Effect of Silhouette Quality on Hard Problems in Gait Recognition

Zongyi Liu and Sudeep Sarkar, *Member, IEEE*

***Abstract*—Gait as a behavioral biometric has been the subject of recent investigations. However, understanding the limits of gait-based recognition and the quantitative study of the factors effecting gait have been confounded by errors in the extracted silhouettes, upon which most recognition algorithms are based. To enable us to study this effect on a large population of subjects, we present a novel model based silhouette reconstruction strategy, based on a population based hidden Markov model (HMM), coupled with an eigen-stance model, to correct for common errors in silhouette detection arising from shadows and background subtraction. The model is trained and benchmarked using manually specified silhouettes for 71 subjects from the recently formulated HumanID Gait Challenge database. Unlike other essentially pixel-level silhouette cleaning methods, this method can remove shadows, especially between feet for the legs-apart stance, and remove parts due to any objects being carried, such as briefcase or a walking cane. After quantitatively establishing the improved quality of the silhouette over simple background subtraction, we show on the 122 subjects HumanID Gait Challenge Dataset and using two gait recognition algorithms that the observed poor performance of gait recognition for hard problems involving matching across factors such as surface, time, and shoe are *not* due to poor silhouette quality, beyond what is available from statistical background subtraction based methods.**

***Index Terms*—Eigen-stance, gait recognition, population hidden Markov model (HMM), segmentation.**

I. Introduction

RECOGNITION of a person from gait has been a recent focus in computer vision. It is a behavioral biometric source that can be acquired at a distance. At this nascent stage of gait recognition research, the pertinent research questions are those related to understanding the limits of gait recognition and the quantitative study of the various factors effecting gait. In the early 1970s, experiments by Cutting and Kozlowski [13] using light point display experiments, pioneered by Johansson [14], indicated that it is possible to identify a person from the manner of walking, i.e. gait. In computer vision, there have been lot of work on modeling human motion, however, the first effort toward *recognition from gait* was probably by Niyogi and Adelson in the early 1990s [15]. Since then there have been a variety of approaches to gait recognition [1], [3], [4], [6]–[8], [11], [16]–[19].

Manuscript received February 27, 2004; revised June 22, 2004. This work was supported by the Defense Advanced Research Projects Agency AFOSR-F49620-00-1-00388.This paper was recommended by Associate Editor C. T. Lin.

The authors are with the Computer Science and Engineering Department, University of South Florida, Tampa, FL 33620 USA (e-mail: zliu4@csee.usf.edu; sarkar@csee.usf.edu).

Digital Object Identifier 10.1109/TSMCB.2004.842251

TABLE I
Summary of Algorithms and Performances Reported by Current Work on Gait-Based Identification

Algorithm	Scene	Data Covariates	Top rank P_I (# of subjects)
Temporal body length vector correlation (CAS) [1], [2]	Indoor	Time[1]	91%(28)
Shape symmetry analysis (U. of Southampton) [3]	Indoor		95%(6)
Sil. region moment (MIT) [4]	Indoor	Temporal[2]	30-60%(24)
Comb. Exemplars and HMM (UMD) [5]	Indoor Outdoor	Speed, view Session	58%(25) 55%(43)
Continous HMM (UMD) [6]	Indoor Indoor	Speed, view Time[3]	58%(25) 30%(24)
Motion based eigenspace analysis (UMD) [7]	Indoor Outdoor	Speed, view Sessions	76%(25) 75%(44)
Body shape and template correlation (CMU) [8]	Indoor Indoor Indoor Outdoor	Speed, view Time[3] Time[1] Sessions	76%(25) 45%(24) 93%(28) 85%(55)
Gait frieze pattern (CMU) [9]	Indoor	Speed	100%(25)
Static body parameters (Georgia Tech.) [10]–[12]	Indoor Indoor Magnetic	View Sessions Speed	>90%(18) 73%(18) 40%(15)

[1] Minutes, [2] Days, [3] Months.

A review of the current studies, summarized in Table I, shows that they report different performances on datasets of varying sizes and with differing variations in conditions, such as time, speed, and viewpoint. The listed performance numbers are the correct identification rates at the top most rank, i.e. fraction of times the correct match to a probe (the input data) is the top ranked match among all the matches of that probe to the complete gallery set (the prestored templates or models). This is a standard performance metric used in biometrics [20] for the identification scenario, where one is interested to find a match to a given probe from whole the gallery set, i.e. one-to-many match. (For the verification scenario, where one is interested in matching one probe to one gallery (one-to-one match), the performance is specified in terms of standard false alarm and detection rates. In general, identification is considered to be a harder problem than verification.) It is difficult to compare and contrast performance among different algorithms tested on different datasets, but two observations that can made from these studies are the following.

1) Performance on indoor sequences [1], [3], [8] generally tend to be higher than on outdoor sequences [5], [7], [8].
2) Gait recognition performance drops when comparing sequences taken at different times. When the difference in time between gallery (the prestored template) and probe (the input data) is in the order of minutes, the identification performance ranges from 91% to 95% [1], [3], [8], whereas the performances drop from 30% to 45% when the differences are in the order of months and days [4], [8], [21] for similar sized datasets.

A. HumanID Gait Challenge Problem

More objective, quantitative measurement of progress, and the characterization of the properties of gait recognition can be made only based on performances on a *common data set* for a set of common experiments. Toward this end, the HumanID Gait Challenge Problem was formulated [22], [23]. The challenge problem consists of a baseline algorithm, a set of twelve experiments (A–L), and a large data set (1870 sequences, 122 subjects, 1.2 TB of data).

The *database*: was collected outdoors and each person in the data set was studied by varying as many as the five following conditions:

1) two camera angles;
2) two shoe types (A and B);
3) two surfaces [grass(G) and concrete(C)];
4) with and without carrying a brief case (BF or NB);
5) two different dates six months apart [May(M) or November(N)].

Attempt was made to acquire a person's gait in all possible combinations, and there are up to 32 sequences for some persons. The full data set consists of 1870 sequences from 122 individuals. This dataset is significantly larger than those that are being used in present studies and is also unique in the number of covariates exercised. It is the only data set to include walking on a grass surface.

There are twelve *challenge experiments* designed to investigate the effect on performance of the five factors, i.e. change in viewing angle, change in shoe type, change in walking surfaces (concrete and grass), carrying or not carrying a briefcase, and temporal differences. The results from the twelve experiments provide an ordering of the difficulty of the experiments. Patterned after FERET (face) evaluations [20], the challenge tasks are specified in terms of gallery and probe sets. In biometrics nomenclature, the gallery is the set of people known to an algorithm or system, and probes are signatures given to an algorithm to be recognized. The signatures in the present case are the gait video sequences. To allow for a comparison among a set of experiments and to limit the total number of experiments, the gallery was fixed as the control. Then, twelve probe sets were created to examine the effects of different covariates on performance. The gallery consists of sequences with the following covariates: Grass, Shoe Type A, Right Camera, No Briefcase, and collected in May along with those from the *new* subjects from November. This set was selected as the gallery because it was one of the largest for a given set of covariates. Table II lists the five key experiments that exercise only the key covariates. We focus on these experiments in this paper. Three observations are worth noting. First, the time covariate implicitly includes a change of shoe and clothing because we did not require subjects to wear the same clothes or shoes in both data collections. Second, except for Experiment K, the gallery and the probe sequences for the other key experiments were collected around the same time. Third, there is difference in background between gallery and probe only for Experiment D, which exercises surface condition change. For Experiment K, the probe background represents the same location as in the gallery, but after six months.

TABLE II
PROBE SET FOR THE KEY GAIT CHALLENGE EXPERIMENTS, EXERCISING A SINGLE COVARIATE. THE GALLERY FOR ALL OF THE EXPERIMENTS IS (G, A, R, NB, M/N) AND CONSISTS OF 122 INDIVIDUALS. THE ABBREVIATIONS USED ARE AS FOLLOWS: C—CONCRETE, G—GRASS, A—SHOEA, B—SHOEB, R—RIGHT VIEW, L—LEFT VIEW, NB—NO BRIEFCASE, BF—CARRYING BRIEFCASE, M—MAY, N—NOVEMBER

Exp.	Probe (Surf., Shoe, View, Carry, Time) (C/G, A/B, L/R, NB/BF, M/N)	Number of Subjects	Diff.
A	(G, A, L, NB, M/N)	122	View
B	(G, B, R, NB, M/N)	54	Shoe
D	(C, A, R, NB, M/N)	121	Surf.
H	(G, A, R, BF, M/N)	120	Carry
K	(G, A/B, R, NB, N)	33	Time

TABLE III
SUMMARY THE TOP RANK RECOGNITION FOR KEY EXPERIMENTS A (VIEWPOINT), B (SHOE-TYPE), D (SURFACE), H (CARRY), AND K (TIME) IN THE GAIT CHALLENGE DATASET. THE NUMBERS FOR THE FIRST TWO COLUMNS ARE AS READ FROM GRAPHS IN THE CITED PAPERS

Alg.	A	B	D	H	K	# subjects in gallery
Fusion (UMD) [21]	52%	40%	18%			71
DTW (UMD) [24]	78%	65%	10%			71
HMM (UMD) [25]	99%	89%	36%			71
Shape (CMU) [26]	87%	81%	21%			71
HMM (MIT) [27]	88%	75%	25%			71
Body (CAS) [2]	70%	59%	34%			71
Baseline (USF) [22]	79%	66%	30%			71
Baseline (USF)	73%	78%	32%	61%	3%	122

The Gait Challenge problem also comes with a baseline gait recognition algorithm based on spatio-temporal template correlation. We provide some more details about this algorithm in Section V-A. The detailed specifications of the experiments, the gait data, the source code of the baseline algorithm, and scripts to run, score, and analyze the challenge experiments are available at www.GaitChallenge.org.

B. Lessons From Performance on Gait Challenge Data

For the key Gait Challenge experiments, Table III lists summary performances that have been so far reported in the literature on an earlier smaller release of the dataset. Results with the baseline gait recognition algorithm are listed on the small data subset used in the first three studies, as well as on the complete dataset. The listed performance numbers are the correct identification rates at the top most rank. We see that performance of the baseline gait recognition algorithm is quite effective and

competitive with other algorithms. Another observation of particular interest, is the significant effect of change in surface type on the identification rate; this effect is consistent across different types of gait recognition algorithms. For the baseline algorithm, we also see the significant effect of time variation of about six months. This effect of time on gait recognition has been, as we mentioned earlier, documented by others too, but on different (all indoor) data sets. Thus, the hard problems in gait recognition have to do with walking surface invariant gait recognition and being able to overcome gait variation of a person over time.

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

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

C. Hypothesis, Method, and Outline

One might speculate that *if only we had a better background subtraction algorithm to generate the high quality silhouettes, we would be able to get better gait recognition performance on the hard problems*. However, we hypothesize that *the fall in performance due to changes in surface and time cannot be explained fully by silhouette quality*. In fact, silhouette quality, beyond what is available from simple background subtraction based methods, probably has little impact on performance across these covariates. It is also interesting to note that poor quality segmentations, especially for outdoor sequences, can also contribute to *higher* recognition especially for sequences that are collected around the same time and against the same background (Experiments A, B, and H). The high recognition might result due the added information from the shapes of the shadows, which would be similar for the same person across probe and gallery, and the correlated background errors, which are never random. In fact, we do observe this effect.

Our observations are based on performance of two gait recognition algorithms on cleaned silhouettes, generated by a novel model based silhouette reconstruction strategy. This strategy is based on a population-based hidden Markov model (HMM), coupled with an eigen-stance model, that corrects for common errors in silhouette detection arising from shadows and background subtraction. The model is trained and benchmarked using manually specified silhouettes for 71 subjects from the recently formulated HumanID Gait Challenge database. Unlike other essentially pixel-level silhouette cleaning methods, this method can remove shadows, especially between feet for the legs-apart stance, and remove parts due to any objects being carried, such as briefcase or a walking cane. We quantitatively establish the improved quality of the silhouettes over simple background subtraction and demonstrate that it is generalizable to different data sets.

In the next section, we summarize the creation of the manual silhouettes, which are used to train the model-based silhouette reconstruction strategy and also used directly for gait recognition. The HMM-based silhouette recognition strategy is presented in Section III. The improved quality of the reconstructed silhouettes is established in Section IV. In Section V we present gait recognition results with these improved silhouettes and directly using the manual silhouettes. We discuss the results and conclude with Section VI.

II. Manual Silhouettes

Manual silhouettes were created for a subset of Gait Challenge dataset. More details about the process and quality checks can be found in [28], [29], here we highlight some salient aspects. Up to 71 subjects from one of the two collection periods (May collection) were chosen for manual silhouette specification. The sequences corresponding to these subjects were chosen from the following:

1) gallery set (sequences taken on grass, with shoe type A, right camera view);
2) probe B (on grass, with shoe type B, right camera view);
3) probe D (on concrete, with shoe type A, right camera view);
4) probe H (on grass, with shoe A, right camera view, carrying briefcase), and probe K (on grass, time).

We manually specified the silhouette in each frame over one walking cycle, of approximately 30-40 image frames. This cycle was chosen to begin at the right heel strike phase of the walking cycle through to the next right heel strike. We attempted to pick this gait cycle from the same three-dimensional (3-D) location in each sequence, whenever possible. In addition, we tried to exclude the portion that included the black and white calibration box with high contrast, which frequently leads to high background subtraction errors.

We did not just mark a pixel as being from the background or subject, but provided more detailed specifications in terms of body parts too. We explicitly labeled the head, torso, left arm, right arm, left upper leg, left lower leg, right upper leg, and right lower leg using different colors. Quality control checks looked for miscolored parts and backgrounds, randomly colored isolated pixels, errors on the boundary of the body, and missed body parts. Some of the difficulties encountered during the creating process include low-image quality due to varying overall

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

Fig. 1 shows the silhouettes of a subject in image frames taken from four different cameras at different distances and surfaces. To remove possible bias in recognition due the use of silhouette height, we normalize the height of the silhouettes to occupy 128 pixels. The bottom row of Fig. 1 shows the height scaled and centered silhouettes of the kind used by gait recognition algorithms. To facilitate the fast computation of similarity measure, following the baseline algorithm [22], we also align the silhouettes in each frame along the horizontal direction so that the centerline of the torso is at the middle of the frame. This centerline is estimated as follows. First, we compute the number of *connected* foreground pixels in each row in the *upper half* of the silhouette. If there are more than one section of connected foreground pixels in a row, e.g., when person's arm move out of torso, we consider the largest one, which is most likely to be the torso portion. For each row we consider the starting column of connected component s_i (front of the torso at each row) and the half of the size of the connected component, l_i (half the width of the torso at a row). The center line is estimated to be at the average of the median of these two distributions. Alternative strategies such choosing the median of the average of the start and end index of the foreground in each row did not result in good centering of the silhouettes, as judged visually.

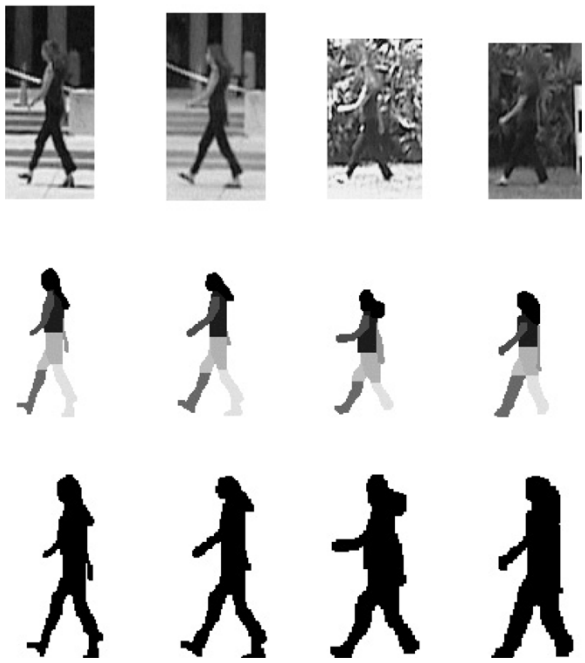


Fig. 1. Top row shows the original images, cropped around the person for four different camera views. The middle row shows the corresponding part-level, manually specified silhouettes. And the bottom row shows the scaled silhouettes of the kind used by gait recognition algorithms.

III. Model–Based Silhouette Reconstruction

Silhouettes, detected by some form of background segmentation, typically involve errors occur due to the following:

1) shadows;
2) inability to segment parts because they fall just below the threshold and are classified as background;
3) moving objects in the background, such as the fluttering tape in the concrete sequences, or moving leaves in the grass sequences, or other moving persons in the background;
4) compression artifacts near the boundaries of the person, which are present in, medium cost, consumer grade cameras.

Our algorithm uses a background subtraction scheme based on Gaussian statistical models of the background color that was used in the baseline gait algorithm to generate silhouettes. We compute the background statistics of the RGB values at each image location, (x, y), in terms of the mean $\mu_B(x, y)$ and the covariances $\Sigma_B(x, y)$ of the RGB values at each pixel location. Using the Mahalanobis distance of a pixel value as the observation, pixels are classified into foreground or background using expectation maximization (EM) with a Gaussian mixture model.

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

The states of the HMM represent a gait stance and the transition probabilities capture the motion dynamics between the states for the subject population. This HMM is learnt based on the manually specified silhouettes for 71 subjects. For each gait stance, we also construct, using the manually specified silhouettes, statistical shape models in terms of the mean silhouette shape and variances of that stance shape. This statistical model, which we call the *eigen-stance gait model* is accomplished by performing principal component analysis (PCA) for each stance.

Each frame in any given sequence is matched onto these stance subspaces using the population based HMMs, statistically describing the gait motion over a subject population. Each silhouette is then reconstructed using the coordinates of the silhouette, found by projecting onto the matched eigen-stance model.

A. Forming Stance Exemplars

The observation model for each HMM state is the most critical aspect of the specification, so we describe it some detail. The observation variables are the distances of a given observed silhouetted from an exemplar set, which we compute by clustering the frames of each given sequence. The particular clustering method employed is constrained K-means clustering. Of course, any clustering method relies on a distance measure, which we define as follows. Let $\mathbf{f_i}$ and $\mathbf{f_j}$ be two, vertically scaled and horizontally aligned (see Fig. 1), silhouette frames,

Fig. 2. Average stances in population exemplars for seven sample states over a gait cycle.

reformatted into row-scanned column vectors, then the similarity between them is

$$S(i,j) = \frac{\mathbf{f_i}^T \mathbf{f_j}}{\mathbf{f_i}^T \mathbf{f_i} + \mathbf{f_j}^T \mathbf{f_j} - \mathbf{f_i}^T \mathbf{f_i}}. \quad (1)$$

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

To create the exemplars, we first partition the frames in one gait cycle into N_S equal segments. We use one full cycle (two strides) so as to retain the asymmetry in gait, i.e. to differentiate stances with left foot forward from those with right foot forward. We group the frames within the jth partition of all people into an exemplar set for the jth gait stance, E_j. Since the gait cycles of the manual silhouettes are aligned, this strategy of corresponding the exemplars from different subjects works.

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

$$\overline{E_j}^{[k]} = \frac{1}{N} \sum_{\mathbf{f_i} \in E_j^{[k]}} \mathbf{f_i} \quad (2)$$

$$E_j^{[k+1]} = \left\{ \mathbf{f_i} | D\left(\mathbf{f_i}, \overline{E_j}^{[k]}\right) < \min\left(D\left(\mathbf{f_i}, \overline{E_{j-1}}^{[k]}\right), D\left(\mathbf{f_i}, \overline{E_{j+1}}^{[k]}\right)\right)\right\}. \quad (3)$$

Note that constraint that frames can only be reassigned to only to neighboring exemplar sets; thus, a frame in E_j can be reassigned to exemplars E_{j-1} or E_{j+1}. We also insist that every exemplar should contain at least one frame from each sequence. We stop when no more reassignments can be done; about ten iterations were enough for our experiments. Fig. 2 shows the mean silhouettes, $\overline{E}_j$, of exemplar sets for seven example stances.

B. Population HMM

We will use an HMM to align any given sequence to generic stance sequences for stance-dependent silhouette reconstruction. An HMM is specified by the possible states, $q_t \in \{1, \ldots, N_s\}$ and the triple $\lambda = (A, B, \pi)$, representing the state transition matrix, observation model, and priors, respectively. The state transition matrix A with entries $a(i,j) = P(q_{t+1} = j | q_t = i)$ is constrained to represent a cyclical version of the left to right Bakis state transition model over N_s states, allowing only for jumps to the next state. The observation model is comprised of the observation models for each state, $B = \{b_j(\mathbf{f_t}) | j = 1, \ldots, N_s\}$, where $b_j(\mathbf{f_t}) = P(\mathbf{f_t} | q_t = j)$, i.e. the conditional probability of the observed silhouette, $\mathbf{f_t}$, at time t given that the state at time t is j. We choose the observation model to be exponential in terms of the Tanimoto distance, D, between any given silhouette, $\mathbf{f_t}$, to the mean of the state exemplars, $\overline{E}_j$

$$b_j(\mathbf{f_t}) = \frac{1}{\mu_j} e^{-\frac{D(\mathbf{f_t}, \overline{E}_j)}{\mu_j}}. \quad (4)$$

The observation model is thus parameterized by the mean μ_j. The HMM structure is somewhat similar to that used in [25] for recognition, but in our case it is designed to model gait dynamics over a population. Differences also exist in the observation model and the state definitions; our model takes into account the gait asymmetry between the two strides over a cycle.

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

$$\mu_j = \frac{\sum_{\mathbf{f_i} \in E_j} D(\mathbf{f_i}, \overline{E}_j)}{|E_j|}. \quad (5)$$

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

$$a^{[0]}(i,j) = \frac{\sum_k \#\text{ of } \mathbf{f_{t+1}^k} \text{ in } E_j \text{ given } \mathbf{f_t^k} \text{ in } E_i}{|E_i|}. \quad (6)$$

We refer the reader to standard texts such as [30] for details regarding the Levinson's method. Here we present just the key equations. Let there be K observation sequences, $\{F^1, \ldots, F^K\}$. However, since for each training sequence we have only one gait cycle, to retain the cyclical property, we extend each sequence by appending its first frame to the tail: $F^i = \{f_0^i, \ldots, f_{N_F}^i, f_0^i\}$. The length of the extended F^k is denoted by T_k. The iterative re-estimate of the population transition probabilities, $A^{[n+1]}$, are given by

$$a^{[n+1]}(i,j) = \frac{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_t^k(i) a^{[n]}(i,j) b_j\left(\mathbf{f}_{t+1}^k\right) \beta_{t+1}^k(j)}{\sum_{k=1}^{K} \frac{1}{P_k} \sum_{t=1}^{T_k - 1} \alpha_t^k(i) \beta_t^k(i)} \quad (7)$$

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

$$\begin{aligned} \alpha_t^k(j) &= P\left(\mathbf{f}_1^k, \ldots, \mathbf{f}_t^k, q_t = j | \lambda\right) \\ &= \frac{b_j\left(\mathbf{f}_1^k\right)}{N_s} \quad t = 1 \\ &= \sum_{i=1}^{N_s} \alpha_{t-1}^k(i) a^{[n]}(i,j) b_j\left(\mathbf{f}_t^k\right) \quad 2 \le t \le T_K \end{aligned} \quad (8)$$

$$P_k = P(F^k|\lambda) = \sum_{j=1}^{N_s} \alpha_{T_k}(j) \tag{9}$$

$$\begin{aligned}\beta_t^k(i) &= P\left(\mathbf{f}_{t+1}^k, \ldots, \mathbf{f}_{T_k}^k | q_t = i, \lambda\right) \\ &= 1 \quad t = T_k \\ &= \sum_{j=1}^{N_s} a^{[n]}(i,j) b_j\left(\mathbf{f}_{t+1}^k\right) \beta_{t+1}^k(j) \;\; t \leq T_k - 1. \end{aligned} \tag{10}$$

Equations (7)–(10) represent the generalization of the Baum-Welch equations for multiple observations and need to be iterated over until the likelihood of the given observations are maximized. The learned transition matrix emphasizes the transitions to forward states, manifesting as high values along the first upper diagonal. We also found high values at the antidiagonal corner, which is because we adopt a cyclical Bakis model.

Model Size Determination: We determine the number of states, N_s, based on the Akaike Information Criterion (AIC) [31], which take both the goodness of fit and generalizability into account

$$\mathrm{AIC} = -2 \sum_{k=1}^{K} \log_2 P(F^k|\lambda) + 2N_{\mathrm{para}} \tag{11}$$

where (λ) is the estimated population HMM model, K is the number of training sequences, and N_{para} is the number of estimated parameters of the model. The estimated parameters include the N_s^2 transition probabilities and the N_s parameters in the observation model. Fig. 3 plots the variation of AIC with the number of states for two different training sets of 71 subjects, one over grass walking surface and the other over concrete walking surface. Based on this plot we choose the round figure of 20 states as being fairly optimal for both the sets of sequences. It is better to err toward the larger number of states so as to retain the shape variations among different individuals.

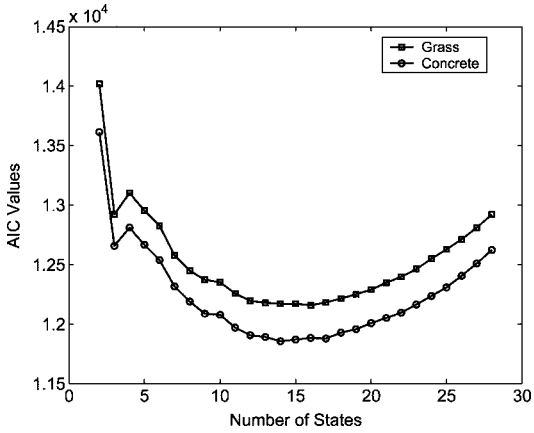


Fig. 3. Variation of AIC with number of states, for models constructed using two different training sets of 71 subjects; one for grass walking surface and the other for concrete walking surface.

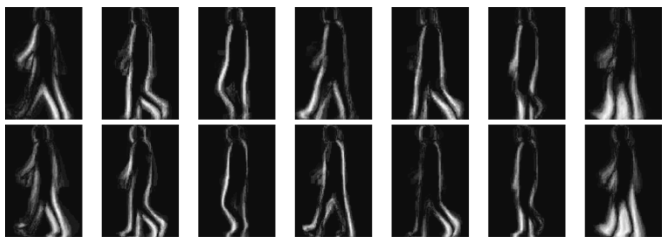


Fig. 4. Samples of the first eigen-stances over one gait cycle, representing the most discriminating directions among persons. The top row was built using silhouettes from the grass walking sequences, and the bottom row is for the concrete walking surface sequences.

C. Eigen-Stance Gait Model

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

Considering the strong impact of walking surface type on gait recognition, we built different eigen-stance gait models, coupled with their own HMMs, for grass and concrete surfaces using manual silhouettes associated with the Gallery set and Probe D set, respectively, from the HumanID Gait Challenge problem. Fig. 4 shows some sample eigen-stances of both spaces. The number of eigenvectors, N_e, is chosen so that at least 80% of the variation is modeled.

D. Stance Matching Using HMM

In order to project and reconstruct silhouette frames in any given sequence, $\{\mathbf{f_1}, \ldots, \mathbf{f_T}\}$, they have to be matched to one of the N_s stances in the population HMM. The dynamic programming based Viterbi algorithm is used for this purpose [32]. It returns the most likely state assignment to the input frames. To reduce the combinatorics of this assignment process, we partition the input sequence into subsequences of roughly one gait cycle length, which is estimated from the periodic variation in the number of foreground pixels in the bottom half of the silhouettes. Note that the starting state of these subsequences need not match the starting HMM state; the cyclical nature of the HMM model can handle this.

E. Reconstruction

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

$$\mathbf{f_i^r} = \boldsymbol{\mu}_\mathbf{j} + \sum_{k=1}^{N_e} \left(\mathbf{e_{j,k}^T}(\mathbf{f_i} - \boldsymbol{\mu}_\mathbf{j})\right) \mathbf{e_{j,k}} \tag{12}$$

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

$$\mathbf{F_i^r}(k) = \begin{cases} \text{Foreground} & \mathbf{f_i^r}(k) > T_{\text{high}} \text{ or } \boldsymbol{\mu}_\mathbf{j}(k) = 1 \\ \text{Background} & \mathbf{f_i^r}(k) < T_{low} \\ \mathbf{f_i}(k) & \text{otherwise.} \end{cases} \tag{13}$$

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

IV. Quality of Reconstructed Silhouettes

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

The raw silhouettes are those produced by the baseline gait recognition algorithm [23] *with some modifications*. The following steps of the baseline silhouette detection algorithm are:

- Step 1) compute the statistics of the individual background pixels in terms of mean and covariance of RGB values;
- Step 2) compute the Mahalanobis distance of a pixel from this background pixel value distribution;
- Step 3) smooth the Mahalanobis distance using a 9×9 triangular window to fill in holes and to join several small pieces;
- Step 4) decide on an optimal threshold to segregate the two classes using EM with the distance values as the observations;
- Step 5) pick the largest connected component.

The smoothing in step 3) above results in thicker silhouettes with high false positive predictive values and have been found to result in poor gait recognition performance for some recognition strategies [27]. So, we *eliminate that smoothing step*. However, this causes the problem of losing body portions because we only pick the largest component in step 5). So, we replace step 5) with a custom proximity based grouping process that assembles disconnected components: first, we morphologically close the silhouette twice in the vertical direction using 3×1 element so as to reconnect body parts; most disconnections happen in the vertical direction, e.g., trunk between leg. Then for each connected component P_i in a frame, we compute two values A_i and B_i in terms of the largest component P_{max}: $A_i = (\text{Area of } P_i / \text{Area of } P_{\max})$ and $B_i = e^{-\theta_i}$, where θ_i is the vertical angle between the center point of $P_{\max}$ and P_i. We decide to group components based on the product of A_i and B_i. Finally we do the morphological closing operation with a 3×3 element in order to fill the holes inside the silhouette.

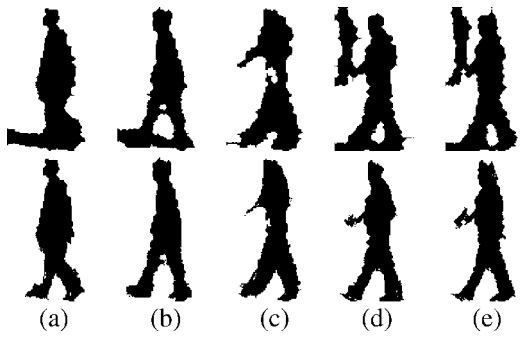


Fig. 5. Top row show some instances of poor quality silhouettes and the bottom row shows the reconstructed silhouettes.

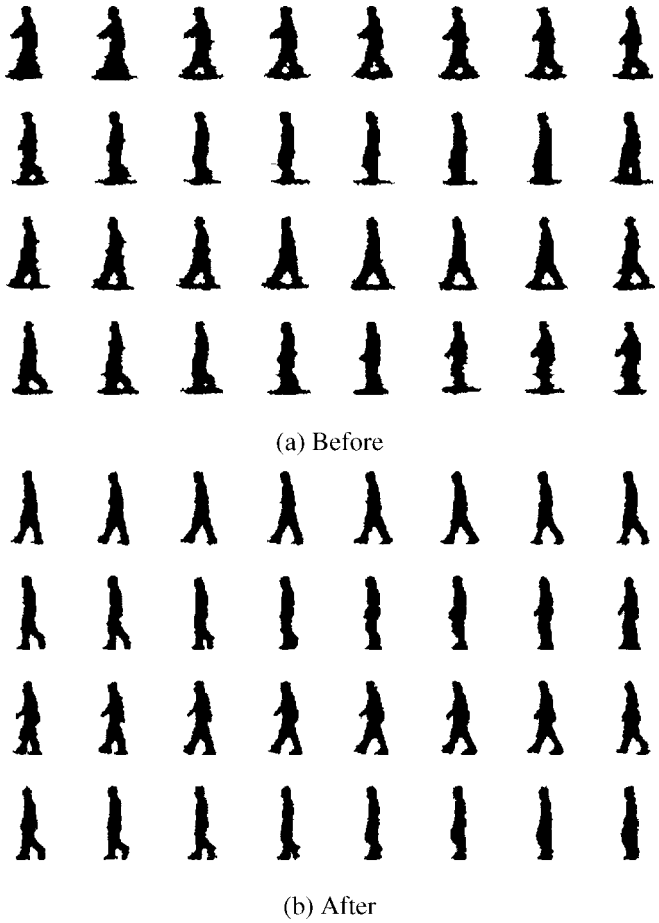


Fig. 6. Silhouettes over one gait cycle (a) before and (b) after reconstruction.

Thus, using simple low-level methods, the quality of the raw baseline silhouettes is enhanced somewhat, however, artifacts do remain. These form the input to the reconstruction process. The eigen-stance based silhouette reconstruction process removes many of the artifacts. Fig. 5 shows some example of the quality of reconstruction (bottom row) for poor quality input silhouettes (top row). Columns (a) and (b) of Fig. 5 show cases where shadows were removed; column (c) shows a case where holes in the foreground were filled in; and (d) and (e) show examples of removal of another person in the background.

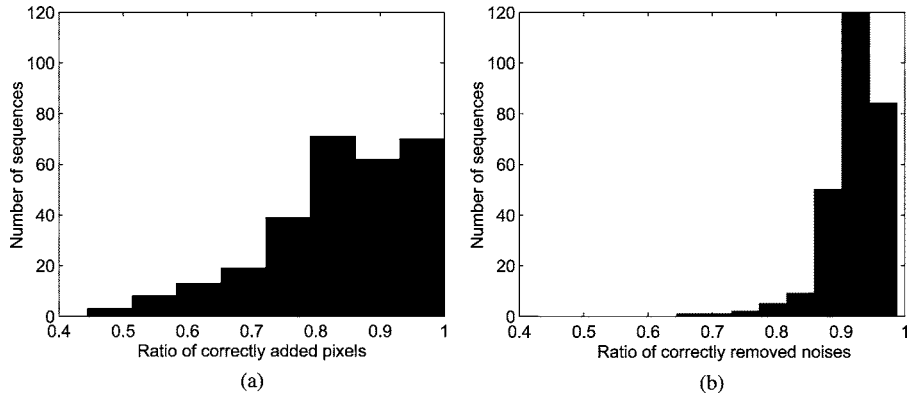


Fig. 7. Histogram of the ratio of (a) *correctly added* foreground pixels to the total number of *added* foreground pixels and (b) *correctly removed* noise pixels to the total number of *removed* pixels.

Fig. 6 shows an example of reconstruction over one gait cycle. We see that most frames have been improved, suggesting our model works well for different gait phases.

A. Pixel Level Quality

One manner to evaluate the eigen-stance model is to analyze the types of pixels that are edited (either removed or added) during the reconstruction process. The measures of performance could be the ratio of *correctly added* foreground pixels to the total number of *added* foreground pixels and the ratio of *correctly removed* noise pixels to the *total* number of removed pixels. Ideally, both these ratios should be one. We compute these two quantities for each frame for which we have manually specified ground-truth image and average them over a sequence. Fig. 7 shows the histogram of the two ratios over all the sequences for which we have manual silhouettes. We see that the histograms are strongly biased toward one. Thus, suggesting that the editing during the reconstruction process is mostly correct.

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

Fig. 8 shows the percentage improvement in pixel level detection ($\Delta P_D = 100((P_D(\text{After}) - P_D(\text{Before}))/P_D(\text{Before}))$) and false positive prediction ($\Delta P_{\overline{PPV}} = 100((P_{\overline{PPV}}(\text{After}) - P_{\overline{PPV}}(\text{Before}))/P_{\overline{PPV}}(\text{Before}))$) with reconstruction. We separately report the results for the grass and concrete sequences since they have different backgrounds. We also show the improvement for sequences with briefcase. Improvement

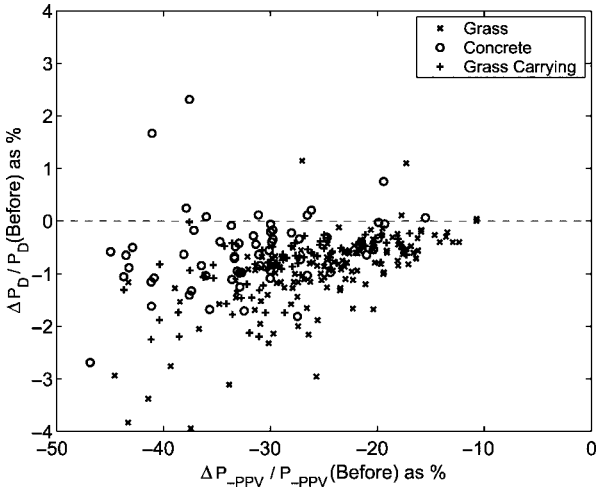


Fig. 8. Scatter plot of percentage change in pixel level detection (ΔP_D) and false predictive values ($\Delta P_{\overline{PPV}}$) after reconstruction. The green circles correspond to the concrete sequences, the red crosses correspond to the grass sequences and the blue stars correspond to the briefcase carrying sequences.

in silhouette quality would be indicated by $\Delta P_{\overline{PPV}} < 0$ and ΔP_D around 0, which is observed in the plots. We see that although the detection rate of the reconstructed silhouettes dropped a little bit by about 1%, the false positive predictive value dropped much more dramatically by about 20%–30%.

B. Robustness With Viewpoint Variation

Since the eigen-stance model is a view-based representation, it is reasonable to ask how effective is the reconstruction process in handling silhouettes viewed from somewhat different viewpoints. The Gait Challenge dataset permits such as study; it includes datasets of the same gait event viewed from two different angles, with the verging angle of roughly 30°. The manual silhouettes, which were used to construct the eigen-stance model, only exist for the sequences viewed from the right camera. We use the sequences viewed from the left camera to test the robustness. However, since we do not have manual silhouettes for these left camera sequences, we can only view the quality subjectively. We show results on ten such sequences in Fig. 9. Samples of both the original and the reconstructed frames are shown. As we have seen before, the model is able to successfully remove most shadows and other background noise artifacts.

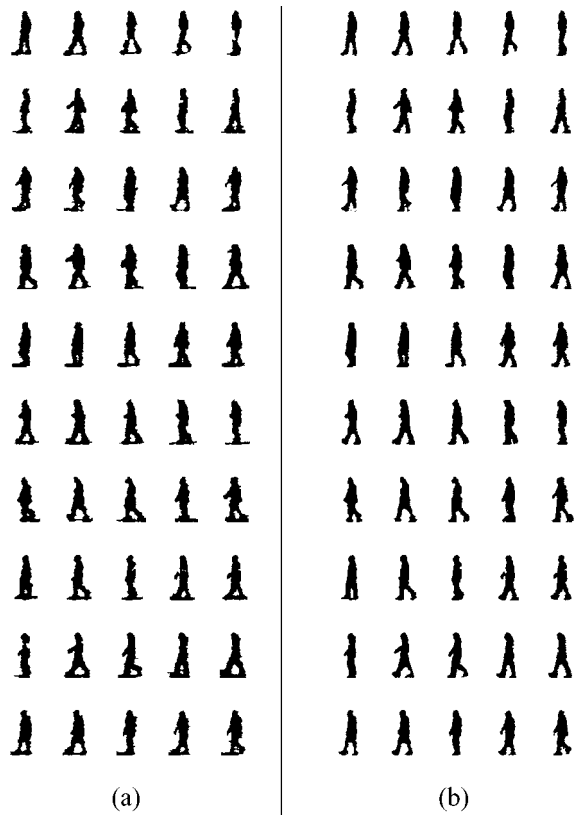


Fig. 9. Reconstruction of silhouettes for a sequence with 30° view angle difference from those used to construct the eigen-stance gait model. Frames in each row are from the same sequence. (a) Original silhouettes. (b) Reconstructed silhouettes.

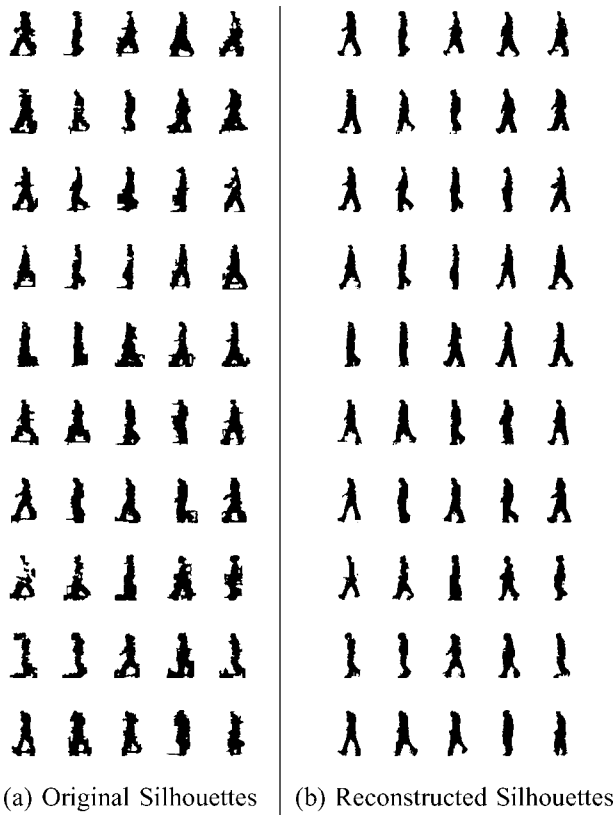


Fig. 10. Samples of (a) original and (b) reconstructed silhouettes over one gait cycle for ten subjects from the Georgia Tech dataset. Each row corresponds to a subject.

C. Generalizability to Different Datasets

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

V. Impact on Gait Recognition

We have illustrated that the eigen-stance model substantially improves the silhouette qualities. But, does the improved silhouettes effect gait recognition performance? One concern is that the model based reconstruction process might result in silhouettes that are more similar to each other, hence bring down recognition performance. To study this, we consider recognition from both (a) the manually specified silhouettes over part of the data set and (b) the reconstructed silhouettes over the whole dataset. Just the manual silhouettes are not sufficient since they are over a limited number of subjects and over only *one gait cycle*. We will quantify gait recognition performance using: 1) the baseline gait recognition algorithm that was specified along with the gait challenge problem [23] and 2) a stance shape based recognition algorithm, whose performance is superior to the baseline algorithm.

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

A. Baseline Gait Recognition Algorithm

The silhouette extraction part of the baseline algorithm was discussed earlier in Section IV. Here we sketch the matching of any given probe sequence to a gallery sequence to arrive at a similarity score. It is based on spatio-temporal correlation. Let the probe and the gallery silhouette sequences be denoted by $\mathbf{S_P} = \{\mathbf{S_P}(1), \ldots, \mathbf{S_P}(M)\}$ and $\mathbf{S_G} = \{\mathbf{S_G}(1), \ldots, \mathbf{S_G}(N)\}$, respectively. First, the probe (input) sequence is partitioned into subsequences, each roughly over one gait period, N_{Gait}. Gait periodicity is estimated based on periodic variation of the count the number of foreground pixels in the lower part of the silhouette in each frame over time. This number will reach a maximum when the two legs are farthest apart (full stride stance) and drop to a minimum when the legs overlap (heels together stance).

Second, each of these probe subsequences, $\mathbf{S_{Pk}} = \{\mathbf{S_P}(k), \ldots, \mathbf{S_P}(k + N_{\mathrm{Gait}})\}$, is cross correlated with the given gallery sequence, $\mathbf{S_G}$ as follows.

$$\mathrm{Corr}(\mathbf{S_{Pk}}, \mathbf{S_G})(l) = \sum_{j=1}^{N_{\mathrm{Gait}}} S\left(\mathbf{S_P}(k+j), \mathbf{S_G}(l+j)\right) \quad (14)$$

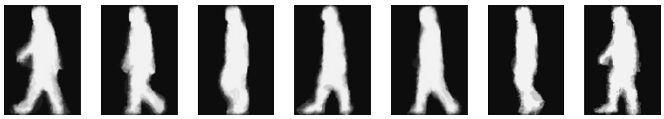


Fig. 11. Time-normalized averaged gait cycle frames for seven stances for one subject.

where, the similarity between two image frames, $S(\mathbf{S_P}(i), \mathbf{S_G}(j))$, is defined to be the Tanimoto similarity between the silhouettes, i.e. the ratio of the number of common pixels to the number of pixels in their union. The overall similarity measure is chosen to be the median value of the maximum correlation of the gallery sequence with each of these probe subsequences. The strategy for breaking up the probe sequence into subsequences allows us to address the case when we have segmentation errors in some contiguous sets of frames due to some background subtraction artifact or due to localized motion in the background as follows.

$$\mathrm{Sim}(\mathbf{S_P}, \mathbf{S_G}) = \mathrm{Median}_k \left(\max_l \mathrm{Corr}(\mathbf{S_{Pk}}, \mathbf{S_G})(l) \right). \quad (15)$$

The above strategy is simple yet effective for gait recognition when compared with performances of more complicated strategies, as we can see in Table III. So, even though we base the conclusions in this section on the performance of the baseline algorithm, we expect the observations to be generalizable to results with other gait recognition algorithms. We have also benchmarked the baseline algorithm on other standard datasets, such as the CMU Mobo [33], and the performances are similar or even better than other approaches for gait recognition that has been proposed in the literature.

B. Shaped Based Recognition Algorithm

Recent gait recognition experiments [8], [34] have shown that silhouette shape, which includes body shape and gait stance shape, has equal, if not more, recognition power than gait dynamics. We designed a new gait recognition algorithm around this idea, exploiting just silhouette shape matching. The details of this algorithm is available in [35], here we present just the rough outline.

Given a silhouette sequence over multiple gait cycles, we estimate the stance of each frame using the Viterbi algorithm, based on the population HMM model described in Section III-B. From these matches we arrive at one time-normalized, averaged gait cycle over a fixed number of stances. Fig. 11 shows one sample such averaged gait cycle. Note that in this representation the gait dynamics, i.e. the transition information between stances, are removed.

Given two averaged gait cycles, the similarity computation process does not have to align the cycles. The corresponding stances can be simply compared and the results summed to arrive at an overall similarity score. However, we consider the distances only for a subset of preselected stances that emphasize the differences between subjects. To select the discriminatory stances, we consider the variation in shape for each stance as reflected in the first and the second eigenvalues associated with the corresponding Eigen-stance model. These are plotted in Fig. 12. We see that states at the ends (states 1–3 and 18–20) and at the middle (9–12) has the largest scatters, indicating that these gait stances carry the bulk of the discriminatory power. These correspond to states near the full stride stances. Let us denote this subset of salient discriminatory states by $\mathbf{S_d}$. To arrive at one similarity score, we compute Euclidean distances between the averaged representation for these stances from the probe sequence I_{P_i} and the gallery sequence I_{G_j} as follows.

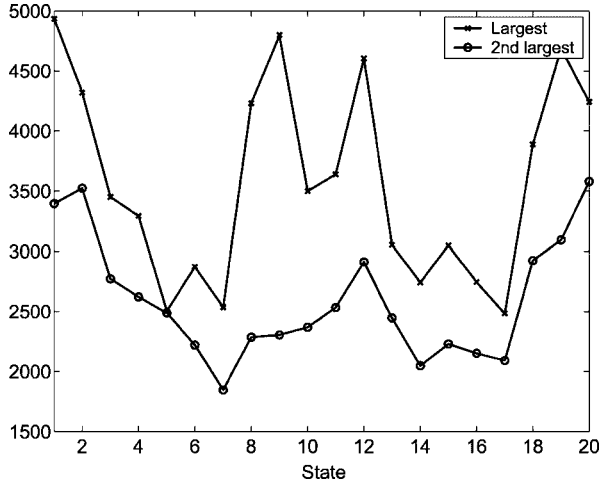


Fig. 12. Variation of largest and second largest eigenvalues associated with each stance shape, as computed in the Eigen-stance model.

$$S\left(I_{P_i}, I_{G_j}\right) = -\sum_{k \in \mathbf{S_d}} \left(I_{P_i}(f_k) - I_{G_j}(f_k)\right)^2. \quad (16)$$

C. Recognition From Manual Silhouettes

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

To compare performance, we consider the key challenge experiments involving shoe, surface, carrying, and time variation between probe and gallery in the gait challenge problem. The gallery and probe sets for the experiments are reduced to contain the sequences for which we have manual silhouettes. Since recognition with manual silhouettes uses just one gait cycle, we

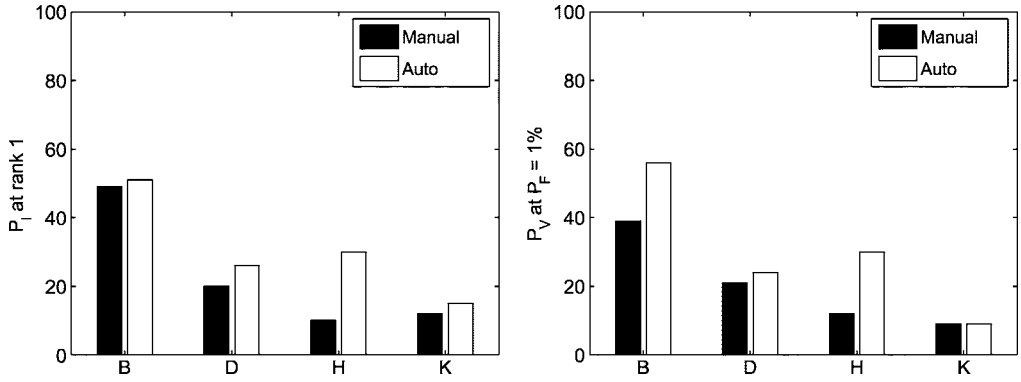


Fig. 13. Recognition performance of the baseline gait recognition algorithm in terms of identification rate at rank 1 and verification rate at a false alarm rate of 1% with manual silhouettes over one gait cycle and with (unreconstructed) automated silhouettes over that same cycle.

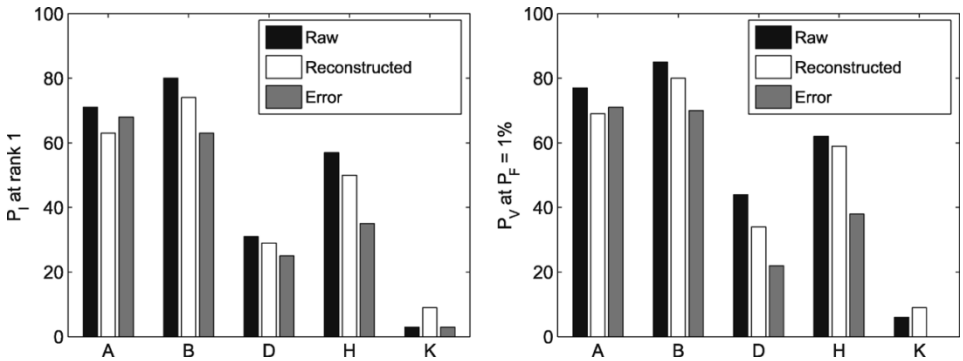


Fig. 14. Performance with baseline gait recognition algorithm. Identification rate (P_I) at rank 1 and verification rate (P_V) at 1% false alarm rate (P_F) with raw silhouettes and after reconstruction, and with error pixels edited (removed or added) during the reconstruction process using the baseline algorithm. Results for the five key experiments are listed: A (viewpoint), B (shoe), D (surface), H (carry), and K (time).

compare the performance with (unreconstructed) automated silhouettes also over the corresponding gait cycles, using a similarity computation strategy same as that for the manual silhouettes. In Fig. 13, we report the performance numbers for both the identification and the verification scenario using the identification rate at top rank and the verification rate for a 1% false alarm values, respectively. We see that the performance with manual silhouette actually drops for experiments that compare sequences involving changes in shoe, carrying condition, and surface. This can be attributed to the removal of shadow or segmentation error correlations between gallery and probe sequences in these experiments. Except for Experiment K, the gallery and the probe sequences for the other key experiments were collected around the same time and with the same clothing. Also, there is difference in background between gallery and probe only for Experiment D, which exercises surface condition change. For Experiment K, the probe background represents the same location as in the gallery, but after six months.

To shed more light on this, we consider the recognition power in: 1) the difference image between the automated silhouette and manual silhouette and 2) the pixels edited (either removed or added) during reconstruction. We found that the identification rates at rank one with these error pixels were indeed high at 28% and 25%, respectively, which roughly make up for the gap between the manual and automated silhouettes based recognition rates. Thus, *the low performance under the impact of surface and time variation can not be explained by the silhouette quality*.

D. Recognition From Reconstructed Silhouettes

We saw that recognition from manual silhouettes over one cycle did not improve recognition. The choice of the one gait cycle might have influenced results. Also, since these manual silhouettes were done by multiple persons they might not be of consistent quality. Do the effects also remain if we use multiple gait cycles and silhouette with consistent quality? For this, we use the reconstructed silhouettes, which were shown to be of better quality than the raw silhouettes.

Fig. 14 summarizes the baseline performances with raw and reconstructed silhouettes for some of the key gait challenge experiments. We see a drop in performance for experiments A (view), B (shoe), and H (carry). This is consistent with the results we obtained for the manual silhouettes. The gallery and probe set sequences of a person for these three experiments were collected with the same background and about the same time. This is particularly true for experiment A (viewpoint) whose gallery and probe sets contain essentially the same temporal event for each person, but taken from two different viewpoints. Thus, there are correlations in the shadows of a subject between the gallery and probe sequence, which possibly contributed to higher rate with the unreconstructed silhouettes. So, we considered just the pixels that were edited during the reconstruction process, i.e., either added or removed. We refer to these as the error pixels. Recall, that we have already established that the edited pixels are mostly error (either false positive prediction or missed detection) pixels (Figs. 7 and 8). We studied the recog-

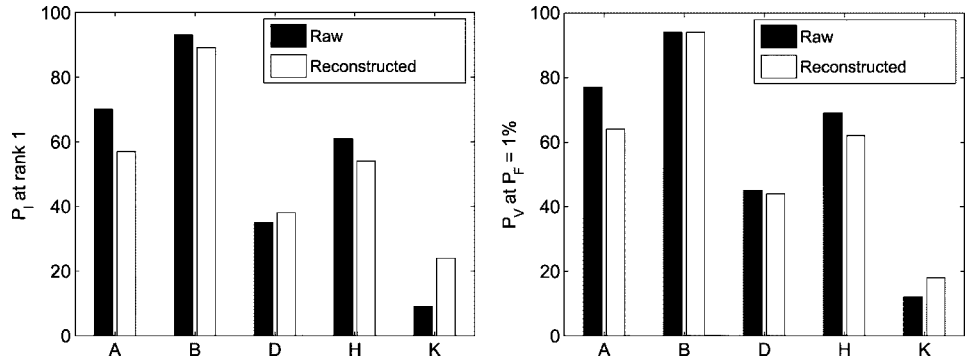


Fig. 15. Performance with stance shape based gait recognition. Identification rate (P_I) at rank 1 and verification rate (P_V) at 1% false alarm rate (P_F) with raw silhouettes and after reconstruction using the shaped based algorithm. Results for the 5 key experiments are listed: A (viewpoint), B (shoe), D (surface), H (carry), and K (time).

nition power from the error pixels. As shown by gray bars in Fig. 14, the recognition from these error pixels is quite high, especially when comparing sequences for a subject collected within a short time duration of each other. For experiments that compare sequences across six months, the error pixels do not have significant recognition.

Fig. 15 shows the identification and verification performance of the stance shape based algorithm on the five key experiments, with the raw and reconstructed silhouettes. We see that: 1) the overall performance of the stance shape based algorithm is much better than the baseline algorithm. It is also better than many of the reported performances (Table III). But more importantly 2) the performance with raw and reconstructed silhouettes are similar. We find that after removing false recognition sources such as shadow pixels and missed detection, possibly due to interaction of clothing texture and background pattern, gait recognition under shoe, surface and view variations did not change, in fact they dropped a bit. The performance across time, for which there is less chance for less correlation, appears to be marginally better with reconstructed silhouettes, but the differences are not statistically significant given the small probe set size of 33, when compared with the probe set sizes of the other experiments.

VI. Conclusions and Discussion

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

Recently, Lee *et al.* [27] also presented a method (here referred to as the MIT-Hp method) for cleaning silhouettes that appears to be similar to that presented here. They also use HMM based time-syncing of sequences and cleanup using template models. However, there are several key differences resulting in demonstrable performance differences. First, our representation of the shape variation model at each HMM state using the *eigen-stance model* allows us to exploit the correlation among the silhouette pixels, whereas, MIT-Hp uses an independent Bernoulli model for the silhouette pixel values. The use of a Bernoulli model is akin to using just the mean images of each stance in our model. Second, our training set consists of manually specified silhouettes, which enables us to *remove shadows*. The existence of shadows in MIT-Hp silhouettes might explain the enhanced performance for the experiments (A–C) on grass. Their performance for experiments where they compared silhouettes across surfaces did not improve to a large extent. Third, MIT-Hp used a two step process involving 1) a cleanup using a population based on mean image constructed by summing all frames for a set of persons and 2) further cleanup using a sequence specific HMM. On the other hand, we have an unified approach that uses a population based HMM model, coupled with population-based stance shape models. The full use of population models lets us overcome many sequence specific segmentation artifacts such as holes due to strange background or foreground texture for a particular person. The power of the use of population models is also evident, to a limited extent, in the work of Lee *et al.*. The performance increase in gait recognition was mostly due to their use of the aggregate population model. The addition of sequence specific HMM did not seem to add to the recognition to a large extent. All of these key differences between the MIT-Hp method and those presented here have impact on actual silhouette quality: MIT-Hp silhouettes have more false positive prediction pixel than those in this paper. Fig. 16 shows the percentage improvement in pixel level detection ($\Delta P_D = 100(P_D(\text{Here})/P_D(MIT-Hp)) - 1$) and false positive prediction between the silhouettes produced here and the HMM-Hp silhouettes ($\Delta P_{\overline{PPV}} = 100(P_{\overline{PPV}}(\text{Here})/P_{\overline{PPV}}(MIT - Hp)) - 1$). We separately report the results for the grass and concrete sequences since they have different backgrounds. Improvement in silhouette quality would be indicated by $\Delta P_{\overline{PPV}} < 0$ and ΔP_D is around 0, which is observed in the plots. The differences in false predictive values are statistically significant (P – value ≤ 0.05, established using paired-t tests) for both concrete and grass sequences. The detection rate, on the other hand, are of the

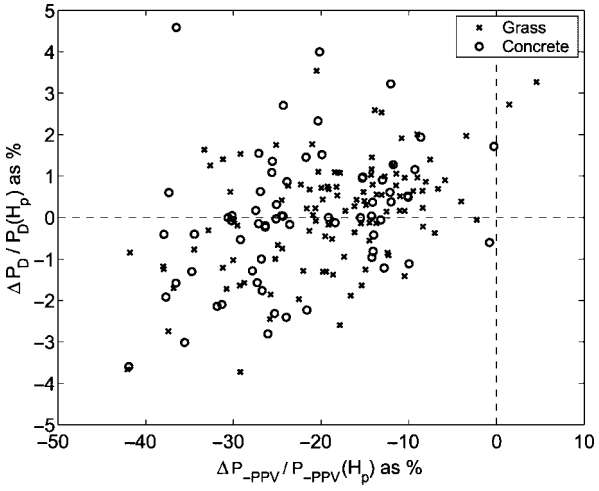


Fig. 16. Scatter plot of percentage improvement in pixel level detection (ΔP_D) and false predictive values ($\Delta P_{\overline{PPV}}$) of the silhouettes produced here and the MIT-Hp silhouettes. The green circles correspond to the concrete sequences and the red crosses correspond to the grass sequences.

same quality for both the grass sequences and the concrete sequences ($P-$ value ≥ 0.05).

Silhouette Quality and Gait Recognition: In the context of gait recognition, we have established that the low performance under the impact of surface and time variation can not be explained by poor silhouette quality. We base our conclusions on two gait recognition algorithms, one exploits both shape and dynamics, while the other exploits just shape. The drop in performance due to surface condition that we observe in the gait challenge problem is *not* due to differences in background. This observation is also corroborated by the performances reported in a fairly recent work by the Lee *et al.* [27]. The observation has implication for future work direction in gait recognition. Instead of searching for better methods for silhouette detection to improve recognition, it would be more productive to study and isolate components of gait that do not change under shoe, surface, or time. One example of this type of study is [12] in which relationship between silhouette shape and speed was studied and then was compensated for by transforming the silhouettes. While it is doubtful whether speed variations can fully explain the drop in performance due to surface or time change, systematic studies such as this would be needed to understand the limitation of gait recognition.

Manual Silhouette Database: Another contribution of this study is the set of part-level manual silhouettes that we have used for 71 subjects over one gait walking cycle, of approximately 40 image frames, spanning variations in shoe-type, surface, and time. This resource can be used in a variety of ways such as learning parameters of a kinematic chain model of gait. Or even learning probabilistic models to label parts of a silhouette. There has been little or no attempt at arriving at part level segmentation of the silhouette. The only works, which we are aware of, that make some effort in this direction are that of [18], where silhouette is segmented into three quadrants, and [4] where ellipses are fit to parts of the silhouettes. Part level segmentation would be great use in gait-recognition, since most of the recognition power from gait lies in the leg and arm region. The existence of part labels would allow one to arrive at more robust high level features in terms of body dimensions and dynamics.

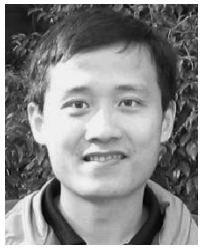
ACKNOWLEDGMENT

The authors wish to thank L. Malave, A. Osuntugun, P. Sudhakar, and C. Bexley for meticulously helping to put together the manual silhouette database. Thanks to L. Malave for running some of the early experiments with manual silhouettes, the gait researchers at CMU, MIT, and Southampton for helping to correct some of the errors in the manual silhouettes, and the MIT gait group for making their silhouettes available.

REFERENCES

[1] L. Wang, W. Hu, and T. Tan, "A new attempt to gait-based human identification," in *Proc. Int. Conf. Pattern Recognition*, vol. 1, 2002, pp. 115–118.
[2] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1505–1518, Dec. 2003.
[3] J. Hayfron-Acquah, M. Nixon, and J. Carter, "Automatic gait recognition by symmetry analysis," in *Proc. Int. Con. Audio- Video-Based Biometric Person Authentication*, 2001, pp. 272–277.
[4] L. Lee and W. Grimson, "Gait analysis for recognition and classification," in *Proc. Int. Conf. Automatic Face Gesture Recognition*, 2002, pp. 155–162.
[5] A. Kale, N. Cuntoor, and R. Chellappa, "A framework for activity specific human identification," in *Proc. Int. Conf. Acoustics, Speech Signal Processing*, vol. 4, 2002, pp. 3660–3663.
[6] A. Kale, A. Rajagopalan, N. Cuntoor, and V. Kruger, "Gait-based recognition of humans using continuous HMMs," in *Proc. Int. Conf. Automatic Face Gesture Recognition*, 2002, pp. 336–341.
[7] C. BenAbdelkader, R. Cutler, and L. Davis, "Motion-based recognition of people in eigengait space," in *Proc. Int. Conf. Automatic Face Gesture Recognition*, 2002, pp. 267–272.
[8] R. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," in *Proc. Int. Conf. Automatic Face Gesture Recognition*, 2002, pp. 366–371.
[9] Y. Liu, R. Collins, and Y. Tsin, "Gait sequence analysis using frieze patterns," in *Proc. Eur. Conf. Computer Vision*, May 2002, pp. 657–671.
[10] A. Johnson and A. Bobick, "A multi-view method for gait recognition using static body parameters," in *Proc. Int. Conf. Audio- Video-Based Biometric Person Authentication*, 2001, pp. 301–311.
[11] R. Tanawongsuwan and A. Bobick, "Gait recognition from time-normalized joint-angle trajectories in the walking plane," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2001, pp. II:726–II:731.
[12] ——, "Performance analysis of time-distance gait parameters under different speeds," in *Int. Conf. Audio- Video-Based Biometric Person Authentication*, 2003, pp. 715–724.
[13] J. E. Cutting and L. T. Kozlowski, "Recognition of friends by their walk," *Bulletin Psychonomic Soc.*, vol. 9, pp. 353–356, 1977.
[14] G. Johansson, "Visual motion perception," *Sci. Amer.*, vol. 232, pp. 75–88, Jun. 1976.
[15] S. Niyogi and E. Adelson, "Analyzing gait with spatiotemporal surfaces," in *Proc. IEEE Workshop Non-Rigid Motion*, 1994, pp. 24–29.
[16] J. Little and J. Boyd, "Recognizing people by their gait: The shape of motion," *Videre*, vol. 1, no. 2, pp. 1–33, 1998.
[17] J. Shutler, M. Nixon, and C. Carter, "Statistical gait description via temporal moments," in *Proc. IEEE 4th Southwest Symp. Image Analysis Int.*, 2000, pp. 291–295.
[18] A. Bobick and A. Johnsson, "Gait recognition using static, activity-specific parameters," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2001, pp. I:423–I:430.
[19] G. Shakhnarovich, L. Lee, and T. Darrell, "Integrated face and gait recognition from multiple views," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2001, pp. I:439–I:446.
[20] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
[21] N. Cuntoor, A. Kale, and R. Chellappa, "Combining multiple evidences for gait recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Processing*, vol. 3, 2003, pp. 113–116.

[22] P. J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer, "Baseline results for the challenge problem of human ID using gait analysis," in *Proc. Int. Conf. Automatic Face Gesture Recognition*, 2002, pp. 137–142.
[23] ——, "The gait identification challenge problem: Data sets and baseline algorithm," in *Proc. Int. Conf. Pattern Recognition*, 2002, pp. 385–388.
[24] A. Kale, B. Yegnanarayana, A. N. Rajagopalan, and R. Chellappa, "Gait analysis for human identification," in *Proc. Int. Conf. Audio- Video-Based Biometric Person Authentication*, 2003, pp. 706–714.
[25] A. Sunderesan, A. K. R. Chowdhury, and R. Chellappa, "A hidden markov model based framework for recognition of humans from gait sequences," in *Proc. IEEE Int. Con. Image Processing*, vol. 2, 2003, pp. 93–96.
[26] D. Tolliver and R. Collins, "Gait shape estimation for identification," in *Proc. Int. Conf. Audio- Video-Based Biometric Person Authentication*, 2003, pp. 734–742.
[27] L. Lee, G. Dalley, and K. Tieu, "Learning pedestrian models for silhouette refinement," in *Proc. Int. Conf. Computer Vision*, 2003, pp. 663–670.
[28] L. H. Malave, "Silhouette Based Gait Recognition: Research Resource and Limits," M.S. thesis, Dept. Comput. Sci. Eng., Univ. South Florida, Tampa, FL, 2003.
[29] Z. Liu, L. Malave, A. Osuntogun, P. Sudhakar, and S. Sarkar, "Toward understanding the limits of gait recognition," in *Proc. SPIE Processings Defense Security Symposium: Biometric Technology Human Identification*, Apr. 2004, pp. 195–205.
[30] L. Rabiner and B. H. Juang, *Fundamental of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[31] H. Akaike, "Information theory as an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Information Theory*, 1973, pp. 267–281.
[32] L. Rabiner and B. H. Juang, "An introduction to hidden markov models," *IEEE ASSP Mag.*, vol. 1, no. 3, pp. 4–16, Jan. 1986.
[33] R. Gross and J. Shi, "The CMU motion of body (MoBo) database," Robotics Inst., Carnegie Mellon Univ., Pittsburgh, PA, Tech. Rep. CMU-RI-TR-01-18, Jun. 2001.
[34] A. Veeraraghavan, A. R. Chowdhury, and R. Chellappa, "Role of shape and kinematics in human movement analysis," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, Washington DC, Jun. 2004.
[35] Z. Liu, "Gait-based human recognition at a distance: Performance, covariate impact, and solutions," Ph.D. dissertation, Dept. Comput. Sci. Eng., Univ. South Florida, Tampa, FL, 2004.



Zongyi Liu received the B.S. degree in business from Shenzhen University, Shenzhen, China, in 1997, the M.S. degree in computer science and application from the University of Electronic Science and Technology of China, Hefei, China, in 2000, and the Ph.D. degree in computer science and engineering from the University of South Florida, Tampa, in 2004.

His research interests are computer-vision-based gait biometrics, pattern recognition, motion, and image segmentation.

Sudeep Sarkar (M'93) received the B.Tech degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, in 1988, and the M.S. and Ph.D. degrees in electrical engineering, on a University Presidential Fellowship, from The Ohio State University, Columbus, in 1990 and 1993, respectively.

Since 1993, he has been with the Computer Science and Engineering Department at the University of South Florida, Tampa, where he is currently a Professor. His research interests include perceptual organization in single images and multiple image sequences, biometrics, gait recognition, color-texture analysis, and performance evaluation of vision systems. He has co-authored one book and co-edited another book on perceptual organization. He served on the editorial board for *Pattern Analysis and Applications Journal* from 2000 to 2001 and currently, *Pattern Recognition Journal*

He is the recipient of the National Science Foundation CAREER Award in 1994, the USF Teaching Incentive Program Award for undergraduate teaching excellence in 1997, the Outstanding Undergraduate Teaching Award in 1998, and the Theodore and Venette Askounes-Ashford Distinguished Scholar Award in 2004. He served on the editorial board for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE from 1999 to 2003, and is currently serving on the with the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, PART—B.