Baseline Results for the Challenge Problem of Human ID Using Gait Analysis

P. Jonathon Phillips[1], Sudeep Sarkar[2], Isidro Robledo[2], Patrick Grother[1], and Kevin Bowyer[3]

[1]NIST, Gaithersburg, MD 20899-8940
[2]Computer Science and Engineering, University of South Florida, Tampa, Florida 33620-5399
[3]Computer Science and Engineering, University of Notre Dame, Notre Dame, Indiana 46556
jonathon@nist.gov, {sarkar, irobledo}@csee.usf.edu, pgrother@nist.gov, kwb@cse.nd.edu

Abstract

Identification of people from gait captured on video has become a challenge problem in computer vision. However, there is not a baseline algorithm or standard dataset for measuring, or determining what factors affect performance. In fact, the conditions under which the problem is "solvable" are not understood or characterized. This paper describes a large set of video sequences (about 300 GB of data related to 452 sequences from 74 subjects) acquired to investigate important dimensions of this problem, such as variations due to viewpoint, footwear, and walking surface. We introduce the HumanID challenge problem. The challenge problem contains a set of experiments of increasing difficulty, a baseline algorithm, and its performance on the challenge problem. Our results suggest that differences in footwear or walking surface type between the gallery and probe video sequence are factors that affect performance. The data set, the source code for the baseline algorithm, and UNIX scripts to reproduce the basic results reported here are available to the research community at `http://marathon.csee.usf.edu/GaitBaseline/`

1. Introduction

Identifying humans from their gait is currently an extremely active area of computer vision (e.g., [1, 2, 3, 6, 7, 5, 4]). To assist the advancement of gait analysis, we introduce the HumanID challenge problem. We describe the data collected to support the challenge problem, provide a baseline algorithm to solve the challenge problem, and present a set of challenge experiments of increasing levels of difficulty. The challenge problem is designed to address the following questions: (1) Under what conditions is gait recognition solvable? (2) What variations in a person's walk affect performance? (3) What directions appear promising for improving the performance of gait recognition? The answer to these questions cannot be provided by the performance figures of one algorithm on a small proprietary database. Rather, the answer will come from detailed analysis of performance statistics of multiple algorithms on a large common data set. This is the framework that the HumanID gait challenge problem provides.

The key to the success of the challenge problem is the database of video sequences collected to support it. The database defines the characteristics and difficulty of the problem(s). The ideal challenge problem includes subproblems that span a range of characteristics and difficulties. These ranges are included in HumanID gait challenge problem because of the number of conditions under which a person's gait is collected, the number of individuals in the database, and the fact that all sequences are taken outside. The database used in the challenge problem is the largest available to date in terms of number of people, number of video sequences, and conditions under which a person's gait is observed. The database consists of 74 individuals, with each individual collected in up to 16 conditions. All the data is collected outside, reflecting the added complications of shadows from sunlight, moving background, and moving shadows due to cloud cover.

The baseline algorithm provides a base for measuring improvement in performance. The infrastructure tools provide the ground work for further investigation. The tools include the scripts for running small and large experiments, processing from intermediate steps, and methods for detailed performance analysis. The small and large experiments will provide a variety of levels for researchers to start investigating gait recognition. The availability of intermediate results will allow researchers to focus on different aspects of the problem. For example, the availability of the silhouette sequences means that a researcher can focus on the recognition part of problem. At the same time, another researcher could focus on the segmentation part of the problem. The analysis tools will provide a basis for determining under what conditions the problem is solvable, identifying

Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition (FGR'02)
0-7695-1602-5/02 $17.00 © 2002 IEEE

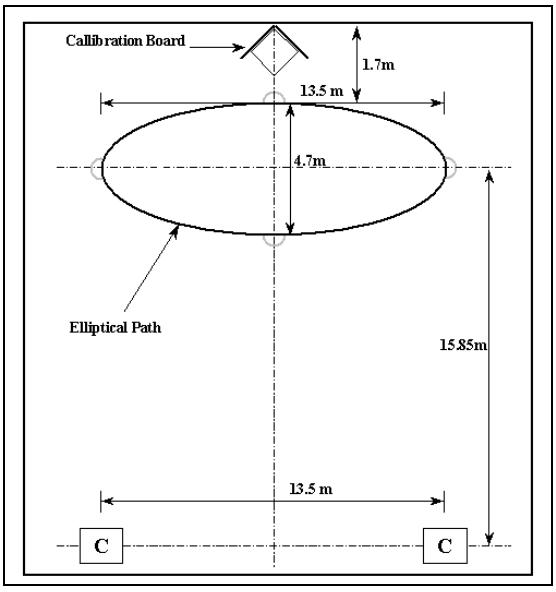

Figure 2. Camera setup for the gait data acquisition.

the underlying reasons for these conditions, and pointing to future directions for research and investigation.

The HumanID gait challenge problem touches on the following computer vision problems: matching and comparing temporal signatures, figure and background segmentation, modeling human motion and dynamics, and occlusion.

Not all of these aspects are included in the baseline algorithm or will be included in every solution. However, improvements in performance over the baseline will touch upon some of these areas. The connection with the challenge problem could serve as bases for developing and improving algorithms in these areas. In addition, the challenge problem can provide a means for measuring the impact of improvements in algorithms from these areas as on a well-defined problem.

2. Data Acquisition

The main goals of the data collection were to acquire data on a larger number of subjects than analyzed in current papers on gait analysis, and to acquire data on a given subject under varied conditions. Each subject walked counterclockwise around each of two similar size and shape elliptical courses. The basic setup is illustrated in Fig. 2. The elliptical courses were approximately 15 meters on the major axis and 5 meters on the minor axis. Both courses were outdoors. One course was laid out on a flat concrete walking surface. The other was laid out on typical grass lawn surface. Each course was viewed by two cameras, whose lines of sight were not parallel, but verged, so that the whole ellipse was just visible from the two cameras. When the persons walked along the rear portion of the ellipse, their view was only approximately fronto-parallel. Although data from one full elliptical circuit for each condition is available, we present the challenge experiments on the data from the rear portion of the ellipse. The gait video data was collected at the University of South Florida on May 21 and 22, 2001.

The cameras were consumer-grade Canon Optura (for the concrete surface) and Optura PI (for the grass surface) cameras. [1] These are progressive-scan, single-CCD cameras capturing 30 frames per second with a shutter speed of 1/250 second and with auto-focus left on as all subjects were essentially at infinity. The cameras stream compressed digital video to DV tape at 25 Mbits per second by applying 4:1:1 chrominance sub-sampling and quantization, and lossy intra-frame adaptive quantization of DCT coefficients.

Subjects were asked to read and sign an IRB-approved consent form when they arrived for the scheduled data acquisition. Information recorded in addition to the video includes sex (75% male), age (19 to 54 yrs), height (1.47 m to 1.91 m), weight (43.1 kg to 122.6 kg), foot dominance (mostly right), type of shoes (sneakers, sandal, etc.), and heel height. Subjects were asked to bring a second pair of shoes, so that they could walk the two ellipses a second time in a different pair of shoes. A little over half of the subjects walked in two different shoe types. In addition, subjects were also asked to walk the ellipses carrying briefcase of known weight (approximately 6 kilograms). Most subjects did walk both carrying and not-carrying the briefcase. Thus there are as many as sixteen video sequences for each subject: (grass / concrete) x (two cameras, L and R) x (shoe A / shoe B) x (briefcase / no briefcase). The current release of the database does *not* include the briefcase carrying condition, which would have doubled the size of the database to about 600 GB. The briefcase sequences would be part of a future release. Table 1 shows the number of sequences for each combination of conditions in the present database.

The imagery was recovered from tape offline. The camera is accessed over its IEEE 1394 Firewire interface using Pinnacle's micro DV 300 PC board. The result is a standalone video file stored using Sony's DV-specific "dvsd" codec in a Microsoft AVI wrapper. This capture from tape does not re-compress and is not additionally lossy. Finally the imagery is transcoded from DV to 24-bit RGB using the Sony decoder and the result is written as PPM files, one file per frame (720x480 PPM file). This representation trades

[1] Commercial equipment is identified in this work in order to adequately specify or describe the subject matter. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the equipment identified is necessarily the best available for this purpose.

Figure 1. Beginning, middle, and another frame of example gait sequence.

Table 1. Number of sequences for each combination of possible surface (G or C), shoe (A or B), and camera view (L or R).

Surface	Concrete (C)		Grass (G)	
Shoe	A	B	A	B
Left Camera	70	44	71	41
Right Camera	70	44	71	41

off storage efficiency for ease of access. The final sequences contain each subject walking several laps of the course. For the gait database we clipped those frames from the last such lap. Sample frames from one sequence appear in Figure 1. These three frames come from the left camera on the grass surface, without the subject carrying the briefcase. This particular sequence is 712 frames in length. Please note that although the database contains frames from one whole lap, the results in this paper are on frames from the rear or back portion (see middle image in Fig. 1). The subject's size in these frames from the rear portion is approximately 100 pixels in height, and 25 to 50 pixels in width.

Because two cameras were used during data acquisition the data is subsequently synchronized by manually aligning the two sequences by inspection of action in successive frames. Given that the cameras do not accept an external trigger, this human-in-the-loop method gives synchronization to no better than 1/15 second. The data should support some level of stereo analysis, although that is not attempted in this paper.

3. Baseline Algorithm

The baseline algorithm, which was designed to be simple and fast, is composed of three parts. The first part semi-automatically defines bounding boxes around the moving person in each frame of a sequence. Using a Java-based GUI, we manually outline bounding boxes in the starting, middle, and ending frames of the sequence. The bounding boxes for the intermediate frames are linearly interpolated from these manual ones. Specifically, the locations of the upper-left and the bottom-right corners are interpolated. This approximation strategy works well for cases where there is nearly fronto-parallel, constant velocity motion, which is true for the experiments reported in this paper. The second and the third parts of the algorithm are silhouette extraction and computation of the similarity measure, which are explained in detail in the next two subsections.

3.1. Silhouette Extraction

We extract the motion silhouette in each frame by background subtraction, but only within the semi-manually defined bounding boxes. In the first pass through a sequence, we compute the background statistics of the RGB values at each image location, (x, y), using pixel values outside the manually defined bounding boxes and all the image frames in the sequence. We compute the mean $\mu_{\mathbf{B}}(x, y)$ and the covariances $\mathbf{\Sigma_B}(\mathbf{x}, \mathbf{y})$ of the RGB values at each pixel location. Fig. 3 shows an example of the estimated mean background image and the associated variances of the RGB channels. Note that the variances are significantly higher in the regions corresponding to the bushes and the grass than other regions.

For pixels within the bounding box of each frame, we compute the Mahanalobis distance of the pixel value from the estimated mean background value. Any pixel with this distance above an user specified threshold D_{Maha} (= 4, in the experiments here) is declared to be a foreground pixel. We have found that if we smooth the difference image using a 9 by 9 pyramidal averaging filter, the resultant silhouette has smooth boundaries. On the difference thresholded image we perform two post processing steps to extract the normalized silhouette. First, we detect small regions, i.e. less than N_{Size} (= 200, in the experiments here) pixels, by connected component labeling and delete them. Second, we scale the remaining foreground region so that its length is 128 pixels so as to occupy the whole length of the 128 by 88 pixels sized output silhouette frame. This scaling offers some amount of scale invariance and facilitates the fast computation of the similarity measure.

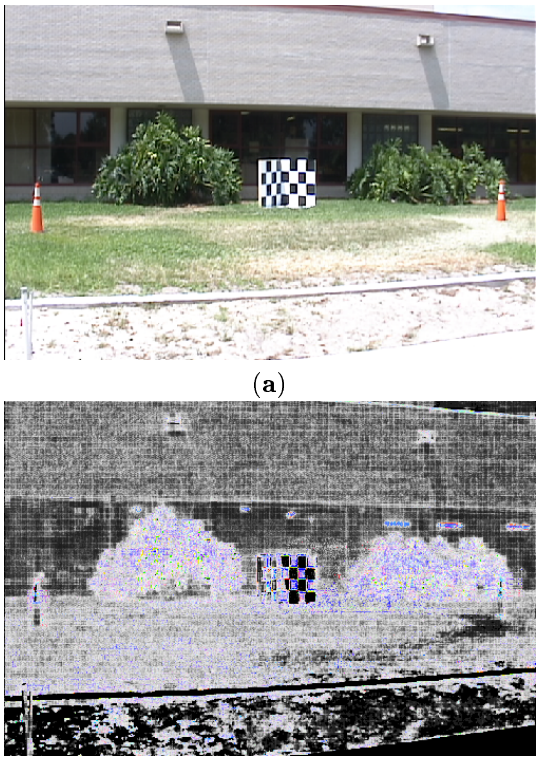

Figure 3. (a) Estimated mean background for the sequences shown earlier. (b) Variance of the RGB channels in the background pixels. The image has been histogram equalized.

3.2. Similarity Computation

Let the probe and the gallery silhouette sequences be denoted by $\mathbf{S_P} = \{\mathbf{S_P}(1), \cdots, \mathbf{S_P}(M)\}$ and $\mathbf{S_G} = \{\mathbf{S_G}(1), \cdots, \mathbf{S_G}(N)\}$, respectively. We first partition the probe sequence into disjoint subsequences of N (= 30, in the experiments here) contiguous frames each, such that each subsequence contains roughly one stride. Let the k-th probe subsequence be denoted by $\mathbf{S_{Pk}} = \{\mathbf{S_P}(k), \cdots, \mathbf{S_P}(k+N)\}$. We then correlate each of these subsequences with the gallery sequence

$$\text{Corr}(\mathbf{S_{Pk}}, \mathbf{S_G})(l) = \sum_{j=1}^{N} \text{FrameSim}\left(\mathbf{S_P}(k+j), \mathbf{S_G}(l+j)\right) \quad (1)$$

The similarity 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

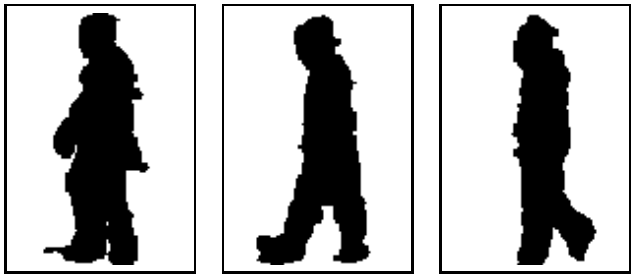

Figure 4. Three silhouette frames.

when we have segmentation errors in some contiguous sets of frames due to some background artifact or localized motion in the background.

$$\text{Sim}(\mathbf{S_P}, \mathbf{S_G}) = \text{Median}_k \left(\max_l \text{Corr}(\mathbf{S_{Pk}}, \mathbf{S_G})(l) \right) \quad (2)$$

At the core of the above computation is, of course, the need to compute the similarity between two silhouette frames, $\text{FrameSim}\left(\mathbf{S_P}(i), \mathbf{S_G}(j)\right)$, which we simply compute to be the ratio of the number of pixels in their intersection to their union. Thus, if we denote the number of foreground pixels in silhouette $\mathbf{S}$ by $\text{Num}(\mathbf{S})$ then we have,

$$\text{FrameSim}(\mathbf{S_P}(i), \mathbf{S_G}(j)) = \frac{\text{Num}(\mathbf{S_P}(i) \cap \mathbf{S_G}(j))}{\text{Num}(\mathbf{S_P}(i) \cup \mathbf{S_G}(j))} \quad (3)$$

3.3. Parameters

There is no calibration requirement. However, the algorithm does have three parameters that need to be chosen. The first parameter, D_{Maha}, is used to threshold the Mahanalobis distance. Since this distance measure is normalized by the covariances, the choice of the threshold tends not to be sensitive to the particular image. We chose it to be 4. The second parameter, N_{Size}, is used to delete small regions and fill in small holes in the thresholded difference image. We chose it to be 200 pixels. The third parameter, N, is the size of each subsequence obtained by partitioning the probe sequence. We chose it to be 30, which is approximately the number of frames for one walk stride. We decided on these chosen values for the thresholds based on visual assessment of the silhouettes from 7 sequences for 3 subjects.

4. Challenge Experiments

In this section we put forward a set of challenge tasks or experiments, of increasing hardness, in gait based recognition and establish a baseline performance for each of them. We structure the challenge tasks in terms of gallery and

Table 2. The probe set for each of challenge experiments. The gallery for all of the experiments is (G,A,R) and consists of 71 individuals. The number of subjects in each subset are in square brackets.

Exp.	Probe	Difference
A	(G, A, L)[71]	View
B	(G, B, R)[41]	Shoe
C	(G, B, L)[41]	Shoe, View
D	(C, A, R)[70]	Surface
E	(C, B, R)[44]	Surface, Shoe
F	(C, A, L)[70]	Surface, View
G	(C, B, L)[44]	Surface, Shoe, View

probe sets, patterned after the FERET evaluations [8], of varying degrees of differences between them in terms of the covariates. Among the four possible covariates, we have so far studied three of them: walking surface, shoe type, and viewpoint. The weight carrying cases are reserved for future exploration, since it is probably the hardest covariate to handle at this point. The USF-NIST data allows for 2 possible values for each of the three covariates, which are: concrete (C) or grass (G) walking surfaces, two shoe types (A and B), and left (L) and right (R) camera viewpoints. Based on the values of these covariates we can divide the dataset into 8 possible subsets: { (G, A, L), (G, A, R), (G, B, L), (G, B, R), (C, A, L), (C, A, R), (C, B, L), (C, B, R) }. Since not every subject was imaged under every possible combination of factors, the sizes of these sets are different (Table 1). We choose one of the large subsets (G, A, R), i.e. (Grass, Shoe Type A, Right Camera), as the gallery set. The rest of the subsets are probe sets, differing in various ways from the gallery. The structure of the challenge experiments is listed in Table 2. More specifically, the gallery and probe sets consist of the frames from the back portion of the elliptical path, where the motion of the subject is mostly fronto-parallel. Detailed specifications of the gallery and probe sets in terms of the exact frame numbers are available at the website mentioned in the abstract.

For each experiment, we compute the similarity of each probe sequence with each gallery sequence. Following the pattern of the FERET evaluations, we measure performance for both identification and verification scenarios, using cumulative match characteristics (CMCs) and receiver operating characteristics (ROCs).

5. Baseline Performance

Fig. 5 plots the CMCs and ROCs of the 7 challenge experiments. Table 3 lists some of the key performance indicators, namely, the identification rate (P_I) at rank 1, the verification rate (P_V) for a false alarm rate of 10%, and the area under the ROC (AUC). There are several observations to be made. First, the identification ranges from 10% to 73% at rank 1, which improves to a range from 45% to 88% at rank 5, which is approximately 7% of the gallery set size. In terms of ROC performances, the detection rates range from 34% to 82% for a false alarm rate of 10%. These are very encouraging performances given the simplistic nature of the baseline algorithm. It is to be expected that more sophisticated algorithms will result in much better performances, for which there is much room.

Second, both the identification rates, as seen in the CMCs, and the detection rates, as seen in the ROCs, fall

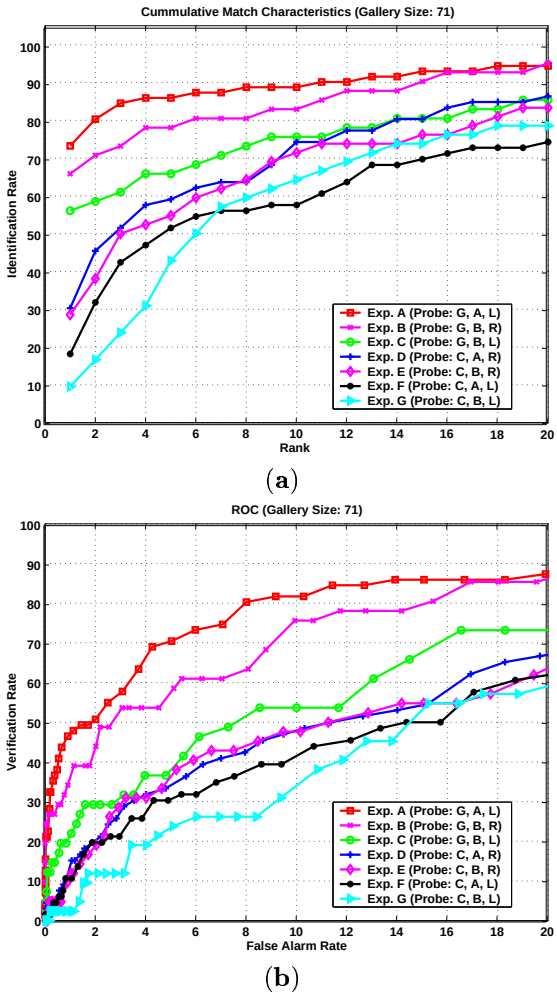

Figure 5. Baseline performances for the challenge experiments (a) CMC curves and (b) ROCs plotted upto a false alarm rate of 20%

Table 3. Baseline performance for the challenge experiments in terms of the identification rate P_I at rank 1, verification rate P_V at a false alarm rate of 10%, and area under ROC (AUC)

Exp.	Difference	P_I	P_V	AUC
A	View	73%	82%	0.90
B	Shoe	66%	76%	0.88
C	Shoe, View	56%	54%	0.82
D	Surface	30%	48%	0.81
E	Surface, Shoe	29%	48%	0.82
F	Surface, View	18%	41%	0.74
G	Surface, Shoe, View	10%	34%	0.77

as one goes from experiment A to G. This offers a natural ranking of the experiments in terms of their challenge nature, i.e. the situation in experiment A, where the difference between probe and gallery is just the viewpoint, is easier to solve than that in experiment G, where the probe is different in terms of all the three covariates.

Third, among the three covariates, view point variation seems to have the least impact and surface type has the most impact based on the drop in the identification rate due to each of these covariates. Apart from the effect of the individual covariates on performance, there also seem to be interactions between their effects. For instance, shoe type (Experiment B) seems to impact performance more than viewpoint (Experiment A) but viewpoint change along with surface change (Experiment F) impacts performance more than shoe type change along with surface change (Experiment E). More detailed statistical studies on larger data sets are needed to quantify these interactions.

6. Conclusions and Discussion

The HumanID gait challenge problem is a valuable and important computer vision problem. Any reasonably general solution will have to address the difficult problems of segmentation and occlusion.

The HumanID gait challenge problem dataset is large and challenging, with subsets representing a range of increasingly difficult problems. The baseline algorithm performance varies from 73% on the simplest case to 10% and 30% on the hardest experiments. The full release of the dataset will be quadrapole the current size, incorporating time and carrying condition and so expanding the range of covariates to be explored. The infrastructure developed around this challenge problem should greatly expand and facilitate research in the area of recognition using gait analysis. Researchers wishing to work on a new algorithm will not have to invest the substantial start-up costs of acquiring a dataset large enough to lend credibility to their results. The varied levels of size and difficulty of problems will allow researchers to enter the research stream in this area at different levels of algorithmic sophistication. Also, the availability of intermediate results will facilitate researchers being able to focus on one sub-problem of the overall problem.

We expect that the availability of this dataset and baseline algorithm will greatly facilitate the reliable evaluation of new algorithms, and make it easier for new researchers to explore their own gait analysis algorithms. Researchers may also be motivated to explore other issues. For example, the problem of automated motion-based segmentation is important and this dataset presents an opportunity to measure the practical impact of advances in this area.

7. Acknowledgment

This research was supported by funds from the DARPA Human ID program under contract AFOSR-F49620-00-1-00388. Stan Janet and Karen Marshall from NIST very meticulously assembled the data for distribution and created the bounding boxes for the sequences. Thanks to the HumanID researchers at CMU, Maryland, MIT, Southampton, and Georgia Tech. for discussion about potentially important covariates for gait analysis. We also thank Dr. Pat Flynn from U. of Notre Dame for testing the baseline algorithm code and scripts before release.

References

[1] C. BenAbdelkader, A. Cutler, H. Nanda, and L. Davis. Eigengait: Motion-based recognition of people using image self-similarity. *3rd International Conference on Audio- and Video-Based Biometric Person Authentication*, June 2001.

[2] J. Hayfron-Acquah, M. Nixon, and J. Carter. Automatic gait recognition by symmetry analysis. *3rd International Conference on Audio- and Video-Based Biometric Person Authentication*, June 2001.

[3] A. Johnsson and A. Bobick. A multi-view method for gait recognition using static body parameters. *3rd International Conference on Audio- and Video-Based Biometric Person Authentication*, June 2001.

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

[5] D. Meyer, J. Posl, and H. Niemann. Gait classification with HMMS for trajectories of body parts extracted by mixture densities. In *BMVC98*, 1998.

[6] H. Murase and R. Sakai. Moving object recognition in eigenspace representation: Gait analysis and lip reading. *PRL*, 17(2):155–162, February 1996.

[7] S. Niyogi and E. Adelson. Analyzing gait with spatiotemporal surfaces. In *Vismod*, 1994.

[8] P. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.