and is more sensitive than the Sign-Test or the Student t-test, especially for small numbers. Using this test, we find that we can easily reject the null hypothesis that the population median of the score changes for each covariate is 0 (with P-values < 0.001), i.e., the score changes for all the covariates are significantly different from zero.

We can also compute the statistical significance for the ordering of the covariate impact ranking by performing pairwise Wilcoxon signed rank test. However, we have to be careful to take into account the multiple comparisons; in general the individual pairwise comparisons must be performed at a tighter significance level than the desired overall significance level. We use the modified Bonferroni significance level based testing of the individual pairwise testing [35]. The individual comparisons, of which we had 10, were rank ordered from most to least significant. So as to achieve an overall significance level of 0.05, for the k-th rank we use a cutoff of $\alpha/(10-k+1)$. Table 3.5 lists which of the pairwise null hypotheses we can reject. Based on the results, statistically speaking, the score changes due to shoe, view, briefcase, and time are similar, whereas the scores changes due to time and surface are similar. Thus, (view, shoe, briefcase, time) $\leq$ (time, surface).

The pairwise statistical tests in Table 3.5 clearly suggest that the impacts due to change in surface type and time are different from the impact of the other covariates. They seem to impact gait at a more fundamental level than other covariates. For example, we have found

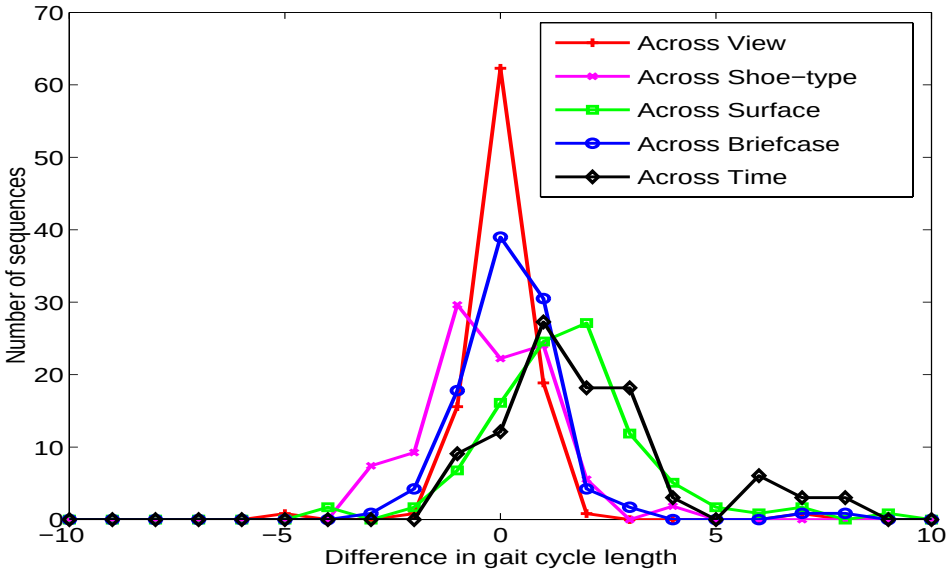


Figure 3.8. Distribution of Period Differences Across Conditions.

that the surface and time covariates impact the gait period more than other covariates. Fig. 3.8 plots the histogram of the differences in gait period for the same subject across views, surface, shoe-type, time, and carrying conditions. If a covariate does not impact the gait period then the histogram should be peaked around zero. However, we notice that for surface-type and time, the histogram spreads to large values, which points to significant differences in gait period. The histogram for the carrying condition (briefcase and no-briefcase) has a peak to the left of that for the surface-type.

3.2.5 Study of Failures

Is there a pattern to the failure in identification? Are there subjects who are difficult to recognize across all conditions? Is there an "easy to recognize" subset of subjects? Answers to these question will help identify the hard sequences to work on in future. To answer such questions, we look at the pattern of failures in identification for each subject across different experiments. To partition the dataset into subsets of subjects who are easy, moderate, and hard to identify, we consider the percentage of the experiments in which a subject was correctly identified. Note, we consider percentages instead of absolute numbers since all

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

Figure 3.9. Samples of Subjects: (a) and (b) are Easy to Identify, (c) and (d) have Moderate Levels of Identification Difficulty, and (e) and (f) are Hard to Identify.

subjects did not participate in all experiments. We consider a subject easy to identify if the subject was identified in more than 80% of the experiments that he/she participated; in our data set there are 12 such subjects. We consider a subject hard to identify if the subject is correctly identified in less than 40% of the experiments; there are 56 subjects in this category. The rest of the subjects are considered moderately difficult to recognize; there are 54 subjects in this category. Fig. 3.9 shows some samples from each class. It is not obvious to us from visually observing the images or the associated silhouettes the reason why some subjects are hard to recognize. There are bad quality silhouettes, e.g. with missing head regions or missing leg regions, in all the classes of subjects. Clothing or shadows also do not seem to play a role. However, to rule out any of these on a firm basis, future in-depth statistical correlation studies will have to be conducted.

3.3 Summary

In this chapter we study the feasibility of gait as a biometrics. We present a parameterless version of baseline algorithm employing the EM classification and spatio-temporal correlation, which is simple but shows strong recognition power. The database used here is the USF/NIST HumanID dataset, which is the largest database consisting a large number of subjects and spanning 32 condition changes of 5 convariates. For a better understanding toward the gait recognition, we also test the CMU Mobo indoor database.

3.3.1 Significant Findings

We investigated two methods for normalizing similarity scores for verification performance. Overall we found that performing normalization significantly increases performance, with the z-norm method being better than the MAD method. For performance on sequences taken on different days, the unnormalized verification rate at a false accept rate of 1% was zero, and 6% after performing z-normalization (experiments K and L). For experiment B, the HumanID dataset, change in shoe type, performance increased from 48% for unnormalized to 87% z-normalized similarity scores.

Focused study of the impact of a covariate on match-score distribution suggests that shoe type has the least effect on performance, but the effect is nevertheless statistically significant. This was followed by either a change in camera view or carrying a brief case. Carrying a brief case does not affect performance as much as one might expect (Section 3.2.4). This effect is marginally larger than changing shoe type but is substantially smaller than a change in surface type. In future experiments, it may be interesting to investigate the effect of carrying a backpack rather than a briefcase, or to vary the object that is carried.

One of the factors that has a large impact is time, resulting in low recognition rates for changes when matching sequences across time. This dependence on time has been reported by others too, but for indoor sequences and for less than 6 months differences. When the difference in time between gallery (the pre-stored template) and probe (the input data) is on the order of minutes, the identification performance ranges from 91% to 95% [93, 30, 15], whereas the performances drop to 30% to 45% when the differences are in the order of months and days [53, 16, 15] for similar sized datasets. Our speculation is that other changes that naturally occur between video acquisition sessions are very important. These include change in clothing worn by the subject, change in the outdoor lighting conditions, and inherent variation in gait over time. For applications that would require matching across days or months, these would most likely be the important variables. However, there

are many applications, such as short term tracking across many surveillance cameras, for which these long term related variations would not be important.

The other factor with large impact on gait recognition is walking surface. With the subject walking on grass in the gallery sequence and on concrete in the probe sequence, rank-one recognition is only 32%. Performance degradation might be even larger if we considered other surface types, such as sand or gravel, that might reasonably be encountered in some applications. The large effect of surface type on performance suggests that an important future research topic might be to investigate whether the change in gait with surface type is predictable. For example, given a description of gait from walking on concrete, is it possible to predict the gait description that would be obtained from walking on grass or sand? Alternatively, is there some other description of gait that is not as sensitive to change in surface type?

3.3.2 Gait vs. Face

One of the open questions is the potential for gait to perform identification. We address this question by comparing our gait results with face recognition. Our analysis provides a rough guide to the current state of gait recognition. Face recognition performance has been well characterized by a number of evaluations, the most recent being the Face Recognition Vendor Test (FRVT) 2002 [40]. Because gallery size is different in the gait challenge problem and FRVT 2002, comparison is made for verification performance at a false accept rate of 1%. Unlike identification, verification performance is not a function of gallery size. Since the gait challenge problem performs recognition from outdoor video, we need to look at face recognition results from outdoor images. In FRVT 2002 there are two results on outdoor facial images. In both cases, the gallery is of indoor full frontal images. In the first result, the probe set consists of outdoor images taken on the same day as the gallery images. Verification performance varied for different systems ranging from 54% to 5%, with a median of 34%. From Table 3.1, gait performance varied from 87% to 20% on the ten experiments where the gallery and probe set sequences were taken on the same day. The

median performance score was 57%. In the second set of outdoor face recognition results, the probe set consists of outdoor images taken on a different day than the gallery image of a person; the median difference in time is about 5 months. Verification performance varied from 47% to 0% for different systems, with a median of 22%. Experiments K and L in the gait recognition problem, which have probes from 6 months later, are comparable to this scenario. The recognition rate for both experiments is 6%. A number of caveats need to be mentioned in this analysis. The FRVT 2002 performance numbers are from a blind evaluation on sequestered data. This is not the case for our gait results. On the other hand, the results in this chapter are for a baseline algorithm at the beginning of intense research of automatic gait recognition. This compares to a decade of intensive development in automatic face recognition. Using the respective performances only as a rough guide, we see that video-based gait as an outdoor at-a-distance biometrics has 1) the potential to be competitive with faces, and 2) as a biometrics to be fused with face.

3.3.3 The Greater Context

Human identification through analysis of gait information extracted from video is an important problem for computer vision. On the practical side, there are valuable potential applications in the area of video surveillance and security. Progress on gait recognition will aid progress on related problems such as characterizing human activity in video. General solutions to the gait problem will address fundamental computer vision problems that include segmentation and handling of occlusion. The process of solving this problem will identify which fundamental problems in computer vision and pattern recognition need further research. In turn, this problem will provide a method for measuring progress on the fundamental computer vision and pattern recognition problems.

The HumanID gait challenge problem provides for a scientific basis for advancing and understanding automatic gait recognition and processing. One aspect of this is that researchers wishing to work on a new algorithm will not have to invest the substantial start-up costs of acquiring a dataset large enough to lend credibility to their results. Advancements

in gait can be quantified by performance on the challenge experiments. The baseline algorithm makes it possible for researchers to focus on developing new techniques for one component of the baseline algorithm. The new component can be substituted for the baseline component and performance computed for the new component. This provides a measure of the effectiveness of the new component to the gait algorithm. As the number of researchers reporting performance results on the challenge problem increases, the potential to understand what are the critical components of gait algorithms work increases. The understanding increases because meta-analysis is possible on the different papers reporting challenge problem results. The more detailed the experimental results presented, the more detailed is the possible meta-analysis, and greater is the understanding. For example, if multiple research groups report results on different silhouettes, the greater the understanding of how silhouettes effect performance. It is this potential from the adoption of this challenge problem that represents a possible revolution in computer vision research methodology.

CHAPTER 4

IMPACT OF SEGMENTATION ON GAIT RECOGNITION

In the previous section we analyzed gait recognition and found that time and surface changes are the hard covariates. Questions arise on how these affect recognition. Are they due to fundamental changes in gait under these conditions? Or are they due to vagaries of low-level processing? Almost all of the approaches to gait recognition are based on the silhouettes of the person, which seems to be the low-level feature representation of choice. This is partly due to its ease of extraction by simple background subtraction; all approaches assume static cameras. Other reasons include the robustness of the silhouettes with respect to clothing color and texture. (It is, however, sensitive to the shape of clothing.) The silhouette representation can also be extracted from low-resolution images of persons taken at a distance, when edge based representation becomes flaky.

It is reasonable to speculate that the quality of the low-level representation is probably at fault. The quality of the silhouettes is dependent on the discriminability between the background and foreground (subject). When comparing sequences taken months apart, differences in clothing and even background would lead to different silhouette qualities. This drop in quality of extracted silhouettes can also be offered as an explanation for the drop in gait-recognition when comparing templates across surfaces because the sequences on grass and concrete also differ with respect to the background. Segmentation of silhouettes in outdoor sequences is hard primarily because of existence of shadow artifacts, changing illumination due to shifting cloud cover, and inevitable movements in the background.

Thus, the hard problems in gait recognition have to do with walking surface invariant gait recognition, being able to overcome gait variation of a person over time, and maybe silhouette quality. One might, however, speculate that *if only we had a better background*

subtraction algorithm to generate the high quality silhouettes, we would be able to get better gait recognition performance on the hard problems. Is the speculation true or not? Toward understanding this, in this chapter we consider the performance with "better" segmented silhouettes, specifically, (i) manually specified silhouettes for a subset of the Gait Challenge dataset and (ii) with silhouettes that have been "cleaned" using a population HMM Eigen-Stance model, for the complete dataset.

4.1 Manual Silhouettes

Manual silhouettes were created for a subset of Gait Challenge dataset. More details about the process and quality checks can be found in [61, 55]. Here we highlight some salient aspects. About 71 subjects from one of the two collection periods (May collection) were chosen for manual silhouette specification. The sequences corresponding to these 71 subjects in the (i) gallery set (sequences taken on grass, with shoe type A, right camera view), (ii) probe B (on grass, with shoe type B, right camera view), (iii) probe D (on concrete, with shoe type A, right camera view), (iv) probe H (on grass, with shoe A, right camera view, carrying briefcase), and probe K (on grass, time). We manually specified the silhouette in each frame over one walking cycle, of approximately 30 to 40 image frames. This cycle was chosen to begin at the right heel strike phase of the walking cycle through to the next right heel strike. We attempted to pick this gait cycle from the same 3D location in each sequence, whenever possible. In addition, we tried to exclude the portion that included the black and white calibration box with high contrast, which frequently leads to high background subtraction errors.

We did not just mark a pixel as being from the background or subject, but provided more detailed specifications in terms of body parts too. We explicitly labeled the head, torso, left arm, right arm, left upper leg, left lower leg, right upper leg, and right lower leg using different colors. Fig. 4.1 shows some examples of part-level ground truth silhouettes corresponding to the images in the top row. Quality control checks looked for miscolored parts and backgrounds, randomly colored isolated pixels, errors on the boundary of the

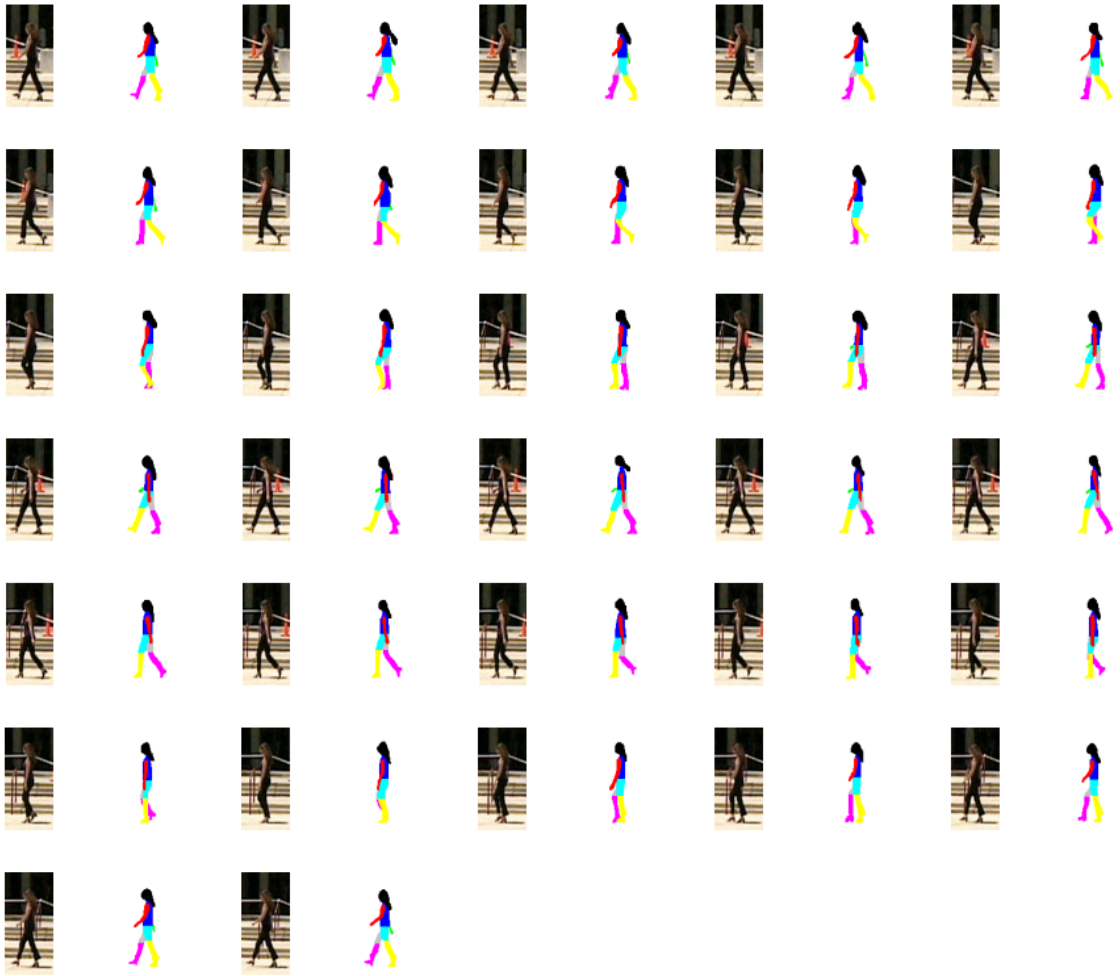


Figure 4.1. Part Level Manual Silhouettes over One Gait Cycle along with the Corresponding Color Images, Cropped Around the Person.

body, and missed body parts. Some of the difficulties encountered during the creating process include low-image quality due to varying overall intensity, occlusion of feet in the grass sequences, similarity of dark skin tones of some subjects with the background, frequent occlusion of the right arm, and the presence of dark or baggy clothing, which made it hard to delineate various body parts. However, despite these difficulties we were able to create pretty consistent quality silhouettes, as judged visually by another subject, across the subjects.

Fig. 4.2 shows the silhouettes of a subject in image frames taken from four different cameras at different distances and surfaces. To remove possible bias in recognition due the use of silhouette height, we normalize the height of the silhouettes to occupy 128 pixels. The bottom row of Fig. 4.2 shows the height scaled and centered silhouettes of the kind used by gait recognition algorithms. To facilitate the fast computation of similarity measure, following the baseline algorithm [42], we also align the silhouettes in each frame along the horizontal direction so that the centerline of the torso is at the middle of the frame. This centerline is estimated as follows. First, we compute the number of *connected* foreground pixels in each row in the *upper half* of the silhouette. If there are more than one section of connected foreground pixels in a row, e.g., when person's arm move out of torso, we consider the largest one, which is most likely to be the torso portion. For each row we consider the starting column of connected component s_i (front of the torso at each row) and the half of the size of the connected component, l_i (half the width of the torso at a row). The center line is estimated to be at the average of the median of these two distributions. Alternative strategies such choosing the median of the average of the start and end index of the foreground in each row did not result in good centering of the silhouettes, as judged visually.

4.2 Model Based Silhouette Reconstruction

Silhouettes, detected by some form of background segmentation, typically involve errors due to: (i) shadows, (ii) inability to segment parts because they fall just below the threshold and are classified as background, (iii) moving objects in the background, such as the fluttering tape in the concrete sequences or moving leaves in the grass sequences, or other moving persons in the background, and (iv) compression artifacts near the boundaries of the person, which are present in medium cost, consumer grade cameras. Fig. 3.3 shows examples of these kinds of errors. The silhouettes were extracted by the baseline algorithm described in Section 3.1.1: we compute the background statistics of the RGB values at each image location, (x, y), in terms of the mean $\mu_B(x, y)$ and the covariances $\Sigma_B(x, y)$ of

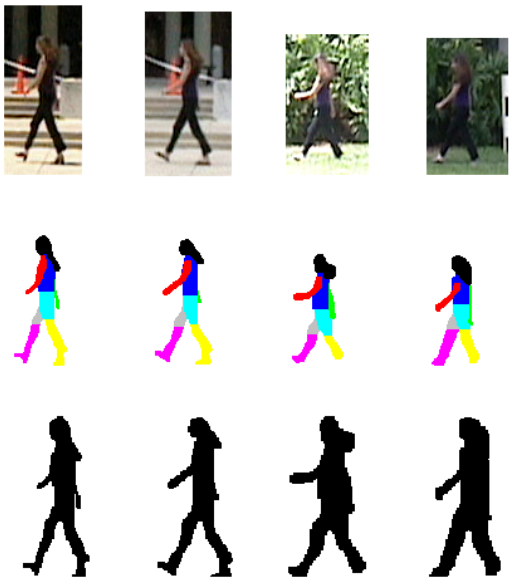


Figure 4.2. Top Row shows the Color Images, Cropped around the Person for Four Different Camera Views. The Middle Row shows the Corresponding Part-Level, Manually Specified Silhouettes. And the Bottom Row shows the Scaled Silhouettes of the Kind Used by Gait Recognition Algorithms.

the RGB values at each pixel location. Using the Mahalanobis distance of a pixel value as the observation, pixels are classified into foreground or background using Expectation Maximization (EM) with a Gaussian mixture model.

In the past, various strategies, mostly based on pixel-based processing of photometric attributes, have been proposed to reduce shadow artifacts. However, these approaches have problems in the presence of strong shadows and, of course, these strategies cannot handle missing body parts or extraneous background moving objects merged with the foreground. We handle these kinds of segmentation problems using prior body shape models, as captured by a population based Hidden Markov Model (pHMM) coupled with an Eigen-Stance gait shape model.

The states of the HMM represent a gait stance and the transition probabilities capture the motion dynamics between the states for the subject population. This HMM is learnt based on the manually specified silhouettes for 71 subjects. For each gait stance, we also

construct, using the manually specified silhouettes, statistical shape models in terms of the mean silhouette shape and variances of that stance shape. This statistical model, which we call the *Eigen-Stance Gait Model* is accomplished by performing principal component analysis (PCA) for each stance.

Each frame in any given sequence is matched onto these stance subspaces using the population based Hidden Markov Models (pHMM), statistically describing the gait motion over a subject population. Each silhouette is then reconstructed using the coordinates of the silhouette, found by projecting onto the matched Eigen-Stance model.

4.2.1 Forming Stance Exemplars

The observation model for each HMM state is the most critical aspect of the specification, so we describe it some detail. The observation variables are the distances of a given observed silhouette from an exemplar set, which we compute by clustering the frames of each given sequence. The particular clustering method employed is constrained K-means clustering. Of course, any clustering method relies on a distance measure, which we define as follows. Let $\mathbf{f_i}$ and $\mathbf{f_j}$ be two vertically scaled and horizontally aligned (see Fig. 4.2), silhouette frames, reformatted into row-scanned column vectors. Then the similarity between them is

$$S(i,j) = \frac{\mathbf{f_i}^T \mathbf{f_j}}{\mathbf{f_i}^T \mathbf{f_i} + \mathbf{f_j}^T \mathbf{f_j} - \mathbf{f_i}^T \mathbf{f_j}} \tag{4.1}$$

Note that for binary silhouettes, with pixels values being just 0 or 1, this similarity is the ratio of the pixels in the intersection of the two overlapped silhouettes to the number of pixels in their union and is also commonly known as the Tanimoto similarity measure. One minus this similarity is the Tanimoto distance metric for binary silhouettes; $D(\mathbf{f_i}, \mathbf{f_j}) = 1 - S(i,j)$. For non-binary silhouettes too, we refer to the above distance (similarity) measure as the Tanimoto distance (similarity).

To create the exemplars, we first partition the frames in one gait cycle into N_S equal segments. We use one full cycle (two strides) so as to retain the asymmetry in gait, i.e. to differentiate stances with left foot forward from those with right foot forward. We group

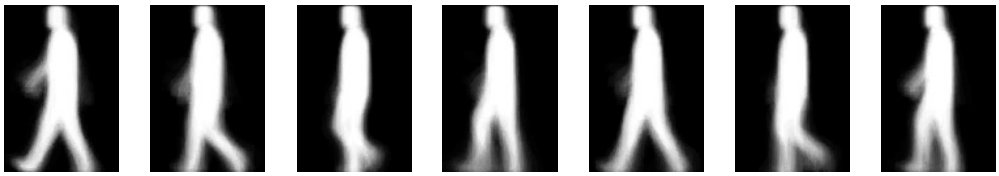


Figure 4.3. Average Stances in Population Exemplars for 7 Sample States over a Gait Cycle.

the frames within the j^{th} partition of all people into an exemplar set for the j-th gait stance, E_j. Since the gait cycles of the manual silhouettes are aligned, this strategy of corresponding the exemplars from different subjects works.

Exemplars for each stance form a set. The initial exemplar sets, $E_j^{[0]}$, are further refined by reassigning the frames, based on the distance, $D(\mathbf{f_i}, \mathbf{f_j}) = 1 - S(i,j)$, by K-means clustering with some constraints. Let $\{E_j | j = 1, \cdots, N_s\}$ represent the set of state exemplars. Then,

$$\overline{E_j}^{[k]} = \frac{1}{N} \sum_{\mathbf{f_i} \in E_j^{[k]}} \mathbf{f_i} \tag{4.2}$$

$$E_j^{[k+1]} = \{\mathbf{f_i} | D(\mathbf{f_i}, \overline{E_j}^{[k]}) < (D(\mathbf{f_i}, \overline{E_{j-1}}^{[k]}), D(\mathbf{f_i}, \overline{E_{j+1}}^{[k]}))\} \tag{4.3}$$

Note that constraint that frames can only be re-assigned to only to neighboring exemplar sets; thus a frame in E_j can be reassigned to exemplars E_{j-1} or E_{j+1}. We also insist that every exemplar should contain at least one frame from each sequence. We stop when no more reassignments can be done; about 10 iterations were enough for our experiments. Fig 4.3 shows the mean silhouettes, $\overline{E}_j$, of exemplar sets for 7 example stances.

4.2.2 Population Hidden Markov Model (pHMM)

We will use an HMM to align any given sequence to generic stance sequences for stance-dependent silhouette reconstruction. A Hidden Markov Model (HMM) is specified by the possible states, $q_t \in \{1, \cdots, N_s\}$ and the triple $\lambda = (A, B, \pi)$, representing the state transition matrix, observation model, and priors, respectively. The state transition matrix A with entries $a(i,j) = P(q_{t+1} = j | q_t = i)$ is constrained to represent a cyclical version of

the left to right Bakis state transition model over N_s states, allowing only for jumps to the next state. The observation model is comprised of the observation models for each state, $B = \{b_j(\mathbf{f_t})|j = 1, \cdots, N_s\}$, where $b_j(\mathbf{f_t}) = P(\mathbf{f_t}|q_t = j)$, i.e. the conditional probability of the observed silhouette, $\mathbf{f_t}$, at time t given that the state at time t is j. We choose the observation model to be exponential in terms of the Tanimoto distance, D, between any given silhouette, $\mathbf{f_t}$, to the mean of the state exemplars, $\overline{E}_j$.

$$b_j(\mathbf{f_t}) = \frac{1}{\mu_j} e^{-\frac{D(\mathbf{f_t}, \overline{E}_j)}{\mu_j}} \tag{4.4}$$

The observation model is thus parameterized by the mean μ_j. The HMM structure is somewhat similar to that used in [84] for recognition, but in our case it is designed to model gait dynamics over a population. Differences also exist in the observation model and the state definitions; our model takes into account the gait asymmetry between the two strides over a cycle.

4.2.2.1 Model Parameter Estimation

We pick equal state priors, i.e. $\pi_i = \frac{1}{N_s}$, since, in practice, any given sequence can begin from any state. However, both the transition matrix and the observation model parameters need to be estimated. Since the exemplar sets have been computed from the given training sequences, we just estimate the observation model parameters for each stance, directly from the corresponding exemplars.

$$\mu_j = \frac{\sum_{\mathbf{f_i} \in E_j} D(\mathbf{f_i}, \overline{E}_j)}{|E_j|} \tag{4.5}$$

The initial estimate of the transitions matrix is also formed from the exemplars and then refined using Levinson's method for training with multiple observation sequences based on iterative Baum-Welch algorithm.

$$a^{[0]}(i,j) = \frac{\sum_k \#\text{ of } \mathbf{f^k_{t+1}} \text{ in } E_j \text{ given } \mathbf{f^k_t} \text{ in } E_i}{|E_i|} \tag{4.6}$$

We refer the reader to standard texts such as [71] for details regarding the Levinson's method. Here we present just the key equations. Let there be K observation sequences, $\{I^1, \cdots, I^K\}$. However, since for each training sequence we have only one gait cycle, to retain the cyclical property, we extend each sequence by appending its first frame to the tail: $I^i = \{f_0^i, \cdots, f_{N_I}^i, f_0^i\}$. The length of the extended I^k is denoted by T_k. The iterative re-estimate of the population transition probabilities, $A^{[n+1]}$, is given by

$$a^{[n+1]}(i,j) = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i) a^{[n]}(i,j) b_j(\mathbf{f}_{t+1}^k) \beta_{t+1}^k(j)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k-1} \alpha_t^k(i) \beta_t^k(i)} \tag{4.7}$$

where $P_k = P(I^k|\lambda)$, the likelihood of k-th observation, and the forward and backward probabilities α^k and β^k, are arrived at by induction as follows.

$$\begin{aligned} \alpha_t^k(j) &= P(\mathbf{f}_1^k, \cdots, \mathbf{f}_t^k, q_t = j|\lambda) \\ &= b_j(\mathbf{f}_1^k)/N_s && \text{for } t = 1 \text{ and } 1 \le j \le N_s \\ &= \sum_{i=1}^{N_s} \alpha_{t-1}^k(i) a^{[n]}(i,j) b_j(\mathbf{f}_t^k) && \text{for } 2 \le t \le T_k - 1 \text{ and } 1 \le j \le N_s \end{aligned} \tag{4.8}$$

$$P_k = P(I^k|\lambda) = \sum_{j=1}^{N_s} \alpha_{T_k}(j) \tag{4.9}$$

$$\begin{aligned} \beta_t^k(i) &= P(\mathbf{f}_{t+1}^k, \cdots, \mathbf{f}_{T_k}^k|q_t = i, \lambda) \\ &= 1 && \text{for } t = T_k \text{ and } 1 \le i \le N_s \\ &= \sum_{j=1}^{N_s} a^{[n]}(i,j) b_j(\mathbf{f}_{t+1}^k) \beta_{t+1}^k(j) && \text{for } t = T_k - 1, \cdots, 1 \text{ and } 1 \le i \le N_s \end{aligned} \tag{4.10}$$

The above equations (Eqs. 4.7-4.10) represent the generalization of the Baum-Welch equations for multiple observations and need to be iterated over until the likelihood of the given observations are maximized. The learnt transition matrix emphasizes the transitions to forward states, manifesting as high values along the first upper diagonal. We also found high values at the anti-diagonal corner, which is because we adopt a cyclical Bakis model.

4.2.2.2 Model Size Determination

We determine the number of states, N_s, based on the Akaike Information Criterion (AIC) [3], which takes both the goodness of fit and generalizability into account:

$$\text{AIC} = -2 \sum_{k=1}^{K} \log_2 P(I^k|\lambda) + 2N_{para} \tag{4.11}$$

where (λ) is the estimated population HMM model, K is the number of training sequences, and N_{para} is the number of estimated parameters of the model. The estimated parameters include the N_s^2 transition probabilities and the N_s parameters in the observation model. Fig. 4.4 plots the variation of AIC with the number of states for two different training sets of 71 subjects, one over grass walking surface and the other over concrete walking surface. Based on this plot we choose the round figure of 20 states as being fairly optimal for both the sets of sequences. It is better to err towards the larger number of states so as to retain the shape variations among different individuals.

4.2.3 Eigen-Stance Gait Model

The goal of the Eigen-Stance gait model is to capture the *shape* variations in the silhouettes for each stance across persons. We model this variation as a multivariate Gaussian distribution, which is estimated from the clustered set of exemplar silhouettes associated with each HMM stance. We use principal component analysis (PCA) to arrive at a compact representation of this distribution. For each stance, k, we have reduced dimensional (with N_e dimensions) shape space, $\Phi(k)$, characterized by the mean, $\mu_{\mathbf{k}}$ and the eigenvectors $\{\mathbf{e_{k,1}}, \cdots, \mathbf{e_{k,N_e}}\}$. Given that the final context is identification, we want this shape space to capture variation across persons. However, we have to be careful to ensure that an equal number of training samples is used for each person so as not to bias the model to any particular subgroup of persons. For instance, persons with slow gait would tend to have more samples in each state exemplar. So, for each stance, we used one sample silhouette per person in the training set. We choose the one closest to the mean of the corresponding

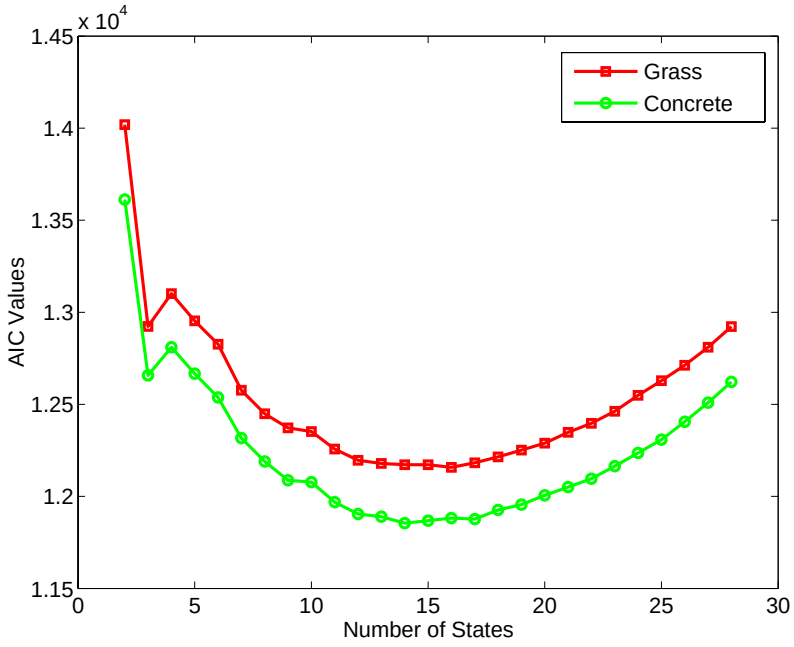


Figure 4.4. Variation of AIC with Number of States, for Models Constructed Using Two Different Training Sets of 71 Subjects; One for Grass Walking Surface(in Green) and the Other for Concrete Walking Surface (in Red).

exemplars, as measured by the Tanimoto distance measure. Notice that this also ensures that the spaces of all $\Phi(k)$'s are constructed with an equal number of training samples.

Considering the strong impact of walking surface type on gait recognition, we built different Eigen-Stance gait models, coupled with their own HMMs, for grass and concrete surfaces using manual silhouettes associated with the Gallery set and Probe D set, respectively, from the USF HumanID database. Fig. 4.5 shows some sample Eigen-Stances of both spaces. The number of eigenvectors, N_e, is chosen so that at least 80% of the variation is modeled.

4.2.4 Stance Matching using HMM

In order to project and reconstruct silhouette frames in any given sequence, $\{\mathbf{f_1}, \cdots, \mathbf{f_T}\}$, they have to be matched to one of the N_s stances in the population HMM. The dynamic programming based Viterbi algorithm is used for this purpose [70]. It returns the most

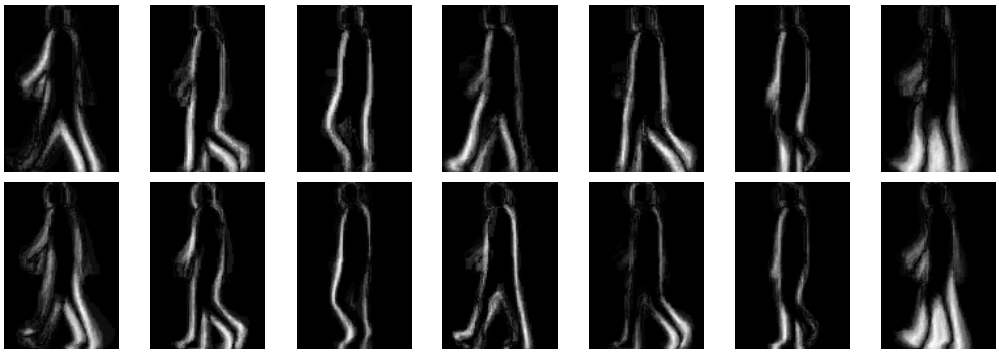


Figure 4.5. Samples of the First Eigen-Stances over one Gait Cycle, Representing the most Discriminating Directions among Persons. The Top Row was Built Using Silhouettes from the Grass Walking Sequences, And the Bottom Row is for the Concrete Walking Surface Sequences.

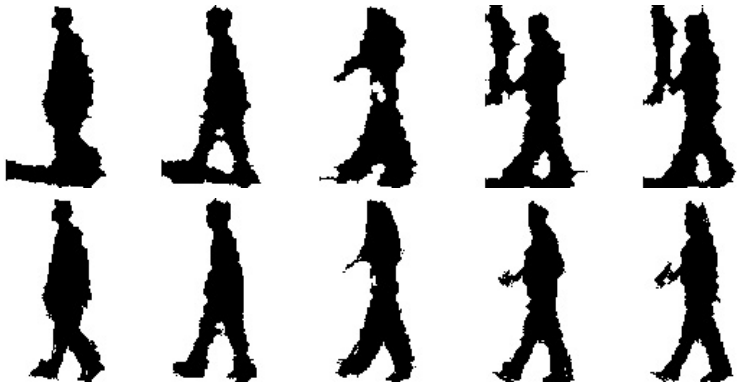
likely state assignment to the input frames. To reduce the combinatorics of this assignment process, we partition the input sequence into subsequences of roughly one gait cycle length, which is estimated from the periodic variation in the number of foreground pixels in the bottom half of the silhouettes. Note that the starting state of these subsequences need not match the starting HMM state; the cyclical nature of the HMM model can handle this.

4.2.5 Reconstruction

After each input frame $\mathbf{f_i}$ is estimated to be at phase j by the HMM, it is projected into the corresponding eigen-space, $\Phi(j) = \{\mu_\mathbf{j}, \mathbf{e_{j,1}}, \cdots, \mathbf{e_{j,N_e}}\}$, and then reconstructed as $\mathbf{f_i^r}$.

$$\mathbf{f_i^r} = \mu_\mathbf{j} + \sum_{k=1}^{N_e} \left(\mathbf{e_{j,k}}^T (\mathbf{f_i} - \mu_\mathbf{j}) \right) \mathbf{e_{j,k}} \tag{4.12}$$

The reconstructed silhouette, $\mathbf{f_i^r}$, has continuous values between 0 and 1 that we threshold to arrive at binary silhouettes. Instead of simple thresholding, we employ a two-level thresholding scheme to minimize the side effect of reconstruction process, which can make silhouettes more similar to each other. We have empirically verified that a single thresholding scheme produces silhouettes that are more similar to the mean silhouettes than the



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

Figure 4.6. The Top Row show some Instances of Poor Quality Silhouettes and the Bottom Row Shows the Reconstructed Silhouettes.

double thresholding scheme given below.

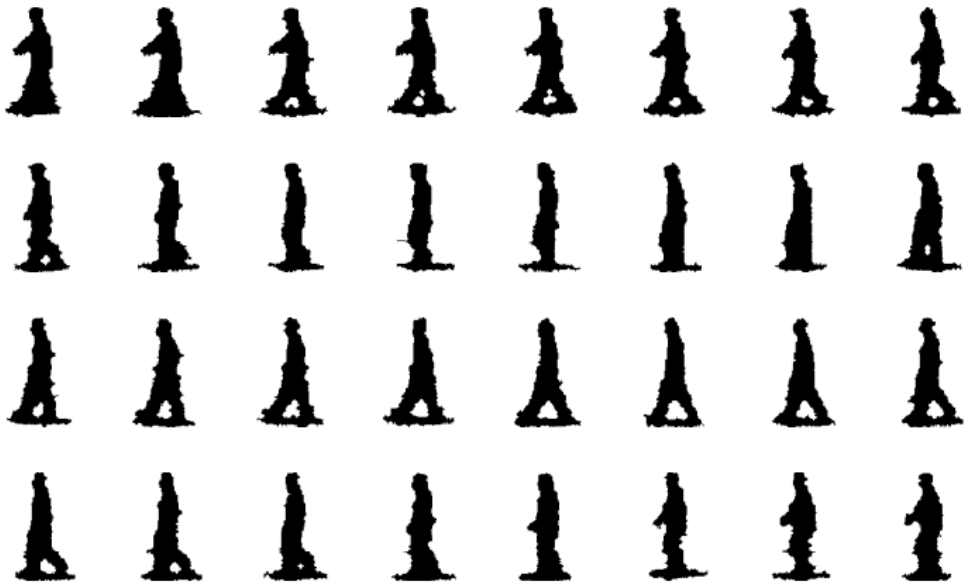
$$\mathbf{F}_{\mathbf{i}}^{\mathbf{r}}(k) = \begin{cases} \text{Foreground} & \text{if } \mathbf{f}_{\mathbf{i}}^{\mathbf{r}}(k) > T_{high} \text{ or } \mu_{\mathbf{j}}(k) = 1 \\ \text{Background} & \text{if } \mathbf{f}_{\mathbf{i}}^{\mathbf{r}}(k) < T_{low} \\ \mathbf{f}_{\mathbf{i}}(k) & \text{otherwise.} \end{cases} \tag{4.13}$$

For the experiments in this dissertation, $T_{low} = 0.2$ and $T_{high} = 0.8$.

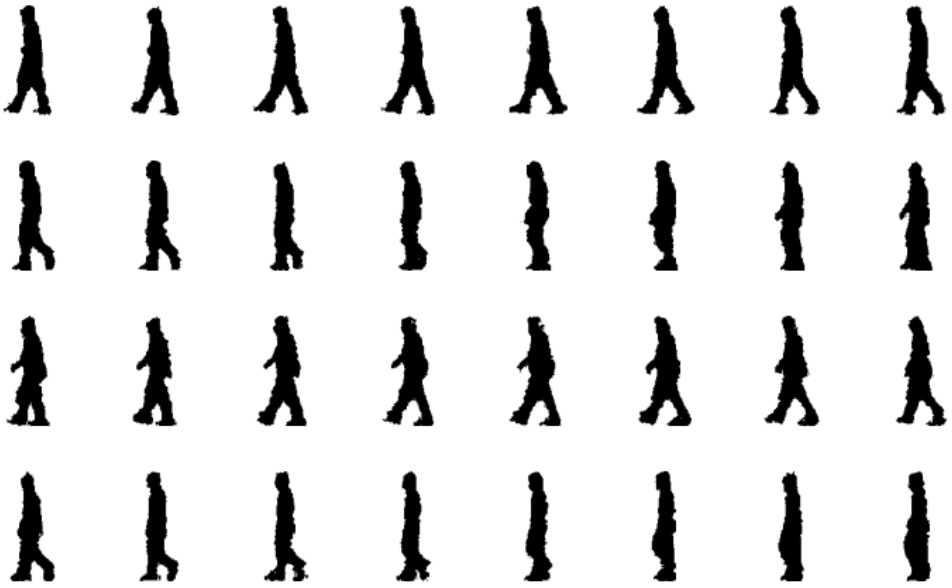
4.3 Quality of Reconstructed Silhouettes

What is the quality of the reconstructed silhouettes? Are the pixels that are removed mostly "noise" pixels? Are any true foreground pixels removed? These questions we address in this section.

The raw silhouettes are those produced by the baseline algorithm *with some modifications*. Steps of the baseline silhouette detection algorithm are (i) compute the statistics of the individual background pixels in terms of mean and covariance of RGB values, (ii) compute the Mahalanobis distance of a pixel from this background pixel value distribution, (iii) smooth the Mahalanobis distance using a 9 by 9 triangular window to fill in holes and to join several small pieces, (iv) decide on an optimal threshold to segregate the two classes using expectation maximization (EM) with the distance values as the observations,

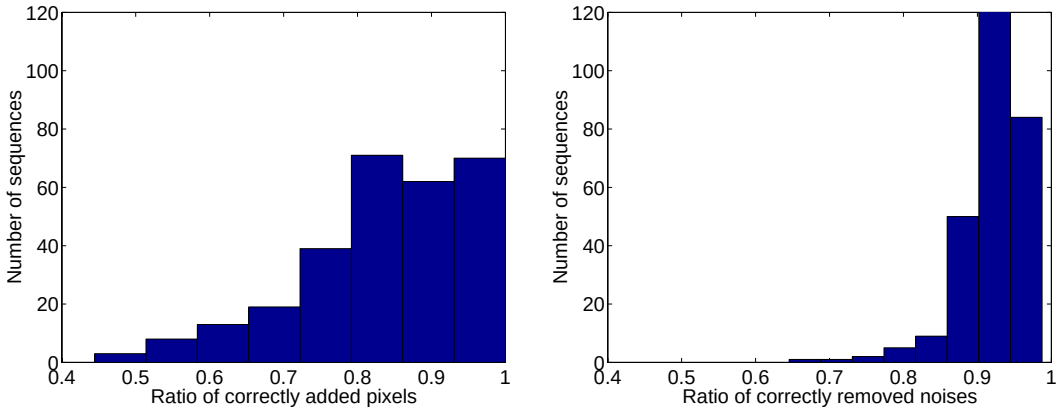


(a) Before



(b) After

Figure 4.7. Silhouettes over One Gait Cycle (a) Before and (b) After Reconstruction.



(a) (b)

Figure 4.8. Histogram of the Ratio of (a) *Correctly Added* Foreground Pixels to the Total Number of *Added* Foreground Pixels and (b) *Correctly Removed* Noise Pixels to the Total Number of *Removed* Pixels.

and (v) pick the largest connected component. The smoothing in step iii above results in thicker silhouettes with high false positive predictive values and have been found to result in poor gait recognition performance for some recognition strategies [52]. So, we *eliminate that smoothing step*. However, this causes the problem of losing body portions because we only pick the largest component in step v. So, we replace step v with a custom proximity based grouping process that assembles disconnected components: first, we morphologically close the silhouette twice in the vertical direction using a 3×1 element so as to reconnect body parts; most disconnections happen in the vertical direction, e.g., trunk between leg. Then for each connected component CP_i in a frame, we compute two values $Para_i(1)$ and $Para_i(2)$ in terms of the largest component CP_{max}: $Para_i(1) = \frac{\text{Area of } CP_i}{\text{Area of } CP_{max}}$ and $Para_i(2) = e^{-\theta_i}$, where θ_i is the vertical angle between the center point of CP_{max} and CP_i. We decide to group components based on the product of $Para_i(1)$ and $Para_i(2)$. Finally we do the morphological closing operation with a 3×3 element in order to fill the holes inside the silhouette.

Thus, using simple low-level methods, the quality of the raw baseline silhouettes is enhanced somewhat. However, artifacts do remain. These form the input to the reconstruction process. The Eigen-Stance based silhouette reconstruction process removes many

of the artifacts. Fig. 4.6 shows some example of the quality of reconstruction (bottom row) for poor quality input silhouettes (top row). Columns (a) and (b) of Fig. 4.6 show cases where shadows were removed; column (c) shows a case where holes in the foreground were filled in; and (d) & (e) show examples of removal of another person in the background. Fig. 4.7 shows an example of reconstruction over one gait cycle. We see that most frames have been improved, suggesting our model works well for different gait phases.

4.3.1 Pixel Level Quality

One manner to evaluate the Eigen-Stance model is to analyze the types of pixels that are edited (either removed or added) during the reconstruction process. The measures of performance could be the ratio of *correctly added* foreground pixels to the total number of *added* foreground pixels and the ratio of *correctly removed* noise pixels to the *total* number of removed pixels. Ideally, both these ratios should be one. We compute these two quantities for each frame for which we have manually specified ground-truth image and average them over a sequence. Fig. 4.8 shows the histogram of the two ratios over all the sequences for which we have manual silhouettes. We see that the histograms are strongly biased towards one. Thus suggesting that the editing during the reconstruction process is mostly correct.

We also evaluate the silhouette quality at pixel level using the measures of false positive predictive value ($P_{\overline{PPV}}$) and and detection rates (P_D). The false positive predictive value is the probability that a pixel classified as foreground is actually from the background. Note that this is different from false alarm probability, which is the fraction of the actual background that is marked as foreground. Since the background is unbounded in an image, computing false alarm probability is not meaningful. The false positive predictive value is also used in epidemiology where data about the negative class is sometimes hard to get. The detection rate is the probability that a foreground pixel is classified as foreground. For each frame with corresponding manual silhouettes we compute the false positive predictive values and detection rates. We then average these quantities over all the frames from one sequence, resulting in one pair of performance numbers for each sequence.

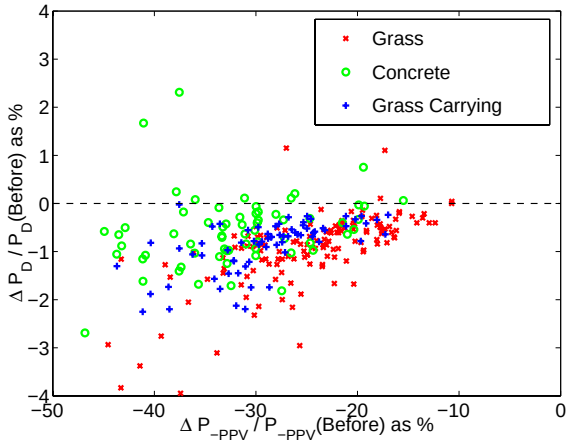


Figure 4.9. Scatter Plot of Percentage Change in Pixel Level Detection (ΔP_D) and False Predictive Values ($\Delta P_{\overline{PPV}}$) after Silhouette Reconstruction. The Green Circles correspond to the Concrete Sequences, the Red Crosses correspond to the Grass Sequences and the Blue Stars correspond to the Briefcase Carrying Sequences.

Fig. 4.9 shows the percentage improvement with reconstruction in terms of pixel level detection:

$$\Delta P_D = 100 \frac{P_D(After) - P_D(Before)}{P_D(Before)}$$

and false positive prediction:

$$\Delta P_{\overline{PPV}} = 100 \frac{P_{\overline{PPV}}(After) - P_{\overline{PPV}}(Before)}{P_{\overline{PPV}}(Before)}$$

We separately report the results for the grass and concrete sequences since they have different backgrounds. We also show the improvement for sequences with briefcase. Improvement in silhouette quality would be indicated by $\Delta P_{\overline{PPV}} < 0$ and ΔP_D around 0, which is observed in the plots. We see that although the detection rate of the reconstructed silhouettes dropped a little bit by about 1%, the false positive predictive value dropped much more dramatically by about 20% to 30%.

4.3.2 Robustness with Viewpoint Variation

Since the Eigen-Stance model is a view-based representation, it is reasonable to ask how effective is the reconstruction process in handling silhouettes viewed from somewhat

different viewpoints. The Gait Challenge dataset permits such study; it includes datasets of the same gait event viewed from two different angles, with the verging angle of roughly 30^o. The manual silhouettes, which were used to construct the Eigen-Stance model, only exist for the sequences viewed from the right camera. We use the sequences viewed from the left camera to test the robustness. However, since we do not have manual silhouettes for these left camera sequences, we can only view the quality subjectively. We show results on 10 such sequences in Fig. 4.10. Samples of both the original and the reconstructed frames are shown. As we have seen before, the model is able to successfully remove most shadows and other background noise artifacts.

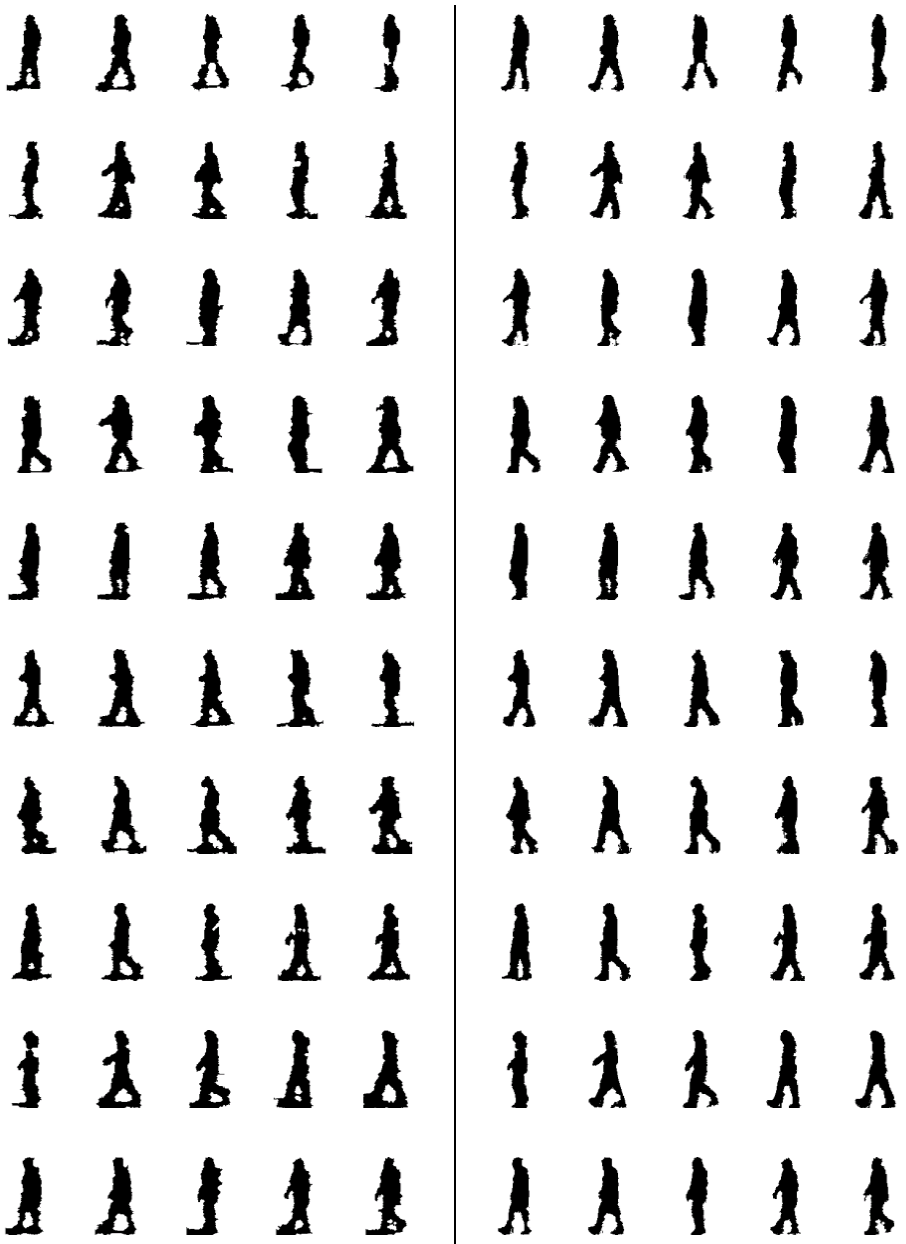
4.3.3 Generalizability to Different Datasets

To evaluate the generalizability of the developed model to other databases, we test it on the Georgia Tech outdoor dataset [1]. As Fig. 4.11 shows, it consists of 20 subjects walking on outdoor concrete surface. We use the modified baseline silhouette extraction algorithm described in Section 4.3 to produce raw silhouettes, which are then processed by the Gait Challenge data based Eigen-Stance model. Fig. 4.12 shows the original and reconstructed Silhouettes over one gait cycle for one subject. And Fig. 4.13 shows the original and reconstructed sample frames from 10 different individuals. We found that the quality of the original silhouettes that we could extract by simple background subtraction is poor due to low contrast and strong outdoor illumination. However, they have been substantially improved by the model, which indicates the applicability of the built Eigen-Stance model beyond the gait challenge dataset.

4.4 Impact on Gait Recognition

We have illustrated that the Eigen-Stance model substantially improves the silhouette qualities. But, do the improved silhouettes affect gait recognition performance? One concern is that the model based reconstruction process might result in silhouettes that

[1] It can be downloaded from `http://www.cc.gatech.edu/cpl/projects/hid/`



(a) Original Silhouettes (b) Reconstructed Silhouettes

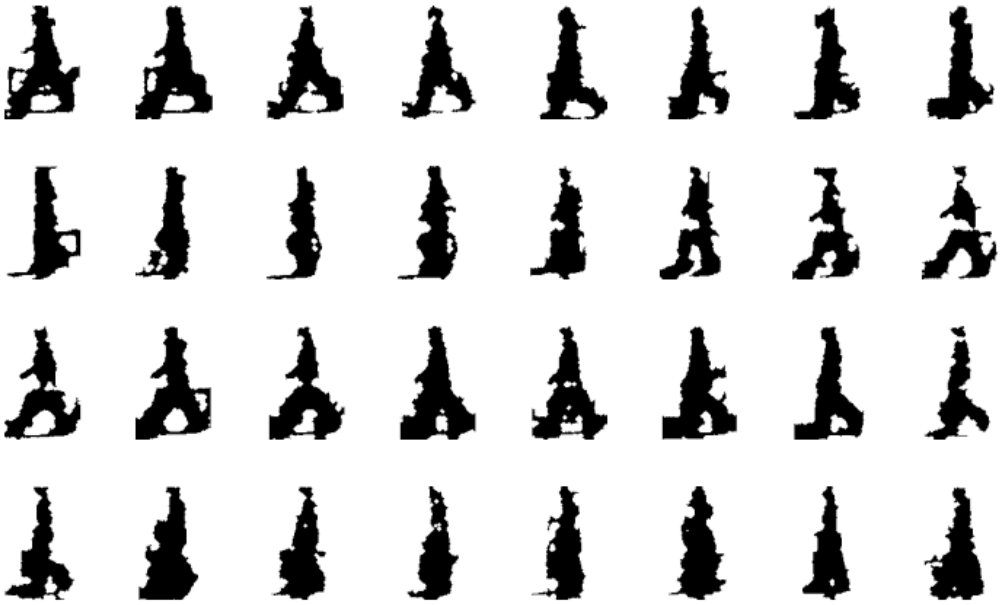
Figure 4.10. The Reconstruction of Silhouettes for a Sequence with 30 Degrees View Angle Difference from those Used to Construct the Eigen-Stance Gait Model. Frames in Each Row are from the Same Sequence.

Figure 4.11. A Sample Frame from the Georgia Tech Outdoor Gait Dataset.

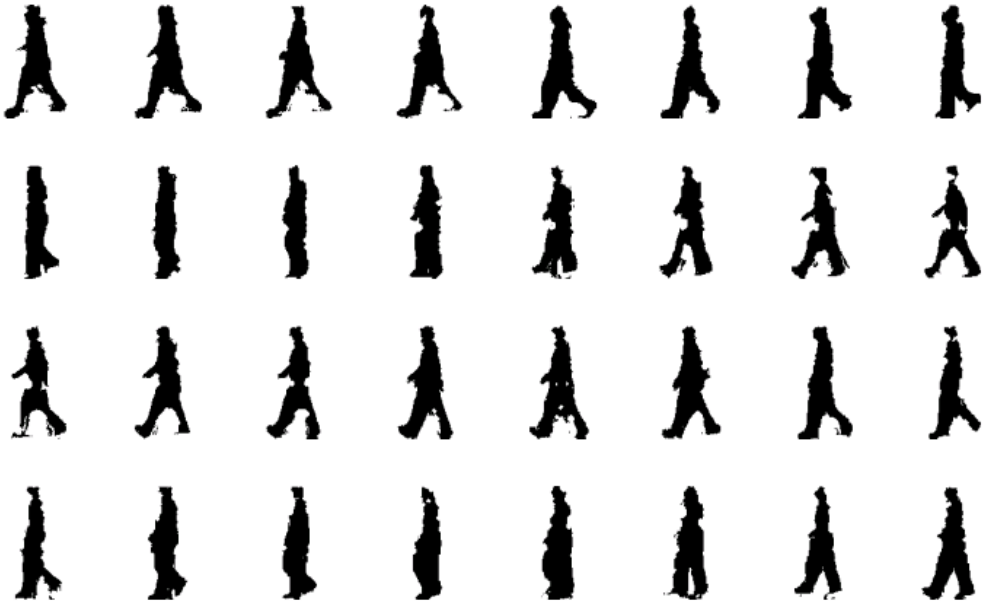
are more similar to each other, bringing down recognition performance. To study this, we consider recognition from both (a) the manually specified silhouettes over part of the data set and (b) the reconstructed silhouettes over the whole dataset. We will quantify gait recognition performance using the baseline algorithms described in Chapter 3.

4.4.1 Recognition from Manual Silhouettes

First, we consider recognition from manual silhouettes using the baseline algorithm. Since the shape based gait recognition algorithm uses a population HMM trained with the manual silhouettes, it does not make empirical sense to also compute recognition rates from the manual silhouettes using it. So, we used just the baseline gait recognition algorithm on the manual silhouettes. However, since manual silhouettes are specified over only one gait cycle, we had to modify the original baseline similarity computation. There is no need for the probe partitioning step and the correlation process. We can simply compute the distance by establishing a mapping between the frames in the two sequences and then summing the corresponding Tanimoto similarities between the matched frames. The fact that all the manual silhouettes start and end in the same stance makes the frame matching process somewhat easy. Some of the strategies include extrapolating the smaller sequence

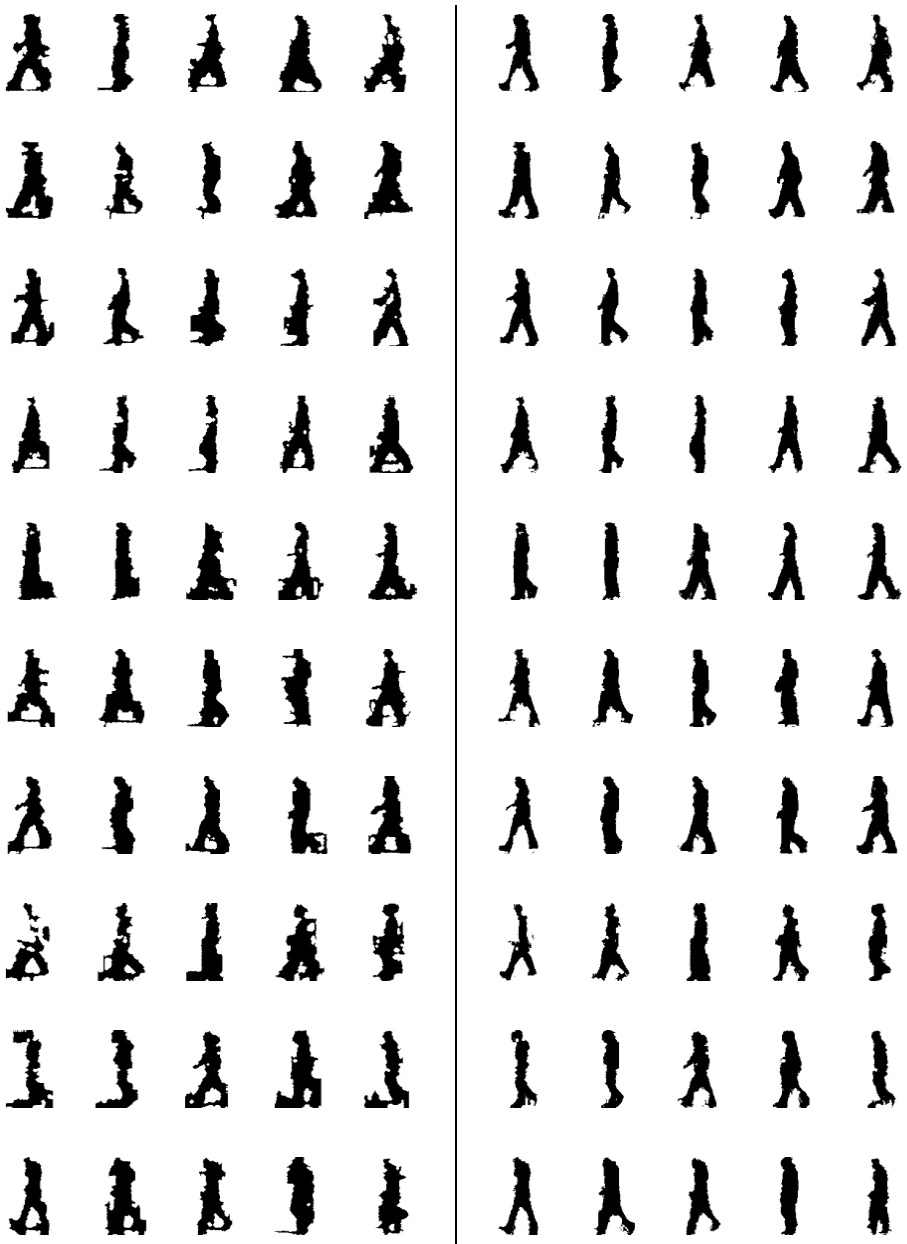


(a) Original



(b) Reconstructed

Figure 4.12. (a) Original and (b) Reconstructed Silhouettes over One Gait Cycle for One Subject from the Georgia Tech Dataset.



(a) Original Silhouettes (b) Reconstructed Silhouettes

Figure 4.13. Samples of (a) Original and (b) Reconstructed Silhouettes over One Gait Cycle for 10 Subjects from the database of Georgia Tech. Each Row corresponds to a Subject.

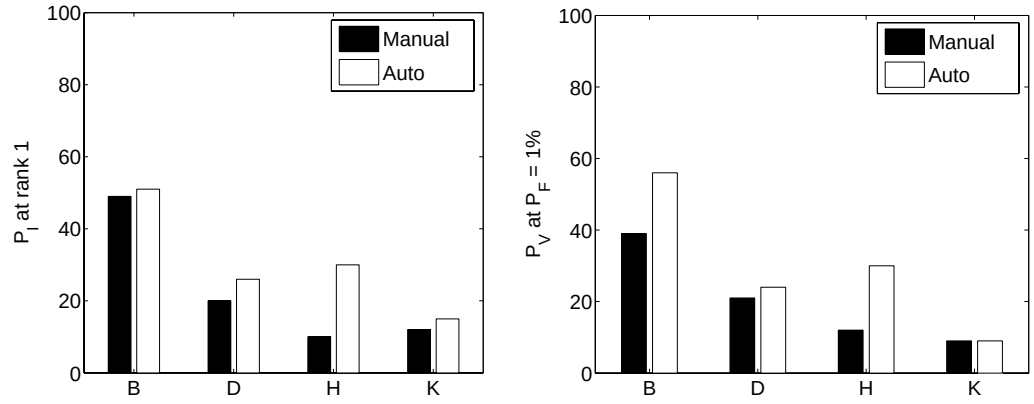


Figure 4.14. Recognition Performance of the Baseline Gait Recognition Algorithm in terms of Identification Rate at Rank 1 and Verification Rate at a False Alarm Rate of 1% with Manual Silhouettes over one Gait Cycle and with (Unreconstructed) Automated Silhouettes over that Same Cycle. Results for four key experiments are listed: B(shoe), D(surface), H(briefcase), and K(Time).

by repeating it, or linearly warping the frames in the smaller sequence to those in the larger sequence, or dropping frames at the beginning or the ending of the larger sequence. Of all the variations, we found that the linear warping strategy produced the best results.

To compare performance, we consider the key challenge experiments involving shoe, surface, carrying, and time variation between probe and gallery in the gait challenge problem. The gallery and probe sets for the experiments are reduced to contain the sequences for which we have manual silhouettes. Since recognition with manual silhouettes uses just one gait cycle, we compare the performance with (unreconstructed) automated silhouettes also over the corresponding gait cycles, using a similarity computation strategy same as that for the manual silhouettes. In Fig. 4.14, we report the performance numbers for both the identification and the verification scenario using the identification rate at top rank and the verification rate for a 1% false alarm values, respectively. We see that the performance with manual silhouette actually *drops* slightly for shoe and the surface variation experiments, possibly due to removal of shadow correlations in these experiments. These drops are not statistically significant according to (non-parametric) McNemar tests. This

suggests that *the low performance under the impact of surface and time variation can not be explained by the silhouette quality.*

On the other hand, significant performance drop is reported in the experiment of carrying briefcase (about 20% at top rank). The briefcase, which weights about 5 kgs, might be changing gait; the high recognition with automated silhouettes might be due to error correlations, i.e. shadow regions or clothing artifacts. To shed more light on this, we consider the recognition power in (i) the difference image between the automated silhouette and manual silhouette, and (ii) the pixels edited (either removed or added) during reconstruction. We found that the identification rates at rank one with these error pixels were 28% and 25%, respectively, which roughly make up for the gap between the recognition rates based on manual and automated silhouettes.

4.4.2 Recognition from Reconstructed Silhouettes

We saw that recognition from manual silhouettes over one cycle did not improve recognition; in fact, the performance dropped. Does this effect also remain if we use multiple cycles? For this, of course, we do not have manual silhouettes. However, we do have the reconstructed silhouettes, which were shown to be of better quality than the raw silhouettes.

Fig. 4.15 summarizes the baseline performances with raw and reconstructed silhouettes for some of the key gait challenge experiments. We see a drop in performance for experiments A (view), B (shoe), and H (carry). This is consistent with the results we obtained for the manual silhouettes. The gallery and probe set sequences of a person for these three experiments were collected with the same background and about the same time. This is particularly true for experiment A (viewpoint) whose gallery and probe sets contain essentially the same temporal event for each person, but taken from two different viewpoints. Thus, there are correlations in the shadows of a subject between the gallery and probe sequence, which possibly contributed a higher rate with the unreconstructed silhouettes. To shed some light on this effect, we considered just the pixels that were edited during the

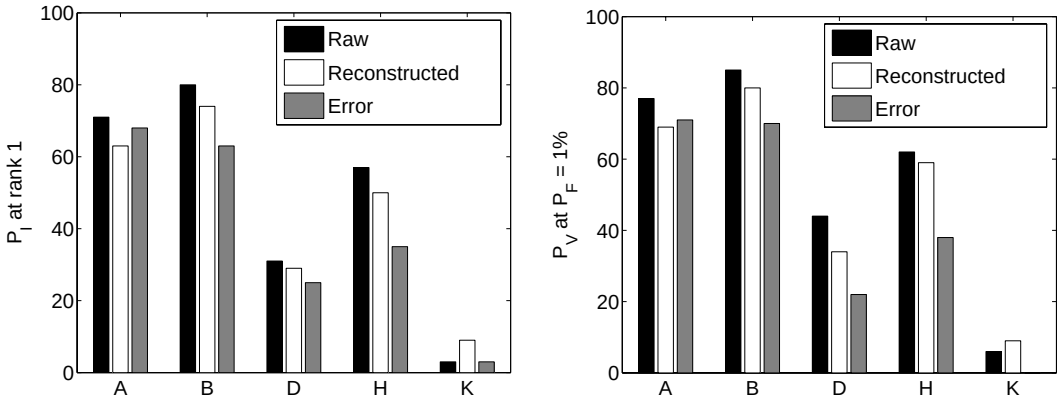


Figure 4.15. Identification Rate (P_I) at Rank 1 and Verification Rate (P_V) at 1% False Alarm Rate(P_F) with Raw Silhouettes and After Reconstruction, and with Error Pixels Edited (Removed Or Added) During the Reconstruction Process Using the Baseline Algorithm. Results for the 5 Key Experiments are Listed: A (Viewpoint), B (Shoe), D (Surface), H (Carry), And K (Time).

reconstruction process, i.e. either added or removed. We refer to these as the error pixels. Recall that we have already established that the edited pixels are mostly error (either false positive prediction or missed detection) pixels (Figs. 4.8 and 4.9). We studied the recognition power from the error pixels. As shown by gray bars in Fig. 4.15, the recognition from these error pixels is quite high, especially when comparing sequences for a subject collected within a short time duration of each other. For experiments that compare sequences across 6 months, the error pixels do not have significant recognition.

To have a better understanding towards the effects of reconstruction, we try another algorithm that employs the pHMM to normalize gait dynamics and bases similarity on the Euclidean distances between corresponding stances. This algorithm is also better than many of the reported performances (Table 3.2), and we should describe it in detail in Section 5.2. Fig. 4.16 shows the identification and verification performance of the stance shape based algorithm on the 5 key experiments, with the raw and reconstructed silhouettes. We see that (i) the overall performance of the stance shape based algorithm is much better than the baseline algorithm. But more importantly, (ii) the performance with raw and reconstructed silhouettes is similar. We find that after removing false recognition sources

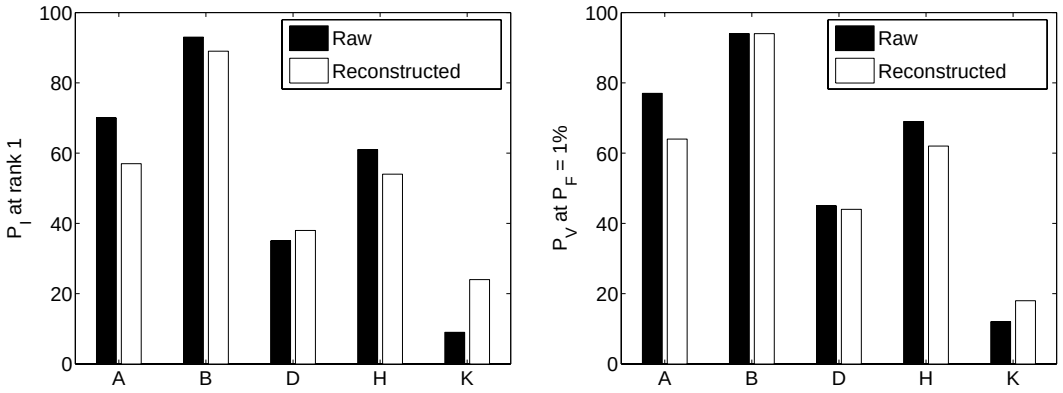


Figure 4.16. Identification Rate (P_I) at Rank 1 and Verification Rate (P_V) at 1% False Alarm Rate(P_F) with Raw Silhouettes and After Reconstruction using the Shaped Based Algorithm. Results for the 5 Key Experiments are Listed: A (Viewpoint), B (Shoe), D (Surface), H (Carry), and K (Time).

such as shadow pixels and missed detection, possibly due to interaction of clothing texture and background pattern, gait recognition under shoe, surface and view variations did not change, in fact they dropped a bit. The identification performance across time appears to be marginally better with reconstructed silhouettes, but the differences are not statistically significant given the small probe set size of 33, when compared with the probe set sizes of the other experiments.

4.5 Summary

In this chapter, we presented and evaluated a template-model based strategy for refining silhouettes in nearly fronto-parallel views. The model consists of an Eigen-Stance model that captures the shape variation of each stance, coupled with a pHMM of the gait dynamics. We empirically established that the quality of the reconstructed silhouettes was better in terms of false positive predictive values and detection rates. We offer this pHMM coupled Eigen-Stance model as a solution to the detection of silhouettes of walking humans, viewed fronto-parallel, within about 30^o view angle variation. Using the Georgia Tech sequences, we showed the model also works for entirely different datasets than used to construct the model.

Recently, Lee *et al.* [52] also presented a method (here referred to as the MIT-Hp method) for cleaning silhouettes that appears to be similar to that presented here. They also use HMM based time-syncing of sequences and cleanup using template models. However, there are several key differences resulting in demonstrable performance differences. First, our representation of the shape variation model at each HMM state using the *Eigen-Stance model* allows us to exploit the correlation among the silhouette pixels, whereas MIT-Hp uses an independent Bernoulli model for the silhouette pixel values. The use of a Bernoulli model is akin to using just the mean images of each stance in our model. Second, our training set consists of manually specified silhouettes, which enables us to *remove shadows*. The existence of shadows in MIT-Hp silhouettes might explain the enhanced performance for the experiments (A through C) on grass. Their performance for experiments where they compared silhouettes across surfaces did not improve to a large extent. Third, MIT-Hp used a two step process involving (i) a cleanup using a population based on mean image constructed by summing all frames for a set of persons, and (ii) further cleanup using a sequence specific HMM. On the other hand, we have an unified approach that uses a population based HMM model, coupled with population based stance shape models. The full use of population models lets us overcome many sequence specific segmentation artifacts such as holes due to strange background or foreground texture for a particular person. The power of the use of population models is also evident, to a limited extent, in the work of Lee *et al.*. The performance increase in gait recognition was mostly due to their use of the aggregate population model. The addition of sequence specific HMM did not seem to add to the recognition to a large extent. All of these key differences between the MIT-Hp method and those presented here have impact on actual silhouette quality: MIT-Hp silhouettes have more false positive prediction pixel than those in this paper. Fig. 4.17 shows the percentage improvement in pixel level detection ($\Delta P_D = 100(\frac{P_D(Here)}{P_D(MIT-Hp)} - 1)$) and false positive prediction between the silhouettes produced here and the HMM-Hp silhouettes ($\Delta P_{\overline{PPV}} = 100(\frac{P_{\overline{PPV}}(Here)}{P_{\overline{PPV}}(MIT-Hp)} - 1)$). We separately report the results for the grass and concrete sequences since they have different backgrounds. Improvement in silhouette

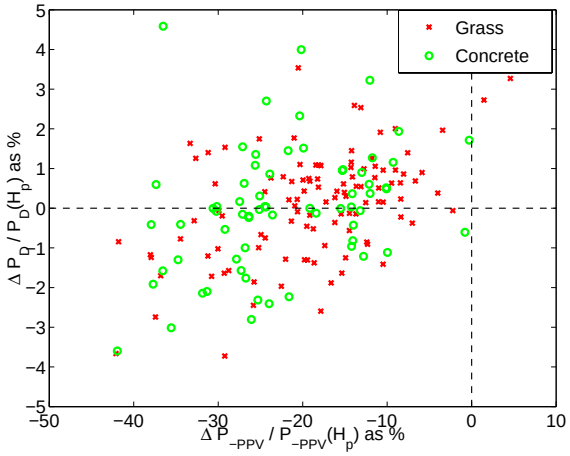


Figure 4.17. Scatter Plot of Percentage Improvement in Pixel Level Detection (ΔP_D) and False Predictive Values ($\Delta P_{\overline{PPV}}$) of the Silhouettes Produced Here and the MIT-Hp Silhouettes. The Green Circles Correspond to the Concrete Sequences and the Red Crosses Correspond to the Grass Sequences.

quality would be indicated by $\Delta P_{\overline{PPV}} < 0$ and ΔP_D is around 0, which is observed in the plots. The differences in false predictive values are statistically significant (P-value ≤ 0.05, established using paired-t tests) for both concrete and grass sequences. The detection rate, on the other hand, are of the same quality for both the grass sequences and the concrete sequences (P-value ≥ 0.05).

In the context of gait recognition, we have established that the low performance under the impact of surface and time variation can not be explained by poor silhouette quality. We base our conclusions on two gait recognition algorithms. One exploits both shape and dynamics, while the other exploits just shape. The drop in performance due to surface condition that we observe in the gait challenge problem is *not* due to differences in background. This observation is also corroborated by the performances reported in a fairly recent work by the Lee *et al.* [52]. The observation has implication for future work direction in gait recognition. Instead of searching for better methods for silhouette detection to improve recognition, it would be more productive to study and isolate components of gait that do not change under shoe, surface, or time. One example of this type of study is [86] in which relationship between silhouette shape and speed was studied and then was compensated

for by transforming the silhouettes. While it is doubtful whether speed variations can fully explain the drop in performance due to surface or time change, systematic studies such as this would be needed to understand the limitation of gait recognition.

CHAPTER 5

INVESTIGATION OF GAIT ALGORITHMS TO IMPROVE RECOGNITION PERFORMANCE

So far, we have studied gait recognition and established the hard problems. We have also illustrated that their impacts can not be explained by the low level representation: silhouette quality. Instead, it is the fundamental gait changes that should be looked at. In this chapter, we investigate approaches to improve recognition performance. Toward this end, three algorithms are proposed: (i) an averaged silhouette based algorithm that deemphasizes gait dynamics; (ii) an algorithm that normalizes gait dynamics by a Population Hidden Markov Model (pHMM), and computes similarity based on Euclidean distance with stance selection; and (iii) an algorithm that also normalizes gait dynamics using pHMM but computes similarity in shape space based on the Linear Discriminant Analysis (LDA), which suppresses within-subject variations affecting recognition, and morphological deformations which removes the within-subject body width differences.

We show that the first algorithm dramatically reduces computation time but achieves similar recognition power as the baseline algorithm. The second algorithm increases the performance of surface and time changes after removing gait dynamics. For the third algorithm, we present results on three different, publicly available, datasets. First, we consider the gait challenge dataset, which is largest gait benchmarking dataset that is available (122 subjects), exercising five different factors, i.e. viewpoint, shoe, surface, carrying condition, and time. The algorithm significantly improves the performance across the hard experiments involving surface change and briefcase carrying conditions. Second, we also show improved performance on the UMD gait dataset that exercises time variations for 55 subjects. Third, on the CMU Mobo dataset, we show results for matching across

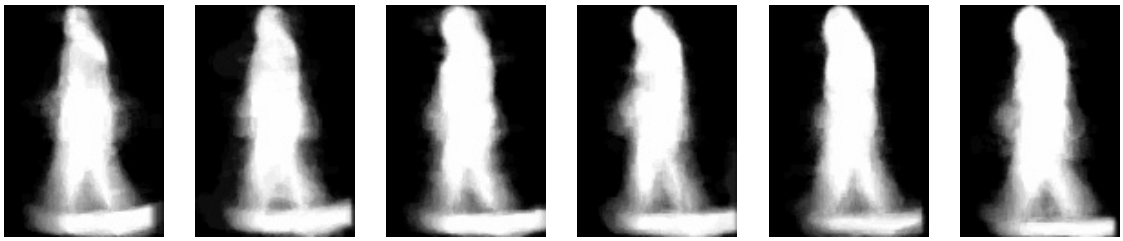


Figure 5.1. Examples of the Averaged Silhouettes of One Subject; Each Averaged over a Different Gait Cycle.

different walking speeds. It is worth noting that there was no separate training for the UMD and CMU datasets.

5.1 Averaged Silhouette Representation Based Algorithm

Our first algorithm is based on the gray level silhouettes averaged from one gait cycle, which is different from most other algorithms that employ traditional binary silhouettes.

There are 3 steps to produce the averaged silhouette representation. The first step is binary silhouette extraction. We use the same technique described in Section 3.1.1 employing the Mahalanobis distance from the background pixel statistics in each frame, and EM classification. The second step is to estimate the gait periodicity, N_{gait}. As mentioned in Section 3.1.2, we model the variation of pixel number in the lower half body in a sequence. The third step is average silhouette computation. Given a sequence of silhouettes, $\mathbf{I} = \{\mathbf{I}(1), \cdots, \mathbf{I}(M)\}$, we partition it into subsequences of gait period length, denoted by $\mathbf{I_{Pk}} = \{\mathbf{I}(k), \cdots, \mathbf{I}(k+N_{Gait})\}$. For each subsequence we average the silhouettes to arrive at a set of average silhouettes, $\mathbf{AI}(i), i = 1, \cdots, \lfloor \frac{M}{N_{Gait}} \rfloor$.

$$\mathbf{AI}(i) = \frac{1}{N_{Gait}} \sum_{k=iN_{Gait}}^{(i+1)N_{Gait}-1} \mathbf{I}(k) \tag{5.1}$$

Fig. 5.1 shows examples of the average silhouette representation for a sequence. Note that this representation implicitly captures the shape of the template and, to a lesser extent, the

temporal dynamics of gait. The time spent at each stance shows up indirectly as intensity in the average silhouette representation.

For gait recognition, we need to compute the similarity between a given probe sequence and a stored gallery sequence. Let the average silhouettes from a probe and a gallery be denoted by $\{\mathbf{AI_P}(i)|i = 1, \cdots, N_P\}$ and $\{\mathbf{AI_G}(j)|j = 1, \cdots, N_G\}$, respectively. The similarity is defined as the negative of the median of the Euclidean distance between the averaged silhouettes from the probe and the gallery.

$$\mathrm{Sim}(\mathbf{AI_P}, \mathbf{AI_G}) = -\mathrm{Median}_{i=1}^{N_p} \left(\min_{j=1}^{N_G} ||\mathbf{AI_P}(i) - \mathbf{AI_G}(j)|| \right) \tag{5.2}$$

With regard to the recognition performance, we use the HumanID database to demonstrate the efficacy of the proposed representation. Similar to the baseline algorithm, we list performance in terms of identification rate (CMCs) and verification rate (ROCs). And we only report the 5 key experiments, which exhibit the impact of individual covariate.

In Fig. 5.2, we plot CMCs for the first 20 ranks and ROCs up to 20% false alarm for the gait baseline algorithm, which uses spatio-temporal correlation of the silhouette, and recognition based on the averaged silhouette representation proposed here. We see that performance on three of the experiments, i.e. A (view), B (shoe), and H (carry), is better with averaged silhouettes. There is some fall in performance for the other two experiments exercising surface (D) and time (K). However, McNemar's test shows that the rank 1 identification rates are not statistically significant (P-value > 0.05). On the other hand, the covariate of surface and time also exhibit strong impact on this algorithm, where we see about 40% drops in the top rank recognition rate.

One contribution of this algorithm is its efficiency on computation time. On a 800 MHz SunFire server it took 4.63s on average to compare two sequences by spatio-temporal correlation as compared to 0.14s on average to compare similarity using the average silhouette; a 30 times improvement in time.

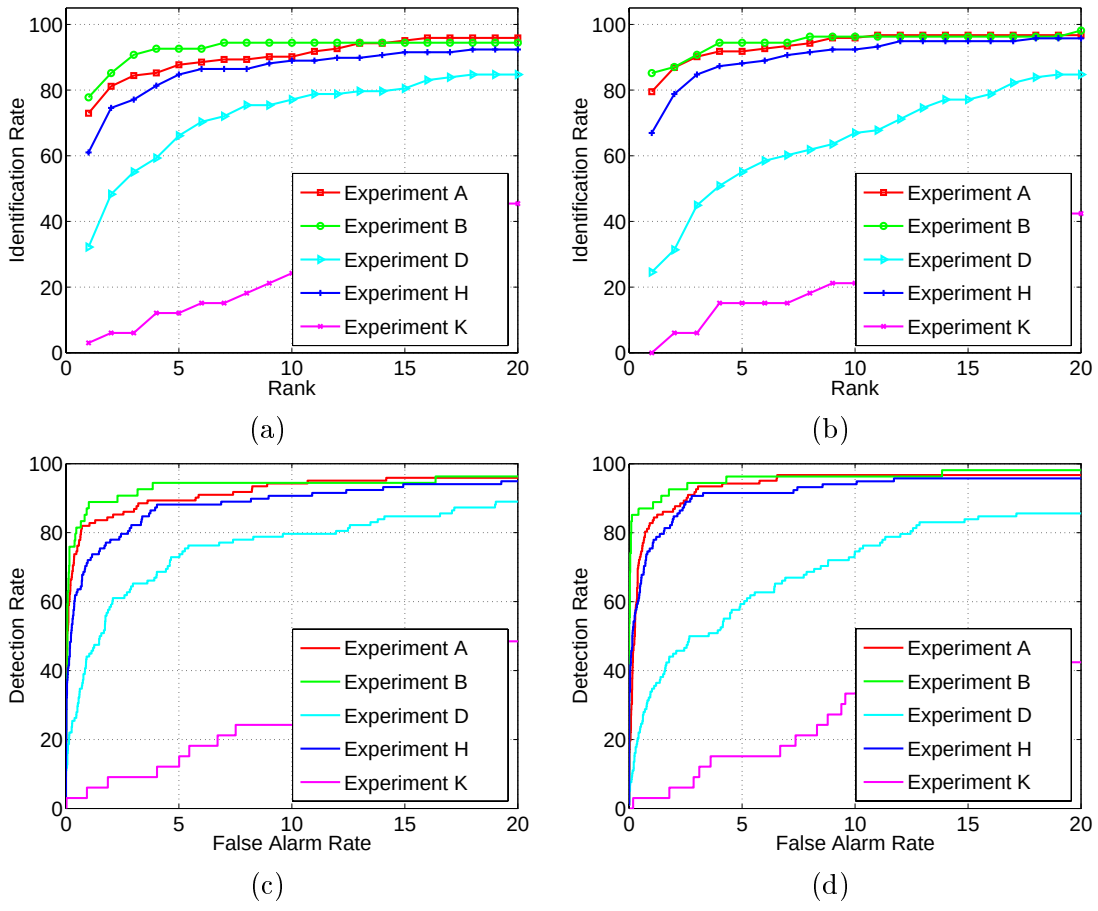


Figure 5.2. Performance on 5 Key Experiments from the USF/NIST HumanID Database in terms of CMCs ((a) and (b)) and ROCs ((c) and (d)) with a Gallery Set of 122 Subjects. the Left Column is the Result Of the Baseline Algorithm Using Individual Silhouette Frames, and the Right Column is the Result of the Averaged Silhouettes.

5.2 Dynamics Normalization with Euclidean Distance and Stance Selection Algorithm

The second gait recognition algorithm is designed to exploit body shape matching. The idea follows from the insights of recent gait recognition works [90, 15] that show that silhouette shape, which includes body shape and gait stance shape, has equal, if not more, recognition power than gait dynamics.

We present an *averaged gait cycle representation* [57]. Each silhouette sequence, typically consisting of multiple gait cycles, is first aligned to form one dynamics-normalized, averaged gait cycle, over a fixed number of stances. This normalization is accomplished by a *Population Hidden Markov Model (pHMM)* based on a subject population. Note this representation averages *means of frames in a same state of multiple cycles*, which is different from the one in the previous section that averages *all frames in different states within one cycle*. With the stances aligned by the pHMM, temporal correlation is no longer needed during similarity computation. Instead, the similarity can be computed by simply comparing images in the corresponding stances, where we choose the Euclidean distance measurement.

5.2.1 Population Hidden Markov Model (pHMM)

In Section 4.2.2 we have described the population Hidden Markov Model in detail. Here let's concisely review it. Like traditional HMM, the Population Hidden Markov Model (pHMM) is specified by the possible states, $q_t \in \{1, \cdots, N_s\}$, which represent gait stances, and the triple parameters $\lambda = (A, B, \pi)$, represent the state transition matrix, observation model, and priors, respectively. The model is built over one gait cycle, which is partitioned into 20 states, based on the Akaike Information Criterion (AIC) [3]. Due to the gait nature, we adopt the left-right cyclic Bakis model, that is, each state can either go to next state or stay unchanged, while the last state can go back to the first state. The model is trained on a set of manually created silhouettes for a set of subjects using Baum-Welch algorithm. Each gait cycle is chosen to begin at the right heel strike phase of the walking cycle through

to the next right heel strike. Fig. 4.2 shows examples of these manual silhouettes. We also vertically normalize and horizontally align these silhouettes (see the third row of Fig. 4.2) to reduce the effects of distances variation of subject to camera.

5.2.2 Dynamics Normalized Gait Cycle

After building the pHMM, we normalize the dynamics for any given gait sequence, $\mathbf{I} = \{\mathbf{f_1}, \cdots, \mathbf{f_N}\}$, by first estimating the stance state for each frame and then averaging the frames mapped to each state to arrive one, dynamics-normalized, gait cycle over N_s frames, denoted by $\mathbf{I_{DN}} = \{\mathbf{g_1}, \cdots, \mathbf{g_{N_s}}\}$. The dynamics normalized gait cycle is computed by averaging frames mapped to the same state. We refer to this averaged representation for each stance, $\mathbf{g_i}$, as the *stance-frame*. The stance estimation of each frame is based on the dynamic programming based Viterbi algorithm [71], which returns the most likely state assignment for each frame. To reduce the combinatorics of this assignment process, we partition an input sequence into subsequences of roughly one gait cycle length, which we can easily estimate from the periodic variation in the number of foreground pixels in the bottom half of the silhouettes. Note that the subsequences can start from any stance because of the cyclical nature of the HMM model.

Fig. 5.3 shows some stance-frames for one subject under different conditions. Notice that the stance-frames for the same stance are similar across different sequences, which indicates that silhouette-to-stance matching is correctly estimated by the Viterbi algorithm.

5.2.3 Similarity Computation

Given two averaged gait cycles, the similarity computation process does not have to align the cycles. The corresponding stances can be simply compared and the results summed to arrive at an overall similarity score. However, we consider the distances only for a subset of pre-selected stances that emphasize the differences between subjects. To select the discriminatory stances, we consider the variation in shape for each stance as reflected in the first and the second eigenvalues associated with the corresponding Eigen-stance model.

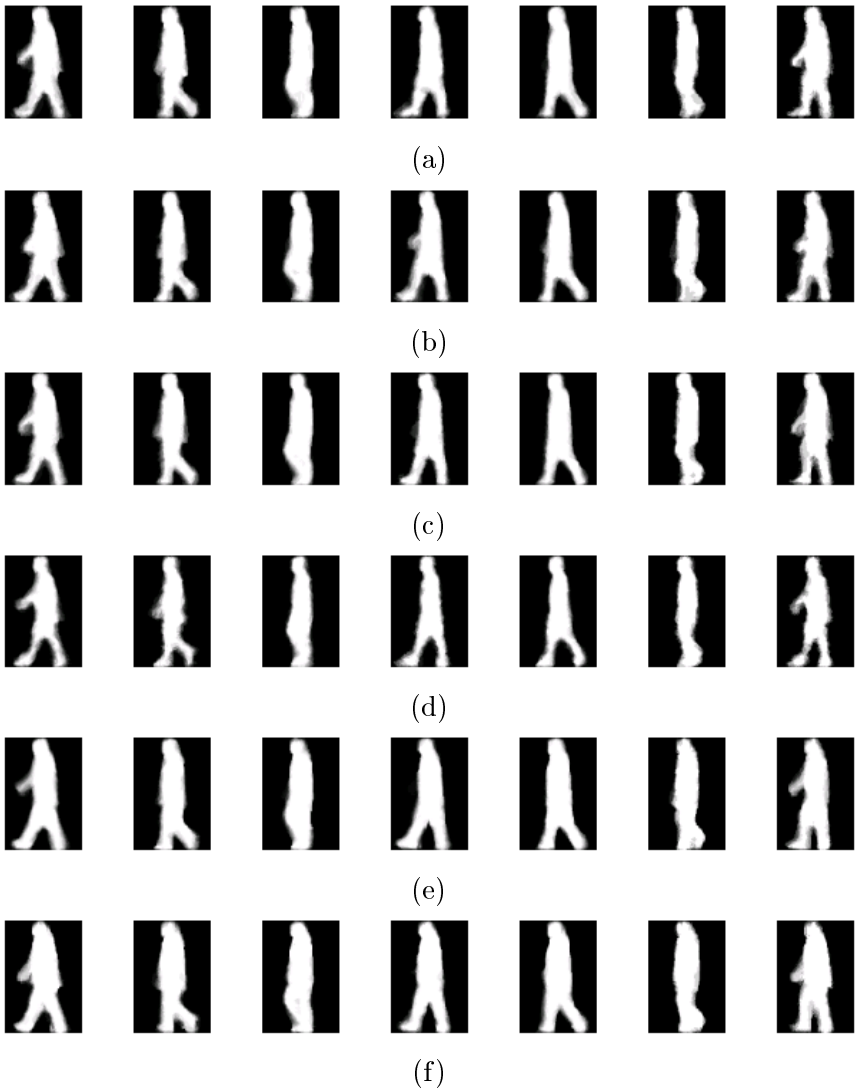


Figure 5.3. Example of Dynamics-Normalized Stance-Frames from One Subject in the (a) Gallery, and the Corresponding Stances in the Probes Corresponding to Changes in (b) View, (c) Shoe-Type, (d) Surface, (e) Carrying Condition, and (f) Time (Six Months).

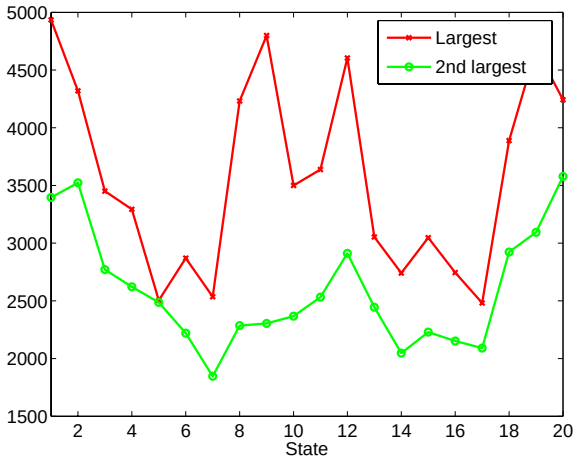
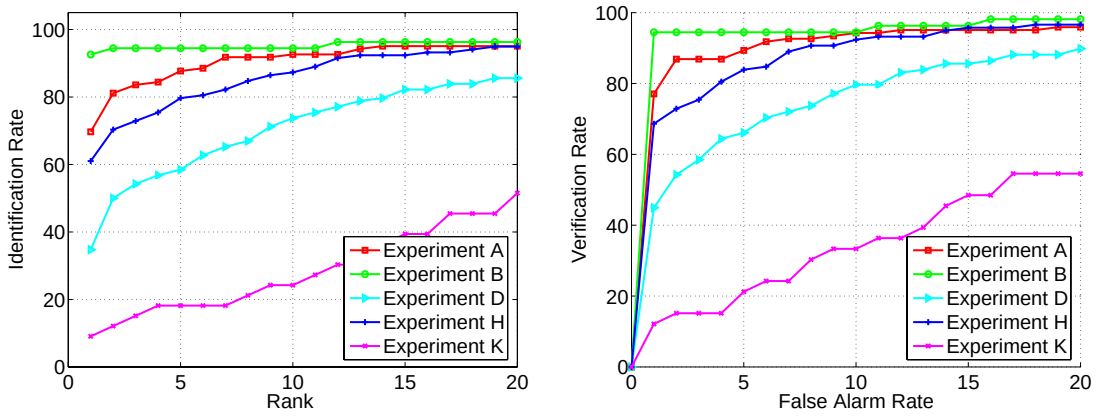


Figure 5.4. The Variation of the Largest and Second Largest Eigenvalues Associated with Each Stance Shape, as Computed in the Eigen-Stance Model.

These are plotted in Fig. 5.4. We see that states at the ends (states 1 to 3 and 18 to 20) and at the middle (9 to 12) have the largest scatters, indicating that these gait stances carry the bulk of the discriminatory power. These correspond to states near the full stride stances. Let us denote this subset of salient discriminatory states by $\mathbf{S_d}$. To arrive at one similarity score, we compute Euclidean distances between the averaged representation for these stances from the probe sequence I_{P_i} and the gallery sequence I_{G_j}.

$$S(I_{P_i}, I_{G_j}) = -\sum_{k \in \mathbf{S_d}} \Big(I_{P_i}(f_k) - I_{G_j}(f_k) \Big)^2 \tag{5.3}$$

As to recognition performance, we also test the algorithm on the 5 key experiments of USF HumanID database. In Fig. 5.5 we plot the corresponding CMCs and ROCs. It indicates that (i) the shape based algorithm slightly outperforms the baseline algorithm for some covariates, e.g., the shoe-type (experiment B), surface (experiment D), and time (experiment K); (ii) more importantly, the dramatic drops on the hard problem are also observed, which is consistent with the results of all other algorithms.



(a) (b)

Figure 5.5. Shape Based Recognition Performances for the 5 Key Experiments of USF/NIST HumanID Database (a) CMC Curves and (b) ROCs Plotted upto a False Alarm Rate of 20%.

5.3 Dynamics Normalization with LDA and Morphological Deformation

Fig. 5.6 shows the flowchart of the third algorithm on computing similarity of two sequences. The different algorithmic modules are shown, along with example inputs and intermediate representations. The inputs consist of silhouette sequences, which can be extracted from raw sequences in a number of ways. We compute the silhouettes using the eigenstance reconstruction model [56, 59], which linearly projects each frame into the eigenstance space corresponding to the mapped state and then reconstructs it. This was shown to significantly reduce the effect of shadows and other segmentation errors.

Similar to the algorithm in Section 5.2, the new algorithm first aligns each silhouette sequence to form one dynamics-normalized, averaged gait cycle, over a fixed number of stances by pHMM. However, in the similarity computation stage, to improve performance, we no longer use Euclidean distance. Instead, we employ the Fisher's Linear Discriminant Analysis (LDA) to suppress the variations due to external conditions, such as surface, shoe, carrying, clothing, and time that effect recognition. To impart some amount of invariance of the similarity value with respect to erosion or dilations of the underlying shape, we

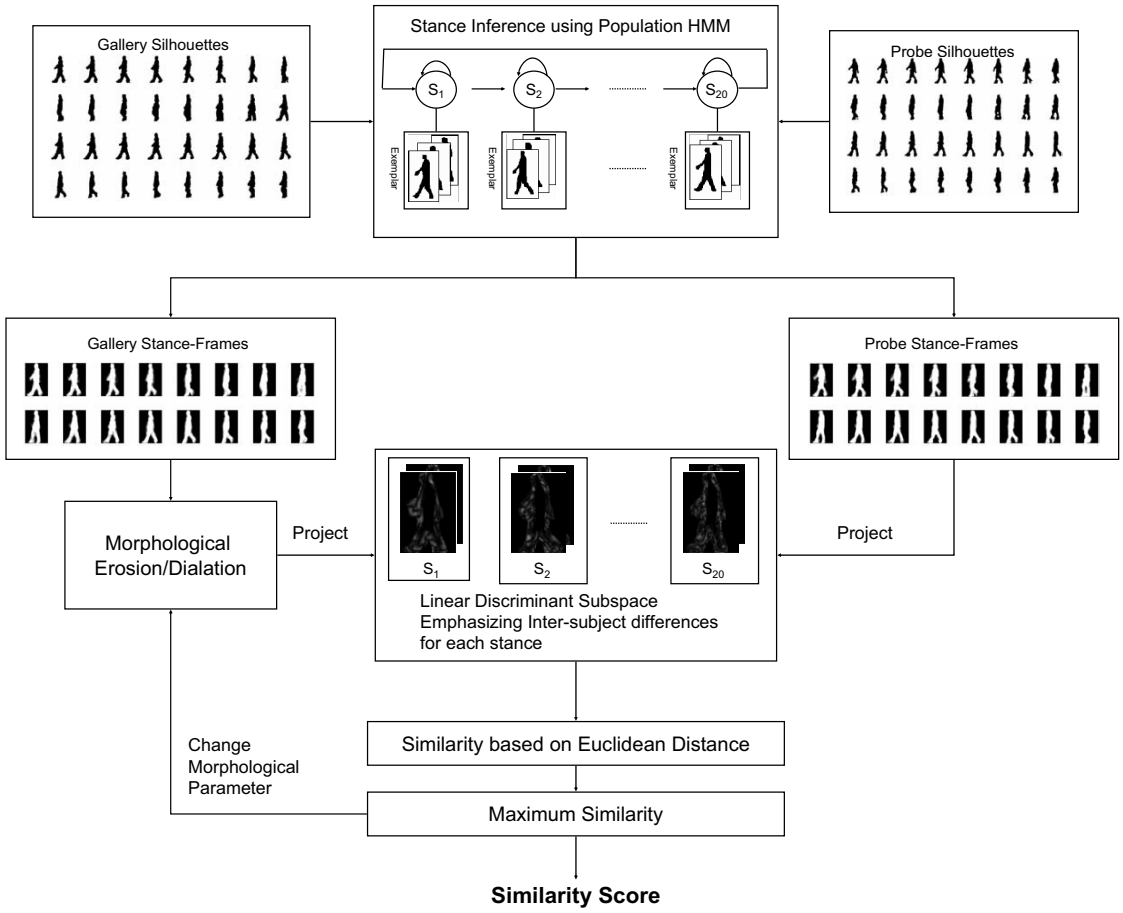


Figure 5.6. The Flowchart of the Gait Recognition Algorithm Based on Gait-Dynamics Normalization.

embed the similarity computation in a maximization loop. This helps to handle output variances of the underlying low-level silhouette detection processes.

5.3.1 Linear Discriminant Analysis (LDA)

A dynamics normalized gait cycle consists of a fixed number of stance frames, which simplifies the similarity computation between two given sequences. A separate alignment process is not needed. We can simply consider the distances between the corresponding stance-frames. Instead of simple Euclidean distances between stance-frames, we compute distances in the Linear Discriminant Analysis (LDA) Space, designed to maximize the differences between frames from different subjects and to minimize the distances between frames from the same subject under different conditions. We used the PCA+LDA formulation that was advocated by Belhumeur *et al.* [5] so as to address the singularity issues that can arise in pure LDA. We present just the outline here.

For *each* of the N_s stances, we construct a linear discriminant space as follows. The set of individuals form the classes, $(I_1, I_2, \cdots, I_c)$. For each individual, I_k, the stance-frames, $\mathbf{g_s^i}$ (s is the stance index), under various different conditions form the samples for that class. The between-class scatter matrix for the s-the stance is

$$S_B = \sum_{i=1}^{c} N_i(\mu_\mathbf{i} - \mu)(\mu_\mathbf{i} - \mu)^T \tag{5.4}$$

the within-class scatter matrix is

$$S_W = \sum_{i=1}^{c} \sum_{\mathbf{g_s^i} \in I_i} (\mathbf{g_s^i} - \mu_\mathbf{i})(\mathbf{g_s^i} - \mu_\mathbf{i})^T \tag{5.5}$$

and the total scatter matrix is

$$S_T = \sum_{i=1}^{N} (\mathbf{g_s^i} - \mu)(\mathbf{g_s^i} - \mu)^T \tag{5.6}$$

where $\mu_{\mathbf{i}}$ is the mean vector of class i, μ is the mean vector of all samples, and N is the total sample number. If S_W is non-singular, then the optimal discriminating space$\mathbf{V}_{opt}$ for classification can be simply computed as

$$\mathbf{V}_{opt} = arg \max_{\mathbf{V}} \frac{\mid \mathbf{V}^T S_B \mathbf{V} \mid}{\mid \mathbf{V}^T S_W \mathbf{V} \mid} \tag{5.7}$$

Specifically, $\mathbf{V}_{opt} = \{\mathbf{v_1}, \mathbf{v_2}, \cdots, \mathbf{v_m}\}$, the set of generalized eigenvectors of S_B and S_W corresponding to the m largest eigenvalues $(\lambda_1, \lambda_2, \cdots, \lambda_m)$, i.e.,

$$S_B \mathbf{v_i} = \lambda_i S_W \mathbf{v_i}, \qquad i = 1, 2, \cdots, m$$

However, when there are more than one class in the training set, S_W is always singular because its rank is at most $N - c$. One solutions is to project the within-class and between-class scatter matrices into a lower dimension space so that the resulting S_W is non-singular. The PCA (Principal Component Analysis) can be employed to reduce the dimension [5].

$$\mathbf{V'}_{opt} = \mathbf{V}_{PCA} \mathbf{V}_{LDA} \tag{5.8}$$

where

$$\mathbf{V}_{PCA} = arg \max_{\mathbf{V}} \mid \mathbf{V}^T S_T \mathbf{V} \mid$$

$$\mathbf{V}_{LDA} = arg \max_{\mathbf{V}} \frac{\mid \mathbf{V}^T \mathbf{V}_{PCA}^T S_B \mathbf{V}_{PCA} \mathbf{V} \mid}{\mid \mathbf{V}^T \mathbf{V}_{PCA}^T S_W \mathbf{V}_{PCA} \mathbf{V} \mid}$$

and $\mathbf{V}_{PCA}$ should keep no more than the largest $N - c$ principal components so that the corresponding S_W is non-singular.

For each stance, s, in the dynamics-normalized gait representation, we create $\mathbf{V^s}_{PCA}$ to model only 90% energy in corresponding total scatter matrix S_T. We have found the number of eigenvectors thus needed is much less than $N - c$. The subsequent $\mathbf{V^s}_{LDA}$ space consist of $c - 1$ non-zero generalized eigenvectors of within subjects and between subjects scatter matrix for the k-th stance. Given two dynamics normalized sequences, $\mathbf{I^a_{DN}}$ and $\mathbf{I^b_{DN}}$, we compute the distance by first projecting each stance-frame, $\mathbf{g^a_k}$ into the

corresponding $\mathbf{V^s}_{LDA}$ space. This negated sum of the Euclidean distances in these LDA stance spaces is a similarity measure, $S(\mathbf{I^a_{DN}}, \mathbf{I^b_{DN}})$. More specifically,

$$\begin{aligned} S(\mathbf{I^a_{DN}}, \mathbf{I^b_{DN}}) &= -\sum_{k=1}^{N_s} ||\mathbf{g_k^{a}}^T \mathbf{V^s}_{LDA} - \mathbf{g_k^{b}}^T \mathbf{V^s}_{LDA}|| \\ &= -\sum_{k=1}^{N_s} (\mathbf{g_k^a} - \mathbf{g_k^b})^T \mathbf{V^s}_{LDA} (\mathbf{V^s}_{LDA})^T (\mathbf{g_k^a} - \mathbf{g_k^b}) \end{aligned} \tag{5.9}$$

5.3.2 Similarity under Silhouette Deformations

In Fig. 5.3 we noticed that the "width" of the stance-frame for the same person varies across different conditions. For instance, there were changes associated with change in surface and time. This type of variation arises because of variabilities of low-level silhouette detection processes, induced by changes in the background statistics. These are hard to completely eliminate. Accepting this constraint, we modify our similarity computation to be some what robust with respect to changes in overall silhouette "widths". The new similarity, S_{WN}, is the maximum possible similarity over possible morphological deformations of the stack-frames in one of the sequences. The morphological deformations of erosion and dilation model the possible variations in widths. Specifically,

$$\begin{aligned} &S_{WN}(\mathbf{F^a_{DN}}, \mathbf{F^b_{DN}}) = \\ &\sum_{k=1}^{N_s} arg\max_{m \in -M, \cdots, M} (Mor(m, \mathbf{g_k^a}) - \mathbf{g_k^b})^T \mathbf{V^s}_{LDA} (\mathbf{V^s}_{LDA})^T (Mor(m, \mathbf{g_k^a}) - \mathbf{g_k^b}) \end{aligned} \tag{5.10}$$

where

$$Mor(m, \mathbf{g_s^i}) = \begin{cases} Dilate(m, \mathbf{g_s^i}) & \text{if} \quad m \geq 0 \\ Erode(m, \mathbf{g_s^i}) & \text{if} \quad m < 0 \end{cases} \tag{5.11}$$

The *Erode* and *Dilate* are gray-level morphological operations, as implemented in Matlab, employing structured object decomposition and bit packing [9].

In Fig. 5.7 shows the morphologically processed stance-frames in the *gallery*, corresponding to the probes in Fig. 5.3, that maximized the overall similarity. We see that eroded forms of some gallery stance-frames, e.g. in 1st and 4th columns, are more similar to the surface-probe shown in Fig. 5.3 than the original gallery stance-frames. Another

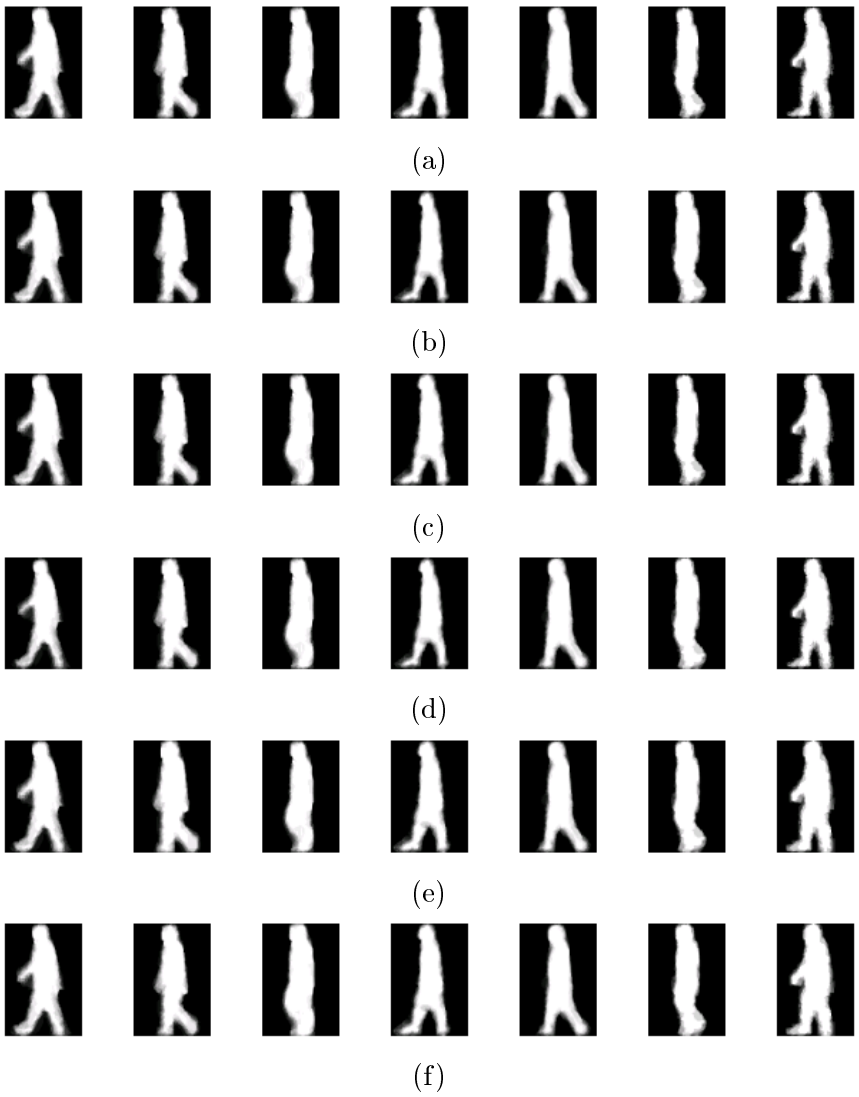


Figure 5.7. Examples of the Average Stances in the Gallery Set (a) Before and After Morphological Operation for Best Shape Match to (b) Probe-View, (c) Probe-Shoe, (d) Probe-Surface, (e) Probe-Briefcase, and (f) Probe-Time in Fig. 5.3.

example is the stance-frames shown in the the 3rd, 6th and 7th columns for the time-probe. This is consistent with our initial observation that the silhouettes in the surface-probe appear to be thinner than the gallery while those in time-probe appear to be thicker. The new deformation invariant similarity measure helps us handle such cases.

5.3.3 Experiments and Analysis

We present results on three different publicly available datasets: the HumanID gait challenge dataset, the UMD outdoor dataset, and the CMU indoor MoBo dataset, which are described in Section 2.2. We study the ability of the proposed algorithm to improve performance for matching across surfaces, time, carrying condition, and walking speed variations. We also present results for varying degrees of separation between training and test sets demonstrating generalizability across individuals, data collection sites, and camera configurations.

5.3.3.1 Training and Test Sets

All training was done using chosen *subsets* of the HumanID Gait Challenge data. The pHMM was trained using the *manual* silhouettes over *one* gait cycle from 71 subjects in the HumanID gait challenge data. Specifically, we choose 71 manual silhouette sequences corresponding to subjects walking on grass, viewed from the right camera, in the May collection. As we will see later, the data corresponding to these manual silhouette form the *gallery* of the gait experiments defined for the gait challenge dataset. None of the data from the probes were used for training the pHMM. In terms of experimental protocols this offers us acceptable separation of train and test conditions for the experiments on the gait challenge dataset. In a biometrics application, the gallery set represents the watch-list and is pre-defined; gallery is akin to the concept of a model-base in object recognition. For experiments with the UMD and CMU datasets, since we do not re-train the pHMM, there is *complete* separation of train and test.

To create the linear discriminant stance spaces, we also need a training set, comprising of stance-frame samples from different subjects under different conditions. For this, we used *subsets* of the HumanID gait challenge data to construct two different training data sets to allow us to experiment with different level of differences between train and test sets. The first training set consists of automated silhouette data from 33 subjects, with 16 sequences per subject, corresponding to the various combinations of changes in the two-possible values for four covariate: view-point, surface-type, carry-condition, and time. The second training set consists of automated silhouette data from 51 subjects collected in November, with no overlap with the May subjects. For each subject, we had 8 sequences corresponding to the various combinations of two-values of view point, surface, and carrying conditions. We did not include shoe-variation in the training sets as its inclusion reduced the number of common subjects for each combination of conditions. This is not of much concern because, as reported results on the Gait Challenge problem indicate, the impact of shoe on gait recognition is the lowest [75]. Both the training sets were further restricted to just the *front* portion of the full elliptical sequence. A sample frame is shown in Fig. 2.2(c).

As we will see in the next sections, we used test sequences with varying degrees from the training set. The first set consists of sequences from 122 subjects, including the 33 training subjects in the first training set, but using the *back* elliptical portions of the trajectories. The second test consists of sequences from the Gait Challenge dataset collected in May and does not include *any* subjects from the second training set. The third and fourth test datasets correspond to the UMD and CMU datasets, respectively. They are not only for different subjects, but were collected at different sites, with different viewing geometry and cameras than the training set.

5.3.3.2 Gait Challenge Problems: View, Shoe, Surface, Carry, Time

Fig. 5.8 shows the identification performance using CMCs (up to rank 5) and ROCs (up to 5% false alarm rate) for the 12 experiments. Fig. 5.9 compares the top rank identification performance achieved with those reported in the literature for the full gait challenge

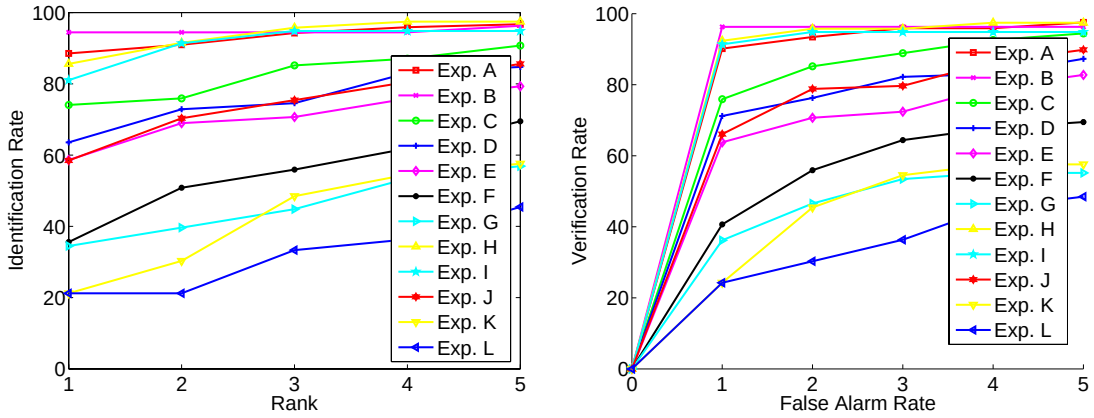


Figure 5.8. Performance of the Dynamics-Normalized LDA Gait Recognition Algorithm for the Twelve Experiments in the HumanID Challenge Problem (with 122 Subjects) (a) CMCs are Shown up to Rank 5 and (b) ROCs Plotted up to a False Alarm Rate of 5%.

problem. Specifically, we compare with the baseline algorithm that came with gait challenge problem [75], UMD's HMM based recognition strategy [90], and UCR's gait energy + learning based strategy [29]. We see that the new dynamics-normalized algorithm achieves the best performance in most experiments. It is slightly low for experiments A (view) and C (view+shoe). Of particular interest is the dramatic improvement for experiments that involve surface change (experiments D, E, F, G), and carrying condition change (experiments H, I, J). There is also some increase in performance for the hardest experiments involving 6-months time-difference (experiments K and L). (Note the UMD dataset-2 is perhaps the better dataset to consider the time-covariate, consisting of data from 55 subjects and time variations over a week. We report performance on this data set in a later section.)

We also experimented with the gait challenge dataset with complete separation of train and test sets in terms the subjects; no subject used for training were part of any probe. We used the second train set, discussed earlier, consisting of sequences from subjects in the November collection who were *not* in the May collection. The test set consisted of sequences from 71 subjects in the May collection and sequences from the repeat subjects in November.

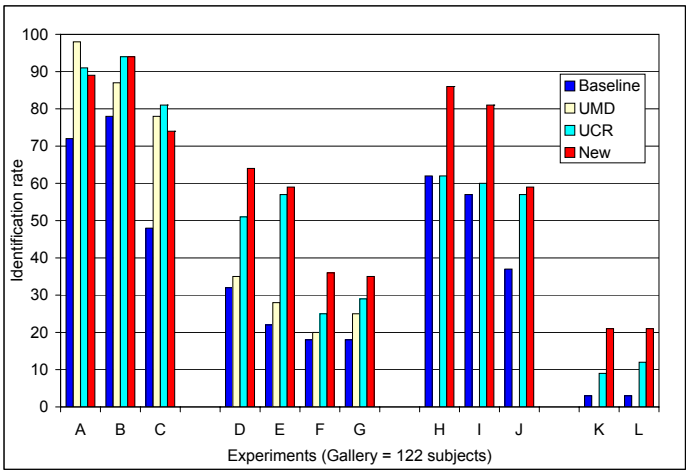


Figure 5.9. Top Rank Recognition Rate Comparison between the Dynamics-Normalized (New) Gait Recognition Algorithm with Results Reported by Other Algorithms: the Baseline Algorithm [75], UMD's HMM Based Algorithm [90], and UCR's Algorithm [29], for the Full HumanID dataset (122 Subjects).

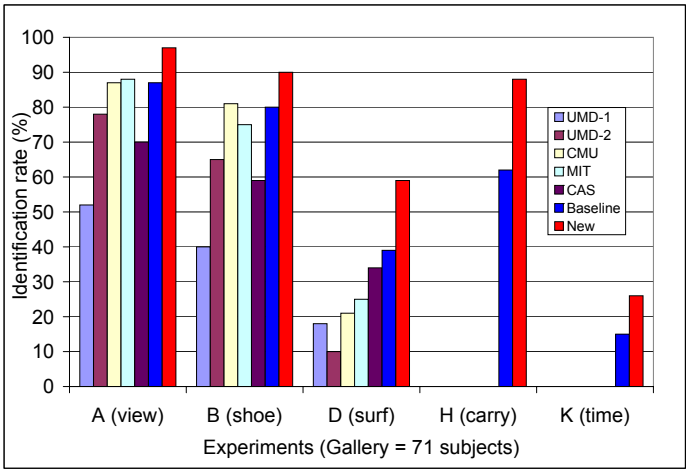


Figure 5.10. Summary of the Top Rank Recognition for Experiments A (Viewpoint), B (Shoe-Type), D (Surface), H (Carry), and K (Time) For the First Release Of HumanID Gait Challenge Dataset (71 Subjects in May Collection). The Algorithms are Based on: Fusion of Width Vectors (UMD-1) [16], DTW (UMD) [44], Silhouette Shape Clustering (CMU) [89], HMM (MIT) [52], Body Shape (CAS) [94], and Baseline (USF) [75, 43]. The Performance for the New Algorithm are Using a Training Set over a Different Set of 51 Subjects from the November Collection.

Of course, the experiment specifications in Table 2.3 has to be reduced to include just the May sequences (M) and exclude all the November sequences with N_1 tag. Incidently, this corresponds to the experiments for the first release of the gait challenge problem [43], on which more groups have reported performance than for the full dataset [75]. Fig. 5.10 shows the performances for the 5 key experiments: A-view, B-shoe, D-surface, H-carrying, and K-time, based on the new algorithm, as well as, those reported by others. We see that dynamics-normalization significantly improves performance, even with complete separation of training and test sets in terms of subjects.

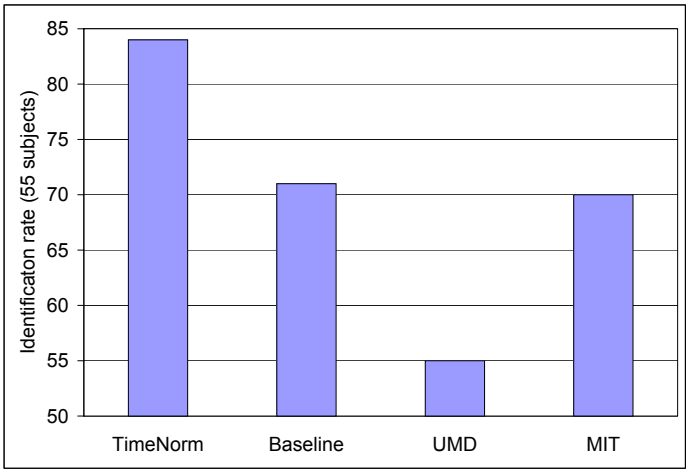


Figure 5.11. The Top Rank Identification Rates on the UMD Dataset (Experiment 1, 55 Subjects): the Dynamics-Normalized Gait Recognition Algorithm, Baseline Algorithm, UMD's HMM Based Algorithm [47], and MIT's Algorithm [51].

5.3.3.3 UMD Database: Time

The UMD dataset-2 offers us an opportunity to test gait recognition with short term (days) time differences for 55 subjects. Specifically, we use the UMD specifications of *experiment 1 for dataset-2*, which compares sequences taken on different days. For more detailed description of the dataset and the experiment specification, please refer to the website *http://degas.umiacs.umd.edu/Hid/data.html*.

There was no separate training of the gait recognition algorithm on this dataset. We use the version trained on the gait challenge data. Fig. 5.11 shows the top rank identification rate for the new dynamics-normalized gait recognition algorithm at 84% is a big improvement over the 71% rate of the baseline algorithm, 55% of UMD's HMM based algorithm [47], and 70% of MIT's algorithm [51]. This improvement can be attributed to the dynamic-normalization process that removed dynamics variabilities over time.

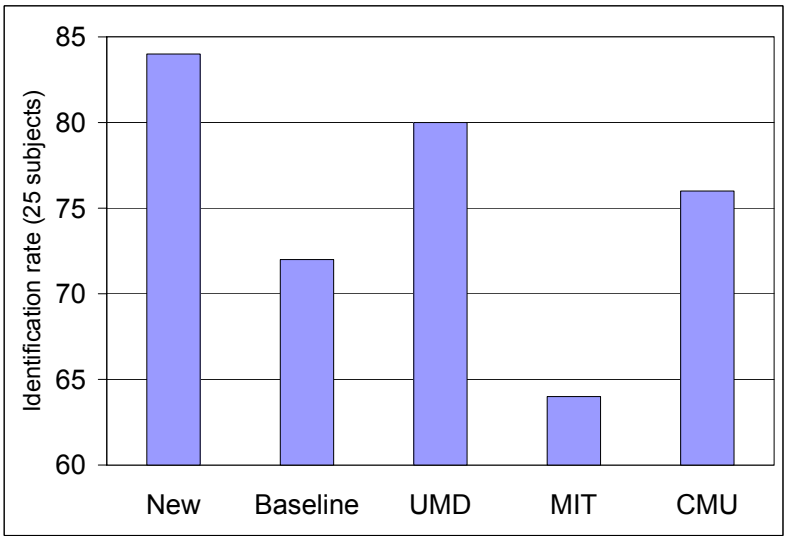


Figure 5.12. The Top Rank Identification Rate on the CMU Mobo Dataset (Experiment 3.1, 25 Subjects): the Dynamics Normalized Algorithm, Baseline Algorithm, UMD Algorithm [90], MIT Algorithm [44], and CMU Algorithm [15].

5.3.3.4 CMU MoBo Database: Speed

The CMU MoBo database, collected indoors on treadmill, supports the study of gait recognition variation with respect to walking speed. Specifically, we use experiment 3.1, defined by CMU (see Section 2.4), to test gait recognition across different speeds, viewed fronto-parallel. This is ideal for testing our dynamics-normalization scheme and benchmark it with respect to performances of other gait recognition approaches that do not normalize dynamics. Like the experiments with the UMD dataset, we *did not* re-train the dynamics-normalization model on this dataset. As Fig. 5.12 shows, the performance with dynamics-normalization is high, when compared with reported performances of four other algorithms: baseline [75, 43], UMD [90], MIT [44], and CMU [15].

5.4 Summary and Analysis

In this chapter, we presented three gait recognition algorithms. The first one employs the averaged silhouettes over one gait cycle representation. It significantly reduces computation time and achieve similar recognition power with the baseline algorithm. The second normalizes gait dynamics, using a population Hidden Markov Model (pHMM), so that a gait sequence is represented as an aligned average gait cycle, so that correlation is no longer necessary. Instead, we only need to compute distance between each stance, where the Euclidean distance is chosen. We demonstrated that this algorithm slightly improves the performance on surface and time. However, no statistically significant difference is reported. The third algorithm also normalizes gait dynamics using pHMM. However, in the similarity computation stage, it replaces the simple Euclidean distance with a Linear Discriminant Analysis based shape space, emphasizing differences in stance shapes between subjects and suppressing differences for the same subject under different conditions. The similarity computation in this space was designed to be robust with respect to "thickening" or "thinning" of silhouettes due to variations in low-level thresholds. Unlike other HMM based gait algorithms [52, 90] that use HMMs for *recognition* we do not use it for recognition, but rather for dynamics-normalization. Consequently, in contrast to other HMM-based gait recognition algorithms that build one HMM for every gallery sequence, we use *one* population HMM.

Based on extensive experimentation on multiple, publicly available databases (HumanID Gait Challenge, UMD, and CMU-Mobo), we can assert that dynamics-normalization greatly improves overall gait recognition performance, especially when comparing across surface, carrying condition, time, and different speed. The approach is also not dependent on the training set choice. It generalizes well not only across different subjects, but also across different datasets with varying imaging geometries. We attribute this significant improvement to gait dynamics-normalization. The other two components: the LDA stance shape space and the morphological operation based distance computation also improve performance, but the former has more impact than the latter. Fig. 5.13 shows the identifi-

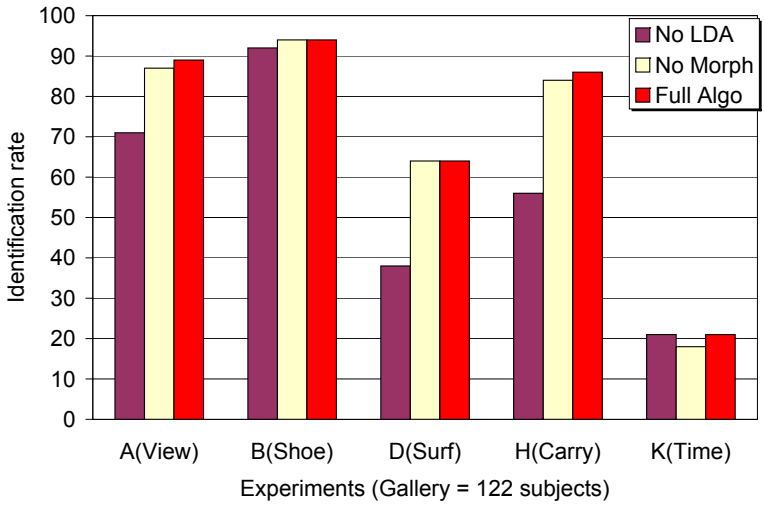
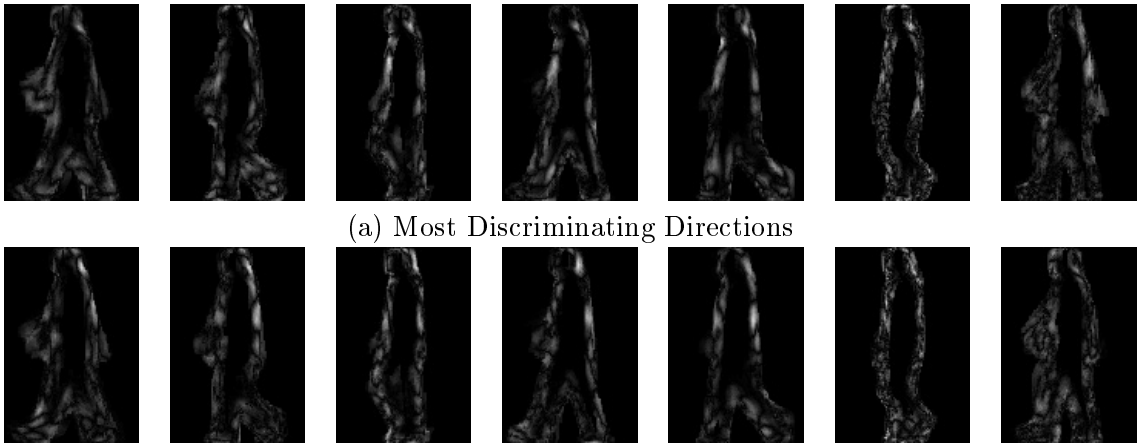


Figure 5.13. The Top Rank Identification Rate on the 5 Key Gait Challenge Experiments on the HumanID Gait Dataset of the Dynamics-Normalization Based Algorithm Based on (a) Euclidean Distances Between Stance Frames, instead of Distances in the LDA Stance Shape Space (No LDA), (b) without Accounting for Silhouette Deformation during Distance Computation (No Morph), and (c) with Both the Parts (Full Algo).

cation rates for the key experiments on the HumanID gait dataset with and without these components. We see that the LDA stance shape space has the most impact.

Efficacy of dynamics normalization suggests that body-stance shape plays a more important role than dynamics in gait recognition. This is also supported by good performance of gait recognition algorithms of Veeraraghavan *et al.* [90] and Collins *et al.* [15], which focus more on body shape than dynamics. To get some insight into the kinds of shape features that seem to be important we consider the top-two most inter-subject discriminating directions for each stance, as found by LDA of the silhouette shapes in the training set used for gait-normalization. Fig. 5.14 shows these directions as images for some of the stances, with brightness proportional to the absolute value of the corresponding eigenvector components. Bright pixels are the important ones. We see that (i) upper body, (ii) knee, and



(a) Most Discriminating Directions

(b) Second Most Discriminating Directions

Figure 5.14. The top Two Most Inter-Subject Discriminating Directions for Each Stance, as Found by LDA of Silhouette Shapes from 33 Subjects for that Stance. Intra-Subject Variations Span 16 Combinations of Variations in View, Carrying Condition, Time, and Surface.

(iii) lower leg during gait swing phases seem to be the important features that are being picked up. For further improvements in gait recognition it would be necessary to model and correct for the stance-shape changes under varying conditions. An excellent start is the work of Tanawongsuwan and Bobick [87] who studied the effect of walking speed on silhouette shapes.

CHAPTER 6

IMPROVING RECOGNITION BY COMBINING WITH FACE

In Chapter 5 we investigated methods to improve performance of gait algorithms. In this chapter we study improvement in human identification by combining gait with face. It has been demonstrated that combination or fusion of biometrics can offer a way to break the barrier of poor individual biometric performance. Lin *et al.* [33] demonstrate that multi-biometric integration does indeed result in a consistent performance improvement. Schiele [76] empirically showed that the more biometrics we combine, the better results we can get. One can talk about inter-modal combination [72, 95, 10, 78, 80, 49, 36, 34], e.g. combination of face with iris, and intra-modal combination [1, 100, 60, 102, 37, 69, 62, 103], e.g., combination of outputs of two biometrics on the same modality, or the combination of outputs of two different sensors, such as IR and visible [14, 13] and visible and 3D [11, 12, 13], on the same modality. In Table 6.1 we summarize the work in computer-vision based multi-modal biometric combination. Fusion can be done at the three levels [72]:

1. The feature extraction level, where data from each sensor are combined to form one feature vector [18, 10].

2. The matching score level, where the similarity scores computed by individual classifiers are fused [72, 60, 37, 36, 1, 95, 102, 78, 80, 49]. The scores from different biometrics are usually first transformed into the same range using linear transformation, polynomial transformation, or logarithm transformation. The normalized scores are then combined using rules such as sum, product, maximum, and minimum.

Table 6.1. Inter- and Intra-Modal Biometric Fusion.

Work	Comb. schemes	Face	Finger print	Hand	Iris	Ear	Gait	Speech
MSU [72]	Score	√	√	√				
MSU [60]	Score	√ √						
MSU [37]	Score		√ √					
MSU [36]	Score	√	√					√
MSU [34]	Decision	√	√					
MSU [69]	Decision		√ √					
U. Bern [1]	Score	√						
CAS & MSU [95]	Score	√			√			
UND & USF[10]	Score	√				√		
HK Polytechnic [102]	Score	√						
MIT [78, 80]	Score	√					√	
U. of Surrey [49]	Score & Decision	√						√
Rutgers [62]	Score & Decision	√ √						
UMD [48]	Score Decision	√					√	
UND [14, 11, 12, 13]	Decision	√ √						

3. The decision level where the each classifier makes its own classification and votes for the final decision [69, 34, 49, 62, 14, 11, 12, 13]. The popular vote rules include rank sum and majority vote.

In this chapter, we show that the combination of gait and face can effectively enhance the performance of outdoor biometrics at a distance. We demonstrate this for conditions that are known to be "hard" in face and gait recognition. Experiments also show that cross modal combination of gait and face is superior to the fusion within modality. Gait and face combination studies have been presented by others [80, 78, 48]. However, unlike previous studies that used either indoor data or outdoor data taken on the same day, resulting in high performance of the individual biometrics to begin with, our study involves *outdoor data, taken months apart*. We show that we can significantly improve recognition at a

distance in outdoor conditions and over time, both of which are hard conditions, using biometric fusion.

6.1 Recognition Algorithms

The primary focus of this paper is to investigate the power of the face and gait biometric fusion. So, the individual biometric algorithms we used are not necessarily the absolute best that are currently available, but they have performances that are close to the best available ones. They beat their corresponding established baseline algorithms by significant amounts.

6.1.1 Face Recognition Algorithm

We use the Gabor features based Elastic Bunch Graph Matching (EBGM) [98] algorithm for face recognition. It is a feature based method for face recognition that has superior performance to other template based methods, such as PCA, LDA, or Bayesian. We used the CSU implementation of the algorithms [1]. The approach first locates landmarks on a face, related to salient points on eyes, nose, and mouth, and then employs the frequency information of the local regions that surround the landmark locations as the landmark features (landmark jet). We did not re-train the algorithm. Instead we used the CSU trained version, which is based on 70 subjects. With regard to distance measurements, we choose the phase similarity, corrected by small displacements [8]:

$$D(J_i, J_i', \overrightarrow{d}) = \frac{\sum_{j=0}^{N_i} a_{ij} a_{ij}' cos(\phi_{ij} - (\phi_{ij}' + \overrightarrow{d}\overrightarrow{k_{ij}}))}{\sqrt{\sum_{j=0}^{N_i} a_{ij}^2 \sum_{j=0}^{N_i} a_{ij}'^2}} \tag{6.1}$$

where J_i and J_i' are the landmark jets of $i-th$ landmark point for graph J and J', N_i is the number of wavelet coefficients in the jet, a and ϕ are the magnitude and phase, $\overrightarrow{d}$ is the estimated displacement vector, and $\overrightarrow{k}$ is a vector pointing in the direction of the wave and having the magnitude equal to frequency of the wave. Obviously, the estimation

[1] It is available at `http://www.cs.colostate.edu/evalfacerec/`

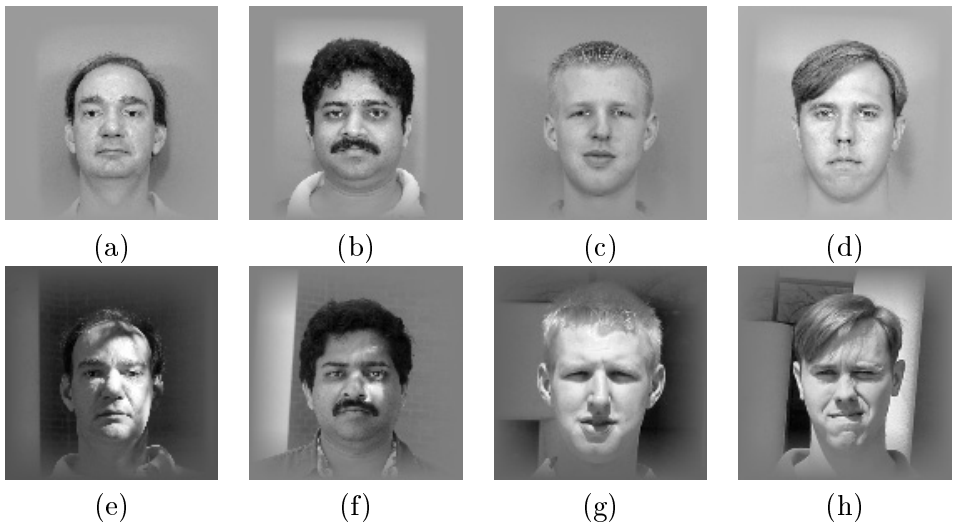


Figure 6.1. Samples of Computed Intermediate Representations Face Biometric that are Matched. (a)-(d) Gallery, (e)-(h) Probes.

of the displacement vector $\overrightarrow{k}$ is very important for Eq. 6.1. In this paper, we use the Displacement Estimation Narrowing Local Search (DENarrowingLocalSearch), which uses a local search method to find an optimum and empirically gives the best performance.

According to the FERET evaluations [67], the EBGM approaches provide the *best* recognition performance. Fig. 6.2 summarizes the reported top rank identification performance (with a gallery size of 1200) on three experiments involving matching (i) across indoor illumination variations, (ii) across 1 year time differences, and (iii) across more than 1 year time difference. EGBM had the top rank among five algorithms, for all the 3 experiments. It outperformed by around 20% the next best algorithm. Also, note the poor performance on datasets that involve comparison over time.

6.1.2 Gait Recognition Algorithm

We select the averaged gait cycle based algorithm as introduced in Section 5.2. Here let's concisely review the procedures: in the silhouette detection stage, it uses the background subtraction technique of the baseline algorithm (Section 3.1.1); then it employs the pHMM (Section 4.2.2) to decode a gait sequence into a set of gait states; and then it

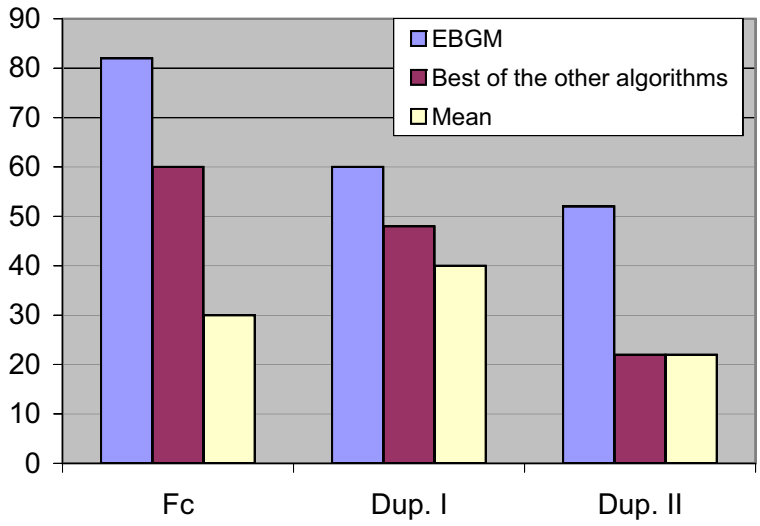


Figure 6.2. Top Rank Identification Performance (On a Gallery Set of 1200) of the EBGM and Four other Face Recognition Algorithms as Reported By FERET-2000 [67]. The Experiment *Fc* Matches across Illumination Variation, the *Dup. I* Experiment involves Temporal Difference within 1 year, and the *Dup. II* Experiment involves Temporal Difference more than 1 Year.

computes the mean image of all frames estimated to be at a same state; in the similarity computation stage, it simply computes the Euclidean distances between the mean images of the corresponding states; in addition, to boost up the performance, it only chooses stances with large scatters in the eigen-spaces. Note that the algorithm only employs the shape as recognition cue. And its performance is close to the best available ones, as Table. 6.2 indicates.

Table 6.2. The Top Rank Identification Rate for the Experiments of the USF HumanID Database involving the "Hard" Covariates of Surface and Time. The Gallery Size is 122 Subjects.

Covariates	Baseline	Algo1	Algo2	Algo3	Algo4	Averaged gait cycle
(Exp D) Surface	32	33	45	19	23	38
(Exp K) Time	3	15	24	3	6	24

6.2 Fusion Schemes

Before combination, scores from each classifier are transformed to a common range. Here we choose the Gaussian model based z-normalization, which was also used in FRVT-2002 [66]. For a given probe I_P, we compute its similarity value to each subjects in the gallery set $(I_{G_1}, I_{G_2}, \cdots, I_{G_{N_G}})$. Then we compute the mean value (μ_{I_P}) and standard deviation (σ_{I_P}) of the similarity values. The similarity value between each I_P and I_{G_j} is normalized as:

$$NormS(I_P, I_{G_j}) = \frac{S(I_P, I_{G_j}) - \mu_{I_P}}{\sigma_{I_P}} \tag{6.2}$$

This normalization not only maps the score onto a common scale, but also removes the dependencies of the scores on the particular probe. It is common in biometrics to observe that the non-match similarity scores are dependent on the chosen probe. This impacts the optimality of the single threshold decision rule chosen for verification in biometric systems.

We experimented with score level and decision level integration.

1. *Score Sum* combination strategy makes a decision simply based on the sum of the similarity scores from the corresponding gait and face scores:

$$CombS(I_P, I_{G_j}) = NormS_1(I_P, I_{G_j}) + NormS_2(I_P, I_{G_j}) \tag{6.3}$$

2. The second score fusion scheme is based on the *Bayesian decision* rule. For a given pair of probe and gallery subjects, the similarity values from the individual modalities form the observation vector, $\mathbf{v}$. The two classes correspond to the match (genuine, ω_m) and non-match (imposter, ω_{nm}) classes. The likelihoods of these two classes are modeled as multi-dimensional Gaussian distribution, which is usually a good choice empirically. Fig. 6.3 shows a 2D histogram representation of the gait and face non-match (imposter) scores.

$$P(\mathbf{v}|\omega_m) = \frac{1}{2\pi|\Sigma_m|^{1/2}} e^{[-\frac{1}{2}(\mathbf{v}-\mu_m)^T \Sigma_m^{-1}(\mathbf{v}-\mu_m)]} \tag{6.4}$$

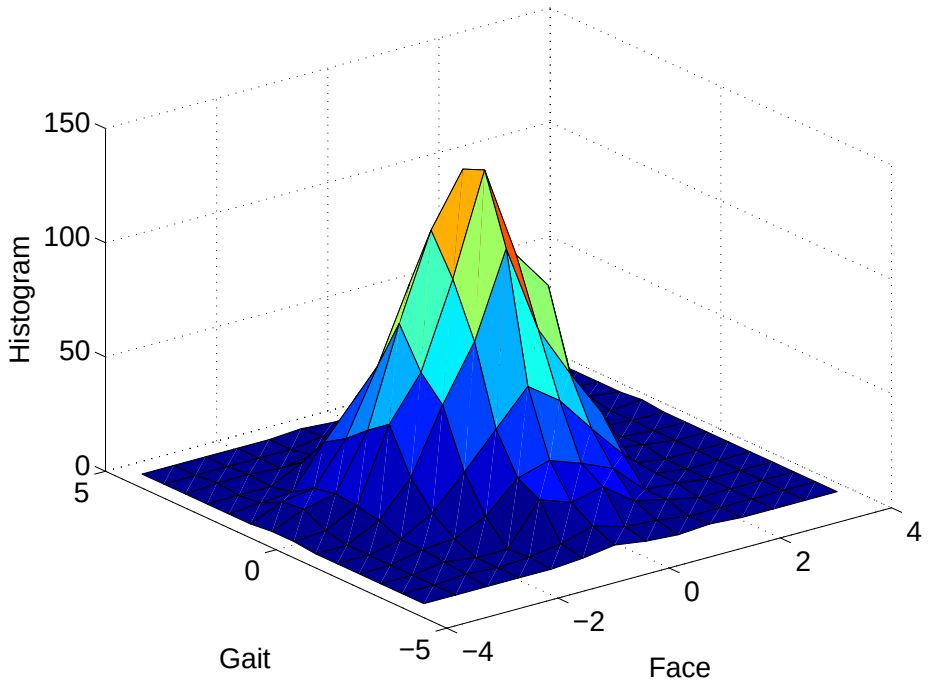


Figure 6.3. The 2D Histogram of Face and Gait Non-Match Scores.

$$P(\mathbf{v}|\omega_m) = \frac{1}{2\pi|\Sigma_m|^{1/2}} e^{[-\frac{1}{2}(\mathbf{v}-\mu_m)^T \Sigma_m^{-1}(\mathbf{v}-\mu_m)]} \tag{6.5}$$

The difference in the posterior probabilities of these two classes forms the combined similarity score.

$$CombS(p, g_j) = P(\omega_m|\mathbf{v}) - P(\omega_{nm}|\mathbf{v}) \tag{6.6}$$

3. The third scheme is the *Confidence Weighted Score Sum* as suggested by the HumanID group at University of Notre Dame. The main idea is that for a given probe subject I_P, we weight its similarity scores in a classifier before combination. The weight is computed from the similarity values of the first few ranks:

$$W_c(I_P) = \frac{S_c(I_P)(1) - S_c(I_P)(2)}{S_c(I_P)(2) - S_c(I_P)(3)} \tag{6.7}$$

(a) (b) (c) (d)

Figure 6.4. The Face Samples under Different Conditions. The Candidates for the Gallery Sets are (a) Regular Expression with Mugshot Lighting, and (b) Regular Expression, Overhead Lighting Images. The Probes are taken Outdoors with (c) Regular Expression, Far View and (d) Regular Expression, Near View.

where $S_c(I_P)(k)$ is the k^{th} largest similarity value of p when compared to the entire gallery set. The score combination is then given by:

$$CombS(I_P, I_{G_j}) = W_1(I_P)S_1(I_P, I_{G_j}) + W_2(I_P)S_2(I_P, I_{G_j}) \tag{6.8}$$

4. In addition to the score level combination schemes mentioned above, we also use a decision level combination: *Rank Sum*. It takes the negated sum of ranks from the face classifier and gait classifier as the similarity value:

$$CombS(p, g_j) = -(Rank_1(I_P, I_{G_j}) + Rank_2(I_P, I_{G_j})) \tag{6.9}$$

A problem of this scheme is that there might be two or more gallery subjects having a same similarity value with a probe. In this dissertation the tie is broken by the sum of the original scores in each classifier.

6.3 Results of Combinations

We conducted a series of studies geared towards answering the following questions in the context of outdoor biometrics:

1. What is the performance of face+gait combination for the same-day data and months-apart data? How does the combination of face and gait compare with single modality?

2. Which combination scheme performs the best?

3. How does the combination of face and gait compare against using multiple samples of the same modality, i.e. face+face or gait+gait?

Answers to the above questions requires careful specification of multiple gallery and probe sets. For faces, the main gallery set (F_{In,Mug,t_1}) consists of 70 faces taken indoors with regular expression and mugshot lighting conditions. The alternate gallery set (F_{In,Ov,t_1}) consists of the corresponding faces taken with overhead lighting. The outdoor images form the probes. Fig. 6.4 shows examples of faces for various lighting conditions. There are four face probe sets, with two probe sets per imaging session. For each imaging session, the near images form one set and the far images form the other set. One pair ($F_{Out,near,t_1}, F_{Out,far,t_1}$) was taken on the same day as the indoors images; there 39 such subjects. And the other pair ($F_{Out,near,t_2}, F_{Out,far,t_2}$) was taken *at least* 3 months apart; there are 21 such subjects.

For gait, the probes and the gallery are constructed from a subset of the USF HumanID gait dataset. The main gallery (G_{Grass,R,t_1}) consists of sequences from 70 individuals walking on grass, outdoors, viewed from the right camera. The alternate gallery set (G_{Grass,L,t_1}) consists of the corresponding sequences taken from the left camera, with a verging angle of approximately 30^o to the right view. Like the face, we consider four different probes. The left and right views of the gait on a different surface condition, i.e. concrete, taking on the same day as the gallery, form two probes ($G_{Concrete,R,t_1}, G_{Concrete,L,t_1}$), respectively. The sample views are listed in Fig. 2.2. Like face, we also consider the time covariate and consider two more probe sets ($G_{Grass,R,t_2}, G_{Grass,L,t_2}$) taken 6 months apart. The sizes of the probe sets match that for the face to allow us to consider biometric combinations.

Based on these gallery and probe, ten experiments were designed, as shown in Table 6.3. The first five experiments deal with same day data and the next five deal with comparing

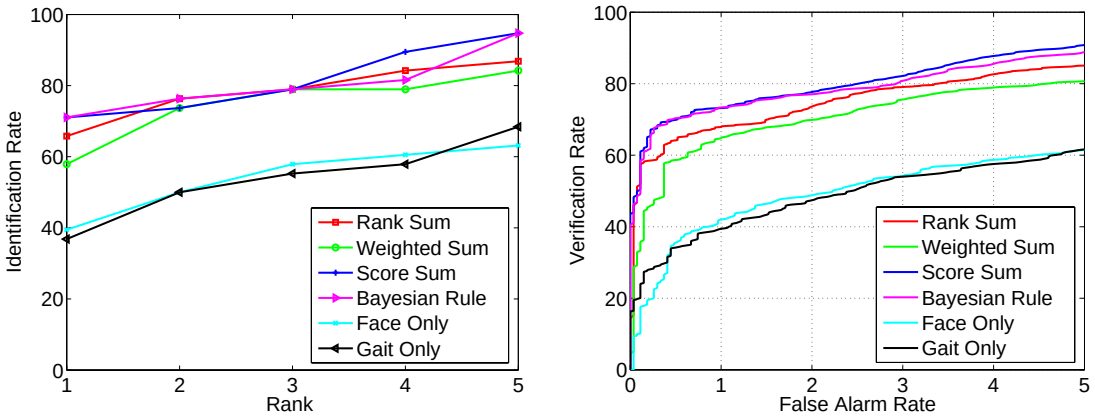
Table 6.3. Gallery and Probe Specifications for the Various Experiments Conducted.

Num	Exp	(Gallery, #)	(Probe, #)	Covariate
S_F	Face	F_{In,Mug,t_1}, 70	$F_{Out,near,t_1}$, 39	In/Outdoor, SameDay
S_G	Gait	G_{Grass,R,t_1}, 70	$G_{Concrete,R,t_1}$, 39	Surface, SameDay
S_{F+G}	Face+	F_{In,Mug,t_1}, 70	$F_{Out,near,t_1}$, 39	In/Outdoor, SameDay
	Gait	G_{Grass,R,t_1}, 70	$G_{Concrete,R,t_1}$, 39	Surface, SameDay
S_{F+F}	Face+	F_{In,Mug,t_1}, 70	$F_{Out,near,t_1}$, 39	In/Outdoor, SameDay
	Face	F_{In,Ov,t_1}, 70	F_{Out,far,t_1}, 39	In/Outdoor, SameDay
S_{G+G}	Gait+	G_{Grass,R,t_1}, 70	$G_{Concrete,R,t_1}$, 39	Surface, SameDay
	Gait	G_{Grass,L,t_1}, 70	$G_{Concrete,L,t_1}$, 39	Surface, SameDay
D_F	Face	F_{In,Mug,t_1}, 70	$F_{Out,near,t_2}$, 21	In/Outdoor, $\geq$ 3 Months Apart
D_G	Gait	G_{Grass,R,t_1}, 70	G_{Grass,R,t_2}, 21	6 Months Apart
D_{F+G}	Face	F_{In,Mug,t_1}, 70	$F_{Out,near,t_2}$, 21	In/Outdoor, $\geq$ 3 Months Apart
	Gait	G_{Grass,R,t_1}, 70	G_{Grass,R,t_2}, 21	6 Months Apart
D_{F+F}	Face	F_{In,Mug,t_1}, 70	$F_{Out,near,t_2}$, 21	In/Outdoor, $\geq$ 3 Months Apart
	Face	F_{In,Ov,t_1}, 70	F_{Out,far,t_2}, 21	In/Outdoor, $\geq$ 3 Months Apart
D_{G+G}	Gait+	G_{Grass,R,t_1}, 70	G_{Grass,R,t_2}, 21	6 Months Apart
	Gait	G_{Grass,L,t_1}, 70	G_{Grass,L,t_1}, 21	6 Months Apart

data taken more than 3 months apart. Each set of five experiments consists of experiments to study face and gait, individually and with inter-modal and intra-modal combinations. Similar to previous chapters, we report recognition performance in terms of identification rate and verification rate.

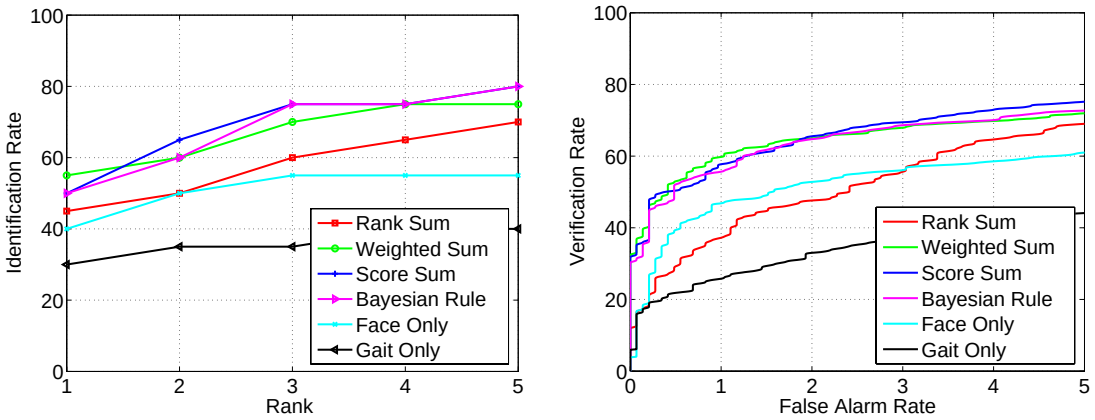
6.3.1 Inter-Modal Combination

Performance of outdoor face (Exp S_F), cross surface gait (Exp S_G), and gait+face (Exp S_{F+G}) on same day data with various combination schemes is shown in Fig. 6.5, which plots the CMC curve up to rank 5 and ROC curve up to 5% false alarm rate. As expected, the recognition from a single biometric is low, specifically, 40% for face and 39% for gait at rank 1. However, the combinations of the two weak biometrics using the four schemes discussed above substantially boosts performance. Particularly, 71% for score sum, 70% for Bayesian rule, 58% for confidence weighted score sum, and 68% for rank sum. As Fig. 6.6 shows, a similar pattern is seen for performance of outdoor face (Exp D_F), cross surface gait (Exp D_G), and gait+face (Exp D_{F+G}) on data taken months apart.



(a) (b)

Figure 6.5. Performance of Outdoor Face (Exp S_F), Cross Surface Gait (Exp S_G), and Gait+Face (Exp S_{F+G}) on Same Day Data with Various Combination Schemes for (a) Identification and (b) Verification Scenarios.



(a) (b)

Figure 6.6. Performance of Outdoor Face (Exp D_F), Cross Surface Gait (Exp D_G), and Gait+Face (Exp D_{F+G}) on Data taken Months Apart with Various Combination Schemes for (a) Identification and (b) Verification Scenarios.

6.3.2 Intra-Modal Combination

The performance of inter-modal combination has to be justified in the context of intra-modal combination. Inter-modal combinations involves the use of different types of sensors resulting in added integration costs. The inter-modal performance gain has to be justified in this context. Inter-modal combination performance has to be greater than intra-modal combination [Kevin Bowyer, Personal Communication]. In the present context, performance of gait and face should be greater than combination of two faces or combination of two gait signatures. For this, we consider the experiments, S_{F+F}, S_{G+G}, D_{F+F}, and D_{G+G}, in Table 6.3.

These intra-modal experiments involve the use of two samples per subject in the gallery *and* in the probe. Each probe is matched against the two gallery samples per person and the maximum similarity score is chosen as the similarity score for that probe. These similarity scores are then combined, as before, using the rules described in Section 6.2.

Fig. 6.7 plots the ROCs of the intra-modal combinations up to a false alarm rate of 5%. Each plot shows the performance with individual probes and their combinations. We see that the intra-modal combination does not seem to improve performance by a significant amount. Fig. 6.8 shows a summary comparison of the inter-modal and intra-model comparison schemes based on the verification rate at a false alarm rate of 5%. We see that face+gait performance is better than than the face+face or gait+gait combinations. This is explained by the strong correlation that exist between the scores for two probes from the same biometric. It is 0.7 for the intra-modal case and is only 0.05 for the inter-modal case. Stronger the correlation between the scores, less is the improvement with combination [88].

6.4 Discussion

Table 6.4 lists the number of subjects that are affected by the combination. It lists the number of subjects who were failed to be recognized by each individual modality or both, but were successfully recognized after combination. It also lists the number of subjects who

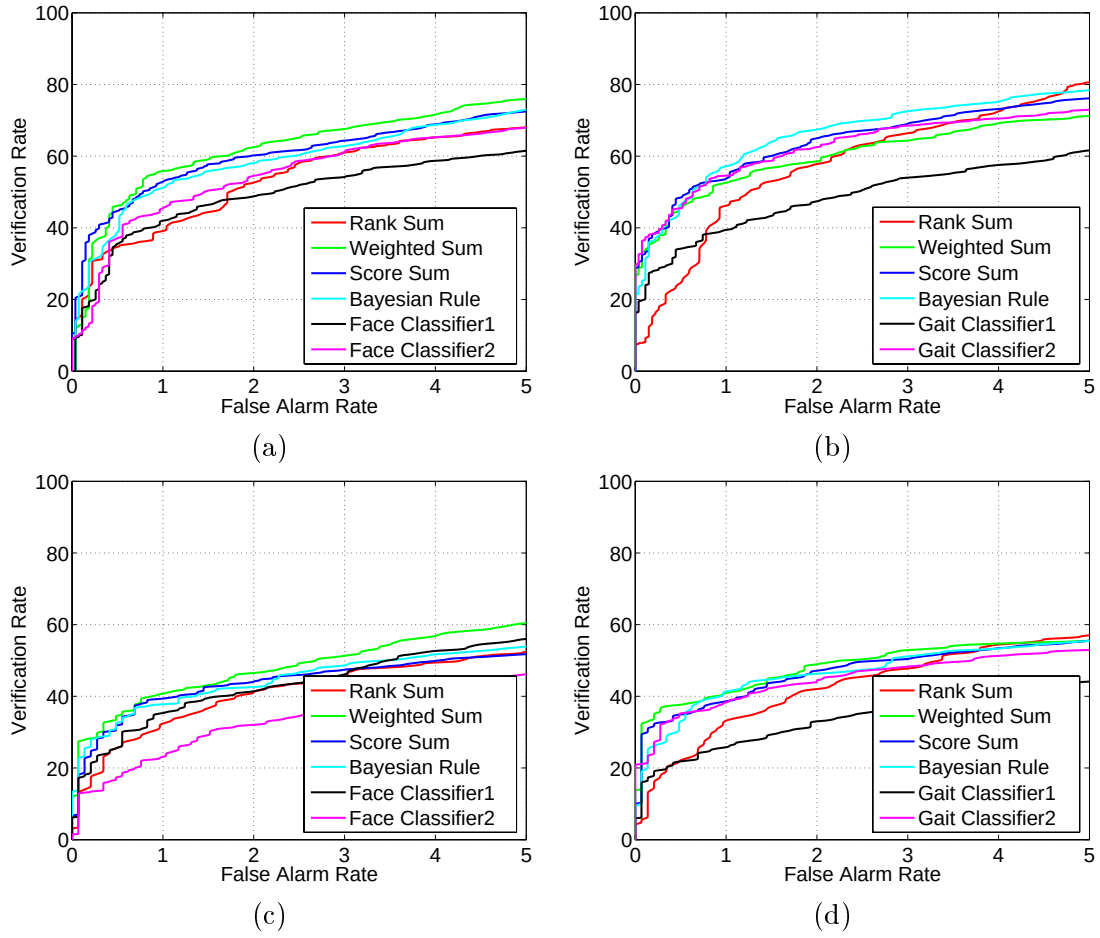
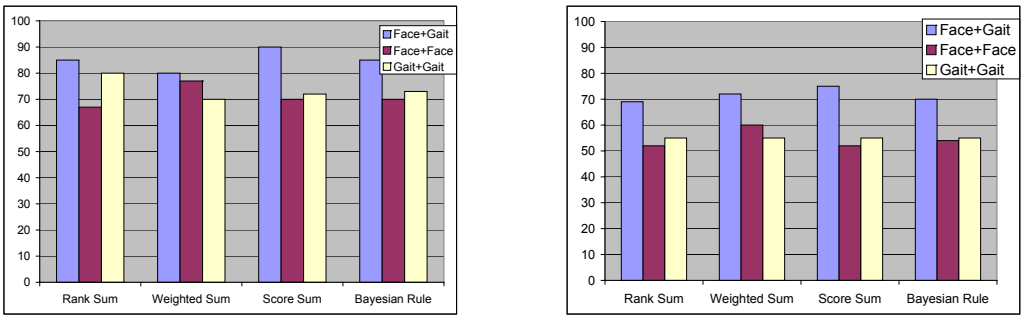


Figure 6.7. Performance of Intra-modal Combination Using Different Strategies. The ROC Curves are Shown for (a) Face+Face, Same Day, (b) Gait+Gait, Same Day, (c) Face+Face, Months Apart, and (d) Gait+Gait, Months Apart. Each Plot shows the Performance with Individual Probes and their Combinations.

were successfully recognized by one modality or both, but their combination resulted in failure. We see that the performance gained by the combination are mostly from subjects who failed only for one of the two biometrics. The combination helps little for subjects who were not correctly identified by both the individual biometrics. On the other hand, we found that fewer subjects were correctly identified in one classifier but failed after combinations. This is especially true for the score sum combination.

To gain some insight into the nature of the face and gait combination, we plot the decision boundary of the experiment for score sum and Gaussian Bayesian fusion at 5%



(a) (b)

Figure 6.8. Bar Plot of Verification Rate at a False Alarm Rate of 5% for Inter- and Intra-Modal Combination of Gait and Face for (a) Same Day Data, and (b) Data Separated by Months.

Table 6.4. Number of Subject Correctly Recognized or Failed to be Recognized by Each Individual Modality or their Combination for the Same-Day Data. The total Number of Subjects is 39.

Combination Scheme	# failed before combination but succeeded after combination			# succeeded before combination but failed after combination		
	Face Only	Gait Only	Both	Face Only	Gait Only	Both
Rank Sum	15	12	4	5	1	0
Confidence Weighted Sum	8	12	0	1	4	0
Score Sum	14	14	3	2	1	0
Bayesian Rule	14	14	3	1	3	0

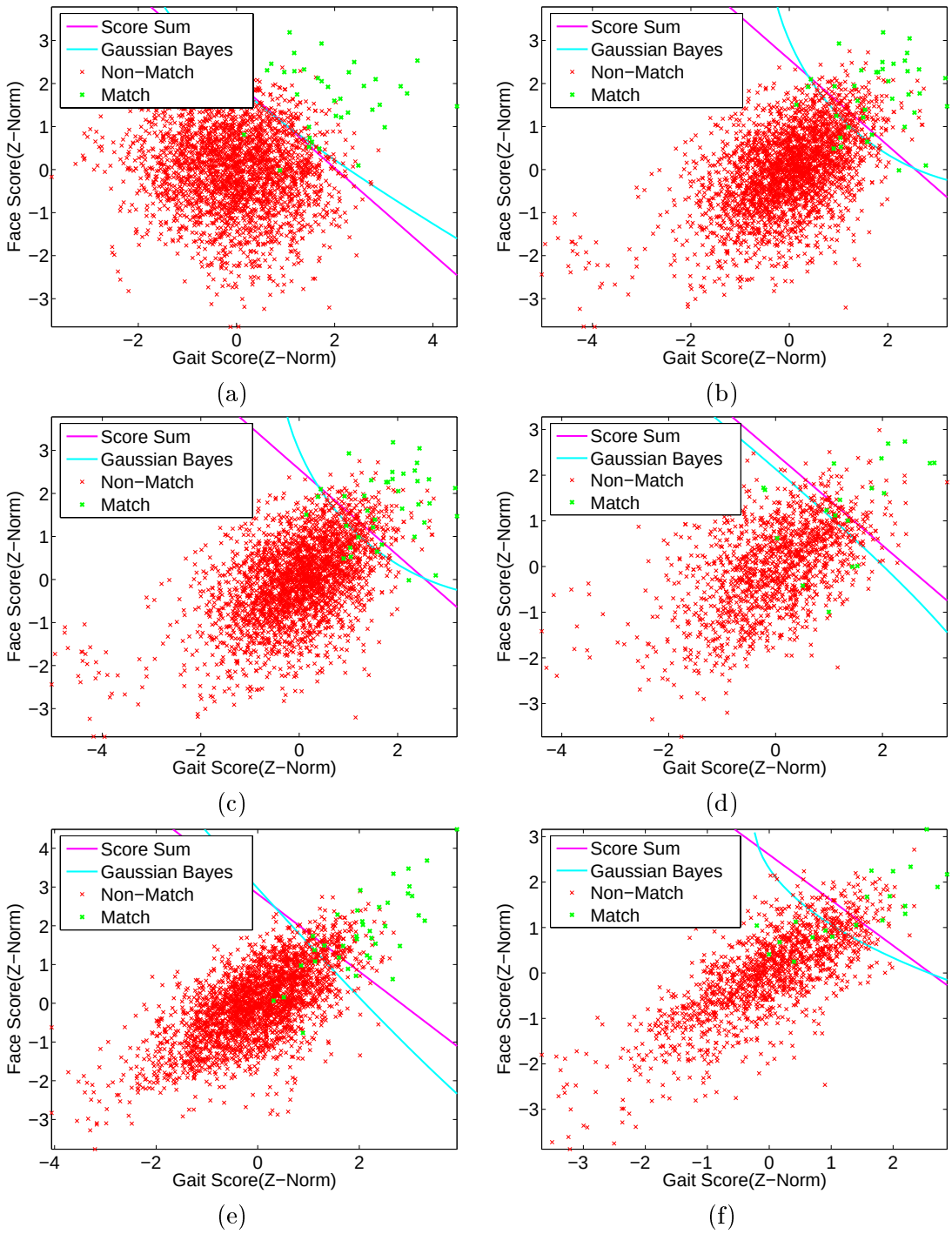


Figure 6.9. The Decision Boundary of the Score Sum and Bayesian Rule Combination Rules at a False Alarm Rate of 5% for (a), (b) Face + Gait Same and Different Days, (c), (d) Face + Face Same and Different Days, and (e), (f) Gait + Gait Same and Different Days, respectively.

false alarm rate in Fig. 6.9. The axes are the *normalized* similarity scores from each modality. We see that (i) the optimal Bayesian decision boundary is roughly linear and is close to the score sum boundary, which explains the high performance with just score sum schemes. And (ii) the non-match scores seem to be uncorrelated forming a nice, symmetric central cluster. This observation would be important for parametric modeling studies. Gaussian models seem to be good for non-match scores.

6.5 Summary of Gait and Face Combination Study

In this chapter we demonstrated an effective strategy for improving overall biometric recognition under "hard" convariates by combining gait with face. Hard problems in face recognition involve comparing across indoor and outdoors conditions and over time (months). We find that the score sum rule of combination offer the best performance. We also find the the inter-modal combination, i.e. face+gait, is better than not only the individual modalities but also combinations of the same modality, i.e. face+face and gait+gait. The inter-modal combination has excellent potential for overcoming the "tough" covariates affecting individual biometrics.

CHAPTER 7

CONCLUSIONS

In this dissertation, we investigated computer vision based gait recognition. We proposed four new gait algorithms, ranging from the simple parameterless baseline algorithm to the dynamics normalized approaches that emphasize shape over dynamics. Based on the performance on the publicly available gait databases, such as the USF/NIST HumanID database, the UMD database and the CMU Mobo database, we (i) established the hard problems in gait biometrics, (ii) investigated the impact of segmentation on recognition, (iii) proposed approaches to improve performance of both gait algorithms, and (iv) explored fusion of gait with face biometrics for outdoor conditions.

We investigated gait recognition with changes in five different conditions and their combinations. To gain an insight of the potential of gait biometrics, we used the simple baseline algorithm. Focused of the study of the impact of a covariate on match-score distribution suggests that shoe type has the least effect on performance, but the effect is nevertheless statistically significant. This is followed by either a change in camera view or carrying a brief case. Carrying a briefcase does not affect performance as much as one might expect (Section 3.2.4). This effect is marginally larger than changing shoe type but is substantially smaller than a change in surface type. The largest effects are due to time and surface.

7.1 Effect of Time on Gait

One of the factors that has large impact is time, resulting in lower recognition rates for changes when matching sequences over time. This dependence on time has been reported by others too, but for indoor sequences and for less than 6 months differences. When the

difference in time between gallery (the pre-stored template) and probe (the input data) is on the order of minutes, the identification performance ranges from 91% to 95% [93, 30, 15], whereas the performances drop to less than 30% when the differences are on the order of months and days [53, 16, 15] for similar sized datasets. Particularly, for the HumanID gait challenge database with 122 subjects, recognition with 6 months apart is lower than 10% for most algorithms. Our speculation is that other changes that naturally occur between video acquisition sessions are very important. These include change in clothing worn by the subject, change in the outdoor lighting conditions, and inherent variation in gait over time. For applications that would require matching across days or months, these would most likely be the important variables. We also found that compared with dynamics, body shape is less sensitive to these variables. This is illustrated by the fact that the dynamics normalization algorithm has substantially improved the recognition across time, specifically, the top rank identification rate increased from 10% to 21% for HumanID database (122 subjects), and increased from 70% to 85% for UMD database (55 subjects).

7.2 Effect of Surface on Gait

The other factor with large impact on gait recognition is walking surface. With the subject walking on grass in the gallery sequence and on concrete in the probe sequence (HumanID challenge experiment D with 122 subjects), rank-one recognition is only 32%. Performance degradation might be even larger if we considered other surface types, such as sand or gravel, that might reasonably be encountered in some applications. However, we also found that the effects can be compensated for only using the cue of body shape. And in Chapter 5 we showed that the recognition rate increases from 32% to 64% after normalizing gait dynamics.

7.3 Segmentation on Gait Recognition

We have established that the low performance under the impact of surface and time variation can not be explained by poor silhouette quality. We base our conclusions on

two gait recognition algorithms, one exploits both shape and dynamics, while the other exploits just shape. The drop in performance due to surface condition that we observe in the gait challenge problem is *not* due to differences in background. This observation is also corroborated by the performances reported in a fairly recent work by the Lee *et al.* [52]. The observation has implication for future work direction in gait recognition. Instead of searching for better methods for silhouette detection to improve recognition, it would be more productive to study and isolate components of gait that do not change under shoe, surface, or time. One example of this type of study is [86] in which relationship between silhouette shape and speed was studied and then was compensated for by transforming the silhouettes. While it is doubtful whether speed variations can fully explain the drop in performance due to surface or time change, systematic studies such as this would be needed to understand the limitation of gait recognition.

7.4 Improving Recognition: Shape over Dynamics

Shape (body shape and stance shape) and dynamics are two components of gait. We tried to separate these two components and studied approaches that emphasize shape, which is more likely to be invariant under changes in covariates, such as surface and time. In this regard, we proposed three gait recognition algorithms: (i) an averaged silhouette based algorithm that deemphasizes gait dynamics, (ii) an algorithm that normalizes gait dynamics and then uses Euclidean distance between corresponding selected stances, and (iii) an algorithm that also normalizes gait dynamics but computes similarity in the Linear Discriminant Analysis (LDA) gait space after morphological deformation. Based on extensive experimentation on multiple, publicly available databases (HumanID Gait Challenge, UMD, and CMU-Mobo), we can assert that dynamics-normalization greatly improves overall gait recognition performance, especially when comparing across surface, carrying condition, time, and different speed. The approach is not dependent on the training set choice. It generalizes well not only across different subjects, but also across different datasets with varying imaging geometries. We attribute this significant improvement to

gait dynamics-normalization. The other two components: the LDA stance shape space and the morphological operation based distance computation also improve performance, but the former has more impact than the latter.

Efficacy of dynamics normalization suggests that body-stance shape plays a more important role than dynamics in gait recognition. This is also supported by good performance of gait recognition algorithms of Veeraraghavan *et al.* [90] and Collins *et al.* [15], which focus more on body shape than dynamics. To get some insight into the kinds of shape features that seem to be important we consider the top-two most inter-subject discriminating directions for each stance, as found by LDA of the silhouette shapes in the training set used for gait-normalization. We plotted these directions in Fig. 5.14 and see that (i) upper body, (ii) knee, and (iii) lower leg during gait swing phases seems to be the important features that are being picked up.

7.5 Gait And Face

One of the open questions is the potential for gait to perform identification. We addressed this question by comparing our gait results with face recognition. Our analysis provides a rough guide to the current state of gait recognition. Face recognition performance has been well characterized by a number of evaluations, the most recent being the Face Recognition Vendor Test (FRVT) 2002 [40]. Because gallery size is different in the gait challenge problem and FRVT 2002, comparison is made for verification performance at a false accept rate of 1%. Unlike identification, verification performance is not a function of gallery size. Since the gait challenge problem performs recognition from outdoor video, we need to look at face recognition results from outdoor images. In FRVT 2002 there are two results on outdoor facial images. In both cases, the gallery is of indoor full frontal images. In the first result, the probe set consists of outdoor images taken on the same day as the gallery images. Verification performance varied for different systems ranging from 54% to 5%, with a median of 34%. From Fig. 5.8 (The best gait algorithm), gait performance varied from 98% to 35% on the ten experiments where the gallery and probe

set sequences were taken on the same day. The median performance score was 70%. In the second set of outdoor face recognition results, the probe set consists of outdoor images taken on a different day than the gallery image of a person; the median difference in time is about 5 months. Verification performance varied from 47% to 0% for different systems, with a median of 22%. Experiments K and L in the gait recognition problem, which have probes from 6 months later, are comparable to this scenario. The verification rate for both experiments is about 21%. A number of caveats need to be mentioned in this analysis. The FRVT 2002 performance numbers are from a blind evaluation on sequestered data. This is not the case for our gait results. Using the respective performances only as a rough guide, we see that video-based gait as an outdoor at-a-distance biometrics has 1) the potential to be competitive with faces, and 2) as a biometrics to be fused with face.

We demonstrated that combining face and gait is an effective strategy to improve overall biometric recognition under hard conditions. Hard problems in face recognition involve comparing across indoor and outdoors conditions and over time (months). We found that the score sum rule of combination offer the best performance, possibly because they appear to be uncorrelated as demonstrated in scatter plot of scores (Fig. 6.9). We also found that the inter-modal combination, i.e. face+gait, is better than not only the individual modalities but also combinations of the same modality, i.e. face+face and gait+gait. The inter-modal combination has excellent potential for overcoming the "tough" covariates affecting individual biometrics.

7.6 Future Research Directions

With respect to recognition performance, our results indicate that investigations should be focused on body shape because gait dynamics is shown to be more sensitive to condition changes. It would be necessary to model and correct for the stance-shape changes under varying conditions. An excellent start is the work of Tanawongsuwan and Bobick [87] who studied the effect of walking speed on silhouette shapes. Another direction is to employ 3D modeling because the three-dimensional morphable models are a technique for

improving recognition of non-frontal images in face recognition, according to the FRVT 2002 report [40]. And it is interesting to discover whether this is true for gait.

Can one predict psychological and physiological properties of carrying weight or age from gait? The gait challenge problem comprises the condition of carrying a briefcase. In future experiments, it may be interesting to investigate the effect of carrying a backpack rather than a briefcase, or to vary the object that is carried. Based on such studies it might be possible to estimate the weight carried by the person. One such interesting work that has started to look at this is the work of Wittman [99], who studied body response to increasing concealed weight.

The large effect of surface type on performance suggests an important future research topic might involve investigating whether changes in gait with surface type is predictable. For example, given a description of gait from walking on concrete, is it possible to predict the gait description that would be obtained from walking on grass or sand? Alternatively, is there some other description of gait that is not as sensitive to change in surface type? The study can be generalized into other within-subject changes, e.g., predict slow gait from fast gait or vice versa. In addition, can gait dynamics be separated from shape so that one can create an artificial sequence involve one subject's shape but using another subject's dynamics?

REFERENCES

[1] B. Achermann and H. Bunke. Combination of face classifiers for person identification. *International Conference on Pattern Recognition*, 1996.

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

[3] H. Akaike. Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281, 1973.

[4] C. D. Barclay, J. E. Cutting, and L. T. Kozlowski. Temporal and spatial factors in gait perception that influence gender recognition. In *Perception and Psychophysics*, volume 23, pages 145–152, 1978.

[5] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. In *IEEE Trans. Pattern Anal. and Mach. Intel.*, volume 19, pages 711–720, July 1997.

[6] C. BenAbdelkader, R. Cutler, and L. Davis. Motion-based recognition of people in eigengait space. In *International Conference on Automatic Face and Gesture Recognition*, pages 267–272, 2002.

[7] A. Bobick and A. Johnson. Gait recognition using static, activity-specific parameters. In *Computer Vision and Pattern Recognition*, pages I:423–430, 2001.

[8] D. S. Bolme. Elastic bunch graph matching. Degree of master of science, Colorado State University, 2003.

[9] V. D. Boomgaard and V. Balen. Image transforms using bitmapped binary images. In *Computer Vision, Graphics, and Image Processing: Graphical Models and Image Processing*, volume 54, pages 254–258, May 1992.

[10] K. Chang, K. W. Bowyer, S. Sarkar, and B. Victor. Comparison and combination of ear and face images in appearance-based biometrics. *IEEE Trans. Pattern Anal. and Mach. Intel.*, pages 1160–1165, September 2003.

[11] K. I. Chang, K. W. Bowyer, and P. J. Flynn. Face recognition using 2d and 3d facial data. In *Workshop in Multimodal User Authentication*, pages 25–32, Santa Barbara, California, December 2003.

[12] K. I. Chang, K. W. Bowyer, and P. J. Flynn. Multi-modal 2d and 3d biometrics for face recognition. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, Nice, France, Oct. 2003.

[13] K. I. Chang, K. W. Bowyer, P. J. Flynn, and X. Chen. Multi-biometrics using facial apperance, shape and temperature. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, Seoul Korea, May 17-19 2004.

[14] X. Chen, P. J. Flynn, and K. W. Bowyer. Visible-light and infrared face recognition. In *The proceedings of Workshop on Multimodal User Authentication*, pages 48–55, Santa Barbara, CA USA., Dec 2003.

[15] R. Collins, R. Gross, and J. Shi. Silhouette-based human identification from body shape and gait. In *International Conference on Automatic Face and Gesture Recognition*, pages 366–371, 2002.

[16] N. Cuntoor, A. Kale, and R. Chellappa. Combining multiple evidences for gait recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003.

[17] J. E. Cutting and L. T. Kozlowski. Recognition of friends by their walk. *Bulletin of the Psychonomic Society*, 9:353–356, 1977.

[18] U. Dieckmann, P. Plankensteiner, and T. Wagner. A biometric person identification system using sensor fusion. *Pattern Recognition Letters*, pages 827–833, 1997.

[19] W. H. Dittrich. Action categories and the perception of biological motion. In *Perception*, volume 22, pages 15–22, 1993.

[20] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley, 2001.

[21] R. B. E.D. Grossman. Brain activity evoked by inverted and imagined biological motion. In *Vision Research*, pages 1475–1482, 2001.

[22] B. Flinchbaugh and B. Chandrasekaran. A theory of spatio-temporal aggregation for vision. *AI*, 17:387–407, 1981.

[23] J. P. Foster, N. M. S., and P.-B. A. Automatic gait recognition using area-based metrics. In *Pattern Recognition Letters*, volume 24, pages 2489–2497, 2003.

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

[25] R. Gross and J. Shi. The cmu motion of body (mobo) database. Tech. report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, June 2001.

[26] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical report, Carnegie Mellon University, 2001.

[27] E. Grossman, M. Donnelly, R. Price, D. Pickens, V. Morgan, G. Neighbor, and R. Blake. Brain areas involved in perception of biological motion. In *Journal of Cognitive Neuroscience*, pages 711–720, September 2000.

[28] J. Grzesa, P. Fonlupta, B. Bertenthalb, C. Delon-Martinc, C. Segebarthc, and J. Decetya. Does perception of biological motion rely on specific brain regions? In *NeroImage*, volume 13, pages 775–785, 2001.

[29] J. Han and B. Bhanu. Statistical feature fusion for gait-based human recognition. In *Computer Vision and Pattern Recognition*, volume II, pages 842–847, June 2004.

[30] J. Hayfron-Acquah, M. Nixon, and J. Carter. Automatic gait recognition by symmetry analysis. In *International Conference on Audio- and Video-Based Biometric Person Authentication*, pages 272–277, 2001.

[31] A. Hilton and P. Fua. Modeling people toward vision-based understanding of a person's shape, appearance, and movement. *Computer Vision and Image Understanding*, 81(3):227–230, March 2001.

[32] D. D. Hoffman and B. E. Flinchbaugh. The interpretation of biological motion. In *Biological Cybernetics*, pages 195–204, 1982.

[33] L. Hong, A. Jain, and S. Pankanti. Can multibiometrics improve performance? *IEEE Workshop on Identification Advanced Technologies*, pages 59–64, 1999.

[34] L. Hong and A. K. Jain. Integrating faces and fingerprints for personal identification. *IEEE Trans. Pattern Anal. and Mach. Intel.*, 20(12):1295–1307, 1998.

[35] J. Jaccard and C. K. Wan. *LISREL Approach to Interaction Effects in Multiple Regression.* Sage Publications, 1996.

[36] A. K. Jain, L.Hong, and Y. Kulkarni. A multimodal biometric system using fingerprints, face and speech. *2nd Int'l Conference on Audio- and Video-based Biometric Person Authentication*, pages 182–187, March 1999.

[37] A. K. Jain, S. Prabhakar, and S. Chen. Combining multiple matchers for a high security fingerprint verification system. *Pattern Recognition Letters*, 20(11-13):1371–1379, 1999.

[38] G. Johansson. Visual motion perception. *SciAmer*, 232:75–88, June 1976.

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

[40] P. Jonathon Phillips, D. Blackburn, M. Bone, P. Grother, R. Micheals, and E. Tabassi. Face recogntion vendor test. *http://www.frvt.org/*, 2002.

[41] P. Jonathon Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. and Mach. Intel.*, 22(10):1090–1104, 2000.

[42] P. Jonathon Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer. Baseline results for the challenge problem of Human ID using gait analysis. In *International Conference on Automatic Face and Gesture Recognition*, pages 137–142, 2002.

[43] P. Jonathon Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer. The gait identification challenge problem: Data sets and baseline algorithm. In *International Conference on Pattern Recognition*, pages 385–388, 2002.

[44] A. Kale, C. B., B. Yegnanarayana, A. N. Rajagopalan, and R. Chellappa. Gait analysis for human identification. In *International Conference on Audio- and Video-Based Biometric Person Authentication*, 2003.

[45] A. Kale, N. Cuntoor, and R. Chellappa. A framework for activity specific human identification. In *International Conference on Acoustics, Speech and Signal Processing*, 2002.

[46] A. Kale, A. Rajagopalan, N. Cuntoor, and V. Kruger. Gait-based recognition of humans using continuous HMMs. In *International Conference on Automatic Face and Gesture Recognition*, pages 336–341, 2002.

[47] A. Kale, A. N. Rajagopalan, A. Sunderesan, N. Cuntoor, and A. R. Chowdhury. Identification of humans using gait. In *IEEE Trans. on Image Processing*, to apper.

[48] A. Kale, A. Roy Chowdhury, and R. Chellappa. Fusion of gait and face for human recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, 2004.

[49] J. Kittler, M. Hatef, R. P. Duin, and J. Matas. On combining classifiers. *IEEE Trans. Pattern Anal. and Mach. Intel.*, 20(3), 1998.

[50] C.-S. Lee and A. Elgammal. Gait style and gait content: Bilinear models for gait recognition using gait re-sampling. In *International Conference on Automatic Face and Gesture Recognition*, pages 147–152, May 2004.

[51] L. Lee. *Gait Analysis for Classification.* PhD thesis, Massachusetts Institute of Technology, June 2003.

[52] L. Lee, G. Dalley, and K. Tieu. Learning pedestrian models for silhouette refinement. In *International Conference on Computer Vision*, 2003.

[53] L. Lee and W. Grimson. Gait analysis for recognition and classification. In *International Conference on Automatic Face and Gesture Recognition*, pages 155–162, 2002.

[54] J. Little and J. Boyd. Recognizing people by their gait: The shape of motion. *Videre*, 1(2):1–33, 1998.

[55] Z. Liu, L. Malave, A. Osuntogun, P. Sudhakar, and S. Sarkar. Toward understanding the limits of gait recognition. In *SPIE Processings of Defense and Security Symposium: Biometric Technology for Human Identification*, pages 195–205, 2004.

[56] Z. Liu, L. Malave, and S. Sarkar. Studies on silhouette quality and gait recognition. In *Computer Vision and Pattern Recognition*, volume II, pages 704–711, June 2004.

[57] Z. Liu and S. Sarkar. Outdoor biometrics over time by fusing gait with face. In *Biometric Consortium Conference*, Washington D.C., September 2004.

[58] Z. Liu and S. Sarkar. Simplest representation yet for gait recognition: Averaged silhouette. In *International Conference on Pattern Recognition*, volume 4, pages 211–214, 2004.

[59] Z. Liu and S. Sarkar. Effect of silhouette quality on hard problems in gait recognition. In *IEEE Transactions on Systems, Man, and Cybernetics (PartB)*, to appear.

[60] X. Lu, Y. Wang, and A. K. Jain. Combining classifiers for face recognition. *IEEE International Conference on Multimedia And Expo*, 3:13–16, July 2003.

[61] L. H. Malave. Silhouette based gait recognition: Research resource and limits. Master's thesis, University of South Florida, 2003.

[62] O. Melnik, Y. Vardi, and C. Zhang. Mixed group ranks: Preference and confidence in classifier combination. *Rutgers Statistics Department Tech Report*, September 2003.

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

[64] M. Murray, A. Drought, and R. Kory. Walking patterns of normal men. In *Journal of Bone and Joint Surgery*, volume 46-A, pages 335–360, 1964.

[65] S. Niyogi and E. Adelson. Analyzing gait with spatiotemporal surfaces. In *Computer Vision and Pattern Recognition*, 1994.

[66] P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002. *http://www.frvt.org*, March 2002.

[67] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. and Mach. Intel.*, 22(10), Oct. 2000.

[68] F. E. Pollick, H. M. Paterson, A. Bruderlin, and A. J. Sanford. Perceiving affect from arm movement. In *Cognition*, pages B51–B61, 2001.

[69] S. Prabhakar and A. K. Jain. Decision-level fusion in fingerprint verification. *Pattern Recognition*, 35(4):861–874, 2002.

[70] L. Rabiner and B. H. Juang. An introduction to hidden markov models. In *IEEE ASSP Magazine*, pages 4–16, 1986.

[71] L. Rabiner and B. H. Juang. *Fundamental of Speech Recognition*. Prentice Hall, 1993.

[72] A. Ross and A. K. Jain. Information fusion in biometrics. *Pattern Recognition Letters*, 24:2115–2125, September 2003.

[73] S. Runeson and G. Frykholm. Visual perception of lifted weight. In *Journal of Experimental Psychology: Human Perception and Performance*, volume 7, pages 733–740, 1981.

[74] S. Runeson and G. Frykholm. Kinematic specificaiton of dynamics as an informational basis for person-and-action perception: Expectation, gender recognition, and deceptive intention. In *Journal of Experimental Psychology: General*, volume 112, pages 585–615, 1983.

[75] S. Sarkar, P. J. Phillips, Z. Liu, I. Robledo, P. Grother, and K. Bowyer. The Human ID gait challenge problem: Data sets, performance, and analysis. In *IEEE Trans. Pattern Anal. and Mach. Intel.*, to appear.

[76] B. Schiele. How many classifiers do i need? In *International Conference on Pattern Recognition*, pages II: 176–179, 2002.

[77] M. Shah and R. Jain (Eds.). *Motion-Based Recognition*. Kluwer Publishers, 1997.

[78] G. Shakhnarovich and T. Darrell. On probabilistic combination of face and gait cues for identification. *Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 2002.

[79] G. Shakhnarovich, L. Lee, and T. Darrell. Integrated face and gait recognition from multiple views. In *Computer Vision and Pattern Recognition*, pages I:439–446, 2001.

[80] G. Shakhnarovich, L. Lee, and T. Darrell. Integrated face and gait recognition with multiple views. *Computer Vision and Pattern Recognition*, 2001.

[81] J. Shutler, M. Nixon, and C. Carter. Statistical gait description via temporal moments. In *4th IEEE Southwest Symp. on Image Analysis and Int.*, pages 291–295, 2000.

[82] S. V. Stevenage, M. S. Nixon, and V. K. Visual analysis of gait as a cue to identity. *Applied Cognitive Psychology*, 13(6):513–526, Dec. 1999.

[83] D. M. Stuart and S. N. Mark. Automatic gait recognition via fourier descriptors of deformable objects. In *Proceedings of Audio Visual Biometric Person Authentication*, pages 566–573, 2003.

[84] A. Sunderesan, A. K. Roy Chowdhury, and R. Chellappa. A hidden markov model based framework for recognition of humans from gait sequences. In *IEEE International Conference on Image Processing*, 2003.

[85] R. Tanawongsuwan and A. Bobick. Gait recognition from time-normalized joint-angle trajectories in the walking plane. In *Computer Vision and Pattern Recognition*, pages II:726–731, 2001.

[86] R. Tanawongsuwan and A. Bobick. Performance analysis of time-distance gait parameters under different speeds. In *International Conference on Audio- and Video-Based Biometric Person Authentication*, 2003.

[87] R. Tanawongsuwan and A. Bobick. Modelling the effects ofwalking speed on appearance-based gait recognition. *Computer Vision and Pattern Recognition*, pages 783–790, June 2004.

[88] D. M. Tax, M. van Breukelen, R. P. Duin, and J. Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33(9):1475–1485, September 2000.

[89] D. Tolliver and R. Collins. Gait shape estimation for identification. In *International Conference on Audio- and Video-Based Biometric Person Authentication*, 2003.

[90] A. Veeraraghavan, A. R. Chowdhury, and R. Chellappa. Role of shape and kinematics in human movement analysis. In *Computer Vision and Pattern Recognition*, Washington D.C., USA, June 2004.

[91] I. R. Vega. *Motion model based on statistics of feature relations: human identification from gait*. PhD thesis, University of South Florida, 2002.

[92] I. R. Vega and S. Sarkar. Statistical motion model based on the change of feature relationships: Human gait-based recognition. In *IEEE Trans. Pattern Anal. and Mach. Intel.*, volume 25, pages 1323–1328, 2003.

[93] L. Wang, W. Hu, and T. Tan. A new attempt to gait-based human identification. In *International Conference on Pattern Recognition*, volume 1, pages 115–118, 2002.

[94] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *IEEE Trans. Pattern Anal. and Mach. Intel.*, 25:1505–1518, Dec. 2003.

[95] Y. Wang, T. Tan, and A. K. Jain. Combining face and iris biometrics for identity verification. *Int'l Conf. on Audio- and Video-Based Biometric Person Authentication*, pages 805–813, June 2003.

[96] J. A. Webb and J. K. Aggarwal. Structure from motion of rigid and jointed objects. *Artificial Intelligence*, 19:107–130, 1982.

[97] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.

[98] L. Wiskott, J.-M. Fellous, N. Kruger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. and Mach. Intel.*, 19(7):775–779, July 1997.

[99] M. Wittman. Visual analysis of the effects of load carriage on gait. In *Biometric Consortium Conference*, Washington D.C., September 2004.

[100] K. Woods, W. Philip Kegelmeyer Jr., and K. Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Trans. Pattern Anal. and Mach. Intel.*, 19(4), 1997.

[101] C. Y. Yam, M. S. Nixon, and J. N. Carter. On the relationship of human walking and running: Automatic person identification by gait. In *Proceedings of International Conference on Pattern Recognition*, pages 287–290, 2003.

[102] J. Zhou and D. Zhang. Face recognition by combining several algorithms. In *International Conference on Pattern Recognition*, pages III: 497–500, 2002.

[103] Y. Zuev and S. Ivanon. The voting as a way to increase the decision reliability. *Foundations of Information/Decision Fusion with Applications to Engineering Problems*, pages 206–210, 1996.

ABOUT THE AUTHOR

Zongyi Liu received his BS degree in business from Shenzhen University in 1997 and the MS degree in computer science and application from University of Electronic Science and Technology of China in 2000. He is currently a PhD candidate at University of South Florida. His research interests are computer-vision based gait biometrics, pattern recognition, motion and image segmentation.