
massachusetts institute of technology — artificial intelligence laboratory

# Gait Analysis for Classification

Lily Lee

AI Technical Report 2003-014 June 2003

# Gait Analysis for Classification

by

Lily Lee

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2002



Certified by: W.E.L. Grimson
Bernard Gordon Professor of Medical Engineering
Thesis Supervisor

Accepted by: Arthur C. Smith
Chairman, Department Committee on Graduate Students

**Gait Analysis for Classification**

by

Lily Lee

Submitted to the Department of Electrical Engineering and Computer Science on June 1st, 2002, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Computer Science and Engineering

## Abstract

This thesis describes a representation of gait appearance for the purpose of person identification and classification. This gait representation is based on simple localized image features such as moments extracted from orthogonal view video silhouettes of human walking motion. A suite of time-integration methods, spanning a range of coarseness of time aggregation and modeling of feature distributions, are applied to these image features to create a suite of gait sequence representations. Despite their simplicity, the resulting feature vectors contain enough information to perform well on human identification and gender classification tasks. We demonstrate the accuracy of recognition on gait video sequences collected over different days and times and under varying lighting environments. Each of the integration methods are investigated for their advantages and disadvantages. An improved gait representation is built based on our experiences with the initial set of gait representations. In addition, we show gender classification results using our gait appearance features, the effect of our heuristic feature selection method, and the significance of individual features.

Thesis Supervisor: W.E.L. Grimson
Title: Bernard Gordon Professor of Medical Engineering

## Acknowledgments

It is with great pleasure that I reflect on my time at MIT about all the people who made this thesis possible. First and foremost, I would like to thank my advisor, Professor Grimson, for his patient, thoughtful guidance and support. The years I have spent at the MIT Artificial Intelligence Lab have been the most exciting of my life, thanks to the intellectual environment fostered by its faculty, Rod Brooks, Tomas Lozano-Perez, Eric Grimson, Berthold Horn, and Trevor Darrell.

I would like to thank past and present members of the WELG group for many wonderful discussions and suggestions. Tina Kapur, Tao Alter, Gideon Stein, and Greg Klanderman guided me through the bootstrapping process of my computer vision research here at lab. Current members Raquel Romano, Chris Stauffer, Janey Yu, Lilla Zollei, Polina Golland, Kinh Tieu, Lauren O'Donnell, Samson Timoner, and Eric Cosman have been invaluable in helping me clarify my sometimes cryptic thoughts. Outside of the group I would like to thank Christian Shelton, John Fisher, Eric Miller, Edward Wang, Ron Dror, Greg Shakhnarovich, Terran Lane, who have been very generous with their time and expertise. My officemates over the years, Jose Robles, Tina Kapur, Oded Maron, Christian Shelton, Gideon Stein, and Raquel Romano, Neil Weisenfeld have each in their own ways (and sometimes with their antics) made life in the lab interesting and exciting.

My best friends, Mark A. Smith and Wing-Sau Young, have helped me through the many trials and tribulations in my personal and academic life, and kept me sane all this while. Mark, in addition to being a friend, has been the best mentor that I could have asked for. I am grateful to My undergraduate advisor, Ellen Hildreth for encouraging me to come to MIT, and my high school computer science teacher, the late Mr. Benjamin, for inspiring me to study computer science.

Finally, I am grateful to my parents, Don Lee and Pei-lei Liang, for making it all possible.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

This thesis explores the topic of recognizing and classifying people by their intrinsic characteristics estimated from video sequences of their walking gait. We have designed an image-based representation for the overall instantaneous appearance of human walking figures that facilitates the recognition and classification of people by their gait. In addition, we have developed a suite of representations that integrate these instantaneous appearance features over time to arrive at several types of gait sequence features that can be used to extract high level characterizations, such as gender and identity, of the walking subjects. These time-integration methods, spanning a range of coarseness of aggregation, are designed to answer the question, "How much information is contained in the time domain of gait appearance?" These gait features are tested on video data we collected to simulate realistic scenarios.

## 1.1 Motivation

Gait is defined as "a manner of walking" in Webster's New Collegiate Dictionary. However, human gait is more than that: it is an idiosyncratic feature of a person that is determined by, among other things, an individual's weight, limb length, footwear, and posture combined with characteristic motion. Hence, gait can be used as a biometric measure to recognize known persons and classify unknown subjects. Moreover, we extend our definition of gait to include the appearance of the person, the aspect ratio of the torso, the clothing, the amount of arm swing, and the period and phase of a walking cycle, etc., all as part of one's gait.

Gait can be detected and measured at low image resolution from

video, and therefore it can be used in situations where face or iris information is not available in high enough resolution for recognition. It does not require a cooperating subject and can be used at a distance. In addition, gait is also harder to disguise than static appearance features, such as the face. Johansson [17] had shown in the 1970's that observers could recognize walking subjects familiar to them by just watching video sequences of lights affixed to joints of the walker. Hence, in theory, joint angles are sufficient for recognition of people by their gait. However, recovering joint angles from a video of walking person is an unsolved problem. In addition, using only joint angles ignores the appearance traits that are associated with individuals, such as heavy-set vs. slim, long hair vs. bald, and particular objects that one always wears. For these reasons, we have included appearance as part of our gait recognition features.

## 1.2 The Challenges

The challenges involved in gait recognition include imperfect foreground segmentation of the walking subject from the background scene, changes in clothing of the subject, variations in the camera viewing angle with respect to the walking subjects, and changes in gait as a result of mood or speed change, or as a result of carrying objects. The gait appearance features presented in this thesis will tolerate some imperfection in segmentation and clothing changes, but not drastic style changes such as pants vs. skirts, nor is it impervious to changes in a person's gait. The view-dependent constraint of our gait appearance feature representation has been removed in a joint project with Shakhnarovich and Darrell [37] by synthesizing a walking sequence in a canonical view using the visual hull [23, 27, 26] constructed from multiple cameras.

## 1.3 Related Work

There have been may studies done in the area of gait recognition and understanding. They fall into two classes, those that examine the human ability to interpret gait, and those that develop computational algorithms for gait recognition and understanding. We will introduce first the psychophysical evidence for gait recognition, followed by a brief summary of computational algorithms for gait recognition. A more thorough discussion of computational gait representation is included in Chapter 7.

### 1.3.1 Psychophysical Evidence

The most recognized and earliest psychophysical study of human perception of gait was done by Johansson [16, 17] using moving light displays (MLD). MLD's are lights affixed to the joints of an active subject to produce visual stimuli for the observing human subject. The initial experiments showed that human observers are remarkably good at perceiving the human motion that generated the MLD stimuli—only 0.2 sec of the MLD stimuli was needed for observers to identify the motion as humans walking. In addition, Maas and Johansson [25] speculated that human observers might be able to identify gender from MLD stimuli.

Given Johansson's early success, Cutting, *et al.* [10] studied human perception of gait and their ability to identify individuals using MLD [9]. The authors used 6 walking subjects and collected the visual stimuli by placing reflective tape on their joints and recording their walking motion. Seven observers (including all 6 walking subjects) who were familiar with one another were asked one month later to identify their friends using the MLD stimuli. The authors reported that the observers correctly identified the walkers between 20% to 58% (chance performance was 16.7%) of the time, with better performance after the observer had gained some experience viewing the MLD. More interestingly, based on responses of introspection by the observers, the authors speculated that the observers consciously designed algorithms for person identification based on MLD stimuli rather than using direct perception—when the observer just "sees" the identity of the walking subject. They further speculated that human observers could be trained to use MLD to identify familiar walking subjects.

Kozlowski and Cutting [22] conducted an initial study of human perception of gender through MLD using a small set (3 men, 3 women) of walking subjects and 30 observers. Their results showed that human observers were able to correctly identify gender using full body joint markers approximately 70% of the time, although some subjects were consistently mis-classified. In addition, the authors discovered that the markers placed on the upper body of the walking subject appeared to be more significant for gender classification than those on the lower body joints. Barclay *et al.* [1] expanded the experiments to include more walking subjects (7 men and 7 women) for gender classification. The authors reported human observers achieved average correct identification rate of 65%. They also demonstrated that shoulder and hip sizes were significant factors for correct gender identification by human observers.

The above mentioned psychophysical studies, while interesting, led us to doubt the utility of using purely joint angles or joint locations for person recognition and gender classification. Particularly in the gender classification case, introspection shows that we can identify the gender of walkers at rates much higher than 65%. It is hence possible that we rely much more on familiarity cues, such as the length of hair, color and style of clothing to identify gender. These familiarity cues are much more readily available to the observer than joint locations. This is the assumption that led us to arrive at our own definition of gait, that is, gait for the purpose of identification and gender classification needs to include the appearance of the walking subject.

### 1.3.2 Computational Approach to Gait Recognition

There has been an explosion of research on gait recognition in recent years. We attempt to give a summary of some examples below, but this listing is by no means intended to be complete. A more in-depth treatment of the representational issues will be presented in Chapter 7.

Given the ability of humans to identify persons and classify gender by the joint angles of a walking subject, Goddard [13] developed a connectionist algorithm for gait recognition using joint locations obtained from moving light displays. Bobick and Tanawongsuwan [41] used joint angles of the lower body obtained from motion capture to determine the extent of identity information present in joint angle data. However, computing joint angles from video sequence is still a difficult problem, though several attempts have been made [4, 39, 12]. Particular difficulties of joint angle computation from monocular video sequence include occlusion and joint angle singularities. Self-occlusion of a limb from the camera view causes difficulties in tracking the hidden limb(s). Rehg and Morris [34] pointed out the singularity in motion along the optical axis of a camera.

There have been a number of appearance-based algorithms for gait and activity recognition. Cutler and Davis [8] used self-correlation of moving foreground objects to distinguish walking humans from other moving objects such as cars. Polana and Nelson [33] detected periodicity in optical flow and used this to recognize activities such as frogs jumping and humans walking. Bobick [3] used a time-delayed motion template to classify activities. Little and Boyd [24] used moment features and periodicity of foreground silhouettes and optical flow to identify walkers. Nixon, *et al.* [30] used principal component analysis of images of a walking person to identify the walker by gait. Shutler, *et al.* [38] used higher-order moments summed over successive images of a

walking sequence as features in the task of identifying persons by their gait. Johnson and Bobick [18] used static parameters of the walking figure, such as height and stride length, to identify individuals.

The correlogram method for differentiating between human and car motion was applied by BenAbdelkader [2] for the identification of individuals. The authors applied principle component analysis to the correlogram and used the principle components for recognition.

The work described in this thesis is most closely related to that of Little and Boyd [24]. However, instead of using moment descriptions and periodicity of the entire silhouette and optical flow of a walker, we divide the silhouettes into regions, compute statistics on these regions, and explore a number of methods to integrate the information in the time dimension. We also further study the capacity of our features in tasks beyond person identification, such as gender classification.

## 1.4 The Road Map

In the following chapters we will discuss the representations used to capture information about gait, present person recognition and classification results using these representations, and discuss alternative representations and future directions of research relevant to gait. The representations of gait include two components, the gait image representation and the gait sequence representation.

We take the view that gait can be directly measured from a video sequence by capturing the independent descriptions for each successive image instance. Chapter 2 presents a scheme to represent gait appearance related features in a silhouette image based on moments computed from localized regions of a video frame. Chapter 3 describes a number of methods to aggregate the gait image representation across time to arrive at compact representations of gait sequences. These aggregation methods vary in their amounts of abstraction of time, from the coarsest abstraction of time to no abstraction at all, and in the amount of abstraction of feature distributions. These aggregated gait representations are tested on a gait dataset we have collected to explore their recognition performances, and the results are presented in Chapter 4. In addition, in Chapter 5 we present gender classification results using the gait sequence representation and discuss the effect of noise on the effectiveness of a gait representation in performing recognition tasks. In Chapter 6 we briefly describe joint work that resolves the view dependent constraint of our gait recognition algorithm. Chapter 7 contains discussions on what has been learned in the process of work-

ing on gait recognition and gender classification from video data: the advantages and the shortcomings of the representations that we had experimented. Finally, we discuss alternative representations, direction of future work, and speculate on integration of gait information into a general recognition and surveillance scenario.

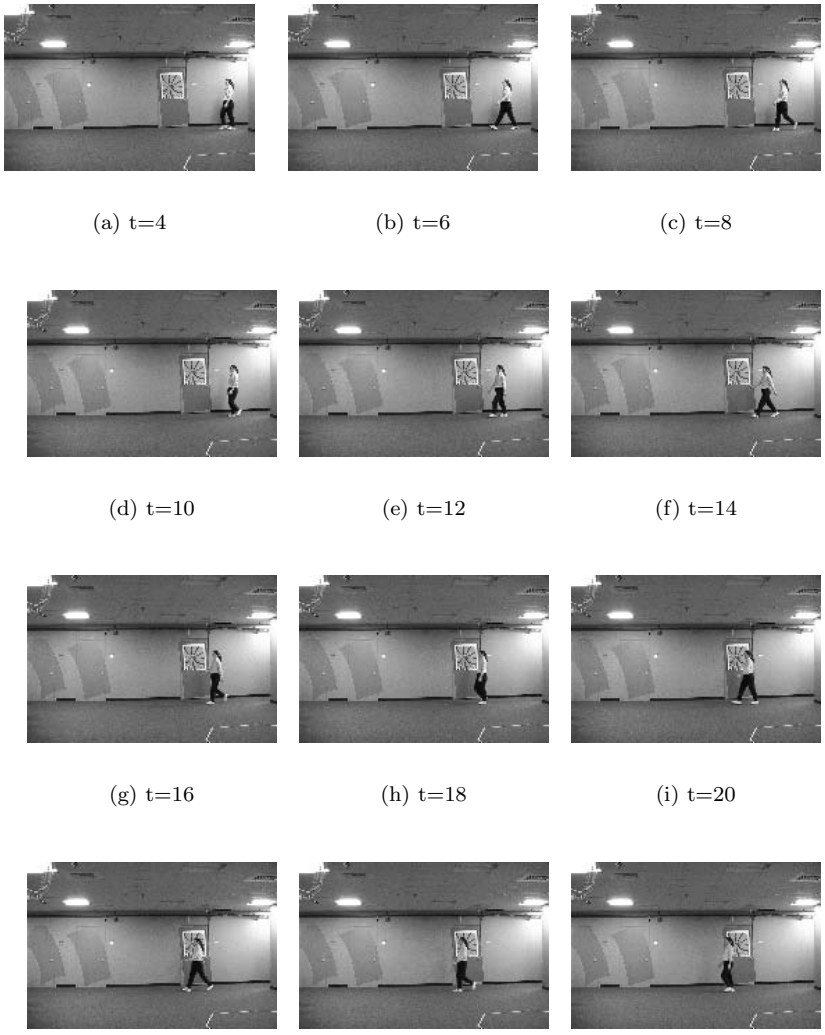
# Chapter 2

# Gait Image Representation

The usual full human walking cycle for one leg consists of the following stages: the initial contact of the foot to the ground, the stance position, to the double support stage, and then the swing of the leg to make the contact for the next step. The majority of people have symmetric walk, *i.e.,* the actions carried out by the two legs are nearly identical and are a half a cycle offset in phase from each other. Recognition by gait, in the traditional sense of joint angle description of gait, is equivalent to detecting the differences in the way individuals carry out the stages of a walking cycle and how they make the transitions between the four stages, *i.e.,* the underlying trajectories used by the limbs to accomplish the walking action. Because we take the position that the appearance of a walker is also indicative of the identity, our representation of the gait image includes descriptions of the appearance.

In this thesis, we are primarily concerned with a view-dependent method of extracting gait appearance features. We consider the canonical view of a walking person to be that which is perpendicular to the direction of walk. Figure 2.1 shows an example of the type of walking video data that our algorithm is designed to process and use to recognize to classify subjects. To simplify the problem of detecting the foreground walking figure, we use gait video data collected using a stationery camera. Also, we assume that only one subject is in the view at a time to simplify the person tracking problem. Finally, we assume that the video is sampled at a known fixed interval so that the time information can be meaningfully recovered.

To remove the effect of changing clothing colors, only the silhou-



(a) t=4 (b) t=6 (c) t=8

(d) t=10 (e) t=12 (f) t=14

(g) t=16 (h) t=18 (i) t=20

(j) t=22 (k) t=24 (l) t=26

Figure 2.1: An example of gait video used for recognition and classification.

ettes of the walking subjects are used in the gait representation. In addition, the silhouettes are scale-normalized to remove the effect of changing depth of the walking subject in the view of the camera. A side effect is that we lose the information about height and size in cases when the subjects are walking at the same depth from the camera. We also assume that the silhouette of the walker is segmented from the background using an existing algorithm (details to follow).

We would like our gait feature vector to have the following properties: ability to describe appearance at a level finer than the whole body; robustness to noise in video foreground segmentation; and simplicity of description and ease of extraction. The walking action involves movements of different components of the body, hence it is reasonable to describe the components separately. Ideally, one would like a description for each of the body components, such as the arms, the torso, and the legs. However, segmenting the silhouette into body components is a difficult problem, especially when the silhouette contains a significant amount of noise. A number of features intuitively come to mind that may measure the static aspects of gait and individual traits. One such feature is the height of an individual, which requires calibrating the camera to recover distances. Other features include the amount of bounce of the whole body in a full stride, the side-to-side sway of the torso, the maximum distance between the front and the back legs at the peak of the swing phase of a stride, the amount of arm and leg swing, etc. We do not use all of these features for various reasons such as inaccessibility (the side-to-side sway of torso) or difficulties in obtaining features, such as detecting the peaks of swing phase when foreground segmentation is noisy and includes shadows. We use a simple yet robust fixed grid system to describe localized silhouette shapes. We will discuss alternative gait image representations in Chapter 7.

Our algorithm for deriving a gait silhouette image representation involves the following steps:

1. The foreground consisting of the walking figure is extracted from a gait video sequence.

2. The silhouette is divided into regions using a fixed grid system.

3. Each region of a silhouette is modeled using a set of ellipse parameters.

4. The set of ellipse parameters from all regions of the silhouette plus one additional global silhouette parameter is concatenated to form a gait image feature vector.

In the following sections we describe in detail the steps taken to arrive at the gait image feature vector.

## 2.1 Preprocessing: Silhouette Extraction

Given a video of a subject walking across the plane of the image, we are only interested in the foreground walking subject. Hence the walking figure needs to be segmented out from the background. To that end, we use an existing adaptive background subtraction algorithm by Stauffer [40], which we summarize below.

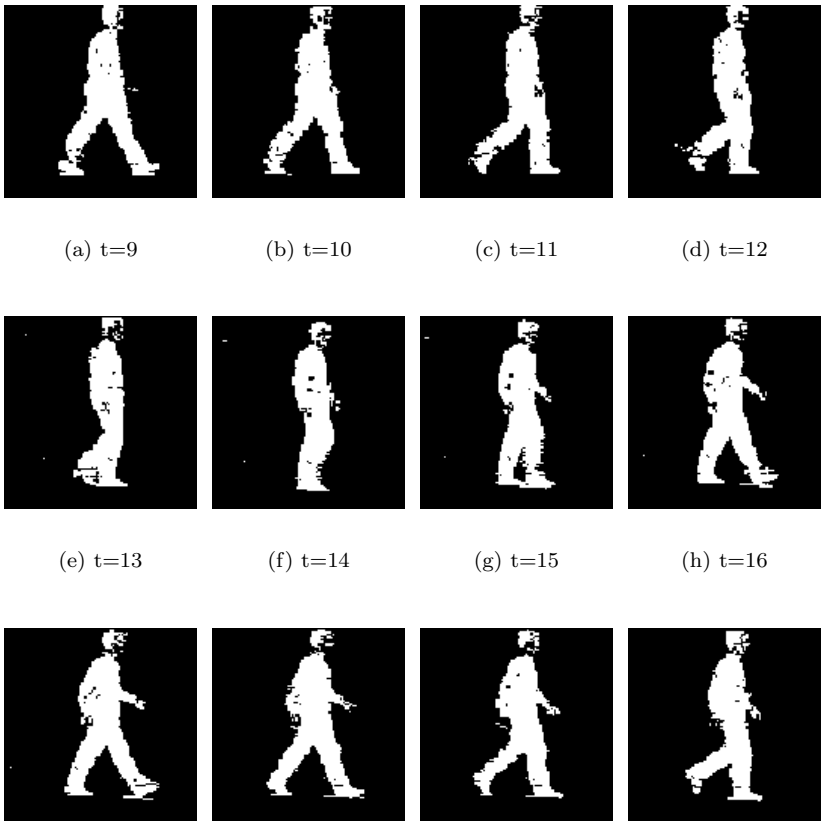
Stauffer developed a real-time adaptive background subtraction algorithm that models the background as a mixture of Gaussians, which are updated online to accommodate changing environmental effects, such as global lighting changes and local changes such as the shimmer of leaves in the wind. Specifically, the background description is a pixel-based model where each pixel is described by a number of Gaussian distributions. Each of these Gaussian distributions models a range of colors that are observed at that pixel. At each pixel, the Gaussian models are weighted by a factor that corresponds to the probability of observing a pixel value that is described by the given Gaussian, assuming the observation comes from the background. The weight of each Gaussian and the parameters of the Gaussian are updated by each new observation based on some rate of adaptation. An observation that agrees with an existing Gaussian model of a pixel increases the weight of that Gaussian model. Conversely, the lack of an observation that agrees with a Gaussian model decreases the weight of the model. A new observation that falls outside of the range described by the multiple Gaussians is considered a foreground pixel, and the least likely of the Gaussian models for the particular pixel is replaced with one that models the current observation. A new pixel value that falls into the range of the newly created Gaussian is still considered a foreground pixel, but each new observation of this kind increases the weight of the newly created Gaussian, until such point when the weight of the Gaussian passed a threshold making it a background color Gaussian model. In other words, an object that moves into a scene and stays put will eventually be considered part of the background. Hence the adaptive nature of this background subtraction algorithm.

Three Gaussians are used to model the background. Because our data was collected from indoor environments—with very little global lighting change such as that caused by moving clouds—the learning rate, *i.e.* the rate of adaptation of the background model, is set to

be very low. In addition, the first 10 frames of a video sequence are assumed to contain only background. Once the foreground image is produced using the algorithm described above, several levels of morphological operators [15] are applied to remove spurious small foreground objects and to connect parts of large foreground objects that became disconnected in the background subtraction process. In the case of our gait data, the remaining largest foreground object is always that of the walking subject. The foreground object is cropped and scaled to a standard size. Because we would like a gait representation that is independent of the color of clothing worn by the subjects, only the silhouette of the walking subject is retained. The color of clothing can be an important indicator of the identity of the subjects under certain circumstances. However, modeling clothing color distribution is not an essential part of this thesis and thus the color information is discarded. These cropped, centered, and scale-normalized silhouettes are the input to our gait recognition and classification algorithm. Figure 2.2 shows an example sequence of silhouettes extracted from a gait video sequence. The quality of silhouettes varied drastically over time, mostly affected by strong ceiling lights and the position of the subject relative to these lights. Figure 2.3 shows several examples of the amount of noise in the silhouettes that are caused by indoor lighting effects and small motions in the background such as the flutter of draperies. These silhouette examples show that our gait representation must be robust to these types and amounts of noise in the silhouettes.

## 2.2 The Image Representation

Our gait appearance feature vector is comprised of parameters of moment features in image regions containing the walking person averaged over time. For each silhouette of a gait video sequence, we find the centroid and proportionally divide the silhouette into 7 parts as shown in Figure 2.4a. The frontal-parallel view of the silhouette is divided into the front and back sections (except for the head region) by a vertical line at the silhouette centroid. The parts above and below the centroid are each equally divided in the horizontal direction, resulting in 7 regions that roughly correspond to: $r_1$, head/shoulder region; $r_2$, front of torso; $r_3$, back of torso; $r_4$, front thigh; $r_5$, back thigh; $r_6$, front calf/foot; and $r_7$, back calf/foot. These regions are by no means meant to segment the body parts precisely. For the present purposes, we are only interested in a method to consistently divide the silhouette of a walking person into regions that will facilitate the person recognition

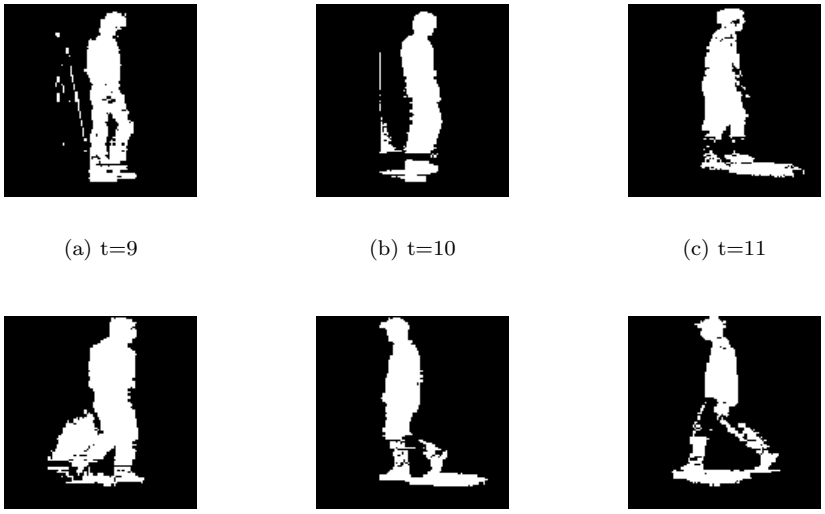


(a) t=9 (b) t=10 (c) t=11 (d) t=12

(e) t=13 (f) t=14 (g) t=15 (h) t=16

(i) t=17 (j) t=18 (k) t=19 (l) t=20

Figure 2.2: Examples of silhouettes extracted using an adaptive background subtraction algorithm by Stauffer [40].

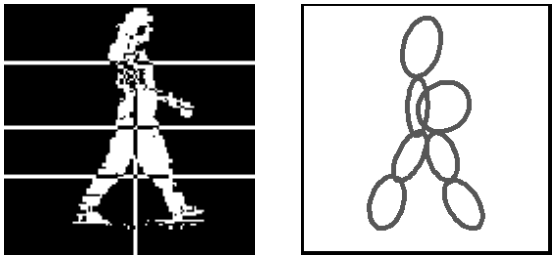


(a) t=9 (b) t=10 (c) t=11

(d) t=12 (e) t=13 (f) t=14

Figure 2.3: Examples of noisy silhouettes.

and gender classification tasks.



(a) (b)

Figure 2.4: The silhouette of a foreground walking person is divided into 7 regions, and ellipses are fitted to each region.

For each of the 7 regions from a silhouette, we fit an ellipse to the portion of foreground object visible in that region (Figure 2.4(b)). The fitting of an ellipse to an image region involves computing the mean and the covariance matrix for the foreground pixels in the region. Let $I(x, y)$ be the binary foreground image of a region to which we want to fit an ellipse. Assume that the foreground pixels are 1 and the background pixels are 0, then the mean $x$ and $y$ positions of the foreground pixels, or the centroid of the region, is

$$\overline{x} = \frac{1}{N} \sum_{x,y} I(x, y)x, \tag{2.1}$$

$$\overline{y} = \frac{1}{N} \sum_{x,y} I(x, y)y, \tag{2.2}$$

where $N$ is the total number of foreground pixels:

$$N = \sum_{x,y} I(x, y). \tag{2.3}$$

The covariance matrix of the foreground region is then,

$$\begin{bmatrix} a & c \\ c & b \end{bmatrix} = \frac{1}{N} \cdot \sum_{x,y} I(x, y) \cdot \begin{bmatrix} (x - \overline{x})^2 & (x - \overline{x})(y - \overline{y}) \\ (x - \overline{x})(y - \overline{y}) & (y - \overline{y})^2 \end{bmatrix}. \tag{2.4}$$

The covariance matrix can be decomposed into eigenvalues, $\lambda_1, \lambda_2$ and eigenvectors $\mathbf{v_1}, \mathbf{v_2}$ which indicate the length and orientation of the

major and minor axes of the ellipse:

$$\begin{bmatrix} a & c \\ c & b \end{bmatrix} \begin{bmatrix} \mathbf{v_1} & \mathbf{v_2} \end{bmatrix} = \begin{bmatrix} \mathbf{v_1} & \mathbf{v_2} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}. \tag{2.5}$$

The elongation of the ellipse, $l$, is given by

$$l = \frac{\lambda_1}{\lambda_2}, \tag{2.6}$$

and the orientation, $\alpha$, of the major axis is given by

$$\alpha = \text{angle}(\mathbf{v_1}) = \arccos\left(\frac{\mathbf{v_1} \cdot \mathbf{x}}{|\mathbf{v_1}|}\right), \tag{2.7}$$

where $\mathbf{x}$ is the unit vector $[1, 0]$. The orientation is only defined modulo $\pi$, so it is chosen to lie in a range of $\pi$ appropriate for each region of the silhouette. In other words, the range of orientation is adapted for each region feature, but the same set of orientation ranges is used for all walking silhouettes.

The ellipse parameters extracted from each region of the silhouette are the centroid, the aspect ratio ($l$) of the major to minor axes of the ellipse, and the orientation ($\alpha$) of major axis which forms the region feature vector $f(r_i)$,

$$f(r_i) = (\overline{x}_i, \overline{y}_i, l_i, \alpha_i), \text{ where } i = 1, \ldots, 7. \tag{2.8}$$

These moment-based features are robust to noise in the silhouettes obtained from background subtraction as long as the number of noise pixels is small and not systematically biased. The features extracted from each frame of a walking sequence consists of features from each of the 7 regions, *i.e.* the frame feature vector $F_j$ of the $j$th frame is,

$$F_j = (f(r_1), \ldots, f(r_7)). \tag{2.9}$$

In addition to these 28 features, we use one additional feature, $h$, the height (relative to body length) of the centroid of the whole silhouette to describe the proportions of the torso and legs. The intuition behind this measure is that an individual with a longer torso will have a silhouette centroid that is positioned lower (relative to body length) on the silhouette than someone with a short torso. The complete set of features extracted from each gait silhouette is summarized in Table 2.1.

$y$ coordinate of the whole body centroid

+

| 7 silhouette regions: | | 4 ellipse parameters: |
|---|---|---|
| head region | | $x$ coordinate of the region centroid |
| chest | $\times$ | $y$ coordinate of the region centroid |
| back | | orientation of the major axis |
| front thigh | | elongation |
| back thigh | | |
| front calf/foot | | |
| back calf/foot | | |

Table 2.1: A summary of the 29 gait image features extracted from each silhouette image.

## 2.3 Characterization of the Silhouette Image Representation

The representation of gait images we have chosen has several properties. The choice of silhouette over color images of the walking figure allows us to ignore the variations in clothing colors of the subject. However, suppose that the purpose of watching people walk is to collect any type of identifying information about the subjects. Then clothing colors may actually be indicative of the subject identity if enough data could be collected to model the distribution well. If a subject has never worn bright-colored clothing in the history of observations, then a walking figure with brightly colored clothing is unlikely to be that subject. Additionally, silhouette images also discards information about hair and skin color. These types of information could be incorporated into a gait appearance feature vector if need be.

Our choice of the silhouette of a walking figure and the point of view—the frontal parallel view—also has other side effects. For example, the majority of humans have fairly symmetric walk; that is, the left step and the right step look roughly the same from a side view. This is true of all the subjects in our data set. Hence there is no distinction between the left step and the right step of a walking cycle. In theory, one could walk in the half-step style—that is, from the double support to the stance stage on one leg and the initial contact to swing phase on the other leg—and the silhouette generated with such a walk would be the same as that generated by a full walking cycle. In reality, if one really tried to carry out a half step walk it would look very different

from a normal walking gait because the dynamics are significantly different: the forward momentum at the swing phase of the walking cycle has to be arrested, hence changing the dynamics.

We made a choice to divide the silhouette into a particular set of regions. The regions are divided into a fixed number, seven, and into a fixed grid. They do not correspond to biologically relevant segmentation of body components. These seven regions are chosen to be very easy to compute in contrast to methods that locate the joints and segment the body at those joints. There are numerous alternative methods to segment the silhouette. We will discuss some of the alternatives in Chapter 7.

# Chapter 3

# Gait Sequence Representation

The gait silhouette images representation described in the previous chapter produces for each gait video sequence a set of time series of all image features. Our goal is to use the information contained in the time series to extract higher level descriptions about the walking subject, such as the gender or the identity. To that end, we would like to answer the following question: "How much information is contained in the feature values that is indicative of gender or identity?"

This chapter discusses the various methods that we have tested in modeling the distribution of instantaneous gait image feature values and in aggregating them over time. In particular, we investigated four methods ranging in coarseness of time aggregation for modeling feature distributions:

1. Averaging gait appearance features across time (which amounts to the zeroth harmonic components of the time series).
2. Histogram of gait appearance features accumulated over time.
3. Fundamental and higher harmonics of the time series.
4. Direct matching of gait appearance features time series.

Each of these methods is intended to test assumptions about the nature of the time series that represents gait. The averaged appearance discards the time dimension, decouples the different parts of the body, and assumes that the distributions of shape appearance features—which are derived from moment ellipses—are completely described by their means

and variances, and that the mean can be optimally estimated by the average of samples. For example, this would be the case for a normal distribution with a given variance. The appearance histogram method differs from the averaged appearance method in that it does not assume any parametric distribution of shape appearance, but it still discards the time dimension and decouples the silhouette components. The relative phase portion of the fundamental harmonic components method preserves the coupling between the silhouette components. The magnitude portion of the fundamental harmonic measures the maximum size of change in appearance. The fundamental period retains some information about time. However, the only way the fundamental harmonic components could completely describe the time series of gait image features is if the time series were perfectly sinusoidal. The addition of higher harmonics to the fundamental harmonic components generalizes the shape of the time series signals, although it still assumes periodicity. The direct matching of gait appearance features acts as a baseline comparison that includes no time aggregation.

## 3.1 Re-cap of Gait Image Appearance Features

To recap, we have 29 features that are measured directly from a silhouette of a walking human figure:

- The relative height of the centroid of the whole body, which captures the intuitive notion of whether a person has long or short torso relative to his or her body length (a single feature).
- Four ellipse parameters: the x, y coordinates of the centroid, the orientation, and the elongation, times each of seven components of the body silhouette: the components that roughly correspond to head region, chest, back, front thigh, back thigh, front calf/foot, and back calf/foot, giving a total of 28 features.

These 29 features are extracted from each frame of a walking silhouette sequence. Hence, each video sequence is reduced to a representation of 29 time series of these features, an example of which is shown in Figure 3.1.

These time series are cumbersome descriptions of the appearance of a person's gait and may not lend themselves to robust generalization. The remaining sections of this chapter describe the various methods we have used to aggregate the 29 time series to generate composite gait sequence features.

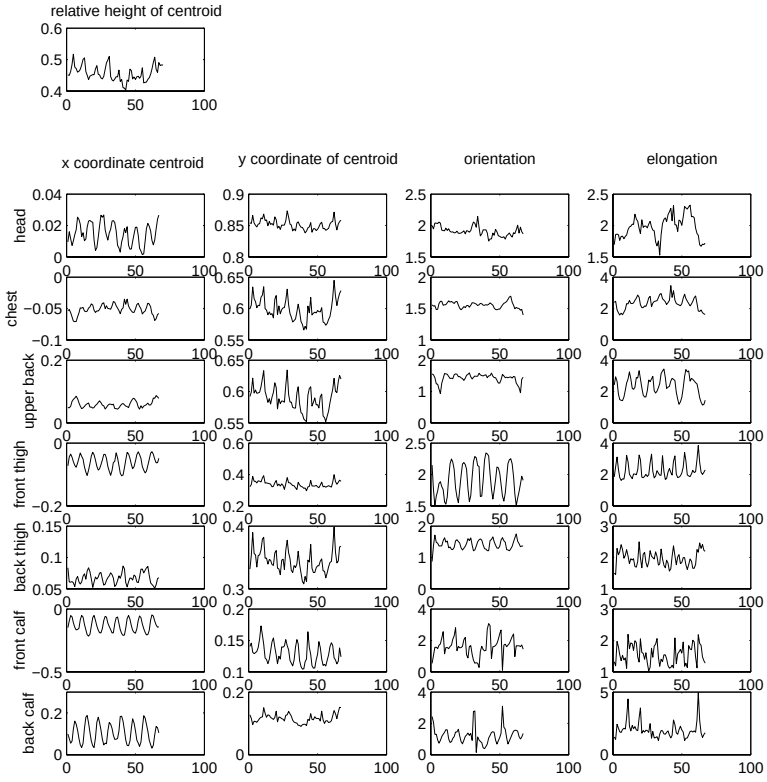



Figure 3.1: An example of the 29 gait appearance features time series from one walking sequence.

## 3.2 Average Appearance Features

The simplest and most compact way to summarize the set of 29 gait feature time series is to assume that all of the features are normally distributed and hence can be represented easily by their means and standard deviations. Specifically, the gait average appearance feature vector of a sequence $s$ is,

$$s = (\text{mean}_j(h_j), \text{mean}_j(F_j), \text{std}_j(F_j)), \tag{3.1}$$

where $j$ is a time index and $j = 1, \ldots,$ last frame, $h_j$ is the relative height of the whole body, and the $F_j$'s are the ellipsoidal descriptions of the 7 silhouette regions, and $s$ is 57-dimensional. This feature set is very simple to compute and robust to noisy foreground silhouettes. Intuitively, the mean features describe the average-looking ellipses for each of the 7 regions of the body; taken together, the 7 ellipses describe the average shape of the body. The standard deviation features roughly describe the changes in the shape of each region caused by the motion of the body, where the amount of change is affected by factors such as how much one swings one's arms and legs. The mean of the relative height of the body centroid is used to capture the intuitive concept of the relative size of the torso to body length. While people generally walk with some amount of bounce in their step cycle, the silhouettes that we use are centered on the centroid of the subject, hence factoring out most of the bounce, or equivalently, the standard deviation of the height of the silhouette centroid.

The Euclidean distance in the 57-dimensional gait average appearance feature space is used to compare two gait video sequences for their resemblance. However, the dynamic ranges of the dimensions differ drastically from one another, resulting in the dimensions with large dynamic range being over-represented in the distance computation. We start with the simplifying assumption that all dimensions of the gait average appearance features are equally significant in capturing the differences between gait video sequences and normalize each dimension by subtracting out the mean of that dimension and then dividing by the standard deviation. The Euclidean distance can either be computed using the normalized features, or weighted by the significance of each dimension in the recognition/classification task. Specifically, the distance $d$ between two gait average appearance features $w$ and $v$ is,

$$d^2 = (w - v)WC_s^{-1}(w - v)^T \tag{3.2}$$

where $C_s^{-1}$ is the covariance of the gait average appearance feature, $s$, and $W$ is a weighting factor for each dimension. We make the sim-

plifying assumption that the different dimensions of the gait average appearance feature are independent, hence both the weight matrix and the covariance matrix are assumed to be diagonal. The weighting factor $W$ is used either to select some of the feature components for distance computation or to weight feature components by their effectiveness in recognition and classification tasks. The details of the weighting factor will be described in the next chapter.

Figure 3.2 shows an example of distances computed using the gait average appearance features between pairs of 40 gait sequences selected from our database. All the sequences for each individual are consecutive, so their similarity is evident in the block structure of the matrix

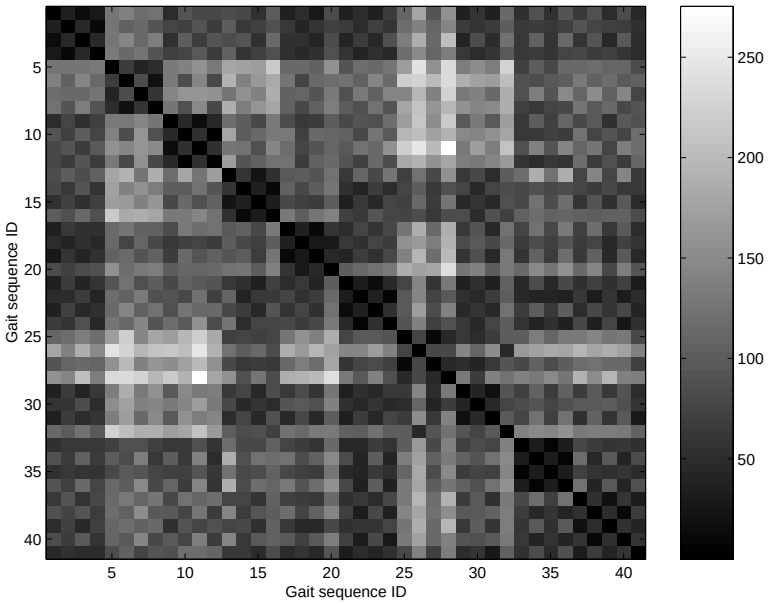



Figure 3.2: An example of pairwise distances between sequences using the gait average appearance feature. The diagonal is self-distance, while the block structures reveal the extent of the consistency of distance as a measure of identification.

## 3.3 Appearance Histogram

The second method of feature time aggregation is the appearance histogram. The means and the standard deviations of the average appearance gait feature are very coarse measures of the distribution of the 29 gait image features over time. While one could model the feature distributions with some other parametric models that are more suitable for the data, we instead simplify the modeling process with a non-parametric distribution, the histogram. The only parameters that need to be globally assigned for the histogram of each feature are the number of bins and size of each bin. The similarity between two gait sequences can be easily measured by comparing their complete sets of histograms.

The range for each feature histogram is based on the mean and the standard deviation of the sequence features, *i.e.*, the average appearance described in the previous section. For each of the 29 features extracted from a gait silhouette, $f_i$, where $i = 1, \ldots, 29$, the edges of the histogram of the feature are given by:

$$\text{left edge}(f_i) = \min_s(\text{mean}_t(f_i(s,t))) - \max_s(\text{std}_t(f_i(s,t))), \quad (3.3)$$

$$\text{right edge}(f_i) = \max_s(\text{mean}_t(f_i(s,t))) + \max_s(\text{std}_t(f_i(s,t))) \quad (3.4)$$

where $s$ is the index for gait sequences in our database, and $t$ is the time frame index for each gait sequence. This range accommodates almost all feature values and, in practice, results in the histogram of each sequence feature spanning less than half of the range.

A good choice for the number of bins in a histogram depends on the amount of data to be used. The gait silhouettes display a periodic nature with a period of between 6 to 9 frames for most people walking in their normal speed (see the next section for the fundamental period of a walking silhouette). The primary gait database used in this thesis contained between 55 to 80 frames for each walking sequence. Balancing between having good resolution in the histogram bins and maintaining a good estimation of the distribution given the number of samples per sequence, we conclude that between 15 to 30 bins will be sufficient. We arbitrarily chose to use 20-bin histograms for all features.

Given the 29 gait image features extracted from each frame of a gait video sequence, we tally the features each into a 20-bin histogram, resulting in a $20 \times 29$ matrix of gait features. Each histogram is normalized to sum to 1, making them probability distribution functions. Figure 3.3 shows an example of the gait appearance histogram for a particular walking video sequence.

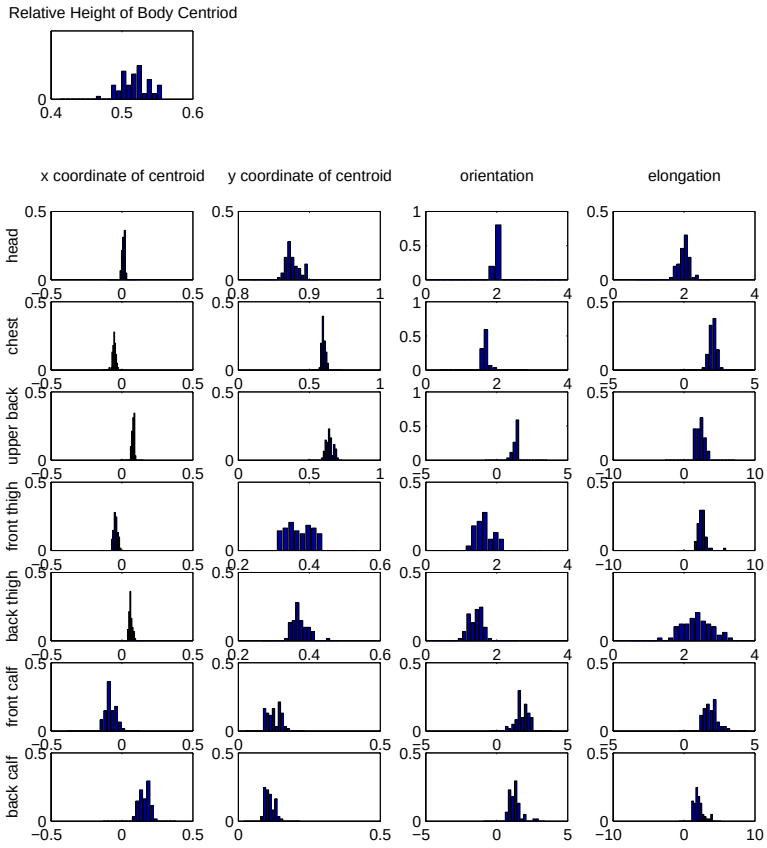



Figure 3.3: An example of gait appearance features binned over time from one walking sequence.

The similarity between two gait sequences $s$ and $t$ is measured by a normalized correlation of the histograms and summed over the 29 features, i.e.,

$$d(s,t) = \sum_{i=1}^{29} W_i h_i(s) \cdot h_i(t) \tag{3.5}$$

where $W_i$ is the weighting factor for feature component $i$. Summing over all 29 inner products serves to increase the signal-to-noise ratio of the similarity measure. The similarity score of pair-wise comparison of sequences looks very much like that shown in Figure 3.2, except with the intensities reversed because this is a similarity score, not a distance.

While the average appearance and the appearance histogram both capture the distribution of the gait image appearance features, they discard information about the correlation between different regions of the silhouette. In other words, one could take a gait video sequence, cut each image into the 7 image regions described in the previous chapter, shuffle each region in time independently of any other region, reassemble the video sequence and the resulting gait sequence would have exactly the same average appearance and the same appearance histogram as the original video sequence. The time dimension has been discarded by both types of features. To recover from this shortcoming, we investigate two additional types of features which do take into consideration the time dimension within each image region and between image regions: (1) the harmonic decomposition and (2) direct sequence matching using dynamic time warping.

## 3.4 Harmonic Decomposition

Walking is a mostly periodic activity [29, 28]. Hence, it seems natural to use the harmonic components as gait features. We use the Fourier decomposition of the time series of the gait image features as the basis from which to extract the fundamental and higher order harmonics. Intuitively, the magnitude measured at the fundamental frequency is a measure of the overall change undergone by the corresponding feature, and the relative phase between different time series is an indication of the time delay between the different features. The higher harmonics measured with respect to the fundamental harmonic describe the non-sinusoidal but still periodic trajectory that a feature undergoes.

A full walking cycle, or a stride, is comprised of two steps, the left step and the right step. However, because we are only using the silhouette of the walking video and our image plane is parallel to the path

of walk, it is difficult to distinguish the left and the right step of each walking cycle assuming that the two steps are symmetric. The only difference stems from a small change caused by perspective distortion. Therefore, the fundamental period of each time series consists of half of a walking cycle, that is, either the left step or the right step. On the other hand, most humans have slightly asymmetric gait between the left step and the right step caused by minor differences in the lengths of the two legs and their weight bearing capabilities.

To determine if the asymmetric gait is detectable from the time series of the gait image features, we computed the Fourier components of the time series to extract the fundamental period (to be discussed in the following section). The majority of humans take slightly under one second for a full stride. Thus, under the video sampling rate of 15 frames per second, that translates to a period of approximately 15 frames. Our analysis shows that the dominant period of sequences to lie between 6 and 9 frames. In addition, the power spectra sometimes shows a dip at the frequency corresponding to the full stride period. When the power spectra do show a high magnitude at the frequency corresponding to the full stride, it is not stable across different sequences of a subject taken on the same day. Hence we conclude that the slight asymmetry in the walk of normal subjects is either not detectable or cannot be accurately detected from our feature set. The Fourier component corresponding to the half stride, on the other hand, is always present. Thus we take the fundamental period to be that of one step.

Because our gait data are short video sequences ranging from 50 to 80+ frames, if we take the Fourier transform directly there is a lack of resolution in the spectral domain. The time series need to be zero-padded so that the spectral domain can be sampled more densely. To make the comparisons between different sequences easier, all sequence signals are zero-padded to the same length. We chose to zero-pad the signals to a sample length of $N = 400$. Given a video sampling rate of 15 frames per second, and the fundamental period of the step at between 6 to 9 frames, a sample length of 400 gives us resolution of approximately 1/6 of a frame in the range of the fundamental period. The discrete periods that we are able to extract are: 9.1, 8.9, 8.7, 8.5, 8.3, 8.16, 8, 7.8, 7.7, 7.5, 7.4, 7.27, 7.14, 7, 6.9, 6.8, 6.67, 6.56 frames per step. As we will discuss later, our gait video sequences contain a minimum of three full strides and a maximum of a little more than four full strides, giving us between six to nine walking steps to estimate the fundamental period of one step.

To distinguish the harmonic decomposition feature from the average appearance features, we remove the mean of all components, thus

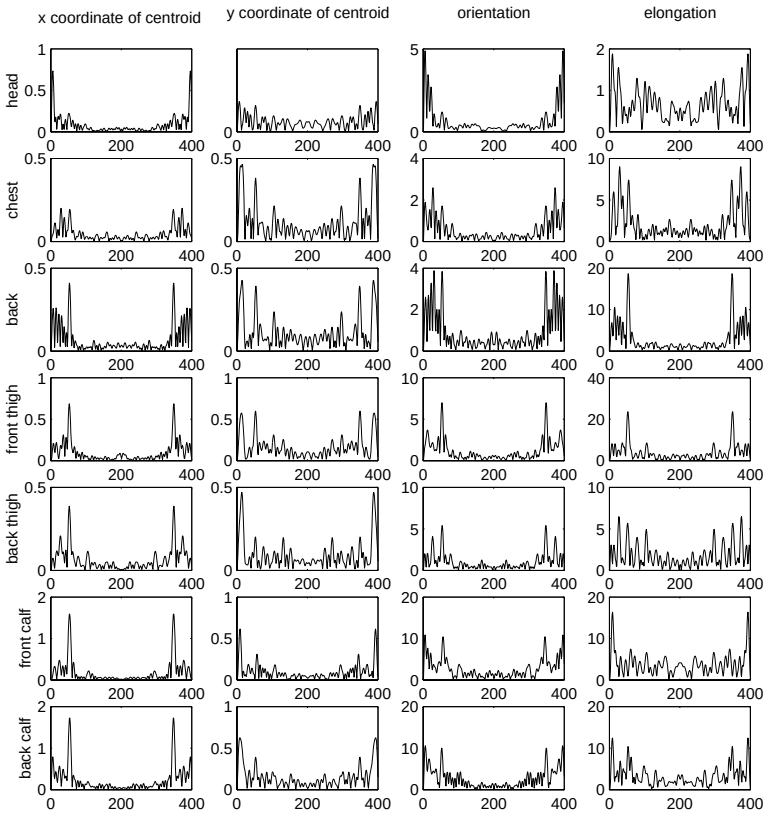



Figure 3.4: An example of the power spectra computed from the 28 gait image feature time series of a walking sequence.

setting the DC component of the Fourier transform to 0. In fact, the mean components of the gait average appearance features are the zeroth harmonic components. The harmonic analysis of time series is only applied to features extracted from the 7 component regions of the silhouette. Thus we are left with 28 gait image feature time series from which we compute the Fourier transform. Figure 3.4 shows an example of the power spectra of the 28 zero-padded signals. Some of the power spectra appear to have dominant peaks, while others lack such peaks. Most of the spectra show a fundamental frequency, some even have sizable magnitude in the second harmonic, but few show any obvious third harmonic. Moreover, there are several reasons we can only reasonably

expect to recover the first and the second harmonic. First, the higher harmonics have lower amplitude and are therefore more susceptible to noise. Second, because our subjects do not have perfectly periodic walks, localization of the fundamental frequency contains some error, which are amplified at the higher harmonics, thus further increasing the amount of noise in the magnitude and phase estimates at the higher harmonics.

### 3.4.1 The Fundamental Harmonic

Our gait fundamental spectral decomposition feature vector for a sequence is

$$t = (\Omega_1, |X_i(\Omega_1)|, \text{phase}(X_i(\Omega_1))), \tag{3.6}$$

where

$$X_i = \text{DiscreteFourierTransform}(F_{j=1\ldots last}(f(r_i))), i = 1 \ldots 28, \tag{3.7}$$

$\Omega_1$ is the fundamental walking frequency of a given sequence—which in the case of silhouettes, corresponds to a single step—and $i$ indicates the type of feature from the four ellipse descriptions of the seven silhouette regions. Intuitively, the magnitude feature components measure the amount of change in each of the 7 regions due to motion of the walking body, and the phase components measure the time delay between the different regions of the silhouette.

Because of noise in the silhouettes, and the fact that subjects do not have perfectly periodic walks, the time series of region features is also noisy. Thus, the power spectra of many region features do not show an obvious dominant peak indicating the fundamental walking frequency. Some even have a component at some very low frequency whose magnitude is much larger than that of the real fundamental walking frequency. This was investigated and it was discovered that this low frequency component corresponds to particularly strong shadows caused by a ceiling light that appear in one of the background from which we collected gait videos on two different days. Even when the peak frequencies were found in a silhouette region feature, they often did not agree between different region features. Therefore, we use a normalized averaging of power spectra of all region features resulting in a much more dominant peak frequency, $\Omega_1$, that is also consistent across all signals:

$$Z = \sum_{i=1}^{28} \frac{|X_i|}{\sum_\omega |X_i(\omega)|}. \tag{3.8}$$

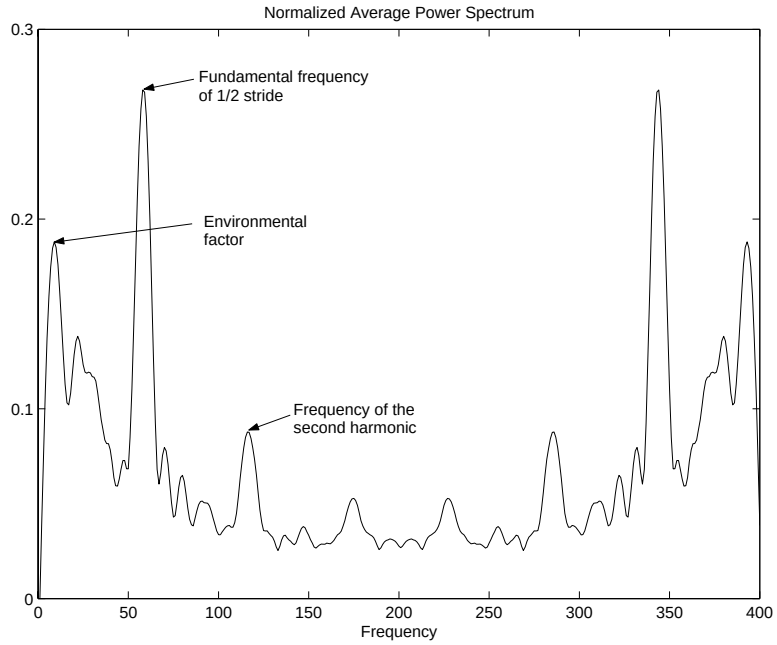



Figure 3.5: The normalized average power spectrum of 28 gait image features. While this particular average power spectrum shows the global peak at the fundamental walking frequency, there are others that show the global peak at a much lower frequency that may corresponding to environmental effects.

Figure 3.5 shows an example of the normalized average spectrum for a gait sequence. Even with the normalized average power spectra $Z$, the global highest peak sometimes still does not correspond to normal walking frequency but to external environmental factors such as strong lighting effects. Hence, we only look for dominant peaks in the region of the normalized average spectrum corresponding to normal walking frequencies of between 5 to 10 frames for a half-stride period.

The magnitude of each region feature at the fundamental frequency $\Omega_1$ can be used directly, but the phase cannot be used directly because each gait sequence is not predetermined to start at a particular point of a walking cycle. Hence the phase is not a stable feature of the gait. Instead, we need a feature related to phase that is translation independent over all gait video sequences and will capture the time

delay between different regions of the gait image feature. We compute the phases of all region features relative to the one particular region feature that is "most stable." The stability of a feature is determined by how closely its gait time series resembles a pure harmonic, because the phase can be more accurately estimated for a pure harmonic signal, given the same spectral sampling rate. The corresponding quality in the frequency domain is the sharpness of the power spectra around the fundamental frequency, which we measure using the 2nd moment of the power spectra about the fundamental frequency:

$$m_i^2(\Omega_1) = \sum_{\omega=1}^{200} \frac{|X_i(\omega)|}{\sum_\nu |X_i(\nu)|}(\omega - \Omega_1)^2, \tag{3.9}$$

where $i = 1, \ldots, 28$, indicates the 28 features extracted from all 7 regions of the silhouette. Because all gait image feature time series are real numbers, the Fourier transform is symmetric, thus we only need to compute the 2nd moment up to half of the frequency, *i.e.*, 200 instead of 400. The second moment about the peak frequency is computed for each sequence in our gait data base and then averaged. The feature with the smallest average $m_i^2(\Omega_1)$ is the feature whose phase is the most accurately estimated overall and is used as the standard phase from which all relative phases are computed. In our case, the phase feature with the highest second moment about the fundamental frequency is that of the $x$ coordinate of the centroid of the front calf/foot region. The sinusoidal purity of this component is apparent in Figures 3.1 and 3.4. The relative phase features preserve the coupling between components of the walking silhouette. The gait fundamental spectral decomposition feature vector has 57-1(the standard phase)=56 dimensions.

Because the fundamental harmonic features are composed of three types, the fundamental period along with the magnitude and phase of the fundamental frequency, the distance between two gait sequences is not a simple Euclidean distance. The fundamental period and the magnitude both reside in Euclidean space, hence the Euclidean distances can be used. However, the phase difference between sequences, $s$ and $t$, of feature $i$ is measured in angular distance:

$$d^2_{\phi_1,i}(s,t) = \min(|\phi_1|, |\phi_1 + 2\pi|, |\phi_1 - 2\pi|)^2, \tag{3.10}$$

where

$$\phi_1 = \text{phase}(X_{i,s}(\Omega_1(s)) - \text{phase}(X_{i,t}(\Omega_1(t)). \tag{3.11}$$

Thus, the overall distance between two gait sequences is taken to be

the sum of the Euclidean and the angular distance, *i.e.*,

$$\begin{aligned} d_1^2(s,t) &= \left(\frac{1}{\Omega_1(s)} - \frac{1}{\Omega_1(t)}\right)^2 \\ &\quad + \sum_i (|X_{i,s}(\Omega_1(s))| - |X_{i,t}(\Omega_1(t))|)^2 \\ &\quad + \sum_i d_{\phi_1,i}^2(s,t). \end{aligned} \tag{3.12}$$

### 3.4.2 The Second Harmonic

While the fundamental harmonic components capture the majority of the information of the time series of the 28 features extracted from gait silhouettes, they do not capture the subtle variations in the dynamics of different features. Higher harmonics are needed to capture these variations. Intuitively, the magnitude of the fundamental frequency together with the magnitude of the second harmonic and the phase of the second harmonic relative to the fundamental frequency provide a translation independent description of a signal that contains only first and second harmonics. We do not look beyond the second harmonic because the sampling rate and the amount of noise in the gait silhouette makes higher harmonic components unstable.

A visual inspection of clinical gait analysis data shows clearly that most time series of gait parameters are not pure sinusoids. Figure 3.6[1] shows an example of the knee joint angle time series of one full stride gait cycle as measured by tracking markers on the joints of a walking subject. The thick dotted band which is the reference knee joint angle for the general population with normal gait clearly shows that the time series of the knee angle contains higher harmonics, at least the second and the third harmonics.

While our silhouette representation cannot capture the amount of detail that is available from a joint angle time series in clinical gait analysis, it is highly likely that the higher harmonics are still present in gait silhouette images. However, it is much less clear if the higher harmonics, in particular the second harmonic, can be easily recovered from the time series of image features which were themselves derived from noisy silhouettes. Therefore, it is questionable whether the second harmonic component that we recover is a meaningful description of gait signature or not. We will answer this question in the next chapter through recognition results.

[1] This example was downloaded from the Clinical Gait Analysis web site at Curtin University of Technology, Australia, *http://guardian.curtin.edu.au/cga/index.html*

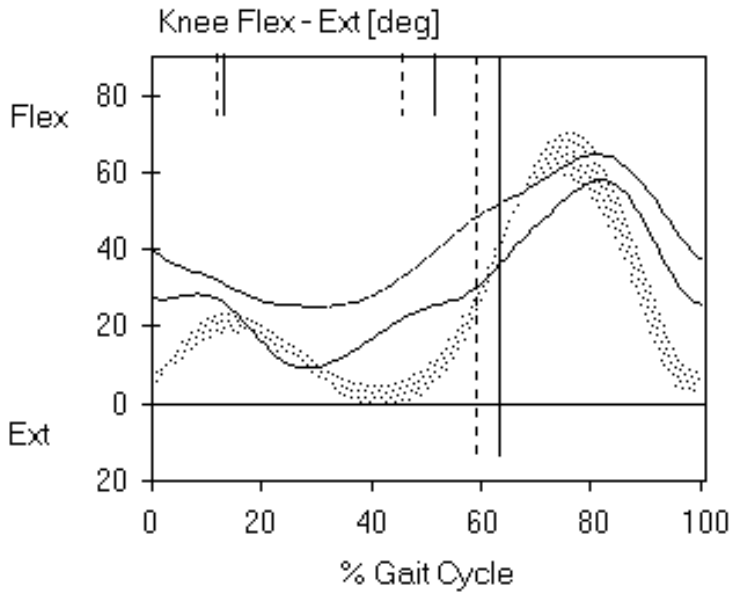



Figure 3.6: An example of knee flex-extension angle in a gait cycle. The thin lines are for the left and right knees of one subject, the dotted band is the range of knee angle time series for the general population with normal gait.

Based on the fundamental frequency computed using the algorithm given in the previous section, the second harmonic is assumed to be at double the frequency of the fundamental frequency, *i.e.*, $\Omega_2 = 2\Omega_1$, even though the local peak may not actually be at that frequency. In most cases, the local peak in the range of the second harmonic occurs in the range $(+1, -1)$ relative to our assumed second harmonic frequency. The magnitude is easily computed, while the relative phase of the second harmonic is measured relative to the phase of the fundamental harmonic as follows:

$$\phi_2 = \text{phase}(X_i(\Omega_2)) - 2 \times \text{phase}(X_i(\Omega_1)). \tag{3.13}$$

The distance between two sequences is computed in the the same way as in the case of the fundamental harmonic, except without the fundamental period component.

## 3.5 Direct Sequence Comparison by Dynamic Time Warping

As a baseline comparison to the gait sequence representations discussed in the previous sections which do contain varying amounts of time aggregation, we compared two sequences directly, without any time aggregation. We choose to use dynamic time warping, which is a method developed in speech recognition [36]. In particular, dynamic time warping (DTW) uses dynamic programming to compare two speech signals which may be uttered at different speeds. Its use in comparing gait image feature time series is appropriate because subjects often vary slightly their walking speed. The details of dynamic time warping are explained in Appendix A.

We use a version of DTW that tries to match two entire sequences (versus piece-wise matching). The gait image feature time series are preprocessed to extract lengths of sub-sequences that are integral multiples of the fundamental periods. These subsequences are aligned in phase based on the phase of the feature with the strongest sinusoidal signal. The fundamental period, $p_1$, was computed in harmonic decomposition. The feature with the purest sinusoidal signal—also derived according to harmonic decomposition in the previous section—is the $x$ centroid of the front calf. We retrieve the phase of this time series and locate the earliest point in time, $t_0$, that corresponds to zero phase. Then the subsequence from $t_0$ to $t_0 + 5 \times p_1$ of all features of the given gait sequence are taken as feature time series for comparison using time

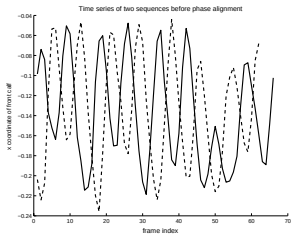
warping. This method of segmenting subsequences for comparison using DTW reduces the possibility that the time series representations of two gait video sequences are excessively penalized because they have different number of periods. At the same time, the phase difference between different features of the same gait video sequence is preserved because the fundamental integral multiples start at the same phase point of a fixed reference feature. Figure 3.7 shows two example time series of the same silhouette feature but from two different gait video sequences (a) in their original signal length and starting point, (b) after they have been cut to integer multiples of the period and aligned at a reference phase, and (c) after the dynamic time warping. The phase alignment in the second stage is not perfect because phase estimation is noisy.

Comparisons are made between two gait sequences, on a feature-by-feature basis, *i.e.*, the comparisons between two gait sequences $s$ and $t$ are,

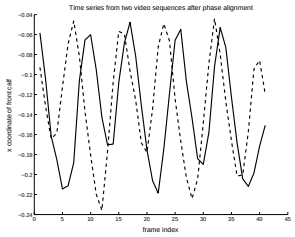
$$\begin{aligned} & dtw(s_1, t_1) \\ & \vdots \\ & dtw(s_i, t_i) \\ & \vdots \\ & dtw(s_{28}, t_{28}), \end{aligned}$$

where $i$ is the feature index. We again use the four ellipse parameters of the seven silhouette regions, requiring matching 28 feature sequences for each gait video sequence.

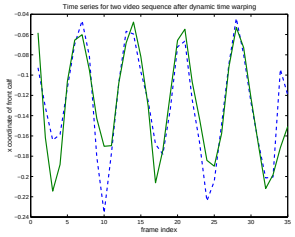
Dynamic time warping produces a warping cost for the pair of sequences being compared. The distance between two gait video sequences compared using dynamic time warping is taken to be the sum of the warping costs of the 28 pairs of feature sequences.




(a)




(b)




(c)

Figure 3.7: Examples of two feature time series (a) in their original length and starting point, (b) after they have been phase-aligned and cut to integer multiples of fundamental period to become DTW input, and (c) after dynamic time warping.

| sequence feature type | appearance-related | time-related |
|---|---|---|
| average appearance | means and standard deviations of image features | none |
| histogram appearance | histogram of image features | none |
| harmonic components | magnitude of Fourier components | fundamental period and relative phases |
| original time series | retains all information | retains all information |

Table 3.1: The four types of gait sequence features.

## 3.6 Summary of Aggregation Methods

We have introduced in this chapter four types of gait sequence features that result from different time-aggregations of the time series of gait image features, as summarized in Table 3.1. The average appearance feature discards the time dimension and uses the coarsest model to describe the underlying distribution of the image features: the means and the standard deviations. The appearance histogram feature also discards the time dimension, but it is a much more accurate model of the underlying distribution of the image features. The fundamental harmonic features capture the magnitude of change for each image feature, and it retains some time information, the fundamental period and the relative phases of the different features. The addition of the second harmonic features to the fundamental harmonic features gives a more precise description of the path traversed by each image feature in time. The baseline gait sequence feature retains all information available from the image features and directly matches the time series using dynamic time warping.

## 3.7 Feature Selection

Some of the features we have described have large dimensions, in particular the average appearance features and the harmonic features. It is likely the case that some of the features are more significant for recognition or gender classification purposes than other features. Ideally we would like to find an entire set of features that are best for identification or gender classification purposes. One method to test if a set of features is significant for gender classification or person recognition is to use analysis of variance (ANOVA). ANOVA is a standard technique

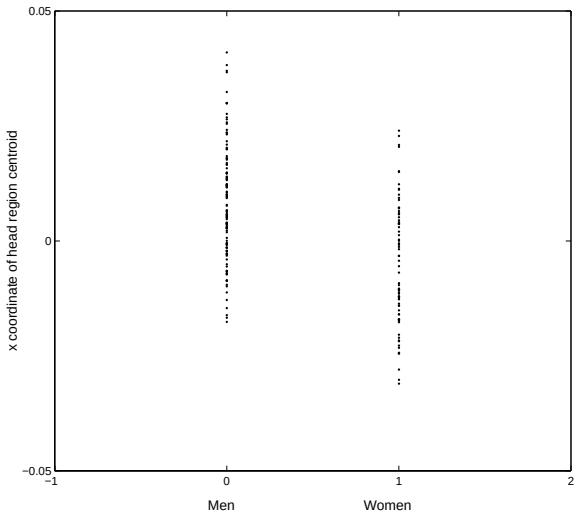



Figure 3.8: The $x$ coordinate of head region centroid for men and women.

for measuring the statistical significance of a set of independent variables in predicting a dependent variable. For a detailed treatment, see *Statistical Inference* by Casella and Berger [6].

ANOVA takes a single feature and the classes associated with the data samples and measures the significance of the class variables in predicting the means of the feature. The measure that ANOVA produces is the $p$-value for the feature set and the class variable. We will illustrate the intuition behind using the $p$-value for purpose of the feature selection with a concrete and simple example. We then apply the test to all features, making the simplifying assumption that individual features are independent.

We consider the problem of deciding whether a particular feature, such as the $x$ centroid of the head region, is useful for discriminating between genders. Figure 3.8 shows the $x$ centroid of head region for men and women. The $p$-value of ANOVA corresponds to the intuitive notion of how unlikely it is to see the two sets of data from men and women if the $x$ coordinate of the head region centroid of both men and women are drawn from the same Gaussian distribution. In this particular case, the $p$-value is numerically indistinguishable from zero, so we conclude that the feature is useful in discriminating between men

and women.

In general, ANOVA allocates the variations of a data set to different sources, which in our case includes either the gender of a subject or the identity of a subject. The total variation is broken into the variation between classes and the variation within classes. The variation within a class is considered random error. The variations between classes are generally not random because these are systematic variations between the genders or between individuals. The ratio of the between class variation to the within class variation is the $F$ statistic. The $p$-value is the probability of observing an $F$ statistic of this magnitude or bigger assuming that the samples for different classes are all drawn from the same Gaussian distribution.

We use the heuristic that if a feature set has a low probability of being drawn from one Gaussian distribution, then this feature set is indicative of the underlying classes. Hence a small $p$ value is an indication of the significance of a feature set for classification or recognition.

To find an entire set of features that are the most significant for recognition or gender classification, one needs to test the significance of all subsets of the features because there may be dependence between different features. The combinatorics makes this a computationally very intensive problem. We again resort to a heuristic method. Making the simplifying assumption that each feature is independent of other features, we could then test for the significance of each feature dimension individually. The best set is assumed to consist of the top few of the individually tested features.

While the ANOVA is suitable for feature sets such as the average appearance and the harmonic components features, it is not obvious how one applies a similar scheme to the appearance histograms and the directly matched sequences. In the case of the appearance histogram, we simply select the feature components by their properties, such as choosing only the centroid-related features or only the orientations, and test the recognition performance using the selected features. We do not use any feature selection on the original sequence. Hence DTW is done with the entire sequence without any ranking or deletion of the features.

# Chapter 4

# Recognition Experiments

Here we apply the four gait features described in the previous chapter: (1) the averaged appearance, (2) the appearance histogram, (3) the fundamental and the second harmonic components, and (4) direct comparison of gait image feature time series using dynamic time warping, to the task of recognizing people by video sequences of their walking gait. Our goal is to test the performance of each set of features under different circumstances.

## 4.1 The Data

We gathered gait data in indoor environments with different backgrounds, on four separate days spanning two months. Two examples of the indoor backgrounds are shown in Figure 4.1. The weather conditions outdoors span from the middle of winter to an unusually hot early spring day, resulting in our subjects wearing clothing ranging from sweaters and long pants to t-shirts, shorts and skirts during the data collection sessions on different days. Moreover, our indoor environment has overhead fluorescent lighting, which cast harsh shadows on the ground when subjects walk under them. Twenty-four subjects, 10 women and 14 men, were asked to walk at their normal speed and stride, back and forth, twice in front of a video camera that was placed perpendicular to their walking path. Because one direction of walk was predefined as the standard walking direction, the walking gait sequences going the opposite direction were modified to produce a walk-

ing sequence in the standard direction. This is achieved by reflecting about the $y$ axis the individual frames of the opposite direction walking sequence. In all, 194 walking sequences were collected, between 4 to 22 sequences for each subject, averaging 8 sequences per subject. A minimum of 3 complete walking cycles were captured, where a complete cycle takes two steps, left-right, or right-left. The videos were recorded using a Sony Digital Handycam VX2000 using the non-interlaced mode, resulting in videos of 720 by 480 pixels at 15 frames per second. We fixed focus and gain control to remove the flicker that may result from auto focus and auto gain control. The camera was positioned at a height of roughly 4.5 feet with the optical axis roughly parallel to the ground plane.

To obtain the silhouette of the walking subjects, we use an adaptive background subtraction algorithm [40] to segment out the walking person as a moving foreground object and scale-normalized it to fit in a $128 \times 128$ pixel binary image. An example of the foreground walking person is shown in Figure 4.2. Note that the foreground segmentation is not perfect: shadows on the ground and in some cases portions of the background are included. However, our gait representation tolerates this amount of noise in the foreground segmentation.

## 4.2 The Recognition Experiments

The most obvious test that can be performed is a simple pair-wise comparison of all sequences to all sequences. Specifically, each sequence in the gait database is treated as a query (or a probe) and compared against all other sequences in the database, which we call the library (or the gallery). Our gait database has the unique characteristic that subjects were filmed on different days; hence they were wearing different clothing, had different hair styles, and might have been in different emotional states at the time of data collection. These differences—in particular, clothing and hair style change—cause significant changes in the appearance of the gait silhouettes that are not present in gait video sequences collected on the same day. Figure 4.3 shows example silhouettes of one subject taken on three different days. Because of the difficulties in measuring one's emotional state, we make the simplifying assumption that it does not seriously affect one's gait under most normal circumstances.

We can exploit the uniqueness of our gait database to test the sensitivity of a gait representation to changes in appearance of the silhouettes caused by static appearance changes of the subject, namely

clothing, hair, and footwear changes, and to kinematic properties of one's gait which do not depend on external appearance changes. To that end, we devised two different tests:

1. The any-day test, where each sequence of the gait database is used as a query against the rest of the database, and
2. The cross-day tests, where gait sequences from one day are compared against sequences taken on other days.

The any-day test is a baseline experiment to examine the capability of a gait representation to capture any informative qualities of a subject's gait. The cross-day test examines the sensitivity of a gait representation to changes in the appearance of a person, such as the changes in clothing and hair style. Given that we have data collected on four different days, there are four sets of cross-day recognition tests, as listed in Table 4.1.

| Cross-day tests | Query sequences | Library Sequences |
|---|---|---|
| xdayA | from day A | from days B, C, D |
| xdayB | from day B | from days A, C, D |
| xdayC | from day C | from days A, B, D |
| xdayD | from day D | from days A, B, C |

Table 4.1: Definitions of the four sets of cross-day recognition tests.

For the remainder of this chapter, the gait representations described in the previous chapter are applied in recognition tests having the following components:

- A probe is a video sequence of one subject walking and its equivalent gait sequence feature representations.
- The library (or gallery) contains representations of individual gait sequences (instead of models of each subject) with a subject identity associated to each sequence representation. A subject has multiple instances represented in the library.
- A probe is correctly identified (at the $k$th retrieval) if, after ranking the library sequences by their distance/similarity to the probe, the $k$th ranked representation is the closest one to have the same subject identity as that of the probe sequence, regardless of which particular instance of the subject in the library is retrieved as the $k$th ranked match.

The classification method used is a nearest neighbor approach.

## 4.3 The Performance Measure

Given the five recognition tests described in the previous section, we need a performance measure to examine the effectiveness of our gait representations in each of the tests. To that end, we employ a standard measure used in the face recognition community, the cumulative match score (CMS), described in [32]. The CMS is a measure of the rate of correct identification as one increases the retrieval rate. We will describe in detail the formulation of the CMS curve. In addition, we provide a baseline comparison CMS produced using a random retrieval algorithm.

### 4.3.1 Cumulative Match Score

The cumulative match score answers the question "Is the correct answer in the top $k$ matches?" It is used in a closed-universe model for recognition, meaning that the correct answer is always in the library. Given a probe gait sequence, the sequences in the library are ranked according to their similarity (or distance) to the probe. Let $\mathcal{P}$ be the number of probes to be scored, and $\mathcal{R}_k$ be the number of these probes correctly identified within the top $k$ matches, the fraction of correct identification, or the CMS, is

$$\text{CMS}(k) = \frac{\mathcal{R}_k}{\mathcal{P}}. \tag{4.1}$$

In the case of the cross-day tests, the probe sequences are a subset of all the sequences taken on one day whose corresponding walking subjects are represented in at least one other day. This paring-down of the probe set is necessary to comply with the restriction of a closed-universe recognition task. The cross-day test library contains all sequences collected on other days. In the case of the any-day test, each probe has its own library/gallery, which is the the rest of the gait database, and the cumulative match score is averaged over all probes and all libraries.

### 4.3.2 Comparison Basis: Random Retrieval

In order to measure the effectiveness of a gait representation for the purpose of recognizing individuals, one needs to know the performance of a completely ineffective algorithm: one that randomly ranks the

sequences in a library. The performance of such an algorithm depends on the number of instances of a probe that are present in the library.

Let $N$ be the number of sequences in the library, and let $m_b$ be the number of instances in the library of a subject with the same identity as that of probe $b$. The probability that probe $b$ is correctly identify by the $k$th retrieval using a random-retrieval algorithm is,

$$P(k,b) = 1 - \frac{\begin{pmatrix} N - m_b \\ k \end{pmatrix}}{\begin{pmatrix} N \\ k \end{pmatrix}} \tag{4.2}$$

*i.e.*, one minus the probability that the probe is not correctly identified by the $k$th retrieval. The theoretical average CMS of the random-retrieval algorithm is thus,

$$\text{CMS}(k) = \frac{1}{\mathcal{P}} \sum_{b} P(k,b) \tag{4.3}$$

where $\mathcal{P}$ is the number of probes. Figure 4.4 shows the CMS curves for the five recognition tests using the random retrieval algorithm. Table 4.2 shows the CMS in text format for easier comparison to recognition results achieved using our gait representations.

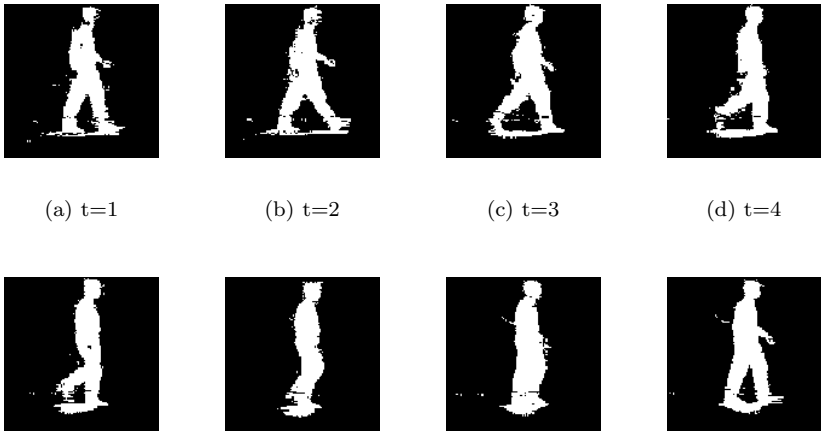
| | 1st | 5 % | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|
| any-day | 5 | 39 | 59 | 81 | 90 | 94 | 97 |
| xdayA | 5 | 32 | 54 | 76 | 88 | 94 | 97 |
| xdayB | 5 | 35 | 57 | 81 | 91 | 96 | 98 |
| xdayC | 5 | 34 | 53 | 77 | 89 | 95 | 98 |
| xdayD | 5 | 31 | 52 | 76 | 87 | 94 | 97 |

Table 4.2: The percentage of correct identification at the given percentage of recall using random retrieval.

(a)

(b)

Figure 4.1: Examples of indoor background for gait data collection.



(a) t=1 (b) t=2 (c) t=3 (d) t=4

(e) t=5 (f) t=6 (g) t=7 (h) t=8

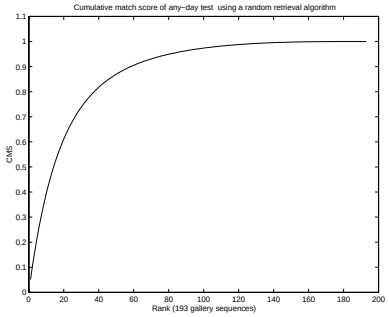
Figure 4.2: A sample sequence of the silhouettes of a subject after background subtraction.

(a) hair up, wearing pants

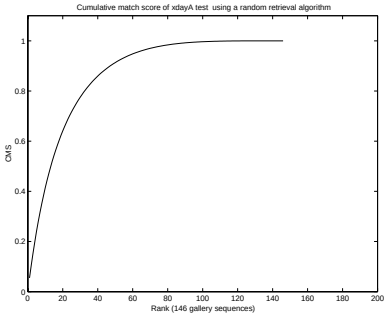
(b) hair down, dress and tights

(c) hair in pony tail, pants

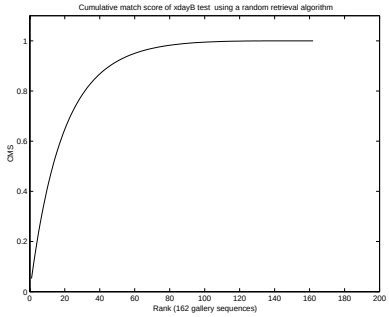
Figure 4.3: Silhouette of one walking subject from video sequences taken on three different days, with three different hair styles and two different types of clothing.



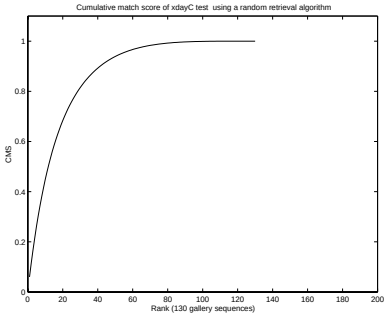

(a) any-day recognition test



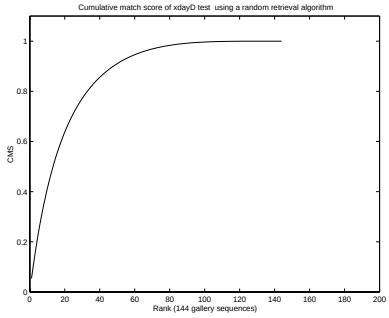

(b) xdayA recognition test




(c) xdayB recognition test




(d) xdayC recognition test




(e) xdayD recognition test

Figure 4.4: Theoretical average cumulative match score curves of five recognition tests using a random-retrieval algorithm for recognition.

## 4.4 The Recognition Results

The recognition results using each of the gait representations described in the previous chapter are presented below. To review, the four gait sequence representations are: (1) average appearance, (2) histogram appearance, (3) harmonic components, and (4) the original feature time series. Each is tested in several variations using different sets of weights on the components of the gait sequence features. The goal of these tests is to explore the capacity of these features to capture information which is significant in the five recognition tests described previously as well as to test the sensitivity of each feature to the changes in the appearance of the silhouettes caused by clothing changes.

### 4.4.1 Average Appearance

The average appearance gait feature vector is used in the five recognition tests with the following eight variations in the weights for each component $k$:

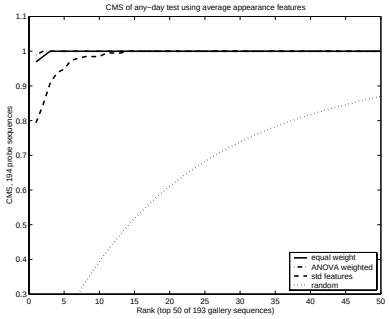
1. Equally weighted full set: $W_k = 1$ for all $k$.
2. ANOVA threshold: $W_k = 1$ if ANOVA results in $p < 10^{-9}$, otherwise $W_k = 0$. There are 41 average appearance feature components that pass this threshold.
3. ANOVA weighted: $W_k = \min(2, -\log_{10}(p_k)/9)$, *i.e.*, each component is weighted in proportion to the log of the reciprocal of the ANOVA $p$-value.
4. Centroid: $W_k = 1$ if feature component $k$ is the mean or the standard deviation of the centroid of a region, and 0 otherwise.
5. Orientation: $W_k = 1$ if feature component $k$ is the mean or the standard deviation of the orientation of a region, and 0 otherwise.
6. Elongation: $W_k = 1$ if feature component $k$ is the mean or the standard deviation of the elongation of a region, and 0 otherwise.
7. Mean components: $W_k = 1$ if feature component $k$ is the mean of any region, any ellipse parameter.
8. Standard deviations: $W_k = 1$ if feature component $k$ is the standard deviation of any region, any ellipse parameter.

In variation 3, the weights based on ANOVA $p$ value are designed so that a component feature is weighted 1 or larger if its corresponding ANOVA $p$ value passes the threshold of $p < 10^{-9}$, and less than 1 if it is larger. The recognition performances of all eight variations of the average appearance measures are listed in Tables 4.3 and 4.4; representative examples are shown in Figure 4.5.

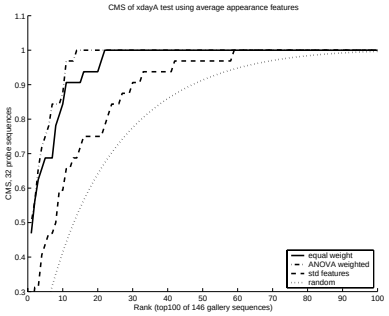
| any-day | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|
| equally weighted full set | 97 | 100 | 100 | 100 | 100 | 100 | 100 |
| ANOVA threshold | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| ANOVA weighted | 99 | 100 | 100 | 100 | 100 | 100 | 100 |
| centroid | 90 | 100 | 100 | 100 | 100 | 100 | 100 |
| orientation | 84 | 99 | 99 | 100 | 100 | 100 | 100 |
| elongation | 84 | 99 | 100 | 100 | 100 | 100 | 100 |
| mean components | 99 | 100 | 100 | 100 | 100 | 100 | 100 |
| standard deviations | 79 | 98 | 100 | 100 | 100 | 100 | 100 |

Table 4.3: Any-day recognition results using variations of average appearance gait features.

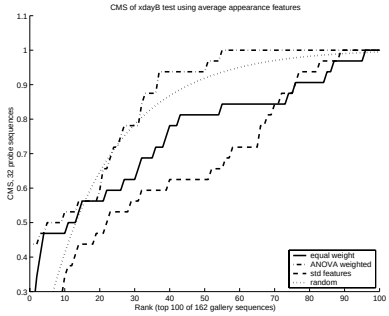
While the recognition rates of the any-day test appear impressive, closer examination shows that for the majority of the variations, the closest correct match for 97% to 98% of the probes is another sequence of the same subject collected on the same day. The exceptions are the orientation and the elongation variations with 94% and 92%, respectively, of the probes identified to another sequence of the subject collected on the same day. This indicates that the average appearance is highly sensitive to the changes in clothing style and background. Moreover, most of the correct matches in the library were of people walking in the same direction as in the probe sequence. We believe this bias is caused by the shadows on the ground casted by ceiling lights. Because we decided that all images of walking subjects should be from the same side of view, the walking sequences collected from the opposite views are reflected about the $y$ axis to make the data set uniform. However, while the silhouettes of the walking figure are symmetric when viewed from the left or right side provided the subject has a symmetric walk, the shadows on the ground do not have the same property, making the left side view of the walking subject significantly different from the right side view. Hence there is a need to test the recognition performance of the features using probe and gallery sequences that are collected on different days where there is a larger variation on the clothing styles of



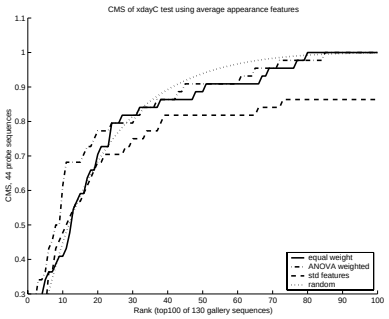

(a) any-day recognition test



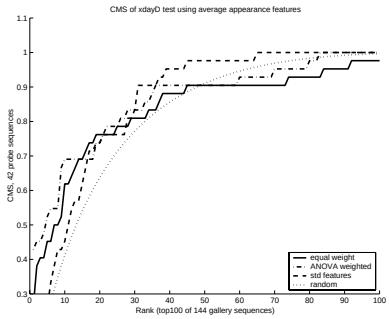

(b) xdayA recognition test




(c) xdayB recognition test




(d) xdayC recognition test




(e) xdayD recognition test

Figure 4.5: The cumulative match score curves of five recognition tests using the Euclidean distance between average appearance features.

the subjects and the backgrounds, including lighting variations.

The overall performance over the five recognition tests appear to favor the two variations of the ANOVA related weights, either the log-weighted or the thresholding of $p$-values. The ANOVA thresholded variation of the average appearance is slightly preferred because it uses fewer feature vector components. Most of the mean components of the average appearance features have small ANOVA $p$-values, and the overwhelming majority of the features that pass the threshold set on ANOVA $p$-values are the mean features. Hence, it is not very surprising that the mean variation of the average appearance features also performed well. All variations of the average appearance representations appear to perform better in the xdayA test than the other cross-day tests. Closer examinations of the ranked matches show that the worse performances, compare to xdayA, in the cross-day B, C, and D tests are the result of lack of similarity between the types of clothing that some of the subjects wore for the day B, C, or D gait sequence data and what they each wore on the other days. In other words, the query sequences show subjects wearing clothing that is substantially different from that in the library sequences. For example, the 7 query sequences with the worst match scores from day B are all from one subject who wore baggy pants and whose only representations in the library were sequences collected on day D when he wore shorts. For the same reason, recognition results of matching day D gait sequences against all other sequences suffer because of lack of a similar appearance model in day B for the same subject. Day C contains sequences of one subject wearing a short dress while the only other sequences in the database show her wearing pants. On the other hand, all the subjects recorded on day A and that also appear on one other day wore similar clothing on the other day(s).

| xdayA | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|
| equally weighted full set | 47 | 69 | 91 | 100 | 100 | 100 | 100 |
| ANOVA threshold | 44 | 78 | 94 | 100 | 100 | 100 | 100 |
| ANOVA weighted | 50 | 84 | 100 | 100 | 100 | 100 | 100 |
| centroid | 22 | 56 | 78 | 94 | 97 | 100 | 100 |
| orientation | 31 | 66 | 81 | 97 | 97 | 100 | 100 |
| elongation | 28 | 63 | 84 | 100 | 100 | 100 | 100 |
| mean components | 50 | 75 | 100 | 100 | 100 | 100 | 100 |
| standard deviations | 22 | 47 | 72 | 88 | 97 | 97 | 100 |
| xdayB | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| equally weighted full set | 25 | 47 | 56 | 69 | 81 | 84 | 91 |
| ANOVA threshold | 47 | 50 | 53 | 88 | 100 | 100 | 100 |
| ANOVA weighted | 44 | 50 | 56 | 84 | 94 | 100 | 100 |
| centroid | 16 | 47 | 50 | 53 | 59 | 69 | 78 |
| orientation | 31 | 75 | 78 | 84 | 94 | 100 | 100 |
| elongation | 13 | 34 | 56 | 84 | 88 | 94 | 97 |
| mean components | 47 | 63 | 69 | 75 | 84 | 97 | 100 |
| standard deviations | 0 | 28 | 44 | 59 | 63 | 72 | 94 |
| xdayC | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| equally weighted full set | 25 | 36 | 55 | 80 | 86 | 91 | 91 |
| ANOVA threshold | 30 | 43 | 55 | 75 | 82 | 89 | 95 |
| ANOVA weighted | 27 | 45 | 68 | 80 | 86 | 91 | 95 |
| centroid | 14 | 43 | 59 | 77 | 82 | 82 | 82 |
| orientation | 18 | 48 | 61 | 73 | 84 | 91 | 91 |
| elongation | 23 | 41 | 57 | 77 | 91 | 100 | 100 |
| mean components | 30 | 48 | 55 | 77 | 89 | 91 | 95 |
| standard deviations | 11 | 36 | 55 | 70 | 82 | 82 | 82 |
| xdayD | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| equally weighted full set | 26 | 50 | 69 | 81 | 88 | 90 | 90 |
| ANOVA threshold | 38 | 55 | 64 | 83 | 88 | 90 | 90 |
| ANOVA weighted | 43 | 55 | 69 | 83 | 90 | 90 | 95 |
| centroid | 26 | 43 | 50 | 62 | 74 | 86 | 90 |
| orientation | 21 | 50 | 67 | 83 | 88 | 88 | 90 |
| elongation | 17 | 48 | 74 | 88 | 90 | 98 | 100 |
| mean components | 33 | 50 | 74 | 86 | 88 | 88 | 90 |
| standard deviations | 5 | 38 | 57 | 81 | 95 | 98 | 100 |

Table 4.4: Cross-day recognition results using variations of average appearance gait features.

### 4.4.2 Appearance Histogram

The following variations of the appearance histograms are examined:

1. All histograms: $W_k = 1$ for all histogram components.
2. Centroid: $W_k = 1$ for centroid-related histogram components, 0 otherwise.
3. Orientation: $W_k = 1$ for orientation-related histogram components, 0 otherwise.
4. Elongation: $W_k = 1$ for elongation-related histogram components, 0 otherwise.

The recognition performance on the five recognition tests are shown in Figure 4.6 for a representative sample of the variations, and in Tables 4.5 and 4.6 for the complete set of histogram feature variations.

| any-day | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|
| all 29 histograms | 100 | 100 | 100 | 100 | 100 | 00 | 100 |
| centroid | 97 | 100 | 100 | 100 | 100 | 100 | 100 |
| orientation | 93 | 99 | 100 | 100 | 100 | 100 | 100 |
| elongation | 95 | 99 | 100 | 100 | 100 | 100 | 100 |

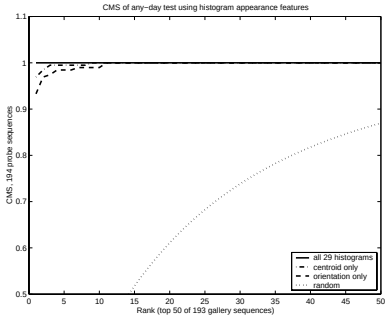
Table 4.5: Any-day recognition results using variations of histogram appearance gait features.

The recognition results based on the percentage of recall in Table 4.5 through Table 4.6 show that the appearance histogram of orientation components performs consistently better than the full set of 57-dimensional averaged appearance features and is better than the ANOVA thresholded 41-dimensional averaged appearance features beyond the 5% recall level. We conclude that the mean and standard deviations of averaged appearance features do not adequately represent the underlying distribution of gait image features, while a histogram approximates the entire distribution. The histogram representation is highly sensitive to changes in the appearance of the gait silhouettes collected on different days. The recognition results on the any-day test show that with the exception of orientation histogram, using the other three histogram measures resulted in 98% of the probes having the first correct match as another sequence of the same subject collected on the same day. The orientation histogram had only 94% of the probes matching to another sequence of the same subject collected on the same

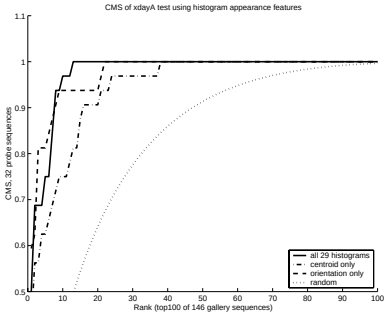
| xdayA | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|
| all 29 histograms | 50 | 84 | 100 | 100 | 100 | 100 | 100 |
| centroid | 44 | 69 | 88 | 97 | 100 | 100 | 100 |
| orientation | 59 | 88 | 94 | 100 | 100 | 100 | 100 |
| elongation | 38 | 75 | 81 | 97 | 100 | 100 | 100 |
| xdayB | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| all 29 histograms | 25 | 75 | 75 | 84 | 94 | 100 | 100 |
| centroid | 22 | 56 | 63 | 72 | 75 | 88 | 91 |
| orientation | 25 | 100 | 100 | 100 | 100 | 100 | 100 |
| elongation | 13 | 53 | 72 | 97 | 100 | 100 | 100 |
| xdayC | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| all 29 histograms | 45 | 70 | 84 | 91 | 91 | 93 | 95 |
| centroid | 32 | 66 | 68 | 84 | 91 | 91 | 91 |
| orientation | 50 | 77 | 86 | 98 | 98 | 100 | 100 |
| elongation | 16 | 55 | 66 | 77 | 84 | 91 | 95 |
| xdayD | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| all 29 histograms | 45 | 60 | 74 | 88 | 95 | 95 | 100 |
| centroid | 33 | 57 | 62 | 71 | 88 | 95 | 95 |
| orientation | 43 | 79 | 95 | 100 | 100 | 100 | 100 |
| elongation | 17 | 40 | 52 | 74 | 81 | 93 | 100 |

Table 4.6: Cross-day recognition results using variations of histogram appearance gait features.

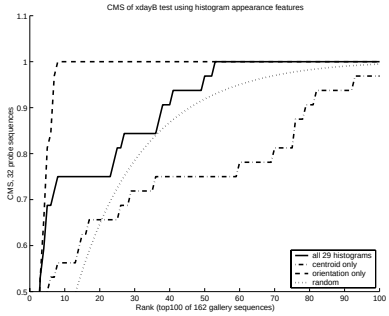
day, which explains why the orientation histogram performs better in the cross-day recognition test.



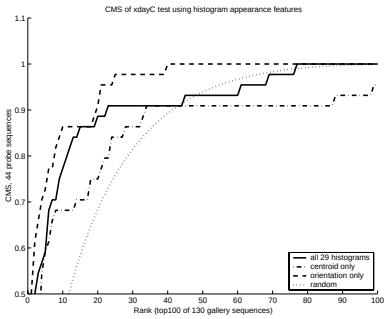

(a) any-day recognition test



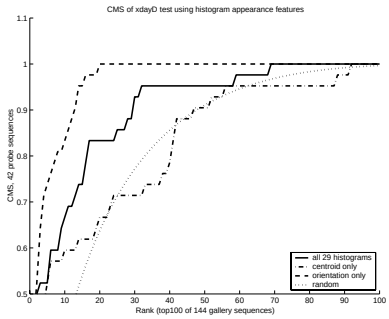

(b) xdayA recognition test




(c) xdayB recognition test




(d) xdayC recognition test




(e) xdayD recognition test

Figure 4.6: The cumulative match score curves of five recognition tests using histogram appearance features.

### 4.4.3 Fundamental Harmonic Components

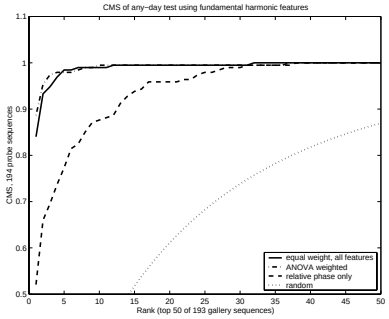
The following weight variations in the fundamental harmonic components are examined:

1. Equally weighted: $W_k = 1$ for all feature components.
2. ANOVA threshold: $W_k = 1$ if ANOVA results in $p_k < 10^{-9}$, otherwise $W_k = 0$. There are 32 fundamental harmonic feature components that pass this threshold.
3. ANOVA weighted: $W_k = \min(2, -\log_{10}(p_k)/9)$, *i.e.*, each component is weighted in proportion to the log of the reciprocal of ANOVA $p$-value.
4. Fundamental period: $W_k = 1$ for the fundamental period only, 0 otherwise.
5. Magnitude: $W_k = 1$ for the magnitude of the fundamental frequency of each feature, 0 otherwise.
6. Relative phase: $W_k = 1$ for the relative phase of the fundamental frequency of each feature, 0 otherwise.

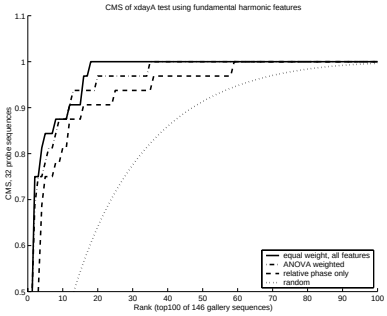
Representative examples of recognition performances are show in Figure 4.7 and the complete set of results in Tables 4.7 and 4.8.

| any-day | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|
| fundamental period | 23 | 80 | 96 | 99 | 100 | 100 | 100 |
| magnitude | 80 | 98 | 99 | 99 | 100 | 100 | 100 |
| relative phase equal weight | 52 | 88 | 96 | 100 | 100 | 100 | 100 |
| 1st harmonic, | 84 | 99 | 99 | 100 | 100 | 100 | 100 |
| 1st harmonic, ANOVA threshold | 89 | 99 | 100 | 100 | 100 | 100 | 100 |
| 1st harmonic, ANOVA weighted | 89 | 99 | 99 | 100 | 100 | 100 | 100 |

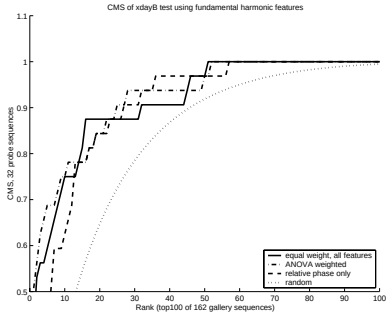
Table 4.7: Any-day recognition results using variations of fundamental harmonic gait features.



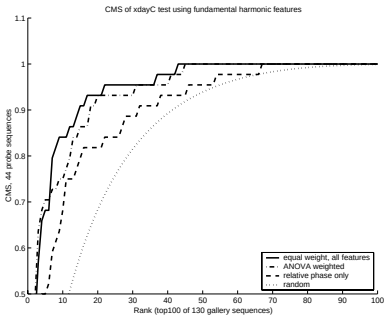

(a) any-day recognition test



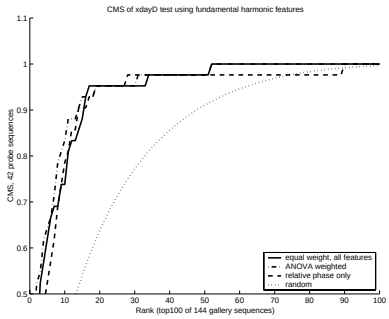

(b) xdayA recognition test




(c) xdayB recognition test




(d) xdayC recognition test




(e) xdayD recognition test

Figure 4.7: The cumulative match score curves of five recognition tests using fundamental harmonic features.

| xdayA | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|
| fundamental period | 9 | 44 | 47 | 72 | 84 | 84 | 100 |
| magnitude | 41 | 84 | 91 | 100 | 100 | 100 | 100 |
| relative phase | 22 | 75 | 88 | 94 | 97 | 97 | 100 |
| 1st harmonic, equal weight | 41 | 84 | 91 | 100 | 100 | 100 | 100 |
| 1st harmonic, ANOVA threshold | 50 | 88 | 94 | 97 | 100 | 100 | 100 |
| 1st harmonic, ANOVA weighted | 44 | 81 | 94 | 97 | 100 | 100 | 100 |
| xdayB | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| fundamental period | 19 | 44 | 66 | 75 | 88 | 91 | 100 |
| magnitude | 28 | 69 | 78 | 88 | 94 | 100 | 100 |
| relative phase | 16 | 59 | 78 | 94 | 97 | 100 | 100 |
| 1st harmonic, equal weight | 41 | 69 | 88 | 91 | 97 | 100 | 100 |
| 1st harmonic, ANOVA threshold | 50 | 72 | 75 | 88 | 94 | 97 | 100 |
| 1st harmonic, ANOVA weighted | 50 | 72 | 78 | 94 | 94 | 100 | 100 |
| xdayC | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| fundamental period | 27 | 52 | 68 | 93 | 95 | 100 | 100 |
| magnitude | 30 | 70 | 84 | 95 | 95 | 100 | 100 |
| relative phase | 36 | 59 | 75 | 84 | 93 | 95 | 98 |
| 1st harmonic, equal weight | 36 | 80 | 86 | 95 | 98 | 100 | 100 |
| 1st harmonic, ANOVA threshold | 45 | 75 | 82 | 93 | 93 | 100 | 100 |
| 1st harmonic, ANOVA weighted | 41 | 73 | 84 | 93 | 95 | 100 | 100 |
| xdayD | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| fundamental period | 14 | 36 | 81 | 90 | 98 | 98 | 98 |
| magnitude | 14 | 67 | 83 | 95 | 95 | 98 | 100 |
| relative phase | 21 | 62 | 90 | 98 | 98 | 98 | 98 |
| 1st harmonic, equal weight | 24 | 69 | 86 | 95 | 98 | 100 | 100 |
| 1st harmonic, ANOVA threshold | 40 | 81 | 95 | 98 | 100 | 100 | 100 |
| 1st harmonic, ANOVA weighted | 43 | 71 | 90 | 95 | 98 | 100 | 100 |

Table 4.8: Cross-day recognition results using variations of fundamental harmonic gait features.

As is evident in the recognition results, the fundamental harmonic decomposition features do not perform as well as the average appearance features or the appearance histogram features in the any-day test; on the other hand, they perform much better than do the average appearance features and are comparable in performance to the appearance histogram features in the cross-day tests. The two ANOVA-related variations of the fundamental harmonic components showed the best recognition performances overall, followed by equally weighting the full set of harmonic features. The ANOVA threshold variation is again preferred because of its smaller set of feature components. Approximately 2/3 of the fundamental harmonic features that pass under the ANOVA $p$ value threshold of $10^{-9}$ are magnitudes of fundamental frequency, and the other 1/3 are the relative phases. The fundamental period also passes the threshold, which suggests that while we intuitively believe people may vary their walking speed, their walking speeds are actually more consistent than we (at least this author) believed. The fundamental harmonic features are less sensitive to silhouette appearance variations in the same subject from gait data collected on different days than both the average appearance features and the appearance histograms. This is quantified by the fraction of the closest correct retrievals that are sequences from the same day in the any-day test as tabulated below:

| fundamental harmonic feature variation | % of probes with closest correct match from the same day |
|---|---|
| fundamental period | 76 |
| magnitudes | 91 |
| relative phase | 74 |
| equal weight, all features | 92 |
| ANOVA threshold | 94 |
| ANOVA -log weighted | 95 |

### 4.4.4 First and Second Harmonic Components

The following weight variations are examined in the fundamental and the second harmonic components:

1. $W_k = 1$ for magnitude of the second frequency of each feature, and 0 otherwise.

2. $W_k = 1$ for phase of the second harmonic relative to the first harmonic of each feature, and 0 otherwise.

3. $W_k = 1$ for all second harmonic feature components.

4. $W_k = 1$ if ANOVA results in $p < 10^{-9}$ of the second harmonic features, otherwise $W_k = 0$. There are 11 second harmonic feature components that pass this threshold.

5. $W_k = \min(-\log_{10}(p_k)/10, 2)$ for all second harmonic components, *i.e.*, each component is weighted by the log of the ANOVA $p$-value.

6. $W_k = 1$ for all first and second harmonic features.

7. $W_k = 1$ if ANOVA results in $p < 10^{-9}$ of the first and the second harmonic features, otherwise $W_k = 0$. There are 43 first and second harmonic feature components that pass this threshold.

8. $W_k = \min(2, -\log_{10}(p_k)/10)$ for all first and second harmonic components, *i.e.*, each component is weighted by the log of the ANOVA $p$-value.

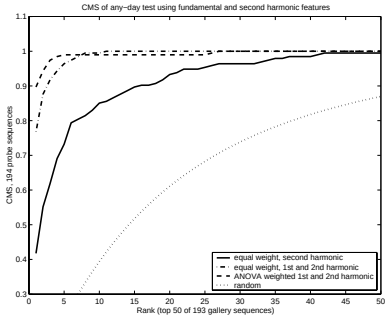
Recognition performances are shown in Figure 4.8 and in Tables 4.9 and 4.10.

The recognition performance shows that the second harmonic components alone are not good features for the recognition tests. Combining the first and second harmonic components performs better than using only the second harmonic. However, there is not a clear advantage to combining the first and second harmonic components over using only the first harmonic components. This may be an indication that the amount of noise in the time series of gait image features may be too high for accurate estimation of the second harmonic components. The second harmonic components alone are not very sensitive to clothing changes resulting from data collected from different days. Approximately 70% of the probes were identified with a sequence of the same subject collected on the same day. The combined first and

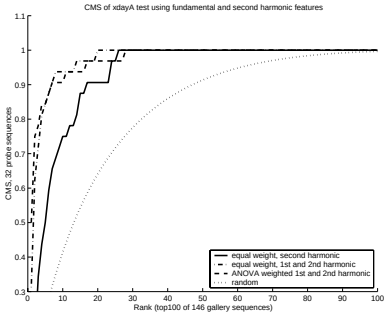
| any-day | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|
| magnitude, 2nd harmonic | 40 | 82 | 91 | 97 | 99 | 99 | 100 |
| relative phase, 2nd harmonic | 28 | 79 | 87 | 93 | 97 | 100 | 100 |
| 2nd harmonic, equally weighted | 42 | 85 | 92 | 98 | 99 | 100 | 100 |
| 2nd harmonic, ANOVA threshold | 39 | 88 | 94 | 98 | 99 | 100 | 100 |
| 2nd harmonic, ANOVA weighted | 52 | 90 | 96 | 99 | 100 | 100 | 100 |
| 1st & 2nd harmonic, equally weighted | 77 | 99 | 100 | 100 | 100 | 100 | 100 |
| 1st & 2nd harmonic, ANOVA threshold | 87 | 100 | 100 | 100 | 100 | 100 | 100 |
| 1st & 2nd harmonic, ANOVA weighted | 90 | 99 | 99 | 100 | 100 | 100 | 100 |

Table 4.9: Any-day recognition results using variations of fundamental and second harmonic gait features.

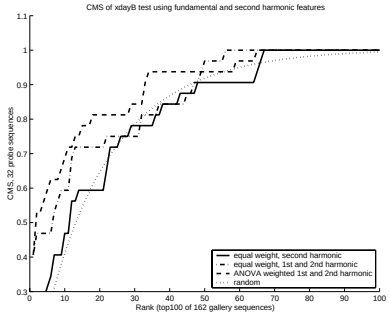
second harmonic features are much more sensitive to silhouette appearance changes resulting from changes of clothing on different days: over 90% of the probes are identified with another sequence of the same subject on the same day.



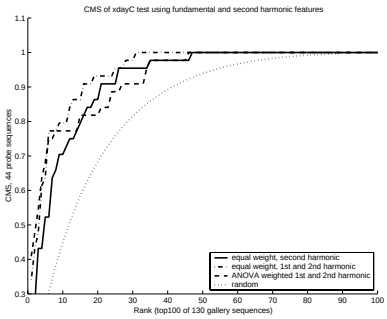

(a) any-day recognition test



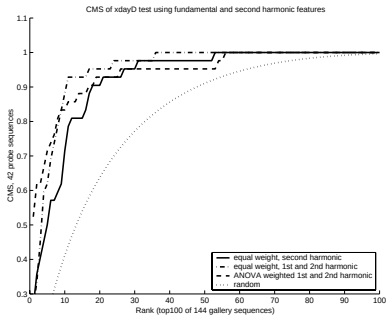

(b) xdayA recognition test




(c) xdayB recognition test




(d) xdayC recognition test




(e) xdayD recognition test

Figure 4.8: The cumulative match score curves of five recognition tests using fundamental and second harmonic features.

| xdayA | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|
| magnitude, 2nd harmonic | 6 | 56 | 84 | 100 | 100 | 100 | 100 |
| relative phase, 2nd harmonic | 25 | 66 | 84 | 84 | 91 | 97 | 97 |
| 2nd harmonic, equally weighted | 6 | 66 | 88 | 100 | 100 | 100 | 100 |
| 2nd harmonic, ANOVA threshold | 25 | 66 | 91 | 97 | 100 | 100 | 100 |
| 2nd harmonic, ANOVA weighted | 28 | 75 | 91 | 100 | 100 | 100 | 100 |
| 1st & 2nd harmonic, equally weighted | 31 | 91 | 97 | 100 | 100 | 100 | 100 |
| 1st & 2nd harmonic, ANOVA threshold | 53 | 88 | 97 | 100 | 100 | 100 | 100 |
| 1st & 2nd harmonic, ANOVA weighted | 50 | 91 | 94 | 100 | 100 | 100 | 100 |
| xdayB | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| magnitude, 2nd harmonic | 9 | 41 | 59 | 78 | 91 | 97 | 100 |
| relative phase, 2nd harmonic | 9 | 47 | 69 | 84 | 97 | 100 | 100 |
| 2nd harmonic, equally weighted | 9 | 41 | 59 | 78 | 91 | 94 | 100 |
| 2nd harmonic, ANOVA threshold | 19 | 53 | 59 | 78 | 94 | 100 | 100 |
| 2nd harmonic, ANOVA weighted | 19 | 50 | 59 | 75 | 81 | 91 | 94 |
| 1st & 2nd harmonic, equally weighted | 41 | 56 | 72 | 81 | 94 | 100 | 100 |
| 1st & 2nd harmonic, ANOVA threshold | 44 | 56 | 72 | 84 | 94 | 97 | 100 |
| 1st & 2nd harmonic, ANOVA weighted | 41 | 63 | 78 | 84 | 94 | 100 | 100 |
| xdayC | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| magnitude, 2nd harmonic | 11 | 59 | 75 | 91 | 98 | 100 | 100 |
| relative phase, 2nd harmonic | 25 | 59 | 70 | 80 | 86 | 93 | 98 |
| 2nd harmonic, equally weighted | 16 | 64 | 75 | 95 | 98 | 100 | 100 |
| 2nd harmonic, ANOVA threshold | 11 | 61 | 70 | 89 | 91 | 93 | 95 |
| 2nd harmonic, ANOVA weighted | 20 | 64 | 75 | 86 | 91 | 93 | 93 |
| 1st & 2nd harmonic, equally weighted | 34 | 77 | 86 | 95 | 100 | 100 | 100 |
| 1st & 2nd harmonic, ANOVA threshold | 36 | 73 | 80 | 89 | 95 | 98 | 100 |
| 1st & 2nd harmonic, ANOVA weighted | 41 | 75 | 77 | 89 | 98 | 100 | 100 |
| xdayD | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| magnitude, 2nd harmonic | 19 | 52 | 83 | 98 | 98 | 100 | 100 |
| relative phase, 2nd harmonic | 10 | 69 | 88 | 90 | 100 | 100 | 100 |
| 2nd harmonic, equally weighted | 26 | 57 | 81 | 95 | 98 | 100 | 100 |
| 2nd harmonic, ANOVA threshold | 17 | 71 | 88 | 95 | 100 | 100 | 100 |
| 2nd harmonic, ANOVA weighted | 21 | 64 | 81 | 93 | 95 | 95 | 95 |
| 1st & 2nd harmonic, equally weighted | 26 | 74 | 93 | 98 | 100 | 100 | 100 |
| 1st & 2nd harmonic, ANOVA threshold | 43 | 90 | 93 | 95 | 100 | 100 | 100 |
| 1st & 2nd harmonic, ANOVA weighted | 52 | 76 | 88 | 95 | 95 | 100 | 100 |

Table 4.10: Cross-day recognition results using variations of fundamental and second harmonic gait features.

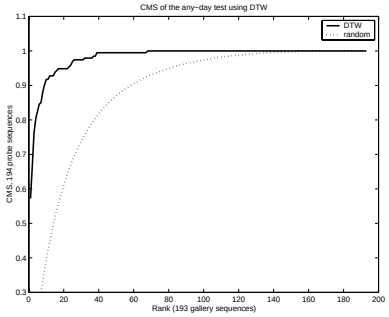
### 4.4.5 Direct Comparison of Time Series

The recognition performance of using dynamic time warping (DTW) to directly compare gait image feature time series is shown in Figure 4.9 and in Table 4.11.

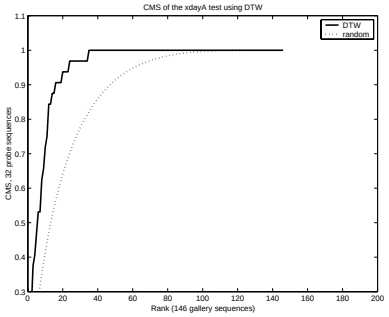
| | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|
| any-day | 57 | 92 | 95 | 99 | 99 | 100 | 100 |
| xdayA | 16 | 53 | 88 | 97 | 100 | 100 | 100 |
| xdayB | 13 | 50 | 59 | 91 | 100 | 100 | 100 |
| xdayC | 11 | 43 | 61 | 82 | 98 | 100 | 100 |
| xdayD | 17 | 50 | 60 | 81 | 90 | 100 | 100 |

Table 4.11: Dynamic time warping recognition results.

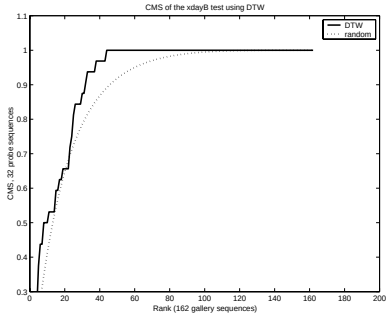
Direct sequence comparison using dynamic time warping is the most computationally intensive of all gait feature comparison methods studied in this thesis. While the recognition performance using DTW on the any-day test is significantly worse than most time-aggregated gait features, the cross-day tests showed results that are not significantly worse. Direct comparison of time series is relatively insensitive to clothing change between data collected on different days: approximately 90% of the probes are identified to sequences of the same subject collected on the same day.



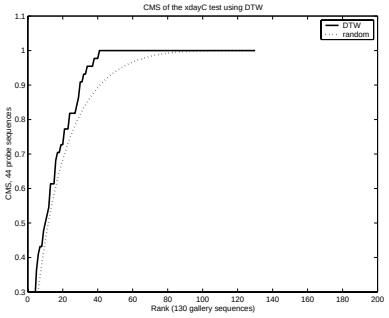

(a) any-day recognition test



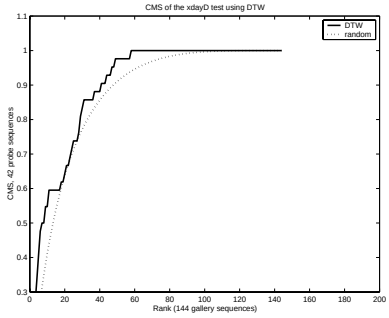

(b) xdayA recognition test




(c) xdayB recognition test




(d) xdayC recognition test




(e) xdayD recognition test

Figure 4.9: The cumulative match score curves of five recognition tests using dynamic time warping to directly compare feature sequences.

## 4.5 Discussion of Recognition Results

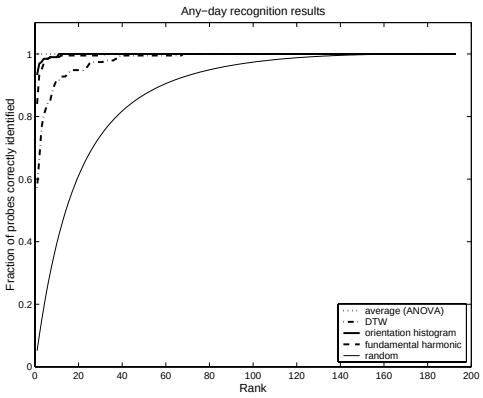
Based on the set of recognition experiments conducted using the four different types of features, we come to the following conclusions:

- The average appearance feature is extremely simple and efficient to compute, but it is not able to provide a detailed enough description of the distribution of the gait image features. It is highly sensitive to clothing changes resulting from data collected on different days.
- The fundamental harmonic features are less sensitive to clothing changes, hence they have better performance on the cross-day recognition tests. The addition of the second harmonic features does not contribute significantly to the recognition performance. It is possible that the amount of noise in the time series precludes the accurate estimation of the second harmonic components.
- The histogram appearance features are sensitive to the silhouette appearance changes resulting from different days of gait data collection. However, it still performs well in the cross-day recognition test.
- The direct comparison of time series using dynamic time warping preserves the most amount of time information and is the most computationally intensive method. It consistently performs better than the random retrieval method, though by relatively small amounts in the cross-day tests.

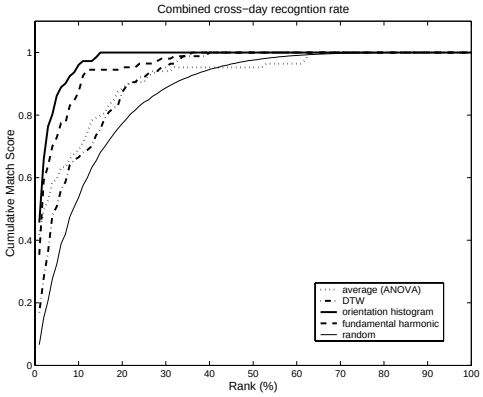
The recognition results of the best performing variation of each aggregation method are displayed in Figure 4.10.

The extreme sensitivity to clothing and background changes of the average appearance feature makes it not the ideal feature set for recognizing walking subjects collected on different days when they may be wearing different clothing. However, we might be able to exploit this sensitivity to detect the clothing model for the walking subject, such as pants vs. shorts vs. skirt, if environmental effects such as shadows could be reduced. In addition, we found that our heuristic method for feature selection—that is, assuming independence of features and using analysis of variance to select for features that highly violate the single Gaussian distribution assumption—not only reduced the dimensionality of the average appearance feature set, but also improved performance.

The recognition results using the fundamental harmonic components showed good performance in the cross-day recognition tests. This




(a) any-day recognition results




(b) combined cross-day recognition results

Figure 4.10: Cumulative match score for the best performing variation of each aggregation method. The four cross-day tests are combined to show percentage of recall.

is consistent with our intuition. Changes in the clothing and hair styles of a subject that occur in multi-day gait data collection causes an overall change in the appearance of the silhouettes. We constructed the fundamental harmonic features by eliminating the means of all image features, hence the resulting harmonic feature only contains the change in each feature over a walking sequence, which is much less sensitive to an overall appearance change. Feature selection using our heuristic method based on the $p$-value of ANOVA reduced the number of fundamental harmonic features from 56 to 31 without adversely affecting the recognition results. We also used the second harmonic components in addition to the fundamental harmonic components. While our recognition results do not support a case for using the second harmonic components for person recognition, we highly suspect that this is caused by the amount of noise in the time series compounded by the low sampling rate in time rather than the lack of second harmonic component features as a biometric feature. Preliminary results by Carter *et al.* [5] showed that the higher harmonic component features were not only present in the silhouette, but that they were useful for identification purposes. They were using video captured at 50 frames per second in addition to having a high resolution view of the walking subject and the chroma-keyed background to produce high quality walking figure silhouettes. They were able to extract not only the second harmonic, but the third, fourth, and even the fifth harmonic components.

Of the four variations of histogram appearance features that we had experimented with, the orientation histogram has the best performance in the cross day recognition tests. This is consistent with our intuitive understanding. The histogram appearance features have roughly the same performance as fundamental harmonic features in the cross-day tests and the best performance in the any-day test. An ideal gait appearance representation should behave in such a way that it is sensitive to the consistency of appearance of data collected on the same day, but not so sensitive that the appearances are over-modeled and appearance-independent gait features are compromised. Hence we conclude that the orientation histogram of gait sequence feature is the best set of the four considered. In addition, we hypothesize that recognition performance could be improved by augmenting the histogram appearance features with other features that are not present in the histogram. Namely, the fundamental harmonic features contain information about relative phases and the fundamental period, which are independent of any feature in the histogram appearance representation.

## 4.6 Better Features for Person Recognition

The histogram of orientation of ellipses in each region is chosen because it is the most compact and the best performing feature of the histogram feature set. The following variations of combining the orientation histogram representation and the fundamental harmonic features are examined:

1. Histogram of orientation alone.
2. Histogram of orientation combined with magnitude of the fundamental frequency components.
3. Histogram combined with relative phase of the fundamental frequency components.
4. Histogram combined with the fundamental period.
5. Histogram combined with relative phase and fundamental period.

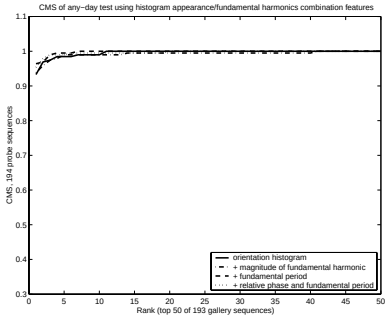
The histogram comparisons result in a similarity score between two gait sequences while the fundamental harmonic features comparisons result in a distance between two gait sequences. Thus the comparison score of the combined histogram/fundamental harmonic feature set needs to resolve this discrepancy. A simple solution is to subtract a multiple of the distance from the histogram similarity score to create a combination score for comparison, *i.e.*,

$$s_c = s_h - Cd_f, \tag{4.4}$$

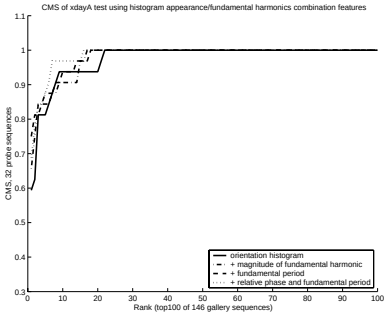
where $s_h$ is the similarity score from histogram comparison, $d_f$ is the distance from components of the fundamental harmonic, and $s_c$ is the new combination similarity score. Because each of these measures have different dynamic ranges, the coefficient $C$ is used to roughly scale the distance measure to match the dynamic range of the histogram score.

Recognition performance using the histogram/fundamental harmonic combinations are shown in Figure 4.11 and in Tables 4.12 and 4.13.

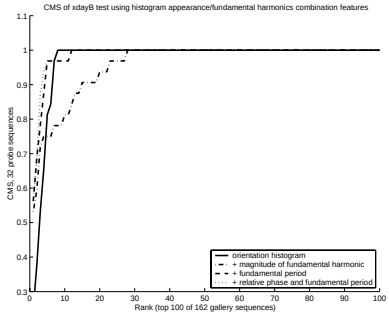
Two of the combination gait features, histogram + fundamental period, and histogram + period + relative phase appear to be the most promising of the features. These combinations of features slightly under-performs in the any-day test but their recognition performances are on average much better than the orientation histogram along almost all points of the cumulative match score curve. and are strictly better than the first harmonic features. Hence we conclude that the histogram + period (+ relative phase) feature sets are better for the recognition tasks.



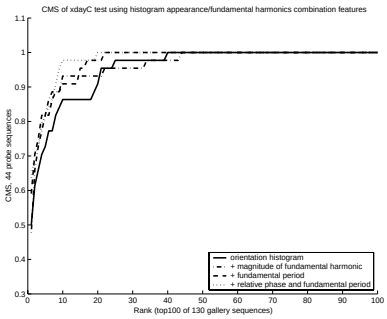

(a) any-day recognition test



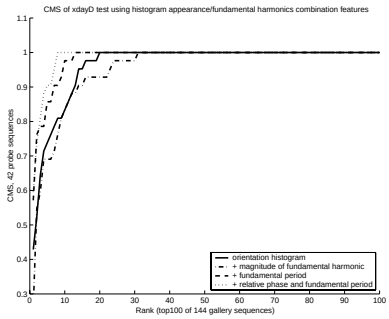

(b) xdayA recognition test




(c) xdayB recognition test




(d) xdayC recognition test




(e) xdayD recognition test

Figure 4.11: The cumulative match score curves of five recognition tests using orientation histogram appearance/fundamental harmonic combination features.

| any-day | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|
| orientation histogram | 93 | 99 | 100 | 100 | 100 | 100 | 100 |
| histogram + magnitude | 94 | 99 | 99 | 99 | 100 | 100 | 100 |
| histogram + relative phase | 86 | 99 | 99 | 100 | 100 | 100 | 100 |
| histogram + fundamental period | 96 | 100 | 100 | 100 | 100 | 100 | 100 |
| histogram + rel. phase + period | 95 | 99 | 99 | 100 | 100 | 100 | 100 |

Table 4.12: Any-day recognition results using combinations of orientation histogram appearance and fundamental harmonic features.

| xdayA | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
|---|---|---|---|---|---|---|---|
| orientation histogram | 59 | 88 | 94 | 100 | 100 | 100 | 100 |
| histogram + magnitude | 66 | 88 | 97 | 100 | 100 | 100 | 100 |
| histogram + relative phase | 53 | 91 | 97 | 100 | 100 | 100 | 100 |
| histogram + fundamental period | 75 | 88 | 97 | 100 | 100 | 100 | 100 |
| histogram + rel. phase + period | 69 | 97 | 97 | 100 | 100 | 100 | 100 |
| xdayB | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| orientation histogram | 25 | 100 | 100 | 100 | 100 | 100 | 100 |
| histogram magnitude | 53 | 78 | 91 | 100 | 100 | 100 | 100 |
| histogram + relative phase | 34 | 97 | 100 | 100 | 100 | 100 | 100 |
| histogram + fundamental period | 56 | 97 | 100 | 100 | 100 | 100 | 100 |
| histogram + rel. phase + period | 56 | 100 | 100 | 100 | 100 | 100 | 100 |
| xdayC | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| orientation histogram | 50 | 77 | 86 | 98 | 98 | 100 | 100 |
| histogram + magnitude | 48 | 89 | 93 | 95 | 98 | 100 | 100 |
| histogram + relative phase | 45 | 84 | 93 | 98 | 100 | 100 | 100 |
| histogram + fundamental period | 59 | 86 | 91 | 100 | 100 | 100 | 100 |
| histogram + rel. phase + period | 64 | 86 | 98 | 100 | 100 | 100 | 100 |
| xdayD | 1st | 5% | 10% | 20% | 30% | 40% | 50% |
| orientation histogram | 43 | 79 | 95 | 100 | 100 | 100 | 100 |
| histogram + magnitude | 26 | 71 | 90 | 98 | 100 | 100 | 100 |
| histogram + relative phase | 33 | 88 | 98 | 100 | 100 | 100 | 100 |
| histogram + fundamental period | 57 | 90 | 100 | 100 | 100 | 100 | 100 |
| histogram + rel. phase + period | 60 | 95 | 100 | 100 | 100 | 100 | 100 |

Table 4.13: Cross-day recognition results using combinations of histogram appearance orientation and fundamental harmonic features.

# Chapter 5

# Other Experiments

In addition to the recognition tests described in the previous chapter, we experimented with gender classification using the gait sequence features and explored through experimentation with different data sets the sensitivity of recognition performance to noise in silhouettes.

## 5.1 Gender Classification

Here we applied the gait average appearance features and the fundamental harmonic component features to the task of gender classification. Specifically, we used the full 57 dimensional average appearance features as described in Section 3, as well as a smaller set of features selected using the $p$-value obtained using analysis of variance. We ranked each of the 57 features based on the $p$-value of ANOVA in separating the genders and set a threshold of $p < 10^{-9}$, which resulted in the best 6 features (Table 5.1) for gender classification. Intuitively, the third

| rank | region | feature type |
|---|---|---|
| 1 | front calf | mean of orientation |
| 2 | back | mean of orientation |
| 3 | head | mean of $x$ coordinate of centroid |
| 4 | head | mean of orientation |
| 5 | back calf | std of $x$ of centroid |
| 6 | back calf | mean of $x$ of centroid |

Table 5.1: Top 6 average appearance features for gender classification

and the fourth ranked features—the mean of the $x$ coordinate of the centroid and the orientation of the head—describe differences in the shape of the profile-view of the head between men and women in addition to posture differences. Women tend to have more hair behind the head than men do, and they also tend to hold up the head slightly more than men do. The mean orientation of the back, ranked second, is another possible indication of the differences in posture between men and women. The first, fifth, and sixth ranked features all relate to stride length (relative to body height) differences between men and women.

A similar process was applied to the fundamental harmonic features for gender classification. We again used two sets of features: (1) the complete set of 56 fundamental harmonic features, and (2) the best features selected based on their significance in indicating gender. Small $p$-values from ANOVA on gender class was used as an indicator for the significance of a feature in gender classification. We set a threshold of $p < 10^{-9}$, which resulted in 5 fundamental harmonic features that are best for gender classification. The five features are listed in Table 5.2.

| rank | region | feature type |
|---|---|---|
| 1 | whole silhouette | fundamental period |
| 2 | back | relative phase of elongation |
| 3 | head region | magnitude of the x of centroid |
| 4 | back thigh | magnitude of $x$ coordinate of centroid |
| 5 | chest | magnitude of the $x$ of centroid |

Table 5.2: Top 5 fundamental harmonic features for gender classification

We trained and tested support vector machines [42] on our gait appearance features under two conditions. Under the random-sequence test, we randomly selected gait feature vectors of approximately half of the sequences, without regard to the identity of the walking subject, and tested on the gait features of the remaining sequences. Under the random-person test, we randomly selected approximately half of our walking subjects, trained the SVM on all sequences from these walkers, and tested on all sequences of the remaining walking subjects. The same subject may appear in both the training and the testing set in the random sequence scenario (though not the same sequence), whereas the a subject never occurs in both the training and the testing set in the random person scenario. Because we saw that people generally "look" like themselves from the recognition experiments of Chapter

4, we expect the random sequence test to be easier than the random person test.

We used an implementation of support-vector machine by Rifkin [35] and experimented with the linear, Gaussian, and the second degree polynomial kernels. The SVM's were trained using the 57 and the 6 gender features and under the random-person vs. random-sequence conditions. The training and testing are conducted as 20 repeated random trials each, *i.e.*, 20 training/testing each of random sequence and random person experiments $\times$ 57 features and 6 best features selected using the $p$-values of ANOVA, resulting in four sets of 20 repeated random trial experiments for each of the three SVM kernels. The exact same set of training and testing experiments (with the training and testing sets replicated) were repeated using the harmonic components features, the full set of 56 features and the subset of 5 selected using a threshold on the $p$-values.

The average results of 20 repeated random trials for the 12 tests conditions using the average appearance features are listed in Table 5.3. Overall, we found that the results for the random-sequence test is better because sequences from the same person, though not the same sequences, are in both the training and the testing set.

| | Random Sequence | |
|---|---|---|
| Kernel type | 57 features | 6 features |
| polynomial(d=2) | 91% | 94% |
| Gaussian | 93.5% | 90% |
| linear | 94% | 88% |
| | Random Person | |
| Kernel type | 57 features | 6 features |
| polynomial(d=2) | 79% | 84.5% |
| Gaussian | 66% | 83.5% |
| linear | 80% | 84.5% |

Table 5.3: SVM gender classification results using the average appearance features.

The random-person test condition is a more accurate representation of how a gender classifier would be used in a real application. The performance of the three kernels in the random-person case show that the linear kernel performed at least as well as the Gaussian and the polynomial kernels. This leads us to believe that the boundary between the genders may be approximately linear using our data set.

The significantly better gender classification performance of the 6 feature set in the random person test than the full 57 average appearance features suggests that the SVM may be fitting the class boundary to variations of the individual subjects that appear in the components that are not highly associated to gender. This probably is a side effect of the small size of training samples. We conjecture that this effect will disappear if many more training samples are used. Alternatively, we conclude that in the case of small dataset feature selection becomes a much more crucial issue.

The same set of 20 repeated random trials and 12 test conditions was repeated using the fundamental harmonic features. The results for gender classification are listed in table Table 5.4.

| | Random Sequence | |
|---|---|---|
| Kernel type | 56 features | 5 features |
| polynomial(d=2) | 82% | 73% |
| Gaussian | 61% | 84% |
| linear | 88% | 78% |
| | Random Person | |
| Kernel type | 56 features | 5 features |
| polynomial(d=2) | 52% | 61% |
| Gaussian | 58% | 72% |
| linear | 74% | 70% |

Table 5.4: SVM gender classification results using the fundamental harmonic features.

The gender classification results using fundamental harmonic features again confirmed that the random sequence test is the easier test. The difference in classification performance in the random person case also shows that the smaller set of features performed better than the full set of 56 features on average. While the linear kernel performed quite well in gender classification. The Gaussian kernel seems to have won by a small margin. A comparison of the results between using the average appearance features and the fundamental harmonic features show that the average appearance features are a better representation for gender classification than the fundamental harmonic features.

## 5.2 The Effect of Background Subtraction Noise on Recognition Performance

The performance of gait recognition using silhouettes of the walking subjects depends heavily on the quality of the silhouettes after background subtraction. We argue that it is not just amount of noise in silhouettes that affect the recognition performance, but also the consistency of noise. In other words, if the noise in silhouettes affects the gait image features consistently across all frames of all sequences, then the impact of noisy silhouettes is minimal on the recognition performance.

In addition to the primary dataset used in the recognition experiments in the previous chapter, three more datasets are employed to test our theory. These three datasets are provided courtesy of Robotics Institute of Carnegie Mellon University, University of Southampton, and University of Maryland. We will refer to these datasets as CMU, Soton, and UMD, respectively. The primary dataset will be referred to as MITAI. Only the frontal-parallel view, or the view closest to the frontal parallel view, of the walking subject was used. Each of these data sets contain gait sequences where each subject was recorded on the single day, hence the cross-day test described in the previous chapter cannot be conducted. The UMD dataset produced the noisiest silhouettes of all four groups. Many of the gait video frames had to be eliminated because the silhouettes produced from those frames were of very poor quality. As a consequence, the silhouettes were not sampled evenly in time, thus removing the possibility of utilizing the harmonic components features and the direct matching of gait image feature time series. For consistency across the different datasets, the recognition tests conducted here involved only data collected on the same day, and only the average appearance feature and the histogram features are used.

### 5.2.1 CMU dataset

The CMU dataset is a multi-view, synchronized capture of walker subjects on a treadmill. Each subject performs three different types of walk: holding a ball, fast walk, and slow walk. Each walking sequence is 10 seconds or more, recorded at 30 frames/second from several angles. We used the foreground silhouettes provided by CMU. Only the frontal-parallel view of subjects were included. These silhouettes were produced using a simple background subtraction of the video frames from a single image of the background. The effect of this simplistic background algorithm is evident in the type of noise in the silhouettes, such as the holes in the torso portion corresponding to background ob-

jects. Figure 5.1 shows several examples of silhouettes from the given dataset. Each sequence is divided into 60-frame subsequences for comparisons between the subsequences. There are a total of 25 subjects, 23 men and 2 women, with 3 types of walk each (with the exception of one subject) and 5 subsequences for each walking type (after dividing into 60-frame subsequences), thus resulting in a total of 370 subsequences.

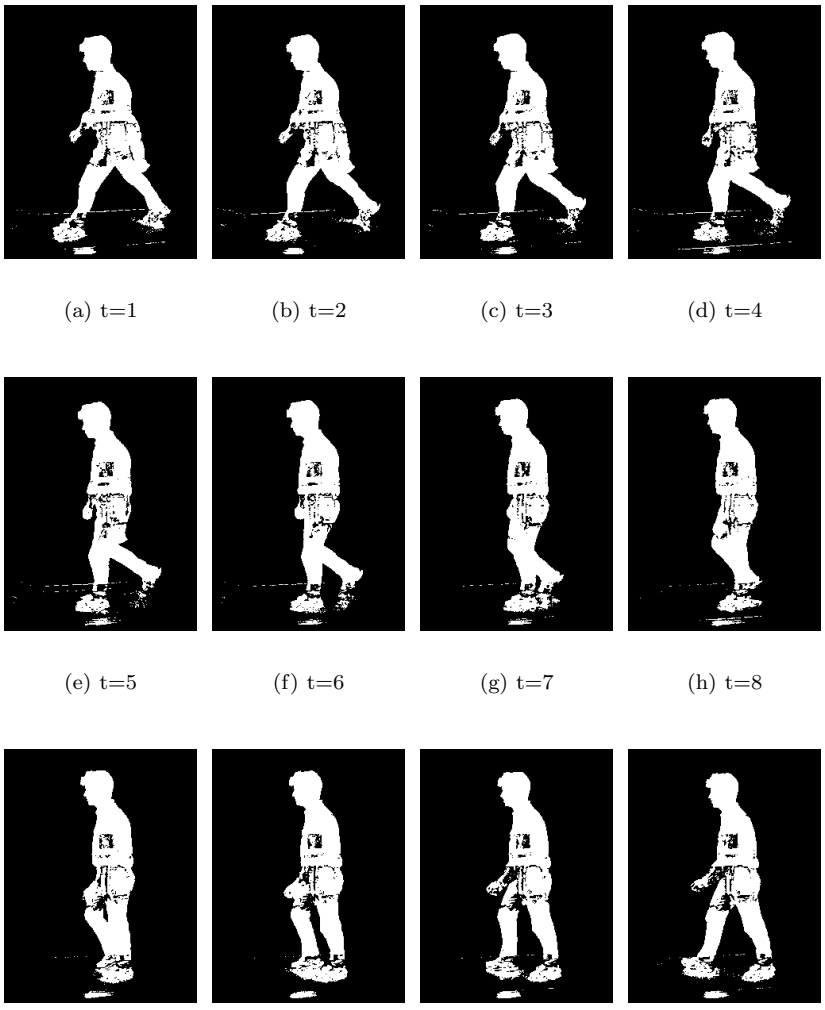
The CMU dataset has the unique characteristic that because the walking subject is fixed at the same location and is thus under fixed environmental lighting conditions, the noise in silhouettes is stable across each sequence, and hence stable between the subsequences that we used for recognition tests. For example, Figure 5.1 consistently shows a shadow below the walker and most of the frames show the edges of the treadmill. In addition, the noise is mostly stable across sequences of different individuals subject to minor shifting of the position of the walker on the treadmill. Figure 5.2 shows sample silhouettes from three different individuals with the same type of silhouette noise as the examples shown in Figure 5.1.

Using the average appearance gait features, all but one of the 370 subsequences are correctly identified at the first recall with another subsequence of the same subject doing the same type of walk. The only exception was caused by one subject coughing and covering his mouth with one hand in one subsequence of his capture session, resulting in a significant change in the appearance of the silhouette. However, this subsequence was still identified with a subsequence of the same subject, only doing a different type of walk.

### 5.2.2 Soton dataset

The Soton dataset was collected in front of a background that was chroma-keyed with green draperies. In addition, the sequences were segmented to include exactly one full stride (or two steps) from heel strike to heel strike. Figure 5.3 shows a typical example silhouette sequence. As should be visually evident from these silhouette images, the amount of noise is very small compared to the other datasets we have used, largely due to the chroma-keyed background.

Using the average appearance gait feature, each silhouette sequence is correctly identified to a sequence of the same individual at the first recall.

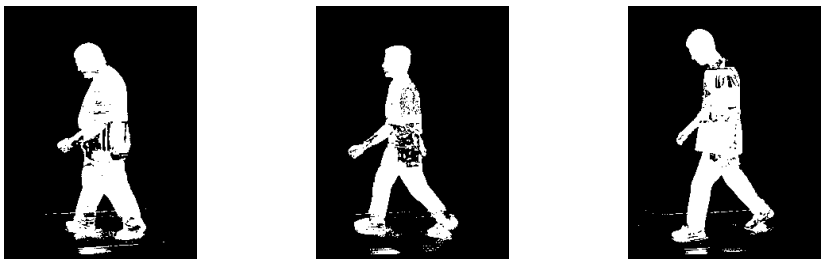


(a) t=1 (b) t=2 (c) t=3 (d) t=4

(e) t=5 (f) t=6 (g) t=7 (h) t=8

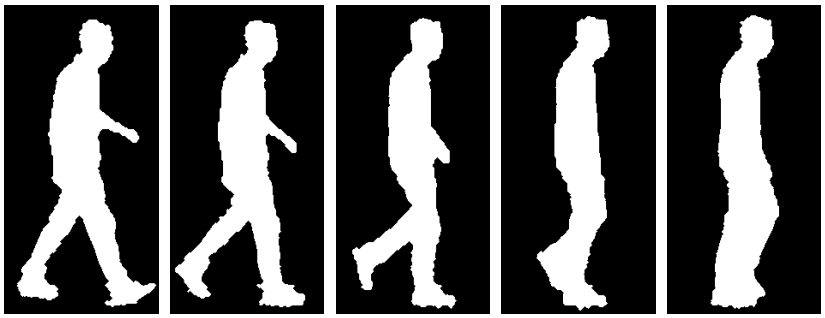
(i) t=9 (j) t=10 (k) t=11 (l) t=12

Figure 5.1: Sample silhouettes from one sequence in the CMU gait dataset.

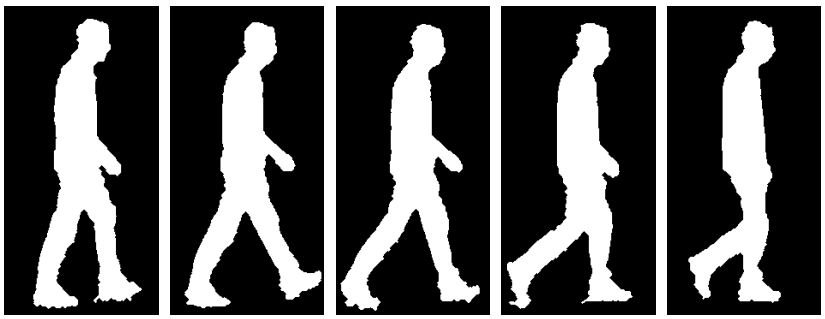


(a) (b) (c)

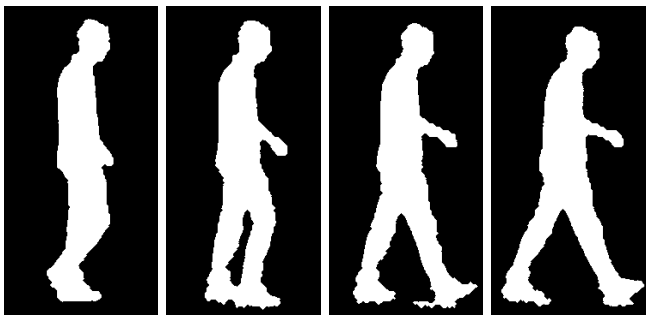
Figure 5.2: Sample silhouettes from three different individual subjects in the CMU gait data set. These silhouettes show consistent noise.



(a) t=1 (b) t=3 (c) t=5 (d) t=7 (e) t=9



(f) t=11 (g) t=13 (h) t=15 (i) t=17 (j) t=19



(k) t=21 (l) t=23 (m) t=25 (n) t=27

Figure 5.3: A typical example silhouette sequence from the Southampton gait dataset.

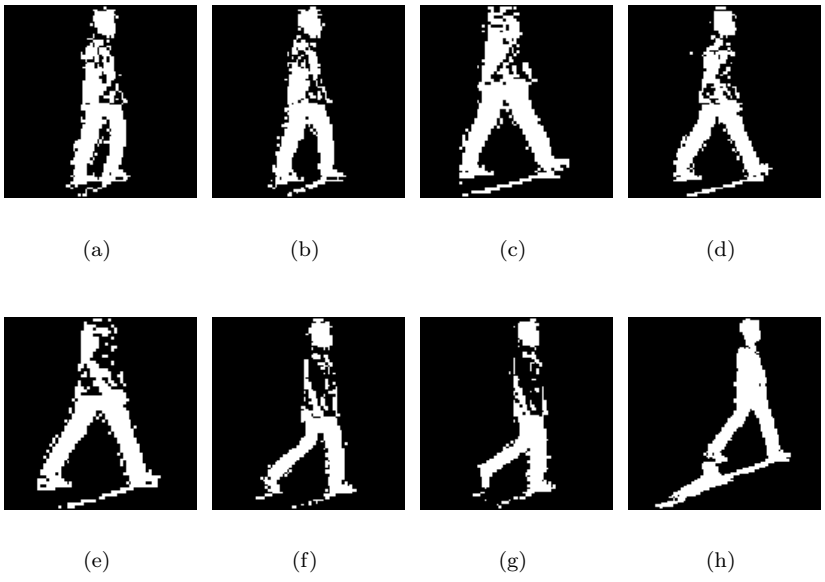



Figure 5.4: A sampling of the silhouettes from one sequence of UMD gait data.

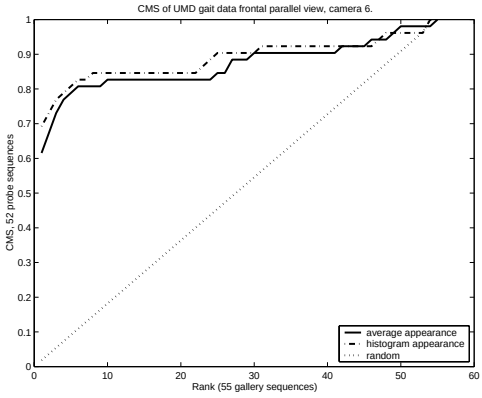
### 5.2.3 UMD dataset

The UMD dataset is the most noisy of all four datasets analyzed in this thesis. The gait data were collected from outdoor environments under different lighting conditions ranging from cloudy to sunny. Furthermore, the videos were captured under interlaced mode. There is some amount of flicker in the brightness of the video frames which we assumed to be caused by the auto gain control of the video camera. Figure 5.4 shows sample silhouettes from one gait video sequence. Not only do these silhouettes appear noisy, they also show drastically different types of noise within the same video sequence, such as shadows on the ground, holes in the body, and heads or feet missing.

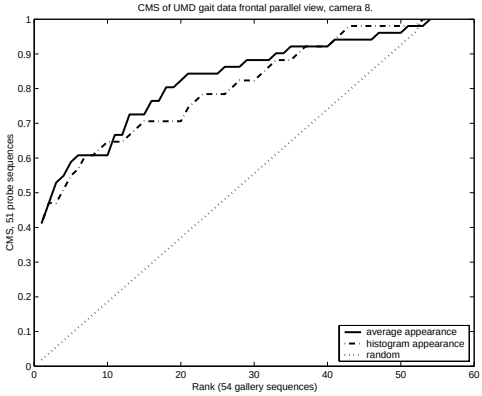
We used two scenarios from the UMD datasets that most closely resemble the conditions under which our primary dataset was collected, one with a camera that was almost parallel to ground capturing a street scene, and one with a camera that pointed at a slightly steeper angle to ground capturing a campus walkway. The recognition results in both scenarios are shown in Figure 5.5 for two gait representations and the

random retrieval algorithm.

The performance of the histogram appearance and the average appearance gait representations are roughly equal, albeit poor.




(a) camera 6




(b) camera 8

Figure 5.5: Recognition performances on UMD data.

### 5.2.4 MITAI data set

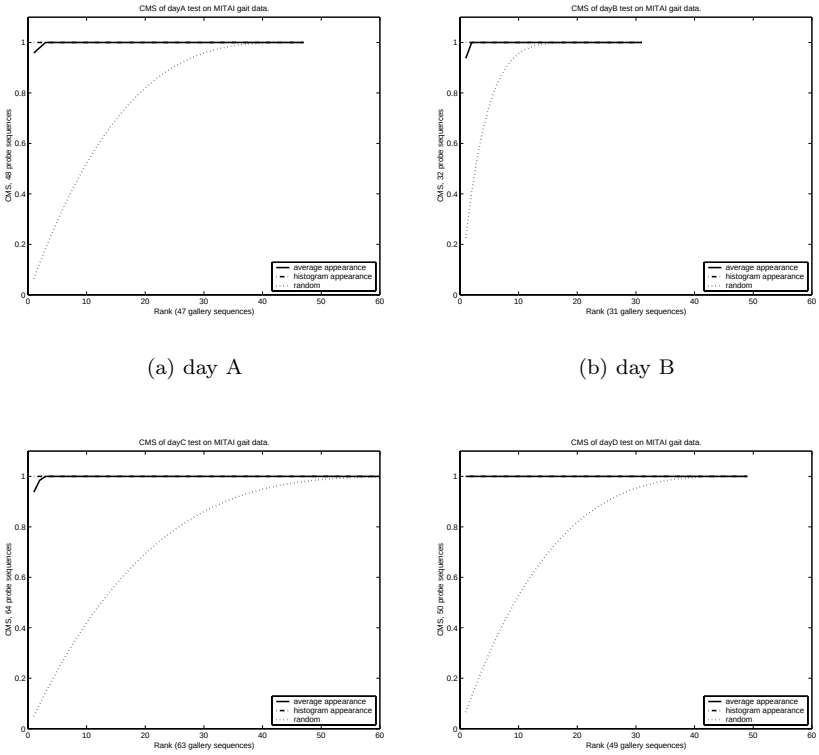
The MITAI dataset was collected on four different days, hence there are four same-day recognition tests. An example of the silhouettes obtained from this dataset was shown in Figure 2.2.

The recognition results using the histogram appearance and the average appearance gait features in the same-day recognition tests are shown in Figure 5.6. The histogram appearance identified each probe correctly at the first recall in each test. The average appearance feature correctly identified each probe after at most the third recall.

### 5.2.5 Summary on the Effect of Noise

A visual survey of the silhouettes from the four datasets show that the silhouettes from the Soton dataset are best in that they have very little noise and no parts of silhouettes missing, followed by the CMU data, the MITAI data, and the UMD data. While the CMU data included much more noise than the Soton data, the noise is very consistent across each sequence and between sequences because the fixed conditions under which the gait data was collected. The performance of both the average appearance and the histogram appearance gait representations were perfect identification at the first recall. The silhouettes of MITAI dataset contain more noise than the Soton dataset, but not significantly more than the CMU dataset. However, the noise in the MITAI silhouettes is much less consistent than the CMU dataset because the localized lighting conditions around the subject changed as the subject walked through a scene. The recognition performance accordingly is lowered for the MITAI dataset. The UMD dataset had the most amount of noise and most varied noise in its silhouettes. Consequently, the recognition performance on this dataset is the worst.

We conclude from the experiments on the four datasets that reducing the amount of noise in the silhouettes improves the gait recognition performance, and that improving the environmental consistency of the silhouettes also improves the recognition performance.




(a) day A

(b) day B

(c) day C

(d) day D

Figure 5.6: Performance of the average appearance and histogram appearance in same-day recognition tests on MITAI gait data set.

# Chapter 6

# Resolving View Dependence

One of the biggest constraints of the silhouette representation of gait presented in Chapter 2 is view-dependence–that is, the camera must capture a frontal-parallel view of the walking subject. In this chapter, I will describe joint work with Shakhnarovich and Darrell [37] that removes the constraint on the placing of the camera with respect to the walking path. While the joint work pertains to the integration of view dependent recognition methods, we will describe the components most related to resolving the view-dependence of our gait representation.

One method of resolving the frontal parallel view constraint is to record the gait video of a subject from many views—as many as one believes is necessary to capture the appearance differences between the different views of a walking subject to accurately retrieve higher level information. However, this method faces a few challenges. One is to determine the view of the walking subject so that the correct model view can be used for comparison. The other is the arbitrariness of the representation and the size of the representation. We do not know how many views are enough for a silhouette description that is good enough for identity or gender classification. In addition, this is a computationally intensive process to represent all possible views of the gait appearance.

We provide an alternative method that uses the visual hull of a walker to synthesize a frontal parallel view of the walking subject, which we call view normalization. The visual hull is the 3D volume carved out of space by the intersection of the silhouettes of an object as seen from cameras with wide base lines. This method can accommodate any view

of a subject walking in an arbitrary path. This is a unified method that requires only one representation that is derived from the synthesized frontal parallel view of the subject. Specifically, our algorithm involves the following steps:

1. A subject is recorded in an environment that is equipped to capture in real-time the visual hull of the walker.
2. The heading of the subject is estimated from 3D volume data at each time instance.
3. A virtual camera is positioned at a fixed length from the walking subject with its optical axis perpendicular to the instantaneous path of the subject.
4. A frontal-parallel view silhouette of a walking subject is synthesized from the view of the virtual camera.
5. The silhouette can be processed in the same way that a real frontal-parallel view of a silhouette is processed to obtain a set of gait sequence features.

The concept of a visual hull was introduced by Laurentini [23]. The basic idea is illustrated in Figure 6.1. From the point of view of each camera, each object in its field of view carves out a 3D volume with its projection of the silhouette onto the image plane of the camera. If multiple cameras are viewing the same object and if the cameras have wide baseline, then the intersection of these silhouettes results in a 3D volume which can be rendered from any point of view. We utilize a real-time visual hull system implemented by Matusik and company [26]. This system uses 4 widely-spaced cameras and can run at close to 14 frames per second, approaching the frame rate that we get by using a real video sequence.

While we can obtain the 3D volume that is the walking subject, we still need to find the side view of the subject. Under the assumption that people generally walk in a direction parallel to the sagittal plane, the side view can be generated by positioning the camera perpendicular to the walking path. We used the centroid of the visual hull as a measure of the path. A Kalman filter was applied to the tracks of the centroid in 3D to smooth out the noise from centroid estimation. Once a smooth path was computed, we positioned the camera perpendicular to the walking path at every time sampling point and synthesized the frontal-parallel view. Figure 6.2 shows an example of a subject walking a curved path as seen from one of the cameras, and the synthesized frontal-parallel view, or the view-normalized silhouette images.

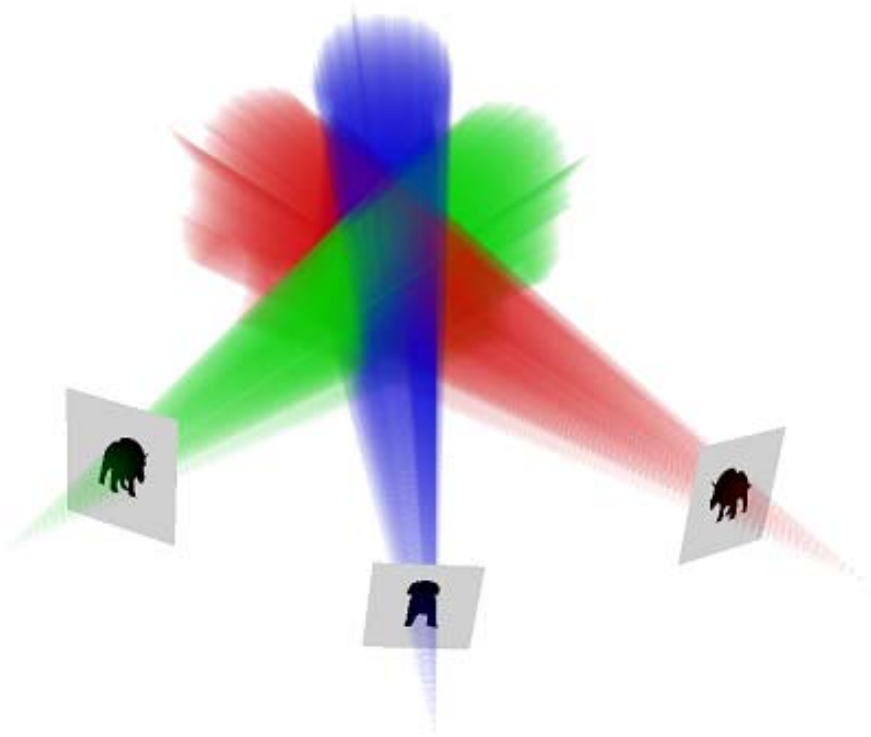


Figure 6.1: Conceptual understanding of the visual hull of an object.

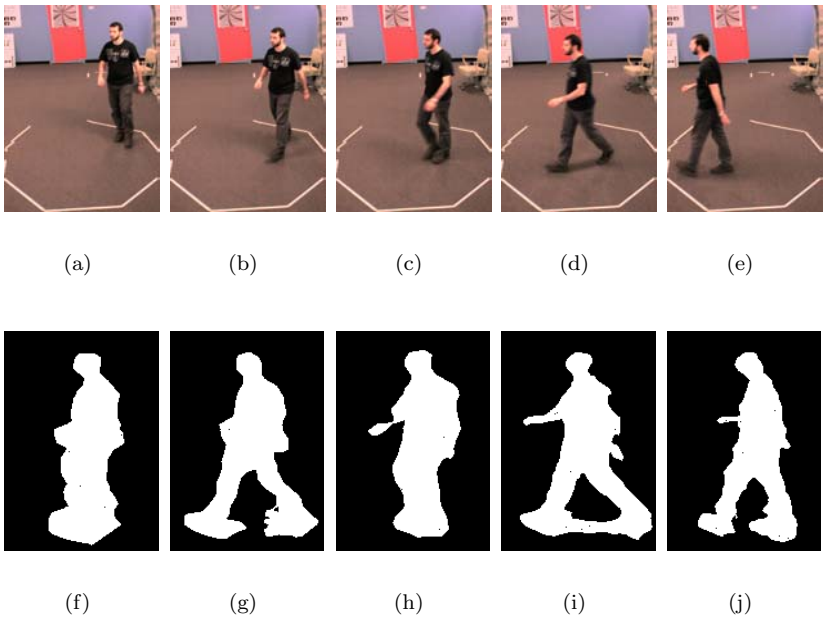



Figure 6.2: Top row: five images in a time series of a subject traveling in a curved path as seen from one camera view. Bottom row: the synthesized view-normalized silhouette of the walking subject as seen from a virtual camera positioned with its optical axis perpendicular to the curved walking path.

We extracted the average appearance features both from the silhouettes as they are extracted from individual cameras and from the view-normalized silhouettes and used these features in the person recognition task. Our database contained 27 individuals, comprising a total of 225 sequences, collected over 3 months. Distances between gait sequences are computed in the same manner as in Chapter 3. To test the effectiveness of view normalization for the purpose of gait recognition, two tests were conducted. In one, we only used gait sequences recovered from each camera, hence we do not have a side view of the walking subject. In the second test, we used the view normalized gait sequences synthesized from the visual hull. In this case we are using the frontal-parallel view of the walking subject. The rate of correct identification at the top match was 23% using the non-view-normalized sequences, and 67% using the view-normalized gait sequences. Clearly the view normalization was effective in stabilizing the view point of the camera with respect to the changing walking path.

# Chapter 7

# Summary, Discussion, and Conclusions

The goal of this thesis was to investigate the information contained in video sequences of human walking and how to extract and represent that information in ways that facilitate tasks such as person recognition and gender classification. We obtained two classes of features from a gait video sequence: (1) the image features and (2) the sequence features. We have an image representation of gait silhouettes that describes the local region shape of a walking figure. The time series of silhouette image region descriptions captured from a gait video sequence are then aggregated over time using four methods to arrive at four sequence representations: (1) the average appearance representation discards the time dimension of the gait sequence and represents the distribution of the image features using means and standard deviations; (2) the histogram appearance features improves on the representation of the image feature distribution by using histograms, but it still does not contain any information about time dependence; (3) the harmonic components features retain time information such as the fundamental period and relative phase information; and (4) the original time series serves as a baseline representation that retains all available information from the image feature time series. This set of gait sequence feature representations explores the amount of detail used in representing the distribution of image features and in representing the time dependence of these features.

We have applied the suite of four gait sequence features to two tasks: person recognition and gender classification. Our experimental results from the recognition test showed that overall, the image repre-

sentation we have chosen is sensitive to static changes in the walking silhouette that are not the direct result of the walking action itself. For example, changes in the clothing or hair styles of the walking subject and changes in the environment that alter the noise characteristics of the silhouette (for example, shadows) will distort the gait image representations, and as a consequence, affect the sequence representations and the performance in recognition tasks. Because we take the view that representations of gait should include the appearance of a subject in addition to the kinematic characteristics of the walking subject, this sensitivity is a desirable characteristic, provided silhouette noise resulting from environmental noise can be eliminated or reduced to a minimal amount. Our results show that the average appearance feature is the most sensitive to external silhouette appearance changes not related to the walking action. The fundamental harmonic features are the least sensitive to changes in the silhouette, and the histogram appearance features are more sensitive than the harmonic features, but less than the average appearance features. We found that given the amount of the noise and the low sampling rate in the time dimension of our gait image features, we were not able to accurately recover the second harmonic features for the purpose of recognition, even though clinical gait analysis shows clear evidence of the second harmonic components in joint angles. We have also explored combining sequence features containing independent dimensions—such as combining the histogram appearance representation with the time-related dimensions of the fundamental harmonic features—to improve the recognition performance. We also experimented with the baseline representation: the original un-aggregated time series of gait image features. Our results show that this representation consistently byt marginally outperforms the random algorithm.

We applied the average appearance features and the fundamental harmonic features to the gender classification task. We found that the average appearance features are better for gender classification than the fundamental harmonic features. In addition, we found that a subset of average appearance features chosen for their significance in gender classification resulted in better gender classification results than using the complete set of features. This subset of gender specific features appeals to our intuitions about the differing appearances between genders; for example, women on average tend to have more hair behind the head and have better upper body posture than men.

The gait image representation chosen in this thesis is clearly view-dependent. We briefly described our prior work on view normalization to overcome this dependence. Our algorithm involves using a real-

time visual hull system to obtain a 3D model of the walking subject, using the heading direction of the walker to choose a synthetic view direction, and finally synthesizing a frontal parallel view of the walker. The newly synthesized view was used in place of silhouettes obtained from adaptive foreground segmentation from a conventionally recorded video sequence.

## 7.1 Alternative Silhouette Representations

One of the areas that could most benefit from further study is the image representation of the silhouettes. While we have shown that localized region descriptions of the shape of silhouette provides a reasonable image representation solution, it is unclear that dividing the walking silhouette in the manner we described in Chapter 2 yields the best or even close to the best results. There are many alternative representations that could be tried.

One solution to the walking silhouette image solution that appeals to the intuition of humans is to divide the silhouette into biologically relevant components. There have been medical studies done on cadavers to measure the size and weight of body segments [11]. One simple alternative to the fixed grid regions as described in Chapter 2 is to divide the silhouette into regions that have a higher correspondence to the average body size parameters as measured from cadavers. A further improvement would be to use a division of the silhouette that is adaptive to each person, or even better, to adapt the region divisions to each frame of a silhouette. Yoo *et al.* [43] showed a method for extracting a stick figure representation of the frontal parallel walking figure by adapting a fixed body plan based on medical studies to segment the body into biologically relevant components. They searched for the joint positions near the expected location of joints using heuristic methods. The effectiveness of an accurate joint locator is likely to be heavily dependent on the quality of the silhouette obtained from background subtraction. The silhouettes used by Yoo *et al.* are the same in quality to the ones we saw in Chapter 5 from the Soton data which, by visual inspection, are the best ones used in this thesis, mainly due to the chroma-keyed background.

Other alternatives of fixed template divisions of the walking figures include the $W^4$ work by Haritaoglu *et al.* [14]. The authors divided the silhouette of the walking figure into fixed template regions, then fitted a cardboard model of the human body based on a scheme similar to that used by Ju, Black, and Yacoob [19]. These fitted cardboard

models were then used to locate the extremities of the silhouette, such as the tips of the arms and feet. While Haritaoglu *et al.* did not use the extremities to identify people, one could imagine that the trajectory of the extremities could be used for identification purposes if they could be accurately recovered.

The silhouettes of the walking figure may be used in person recognition tasks without extracting local descriptions of the silhouette. The method used by Little and Boyd [24] is one such example. The authors computed moments of the whole silhouette of the walking figure in addition to moments based on optical flow in the silhouette regions and used these features for gait recognition. They achieved good recognition results on small number of subjects. Other whole silhouette features, such as a measure of the symmetry of a silhouette, may also be used as a gait signature.

## 7.2 Non Silhouette-based Representations

Niyogi and Adelson [31] took a horizontal slice through different points of the body and across the time dimension of a video sequence to detect twisted pairs of signals in the $X-T$ plane generated by human walking action. These twisted pairs of signals were modeled with snakes [21] and used to distinguish different walkers.

Joint locations recorded using a motion capture system have been used by Tanawongsuwan *et al.* [41]. The authors used the joint location trajectories for direct matching of sequences using dynamic time warping. The authors showed that while recognition results are excellent for data recorded on the same day, the results were significantly worse for data collected on different days. They attributed the degradation to inconsistencies in the placement of markers on the body. This result points to a flaw in using joint locations and joint angles directly: they are very fragile and sensitive to experimental conditions.

While finding joints in silhouettes is a difficult problem, it may be much easier if a real-time 3D model of the walking figure were available. As we described in the previous chapter, we were able to use a real-time visual hull system to synthesize a frontal-parallel view of the walking gait. Moreover, the visual hull system has a 3D volumetric representation of the walking figure constructed by intersecting the pyramids projected in space by the silhouettes from each camera view. The three-dimensional data can remove the ambiguity that results from using silhouettes in frontal0parallel view, and it contains much more information about the shapes of the body components.

## 7.3 Alternative Sequence Representations

We presented four types of gait sequence features, the average appearance, the histogram appearance, the harmonic features, and the original time series of our image features. Two of our sequence features contain no time information, the average appearance features and the histogram appearance features. The remaining two contain information about the time dimension. There are many alternative representations of the time dimension of gait sequence features.

The time dimension of the gait sequence may be discarded in a gait sequence representation. For example, Collins *et al.* [7] detected key frames of a gait video silhouette sequence and used the silhouettes directly to compare the appearances of walking subjects.

Other methods retain much more information in the time dimension of the gait sequence data. The frequency and phase components of gait sequences have been encoded in self similarity template images [8] which were then used for person recognition [2]. Gait frequency and phase information can also be used directly for recognition, as shown by Yoo *et al.* in [44]. Gait time dependence may also be encoded in a hidden Markov model and used for person recognition by gait [20].

## 7.4 High-Level Characterization from Gait

We have demonstrated in this thesis that gait features extracted from silhouettes of walking figures contain identity and gender related information. While the task of identifying walking subjects is interesting, it is not a general scenario. In most surveillance situations, an automatic visual surveillance system does not know all the walking subjects and hence cannot identify the individuals. In these cases, it is more appropriate to provide descriptive information about walking subjects, such as gender. A further question to ask is, "Does gait contain any other informative characteristics?"

We argue that gait contains much more information than identity and gender. One such example is the size of a person. A rounded silhouette may indicate that the walking subject may be over-weight, or is wearing thick clothing, but a thin silhouette indicates that the subject is slim. In addition, the height of a walking subject may be estimated by having a calibrated environment, as shown by Bobick and Johnson [18]. Asymmetry in the gait of a person is also a detectable characteristic. One may also ask if the walking subject has a hunched back, exaggerated arm movements, a long/short torso, or has stride

lengths that are long, or short, compared to the body length. In addition, a gait characterization system could report information about how "distinct" a walking subject is with respect to a particular population.

We have done preliminary studies which indicate that some of the subjects in our gait database are very distinguishable from the rest of those in the database because they have characteristics that "stand out" among a population, while other subjects in our gait database are very "average-looking" and hence easily mistaken for one another. These preliminary results lead us to believe that gait appearance by itself may not be enough to characterize some subjects, but combined with other modalities we may have a much more powerful person characterization tool. For example, a person with a very average gait appearance may have a very distinct face or a unique clothing style. Our grand vision of an automatic person surveillance system combines gait, face, clothing color and style, daily routine, as well as other aspects of appearance or activity, and fuses these pieces of information together to arrive at a customizable description of a walking subject.

# Appendix A

# Dynamic Time Warping

Dynamic time warping (DTW) was developed in the speech recognition community to compare two speech signals uttered at different speeds [36]. The warping is based on dynamic programming methods. We give a brief description of the algorithm as follows.

Given two sequences, $Q$ and $C$, of lengths $m$ and $n$, where

$$Q = q_1, q_2, \ldots, q_m;$$
$$C = c_1, c_2, \ldots, c_n;$$

DTW constructs an $m \times n$ matrix that contains the distances between samples $q_i$ and $c_j$ using, for example, the Euclidean distance,

$$d(q_i, c_j) = (q_i - c_j)^2.$$

A warping function, $W$, is a function mapping one sequence to the other,

$$W = w_1, w_2, \ldots, w_k; \text{where, } \max(m, n) \leq K < m + n - 1.$$

The warping function has the following properties:

1. $w_1 = (1, 1)$ and $w_k = (m, n)$, *i.e.*, the warping function must map the beginnings of the sequences together, and the ends of the sequences together.

2. If $w_k = (a, b)$, then $w_{k+1} = (a', b')$, where $0 \leq a' - a \leq 1$ and $0 \leq b' - b \leq 1$, *i.e.* each warping step may only map to the adjacent cells and the warping must monotonically increase in time.

DTW uses dynamic programming to minimize the cost of warping sequence $Q$ and $C$ together. In other words, DTW minimizes the following function,

$$\mathrm{DTW}(Q, C) = \text{minimize} \left( \frac{\sqrt{\sum_{k=1}^{K} w_k}}{K} \right).$$

The similarity between the sequences $Q$ and $C$ can be measured with the above cost function, which in our case is used directly as a similarity measure between sequences.

# Bibliography


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

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

[3] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(3), March 2001.

[4] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. of IEEE CVPR*, 1998.

[5] J. Carter. Personal communication. May 2002.

[6] G. Casella and R. Berger. *Statistical Inference*. Duxbury, 1990.

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

[8] R. Cutler and L. Davis. Robust real-time periodic motion detection, analysis, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):781 –796, 2000.

[9] J. Cutting and L. Kozlowski. Recognizing friends by their walk: gait perception without familiarity cues. *Bull. Psychonometric Soc.*, (9):353–356, 1977.

[10] J. Cutting, D. Proffitt, and L. Kozlowski. A biomechanical invariant for gait perception. *Journal of Experimental Psychology: Human Perception and Performance*, 4(3):357–372, 1978.

[11] W. Dempster and G. R. L. Gaughran. Properties of body sements based on size and weight. *American Journal of Anatomy*, (120):33–54, 1967.

[12] D. DiFranco, T. Cham, and J. Rehg. Recovery of 3d articulated motion from 2d correspondences. Technical report, Compaq Cambridge Research Lab, 1999.

[13] N. Goddard. *The Perception of Articulated Motion: Recognizing Moving Light Displays*. PhD thesis, University of Rochester, 1992.

[14] I. Haritaoglu, D. Harwood, and L. Davis. W4: Who? when? where? what? In *International Conference on Automatic Face and Gesture Recognition*, 1998.

[15] B. K. P. Horn. *Robot Vision*, pages 66–69, 299–333. The MIT Press, 1986.
[16] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophisics*, (14):201–211, 1973.
[17] G. Johansson. Visual motion perception. *Scientific American*, (232):76–88, 1975.
[18] A. Johnson and A. Bobick. Gait recognition using static, activity-specific parameters. In *CVPR*, 2001.
[19] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. In *International Conference on Face and Gesture Analysis*, 1996.
[20] A. Kale, A. Rajagopalan, N. Cuntoor, and V. Kruger. Gait-based recognition of humans using continuous hmms. In *Proc. Int. Conf. Automatic Face and Gesture Recognition*, pages 336–341, 2002.
[21] M. Kass, A. Widkin, and D. Terzopoulos. Snakes: Active countour models. *Intern. J. Computer Vision*, 1987.
[22] L. Kozlowski and J. Cutting. Recognizing the sex of a walker from a dynamic point-light display. *Perception and Psychophysics*, 21(6):575–580, 1977.
[23] A. Laurentini. The visual hull concept for silhouette-based image understanding. *PAMI*, 16(2):150–162, 1994.
[24] J. Little and J. Boyd. Recognizing people by their gait: The shape of motion. *Videre*, 1(2), 1998.
[25] J. Mass and G. Johansson. Motion perception i: 2-dimensional motion perception. Boston: Houghton Mifflin. film, 1971.
[26] W. Matusik, C. Buehler, and L. McMillan. Polyhedral visual hulls for real-time rendering. In *Eurographics Workshop on Rendering*, 2001.
[27] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan. Image-based visual hulls. In *SIGGRAPH Computer Graphics Proceedings*, pages 369–374, 2000.
[28] M. P. Murray. Gait as a total pattern of movement. *American Journal of Physical Medicine*, 46(1):290–333, 1967.
[29] M. P. Murray, A. B. Drought, and R. C. Kory. Walking pattern of normal men. *Journal of Bone and Joint Surgery*, 46A(2):335–360, 1964.
[30] M. Nixon, J. Carter, J. Nash, P. Huang, D. Cunado, and S. Stevenage. Automatic gait recognition. In *Motion Analysis and Tracking (Ref. No. 1999/103), IEE Colloquium on*, pages 3/1–3/6, 1999.
[31] S. Niyogi and E. Adelson. Analying and recognizing walking figures in xyt. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 469–474, 1994.
[32] P. Phillips, H. Moon, P. Rauss, and S. Rizvi. The feret evaluation methodology for face-recognition algorithms. In *CVPR*, pages 137–143, 1997.
[33] R. Polana and R. Nelson. Detecting activities. In *CVPR*, 1993.
[34] J. Rehg and D. Morris. Singularities in articulated objtect tracking with 2-d and 3-d models. Technical Report CRL 98/8, Compaq Cambridge Research Lab, 1997.

[35] R. Rifkin. *SvmFu*. http://five-percent-nation.mit.edu/SvmFu.
[36] H. Sakoe and R. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoustics, Speech, Signal Processing*, 26, 1978.
[37] G. Shakhnarovich, L. Lee, and T. Darrell. Integrated face and gait recognition from multiple views. In *CVPR*, 2001.
[38] J. Shutler, M. Nixon, and C. Harris. Statistical gait recognition via velocity moments. In *Visual Biometrics, IEEE*, pages 11/1–11/5, 2000.
[39] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3d human figures using 2d image motion. In *European Conference on Computer Vision (ECCV)*, volume 2, pages 702–718, 2000.
[40] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. In *CVPR*, 1999.
[41] R. Tanawongsuwan and A. Bobick. Gait recognition from time-normalized joing-angle trajectories in the walking plane. In *CVPR*, 2001.
[42] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.
[43] J.-H. Yoo, M. Nixon, and C. Harris. Extracting human gait signatures by body segment properties. In *Proc. IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 35–39, 2002.
[44] J.-H. Yoo, M. Nixon, and C. Harris. Extraction and description of moving human body by periodic motion analysis. In *Proc. ISCA 17th International Conference on Computers and Their Applications*, pages 100–113, 2002.