Identification of Humans Using Gait

A. Kale¹, A.N. Rajagopalan², N. Cuntoor ¹ V. Krueger ¹ and R. Chellappa¹ *

¹ Center for Automation Research
University of Maryland at College Park
College Park, MD 20740
² Department of Electrical Engineering
Indian Institute of Technology Madras
Chennai-600 036, India

August 2, 2002

EDICS 2-ANAL (Image/Video Processing:Analysis)

Abstract

In this paper we propose a view based approach to recognize humans using gait. The width of the outer contour of the binarized silhouette of a walking person is chosen as the image feature. A set of exemplars that occur during a walk cycle is chosen for each individual. Using these exemplars a lower dimensional Frame to Exemplar Distance (FED) vector is generated. A continuous HMM is trained using several such FED vector sequences. This methodology serves to compactly capture structural and dynamic features that are unique to an individual. The statistical nature of the HMM renders overall robustness to representation and recognition. Human identification performance of the proposed scheme is illustrated using outdoor video sequences.

*Supported by the DARPA/ONR grant N00014-00-1-0908.
1 Introduction

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

When person identification is attempted in natural settings such as those arising in surveillance applications, it takes on a new dimension. Biometrics such as fingerprint or iris are no longer applicable. Furthermore, night vision capability (an important component in surveillance) is usually not possible with these biometrics. Even though an IR camera would reveal the presence of people, the facial features are far from discernible in an IR image at large distances.

A biometric that can address some of these shortcomings is human 'gait' or the walking style of an individual. The attractiveness of using gait as a biometric is that it is non-intrusive and also typifies the motion characteristics specific to an individual. Anecdotal evidence suggests that people often recognize others by simply observing their gait. Gait can serve as a cue for recognizing people if the database is small. But for large databases, gait information, by itself, may not be sufficient to recognize an individual. In fact, we must realize that the gait recognition capability of even humans is limited. However, it is interesting to study the feasibility of using gait as a cue to narrow down the list of potential matches.
Approaches in computer vision to the gait recognition problem can be broadly classified as being either model-based or model-free. Both methodologies follow the general framework of feature extraction, feature correspondence and high-level processing. The major difference is with regard to feature correspondence between two consecutive frames. Methods which assume a priori models match the 2-D image sequences to the model data. Feature correspondence is automatically achieved once matching between the images and the model data is established. Examples of this approach include the work of Lee et al. [1], where several ellipses are fitted to different parts of the binarized silhouette of the person and the parameters of these ellipses such as location of its centroid, eccentricity etc. are used as a feature to represent the gait of a person. Recognition is achieved by template matching. In [2], Cunado et al. extract a gait signature by fitting the movement of the thighs to an articulated pendulum-like motion model. The idea is somewhat similar to an early work by Murray [3] who modeled the hip rotation angle as a simple pendulum, the motion of which was approximately described by simple harmonic motion. In [4] activity specific static parameters are extracted for gait recognition. Model-free methods establish correspondence between successive frames based upon the prediction or estimation of features related to position, velocity, shape, texture and color. Alternatively, they assume some implicit notion of what is being observed. Examples of this approach include the work of Huang et al. [5], who use optical flow to derive a motion image sequence for a walk cycle. Principal components analysis is then applied to the binarized silhouette to derive what are called eigen gaits. Benabdellkader et al. [6] use image self-similarity plots as a gait feature. Little and Boyd [7] extract frequency and phase features from moments of the motion image derived from optical flow and use template matching to recognize different people by their gait.

If a common scale is chosen, then a careful analysis of gait would reveal that it has two important components. The first is a structural component that captures the physical build of a person e.g. body mass, length of limbs etc. The second component is the motion dynamics of the body during a gait cycle. Our effort in this paper is directed towards deriving and fusing information from these two components. We propose a systematic approach to gait recognition by building representations for the structural and
dynamic components of gait. The assumptions we use are: (i) the camera is static and the only motion within the field of view is that of the moving person, (ii) the subject is monitored by multiple cameras so that the subject presents a side-view to at least one of the cameras. This is because the gait of a person is best brought out in the side-view. The image sequence of that camera which produces the best side view is used. Our experiments were set up in line with the above assumptions. We choose the width of the outer contour of the binarized silhouette as our basic image feature. A set of exemplars that occur during a gait cycle is derived for each individual. Using these exemplars, the high-dimensional width vector is transformed to a lower-dimensional space by generating what we call the Frame to Exemplar (FED) distance. The FED vector nicely captures both structural and dynamic traits of each individual. For compact and effective gait representation and recognition, the gait information in the FED vector sequences is captured in a hidden Markov model (HMM) and an HMM is built for each individual. The performance of the proposed method is tested on different databases including one of our own. For a modest-sized database of 44 people captured in natural settings, the method performs reasonably well.

The organization of the paper is as follows. In Section 2 we describe the basics of gait. Section 3 explores the issue of feature selection. In Section 4, we describe exemplars for gait. In Section 5, we discuss the proposed algorithm for gait representation and recognition. Section 6 describes experimental results while Section 7 concludes the paper.

2 What is gait?

As mentioned earlier, gait characterizes the walking style of an individual. A gait cycle refers to the time interval between successive instances of initial foot-to-floor contact for the same foot. A half gait cycle refers to the time interval between instances of foot-to-floor contact of one foot and the foot-to-floor contact of the other foot. A full gait cycle comprises of two half gait cycles. Each leg has two distinct periods: a stance phase, when the foot is in contact with the floor, a swing phase when the foot is off the floor and the phase of moving forward to the next step. As shown in Figure 1, the cycle begins with the heel strike of one foot which marks the start of the stance phase. The ankle flexes to bring
the foot flat on the floor and the body weight is transferred to it. The other leg swings through in front as the heel lifts off the ground. As the body weight moves on to the other foot, the supporting knee flexes. The remainder of the foot, which is now behind, lifts off the ground ending the stance phase. The start of the swing phase is when the toes of the foot leave the ground. The weight is transferred onto the other leg and the leg swings forward to meet the ground in front of the other foot. The gait cycle ends with the heel strike of the foot. In a gait cycle, the stride length refers to the linear distance in the plane of progression between successive points of contact of the same foot while step length is the distance between the successive heel strikes of opposite feet. Concurrently, it is also possible to look at the swing of hands in a gait cycle.

Interestingly, studies in psychophysics [8] reveal that humans have the capability for recognizing people using impoverished displays of gait, indicating the presence of identity information. Early medical studies [3, 9] suggest that if all gait movements are considered, gait is unique. In all, it appears that there are 24 different components to human gait. However, from an image processing perspective, it is quite difficult to accurately extract these components. Precise extraction of body parts and joint angles in real visual imagery is a very cumbersome task and is often unreliable [10]. Hence, the problem of representing and recognizing gait turns out to be a challenging one.

3 Feature Selection

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

1. Use the entire silhouette.
2. Use only the outer contour of the silhouette.

We propose to use only the outer contour as we believe that it contains adequate information for recognizing gait. Of course, it can be argued that information inside the silhouette may also be relevant but we choose to ignore it for computational simplicity. The success of any methodology depends on a careful choice of the feature vector. For the gait problem, we choose the width of the outer contour of the silhouette as our feature vector. As we shall demonstrate, this feature captures both structural and dynamic information for an individual. In Figure 2 we show plots of the width profiles of two different individuals for several gait cycles. Since we use only the extremities of the silhouette, the two halves of the gait cycle are almost indistinguishable. From hereon, we refer to half cycles as cycles, for the sake of brevity. In Figure 2, the x-axis denotes the frame index while the y-axis denotes the index of the width vector. The \( i \)th horizontal line in the image shows the variations in the \( i \)th element of the width vector as a function of time. A brighter gray-scale indicates a higher value of the width. From the figure, we observe that within each half cycle, there is a systematic temporal progression of width vector magnitude, viz. the dynamics. A similar observation has been made in [11] where the gait patterns are analyzed as Frieze patterns.

For the two width profile plots shown in Figure 2, the differences are quite visible and interesting. For instance, by observing the bright patterns in the upper region of the two images we see that the brightness is more pronounced in the first image as compared to the second. This area of the plot corresponds to the swings of the hand. Secondly, note that the brightness gradient (which translates to velocity in the video sequence) in the lower part of the images is more pronounced for Person 1 as compared to Person 2. This part of the plot corresponds to the swings of the extremities of the foot. Additionally, note that the height of the two persons is different as also, the visibility of the neck part. It must be pointed out, however, that the differences need not be so pronounced for all individuals. Thus, the width profile contains structural and dynamic information peculiar to each individual. Besides this, the use of the width feature imparts uniformity to feature representation across different individuals. Also, by choice, the width vector is translation-invariant. Hence, the width of the outer contour of the
silhouette is indeed a potentially good candidate as a feature. However, it is high dimensional and can be noisy.

Given the image sequence of a subject, the width vector is generated as follows:

- Background subtraction as discussed in [12] is first applied to the image sequence. To remove spurious noise, a standard $3 \times 3$ low-pass filter is applied to the resultant motion image.

- A bounding box is then placed around the part of the motion image that contains the moving person. The size of the box is chosen to accomodate the extreme cases of individuals in the database as regards height and girth. Further operations are carried out on this ‘box’.

- Because we are interested in only the width of the outer contour of the body, the left and right boundaries of the body are traced by examining the intensity of the pixel values with a weighted low pass filter from the left-most and right-most ends of the image. The width of the silhouette along each row of the image is then stored. The width along a given row is simply the difference in the locations of the right-most and the left-most boundary pixels in that row.

4 Exemplars for Gait

A closer examination of the physical process behind generation of gait reveals that, during a gait cycle, it is possible to identify certain distinct phases or stances. In Figure 3, we show five frames that we have picked from a gait cycle for two individuals. In the first stance, the person is at rest. In the second, he is just about to start and his hand is slightly raised. In the third stance, the hands and the feet are separated, while in the fourth, the hands and feet are displaced to a maximum. Finally, in the fifth stance, the person is returning to the rest state. Clearly, every person transits among these successive stances as he/she walks. Although, these stances are generic, there exist differences not only in their image appearance based on the physical build of an individual but also in the way an individual transits across these stances as he/she walks which represents the gait dynamics of the individual.
A reasonable way to build a structural representation for a person is to pick $N$ exemplars (or stances) $\mathcal{E} = e_1, \ldots, e_N$ from the pool of images that will minimize the error in representation of all the images of that person. If the overall average distortion is used as a criterion for codebook design, the selection of the $N$ exemplars is said to be optimal if the overall average distortion is minimized for that choice. There are two conditions for ensuring optimality. The first condition is that the optimal quantizer is realized by using a nearest neighbor selection rule.

$$q(x) = e_i, \iff d(x, e_i) \leq d(x, e_j), j \neq i, 1 \leq i, j \leq N$$

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

$$e_i = \arg\min_{e_i} E[d(x, e_i) | x \in C_i]$$

where the $C_i$s represent the Voronoi partitions [13] across the set of training images. To iteratively minimize the average distortion measure, the most widely used method is the $k$-means algorithm [13, 14]. We consider an arbitrary gait cycle for an individual and pick width vectors corresponding to $N$ images in the gait cycle as the initial centroids. The usual steps of classification, codebook updating and termination in the $k$-means algorithm follow resulting in an optimum exemplar set $\{e_1^*, \ldots, e_N^*\}$.

Of course, there is the issue of picking $N$. This is the classical problem of choosing the appropriate dimensionality of a model that will fit a given set of observations e.g. choice of degree for a polynomial regression and the choice of order for a multi-step Markov chain. The notion of ‘best fits’ can be precisely defined by objective functions involving a penalty for the model complexity. Examples include minimum Bayes information criterion, minimum description length etc. Similar problems exist for the case where there exists no parametric model for the data set e.g. vector quantization. In problems like image compression, it is common practice to look at the rate-distortion curves to examine the marginal reduction in the distortion by an increase in the bits per pixel. We take a similar approach here. In Figure 4, we plot the average distortion as a function of the number of centroids for one of
our databases. It can be observed that beyond five centroids, the average distortion does not decrease as rapidly as the increase in the number of centroids. Similar results were observed for the other two databases considered in this paper. Hence, we chose the number of exemplars for our problem as five.

The exemplars that we extract from the gait sequence can themselves be used in a naive fashion for recognition of an unknown person \( u \) as

\[
\mathbf{u} = \arg \min_i \sum_k \min_{\mathbf{e}_k \in \{1, \cdots, N\}} d(x_u(i), e_k^j)
\]

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

\[
\mathbf{A} = [p(e_i(t)|e_j(t-1))]
\]

(1)

for \( i, j \in \{1, \cdots, N\} \). The matrix \( \mathbf{A} \) encodes the dynamics in terms of state duration densities and transition probabilities.
5 The Proposed Algorithm

In this section, we propose an algorithm for gait representation and recognition based on the discussions in Section 3 and 4. The problem at hand is as follows:

Given \( m \) gait cycles for person \( j \), \( \mathcal{X}^j_m = \{ x_{11}^j, x_{12}^j, \ldots, x_{1T}^j \}, \ldots, \{ x_{m1}^j, x_{m2}^j, \ldots, x_{mT}^j \} \) where \( x_{ij}^j \) represent a suitable image representation such as intensity array, edge map etc., build a model for the gait of person \( j \) and use it to recognize this person from \( P \) different people in the database.

5.1 Gait Representation

Given feature vectors extracted from several gait cycles of a person, the gait representation problem involves effectively capturing the information in the feature vectors in a set of compact parameters. Often, in a practical situation, only a finite amount of training data is available and modeling can be difficult if the feature dimensionality is high. As discussed in Section 3, the width of the outer contour was found to be a good feature that captured the structural and dynamic information. But even the smallest dimension of the width vector of the silhouette is approximately 100. Therefore, the possibility of directly using the width feature vector is ruled out. A more compact way of encoding these observations, while retaining all relevant information, is required. We propose a novel way to address this issue:

Let \( x(t) \) denote the width vector corresponding to an image at time \( t \). The distance of \( x(t) \) from the width vectors corresponding to the exemplars \( e_i \in \mathcal{E} \) can be computed to build a Frame to Exemplar Distance (FED) \( f(t) \) which serves as a lower (5-)dimensional representation of the image at time \( t \). For instance for the \( j \)th individual we compute

\[
f_j^i(t) = d(x_i(t), e_j^i)
\]

(2)

where, \( t \in \{1, \ldots, T\}, e_j^i \) denotes the \( l \)th exemplar for the \( j \)th person and \( l \in \{1, \ldots, 5\} \). Thus, \( f_j^i(t) \) constitutes an observation vector for person \( j \). Similarly, \( f_j^i(t) \) represents the observation sequence of the person \( i \) encoded in terms of the exemplars of person \( j \). Note that the dimension of \( f_j^i(k) \) is only 5.
These observations can be derived for several such gait cycles in the database.

The elements of the new 5-D vector given by equation (4), measures the similarity between the observed image and the five stances and has the following significance: Let $ME_i^j$ denote the arithmetic mean of the norm of the new observation sequence across frames, where we have encoded the observation sequence of the person $i$ in terms of the stances of the person $j$. Thus,

$$ME_i^j = \frac{1}{T} \sum_{t=1}^{T} ||f_j^i(t)||$$  \hspace{1cm} (3)

where $T$ is the number of frames. In equation (3), $ME_i^j$ is a measure of the overall structural similarity between persons $i$ and $j$. It is natural to expect that $ME_i^j$ will be lower than $ME_j^j$ for $j \neq i$ since an observation sequence for person $i$ will be closer to the stances of person $i$ than those of person $j$. Hence, the structural information is clearly embedded in the values of the new observation vector.

Next, we examine the temporal information contained in a sequence of such FEDs. It is clear that as we examine a gait cycle, the proximity of a frame from each of the stances changes with time. Correspondingly, the elements of the vector $f_j^i(k)$ would reflect this feature. To elaborate this further, note that for a frame at the start of the gait cycle, the first element of the observation vector will be smaller in magnitude as compared to the remaining four elements. This is due to the proximity of this frame to the first stance. As time progresses, the first element will increase in magnitude because the frame moves closer to the second stance. The magnitude of the second element will decrease as long as the frame is close to the second stance and then it will start to increase as well. A similar behavior is observed in the rest of the elements of the vector. The duration for which an element of this vector stays low encodes the stance duration density as also the probability of transition to another stance. Figure 5 shows the evolution of the different components of the FED vector $f_j^i(t)$ for a half-gait cycle. As can be seen, there is a systematic succession of troughs for the different FED vector components across time.

We have considered the width of the silhouette as a feature for its simplicity. However, it is possible to view the image sequence in other direct representations such as edge maps. In particular, it is possible to eliminate the mapping of each image to an intermediate feature such as width. The distortion measure
in that case will involve a norm of image metrics such as Hausdorff distance [15] or shuffle distance [16] which seek to directly compare the edge-maps of two images.

As described in Section 4, it is possible to model transition across exemplars by a Markov matrix. For the person \( j \), it is possible to look upon the FED vector sequence \( f_j(t) \) as the observed manifestation of the transition across exemplars (a hidden process). A hidden Markov model (HMM) is the appropriate model for such a signal. HMMs use a Markov process to model the changing statistical characteristics that are only probabilistically manifested through the actual observations. The state sequence is hidden, and can only be observed through another set of observable stochastic processes. Each hidden state of the model is associated with a set of output probability distributions which can be either discrete probability mass functions or continuous probability density functions. Details on HMMs can be found in [17]. For the gait problem, the exemplars can be considered as analogues of states of the HMM while the FED vector sequence can be considered as the observed process. In the proposed model for gait, the primary HMM parameters of interest are the number of states, the initial probability vector \( \pi \), the transition probability matrix \( A \) and the output probability distribution \( B \) which we model as a continuous probability distribution. In this paper, \( \lambda = (A, B, \pi) \) will be used to compactly represent an HMM.

5.2 Recognition

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

To begin with, the given image sequence of the unknown person \( u \) is subjected to the same image processing operations as the training image sequence i.e., the width vector \( x^u(t) \) of this person is
generated for each frame and the FED vector \( f_j^u(t) \) is computed for all \( j \in \{1, \ldots, P\} \) using Equation (2). We wish to compute the likelihood that the observation sequence \( f_j^u \) was generated by the HMM corresponding to the \( j \)th person. This can be deciphered by using the forward algorithm [17] which computes this log probability as

\[
P_j = \log(P(f_j^u|\lambda_j))
\]

(4)

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

1. The Euclidean distance between \( X \) and the stances of person \( m \) will be smaller than that between \( X \) and any other person.

2. The pattern of transitions between stances/states for \( X \) will be closest to that for person \( m \).

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

6 Experimental Results

In this section, we demonstrate the performance of the proposed gait recognition system on real image sequences. Our experiments are aimed at finding how our methodology performs with respect to several different variations such as size of database, speed of walking, clothing, illumination etc. We have considered normal walk as well as treadmill data for analysis. Our video sequences were taken from
(i) Little and Boyd’s database [7] (ii) Carnegie Mellon University (CMU) database and (iii) the University of Maryland (UMD) database. The training and evaluation schemes are outlined below.

**Training:**

Silhouettes of the walking person are extracted using the silhouette extraction procedure described in Section 3. Given the training images for the person, we generate the width vectors corresponding to that person. $K$-means clustering with $N = 5$ is used to obtain the exemplar set for that person. To generate temporally aligned training cycles, we parse the video sequence into distinct half-cycles. For this purpose, we compute the width vector for each frame and identify the start frame of each gait cycle. This is easy to do since walking is a periodic activity and the norm of the width changes from a minimum to maximum. By identifying these local minima, the start frames are approximately identified. Next, the FED vector sequences are computed using the exemplars from Equation 2 for different gait cycles. For the Little and Boyd database about 22 cycles were available for training. For the CMU database 8 cycles were available while for the UMD database 10 cycles were available for training. The observed FED vector sequences corresponding to a person are used to learn an HMM model ($\lambda$) for that person. Note that each gait cycle may be of different lengths e.g. during a natural walk a person may change his speed from cycle to cycle. Through our use of HMMs, we are capable of directly handling such variations and do not require an explicit time normalization like [18].

**Testing:** Assume that we are given the gait cycles of an unknown person $u$ and HMM models $\lambda_i$, $1 \leq i \leq P$. We again parse the video sequence of the person similar to what we did for training and get the width vector sequence for each gait cycle. We compute the FED vector sequences from the gait cycles of the test person and the exemplar set $\mathcal{E}^j$ of every individual $j \in \{1, \ldots, P\}$. For each of the FED vector sequence thus generated the log-probability for each gait cycle is then computed using the Viterbi algorithm [17] as

$$\max_s \{P(x, S|\lambda_j)\}$$
while

\[ P(x|\lambda_j) = \sum_{\forall S} P(x, S|\lambda_j) \]

The Viterbi algorithm is efficient since it can operate in the logarithmic domain using only additions. For every gait cycle, we rank order the probabilities and corresponding person indices in descending order. We then evaluate performance by letting each person in the database be the unknown \( u \) and plot the fraction of times that the right person is within the top \( n \) matches as a function of \( n \). This curve known as the cumulative match score (CMS) characteristic was first proposed in the context of face recognition by Philips et al. [19].

### 6.1 The Little and Boyd Database

This is a small database \(^1\) of five people, captured using a tripod-mounted camera. The number of gait cycles for a given subject is about 44. We used half of the cycles for training and the other half for testing. The size of the image was 100 × 50. The small size of the database facilitates an easier qualitative description of the results. The proposed method was used to identify gait of the five people in this database. The results are given in Table 1. As can be seen, recognition is very good. The unknown person has been identified correctly in each case. In the database, persons 2, 3 and 4 have similar structures and it is interesting to note that the false alarms are also somewhat predominant for these 3 subjects (see Figure 6). There is no confusion in recognizing the gait of person 5, as is to be expected. Our technique has recognized each person by his/her gait correctly and with sufficiently good confidence.

The recognition capability of the proposed method was also tested for different structures of HMMs and different number of states. Both ergodic and left to right models were tried.

From Tables 1 and 2, it is clear that the ergodic model yields better classification as compared to the left-to-right model. We believe that this is due to the fact that left-to-right models impose a hard

---

\(^1\) More details about the data are available at the following URL: http://www-mitpress.mit.edu/e-journals/Videre/001/articles/Little-Boyd/gait/gait.html.
constraint whereas the ergodic model fills the entries in the transition probability matrix by deriving information from the given training sequence.

The issue of the number of states deserves attention special here. In Section 5 we had conjectured that the states and stances in a gait cycle are likely to be related. Note that the states of an HMM can be abstract quantities and it is not necessary that they must correspond to physical features of the underlying process. However, it would definitely be interesting if the physical phenomenon can guide the choice of the number of states in an HMM. We studied the performance of our gait recognizer as a function of the number of states. We experimented with 3, 5 and 8-state HMMs. The merit of a model is decided by examining the worst case performance of the HMMs for different cardinality of the states. The results for a 3 and 8-state HMM are given in Tables 3 and 4, respectively. We note that the recognition results, for both these cases are bad as compared to the five-state case. Thus, it appears that a five-state HMM is best suited for our experiments thereby confirming our conjecture relating the stances and the states. If there are $N$ distinct stances, then choosing less than $N$ states for the HMM results in under-modeling of the problem while choosing more than $N$ results in over-modeling of the problem. Interestingly, Samaria [20] had also made a similar observation in the context of face recognition wherein the states were related to facial features.

Finally, for different state cardinalities, we attempted to compute the optimal state sequence to uncover any hidden information. The Viterbi algorithm was used for this purpose. It was found that for the five-state model, the transitions in the states occur approximately at the same time instants that the shift in stances occurs in the observation sequence. On the other hand, the state sequence for 3-state and 8-state models did not lead itself to a ready physical interpretation, as expected.

### 6.2 The CMU Database

This database\(^2\) has 25 people walking at a fast pace and slow pace on a treadmill and a sequence of people walking while carrying a ball. There are about 16 cycles in each sequence. Half of the cycles

\(^2\)More details about the data are available at the URL http://hid.ri.cmu.edu/HidEval/evaluation.html.
were used for training and the other half for testing. The size of the image was $640 \times 480$. We did the following experiments on this database: (i) train on slow-walk and test on slow-walk, (ii) train on fast-walk and test on fast-walk, (iii) train on slow-walk and test on fast-walk, (iv) train on fast-walk and test on slow-walk (v) train on walk carrying a ball and test on walk carrying a ball. In the cases (i), (ii) and (v), for each person, the sequences were divided into two halves, one half used for training and the other for testing, while in the cases (iii) and (iv), the entire slow/fast sequence was used for training and the other fast/slow sequence was used for evaluation.

The results obtained using the proposed method are given in Figures 7 and 8. It can be seen that the right person in the top 3 matches 90% of the times for the cases where the training and testing sets correspond to the same case. Observe that the results on CMU database when the HMM is trained using cycles from slow walk and tested using cycles from fast walk, the result is poor compared to the situation when the training and testing scenarios are reversed. In an effort to understand this, we ran an experiment whereby we artificially increased the number of frames per activity cycle using interpolation and observed the resulting HMM. It was seen that the $A$ matrix tends towards diagonal dominance. This occurs on account of the fact that the HMM does not provide adequate representation of extreme temporal durations of activity. The probability of $t$ consecutive observations in state $i$ can be written as

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

where $d_i(t)$ is the probability of taking a self-loop at state $i$ for $t$ times viz. a geometric distribution. As the duration of the activity $T \to \infty$ for a fixed $N$, this causes $a_{ii} \to 1$ and $a_{ij} \to 0$. Clearly, the geometric distribution does not represent a realistic description of the state duration density in our gait-modeling problem. Similar issues have been raised in the context of speech recognition and a solution is to explicitly model the distribution of state duration as has been done by Russell [21]. For the case of training with fast-walk and testing on slow-walk, the dip in performance is caused due to the fact that for some individuals as biomechanics suggests, there is a considerable change in body dynamics and stride length as a person changes his speed. For example, observe Figure 9 which show a few frames in
the gait cycles of a person in the two scenarios. As is apparent from the figure, the posture as well as hand swings for the person are quite different in the cases of fast-walk and slow-walk.

When the subjects are walking with a ball in their hands, most of the gait dynamics are confined to the leg region. For this experiment i.e. case (v), we observe from Figure 10, that the top match is the correct match 90% of the time which is higher than the top match score (around 70%) in the normal walk cases. This suggests that for the purpose of recognition, certain parts of the body may be more favourable than others. In particular, the leg motion provides more discriminating evidence as compared to evidence provided by hand and leg motion together.

6.3 The UMD Database

It would be very useful to evaluate the utility of gait as a biometric in more realistic situations than those prevailing in the previous two databases. For example, Little and Boyd’s database consists of just five people who were captured by a tripod mounted camera while the CMU database consists of individuals walking indoors on a treadmill. To get a more realistic evaluation of gait, we designed our own experiment at the University of Maryland. Unlike previous databases, we used outdoor surveillance cameras (Philips G3 EnviroDome camera system) at a height of 15 ft to capture data. The subjects were made to walk along a T-shaped path so that they present a side view to the surveillance cameras. This is in accordance with our basic assumption in Section 1. We collected gait sequences of 44 individuals. For most individuals, the training and test video sequences were collected on different days. The database\(^3\) had diversity in terms of gender, age, ethnicity etc. Moreover, there was a change in clothing of the people across different days as well. This and the fact that the data was collected outdoors under uncontrolled illumination provides a realistic scenario for gait analysis. Each video sequence has approximately 10 cycles. One sequence was used for training and the other for evaluation. The size of the image was 150 × 75. The result using the proposed method is shown in Figure 11 and it can be seen that the performance of the method does not degrade with an increase in the database size. The slight

---

\(^3\) More details about data are available at the URL: http://degas.umiacs.umd.edu/hid/
drop in performance can be attributed to drastic changes in clothing conditions of some subjects and changes in illumination resulting in very noisy binarized silhouettes.

7 Conclusion

In this paper, we have presented a new approach to represent and recognize people by their gait. A low-dimensional observation sequence (FED) is derived from the silhouettes during a gait cycle and an HMM is trained for each person. Gait identification is performed by evaluating the probability that a given observation sequence was generated by a particular HMM model. The method was tested on three different databases. In general, the recognition rates were encouraging. As anticipated, drastic changes in clothing adversely affect recognition performance. The method is also reasonably robust to changes in speed. In some cases, the stride length changed appreciably with walking speed causing a slight drop in the recognition performance.

Acknowledgement

The authors would like to thank Dr. B. Yegnanarayana, Indian Institute of Technology, Madras for helpful discussions on hidden Markov models.
References


<table>
<thead>
<tr>
<th>$T \setminus R$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>19</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21</td>
</tr>
</tbody>
</table>

Table 1: Confusion matrix for five-state ergodic model (Little and Boyd database).

<table>
<thead>
<tr>
<th>$T \setminus R$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>6</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21</td>
</tr>
</tbody>
</table>

Table 2: Confusion matrix for Left-to-right model (Little and Boyd database).

<table>
<thead>
<tr>
<th>$T \setminus R$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>6</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>6</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrix for 3-state model (Little and Boyd database).

<table>
<thead>
<tr>
<th>$T \setminus R$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>6</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>21</td>
</tr>
</tbody>
</table>

Table 4: Confusion matrix for 8-state model (Little and Boyd database).
Figure 1: A typical gait cycle for two individuals.
Figure 2: Width vector profile for several gait cycles of (a) Person 1 and (b) Person 2.
Figure 3: Stances corresponding to the gait cycle of (a) Person 1 and (b) Person 2.
Figure 4: Rate-distortion curve for number of exemplars vs distortion.

Figure 5: FED vector components plotted as a function of time.
Figure 6: Stances of the different people in the Little and Boyd database.
Figure 7: Performance on the CMU database (a) training set: slow-walk, testing set: slow-walk (b) training set: fast-walk, testing set: fast-walk.

Figure 8: Performance on the CMU database (a) training set: slow-walk, testing set: fast-walk (b) training set: fast-walk, testing set: slow-walk.
Figure 9: Sample images of a person corresponding to (a) slow-walk and (b) fast-walk.

Figure 10: Comparison of results for normal walk and walk when carrying an object.
Figure 11: Results for the UMD database (44 people).