Silhouette-based Human Identification from Body Shape and Gait *

Robert T. Collins, Ralph Gross and Jianbo Shi
Robotics Institute, Carnegie Mellon University
Email: {rcollins,rgross,jshi}@cs.cmu.edu

Abstract

Our goal is to establish a simple baseline method for human identification based on body shape and gait. This baseline recognition method provides a lower bound against which to evaluate more complicated procedures. We present a viewpoint dependent technique based on template matching of body silhouettes. Cyclic gait analysis is performed to extract key frames from a test sequence. These frames are compared to training frames using normalized correlation, and subject classification is performed by nearest neighbor matching among correlation scores. The approach implicitly captures biometric shape cues such as body height, width, and body-part proportions, as well as gait cues such as stride length and amount of arm swing. We evaluate the method on four databases with varying viewing angles, background conditions (indoors and outdoors), walk styles and pixels on target.

1. Introduction

Although the basic pattern of bipedal locomotion is similar between healthy humans, gaits do vary between individuals. A person's gait depends on a multitude of factors including physical build and body weight, shoe heel height, clothing and emotional state of mind. There is ample anecdotal evidence about people being able to identify acquaintances based only on their manner of walking.

There is a rich body of work describing computer vision systems for modeling and tracking human bodies (see [4] for a review). However, the vision research community has only recently begun to investigate gait as a biometric [1, 7, 8, 9, 11, 13, 14]. We have developed a simple method for identifying walking humans based on body shape and gait. The method is based on matching 2D silhouettes extracted from key frames across the gait sequence. The benefits of the approach are 1) it is easy to understand and implement, 2) it can tolerate noisy video data, 3) gait sequences as short as one stride can be used, 4) the method is insensitive to clothing color and texture, and 5) it appears to generalize well across different walking gaits. The main drawback to the method is that it is view dependent – since it is based on matching 2D shape silhouettes it cannot classify test subjects viewed from significantly different angles than the training subjects. However, even in this respect the method is analogous to state of the art approaches in face recognition, which are also applicable only over limited viewpoints, namely frontal or "mug-shot" views [6, 12].

Section 2 outlines the method, while Section 3 presents an evaluation on four datasets collected by different computer vision groups. Weaknesses of the approach and ideas for improvements are discussed in Section 4.

2. Method

This section presents a simple method for identify walking humans based on template matching of a sequence of body silhouettes. Key frames from a probe sequence are compared to training frames using normalized correlation, and classification is performed by nearest neighbor matching on correlation scores. Steps in the algorthm are 1) silhouette extraction, 2) gait cycle analysis to identify key frames, 3) template extraction, 4) template matching via normalized correlation, and 5) nearest neighbor classification using combined scores from templates across a full gait cycle.

2.1 Silhouette Extraction

In our experiments, body silhouette extraction is achieved by simple background subtraction and thresholding, followed by a 3x3 median filter operator to suppress isolated pixels. Figure 3 shows sample results. Note that the extracted silhouettes can contain holes, the silhouette boundary can be interrupted, and static background pixels can be mistakenly included (mainly due to shadows). This is typical of the mistakes made by such algorithms in practice.

2.2 Gait Cycle Analysis

Direct comparison of body silhouette images is not possible since 2D body shape changes nonrigidly throughout the gait cycle as the limbs move. We first process each sequence of silhouette images to extract key frames representing landmark poses within the gait cycle. These landmarks are identified by examining periodic signals computed from the sil-

*This work is supported by ONR contract N00014-00-1-0915.

houette sequence. Gait cycle analysis serves two important functions. First, it determines the frequency and phase of each observed gait sequence, allowing us to perform dynamic time warping to align sequences before matching. Secondly, it provides data reduction by summarizing the sequence with a small number of prototypical key frames.

Since we want a recognition algorithm that is robust and efficient, we do not want to base key frame selection on first estimating limb positions in either 2D or 3D. Instead, consider silhouette width as a function of time. For side views of a walking person, we see that this is a periodic function with distinct peaks and valleys (Figure 1). [1] The bounding box alternatively expands (peaks) and contracts (valleys) over time as the person's legs spread and come back together again during the gait cycle. Selecting key frames at the peaks and valleys of this signal results in four key frames summarizing a single stride. The frames extracted correspond roughly to the following physiological gait labels: right double support (both legs spread and touching the ground, right leg is in front), right midstance (legs are closest together with the swinging left leg just passing the planted right foot), left double support, and left midstance.[2] Although we can't disambiguate between right and left phase, this simple method DOES determine double support and midstance frames quite reliably.

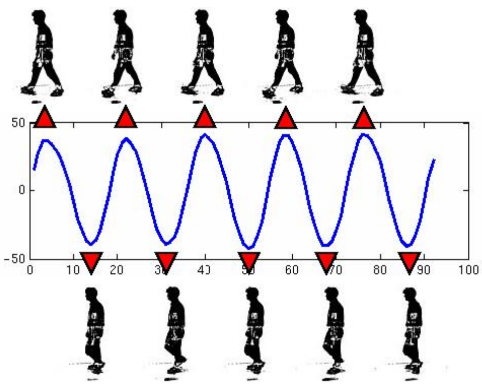


Figure 1: Extracting double support and midstance frames. The Y-axis represents silhouette width (centered at zero), and the X-axis represents time.

For frontal views, silhouette width is less informative, but silhouette height as a function of time plays an analogous role in that its peaks and valleys indicate double support and midstance gait frames, assuming the viewpoint is slightly elevated (for example, from a camera mounted on the ceiling looking down a hallway). As a person's front leg gets closer to the viewer, the silhouette appears to elongate down the rows of the image, resulting in a greater apparent height than when the two legs are close together.

Figure 2 shows periodic width and height signals over time for silhouettes viewed from six widely-space viewpoints of the CMU Mobo database (The Mobo database is illustrated in Figure 3a. Views 1 to 6 of Figure 2 correspond to the views shown left-to-right in Figure 3a). In choosing whether to use the width or height signal, we always use the signal with the highest amplitude, since this signal should have a better signal-to-noise ratio. In Figure 2 this means that we identify key frames using width for viewpoints 1, 4 and 5, and height for viewpoints 2, 3 and 6. Cameras in the Mobo database were synchronized, so that frames for this sequence are precisely aligned temporally. By comparing frame numbers selected by choosing signal peaks and valleys across the viewpoints shown, we find that the average temporal difference between key frames chosen in viewpoint 1 versus the five other views is 1.9 frames, or 0.06 seconds at 30 fps. This indicates that the method has applications to aligning gait sequences from unsynchronized cameras with widely-space viewpoints.

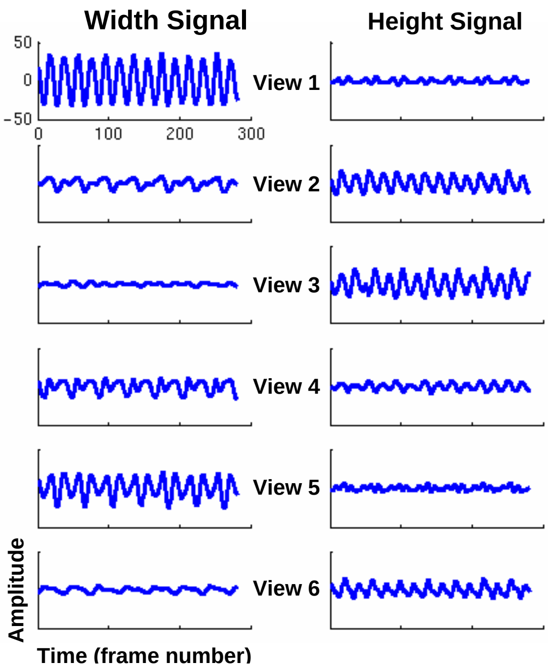


Figure 2: Periodic gait signals based on silhouette width (left) and height (right), compared for the six different viewpoints in the CMU Mobo database. All subplots have the same axis limits. For each viewpoint, the signal with the largest amplitude should be used to extract key frames.

[1] We first filter the raw response with a bandpass filter to suppress noise and accentuate the periodic structure.

[2] The term "midstance" is more precisely the point of transition between the end of the midstance phase and the beginning of the terminal stance phase within each half of the gait cycle. See [3].

2.3 Template Extraction (Training)

After locating gait sequence key frames, we create a template for each by scaling and cropping the key frame silhouette. The silhouette is scaled so that the person is 80 pixels tall, centered within a template 80 pixels wide by 128 pixels high. This leaves a roughly 20 pixel border of zero pixels around the silhouette, which is important when doing shift-invariant correlation using FFT (next section), since circular shifting is being done. Templates are labeled according to whether they come from double support or midstance key frames. A training "gallery" is formed consisting of all key frame templates extracted from each training gait sequence. In the future, we may attempt to summarize long sequences by retaining a smaller set of prototypical templates throughout the length of the sequence.

2.4 Template Matching (Testing)

We compare templates from a test subject (the "probe" set) with templates in the training gallery using normalized correlation. Let a and b be two templates to compare, the match score C between a and b is computed as

$$C(a,b) = \frac{\max(\hat{a} * \hat{b})}{\max(\hat{a} * \hat{a})\max(\hat{b} * \hat{b})}$$

where $\hat{v} = (v - \mathrm{mean}(v))/(\mathrm{std}(v)$ is a normalized vector and the $*$ operator signifies cross-correlation. An FFT-based implementation of cross-correlation is used that computes the correlation at all pixel displacements between the two templates. The maximum correlation value over all shifts is chosen as the template match score $C(a,b)$.

Let the gallery of double support templates be denoted as $\left\{P_k^s\right\}$ with index s ranging over all subjects and k ranging over all double support key frames extracted for subject s. Similarly, $\left\{V_l^s\right\}$ is the set of midstance gallery templates. Extracting templates from a probe sequence produces template sets $\{p_i\}$ and $\left\{v_j\right\}$, with i and j ranging over all double support and midstance key frames, respectively. In preparation for nearest neighbor classification, we compute all match score pairs $\left\{C(p_i,P_k^s)\right\}$ and $\left\{C(v_j,V_l^s)\right\}$, where indices i,j,k,l range over their applicable values. Note that we only match double support probe templates with double support gallery templates, and similarly for midstance templates. We do not, however, distinguish between the right and left phases of double support and midstance, since at present we cannot reliably determine this phase when extracting key frames.

2.5 Nearest Neighbor Classification

After key frame template matching, we have correlation scores between each template in the probe sequence and each relevant template in the training gallery. Rather than performing nearest neighbor classification directly on individual templates, we prefer to combine template scores to form a score for each complete gait cycle. Recall that key frame extraction results in four frames per gait cycle, corresponding to right double support, right midstance, left double support and left midstance. We therefore can form probe quadruplets $\{p_i, v_i, p_{i+1}, v_{i+1}\}$ that contain key frames from each complete stride in the test sequence.

Let $R(p_i, s_0)$ be the relative likelihood that template p_i corresponds to subject s_0, computed as

$$R(p_i,s_0) = \frac{\max\left\{C(p_i,P_k^s)|s = s_0\right\}}{\max\left\{C(p_i,P_k^s)\right\}} \quad .$$

That is, the maximum correlation over templates from subject s_0 divided by the maximum correlation over all gallery double support templates. A similar measure is defined over midstance templates v_i. Our classification s^* for the gallery subject that best matches a probe quadruplet is then

$$s^* = \underset{s}{\mathrm{argmax}} \; \left[R(p_i,s) + R(v_i,s) + R(p_{i+1},s) + R(v_{i+1},s)\right] \quad .$$

The best match is the subject with the highest total relative match score over four adjacent key frames forming one stride. To perform K-nearest neighbor variants, subjects are ranked by decreasing total relative match score.

3. Algorithm Evaluation

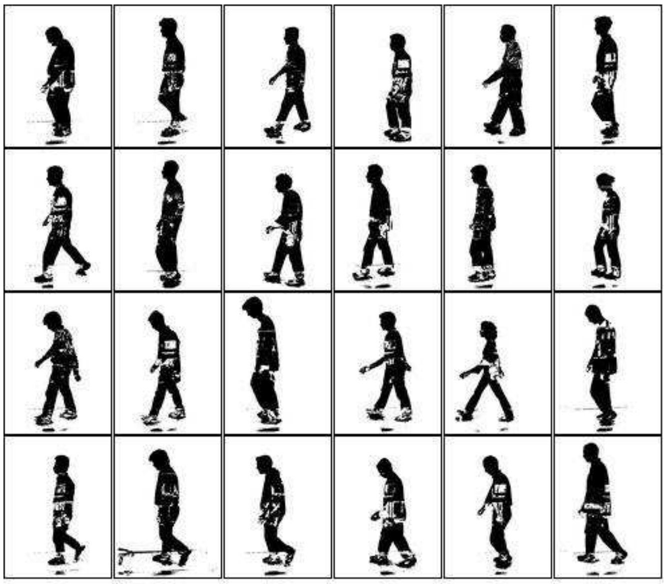
We evaluate our algorithm on four large gait databases collected by Carnegie Mellon University, University of Maryland, University of Southampton and Massachusetts Institute of Technology, within the DARPA Human Identification (HID) program. The databases contain raw image sequence data and foreground masks computed by each collecting institution. Table 1 gives an overview of the different database conditions. Figure 3 lists data collection specifics of each database and shows sample images and silhouettes.

	CMU	MIT	UMD	USH
Walking Location	Indoor Treadmill	Indoor floor	Outdoor ground	Indoor floor
Subjects	25	24	55	28
Views	6	1	2	1
Synchronized	Y	N/A	N	N/A
Walk styles	4	1	1	1
Pixel height	500	100	150	300
Frame rate [fps]	30	15	20	25

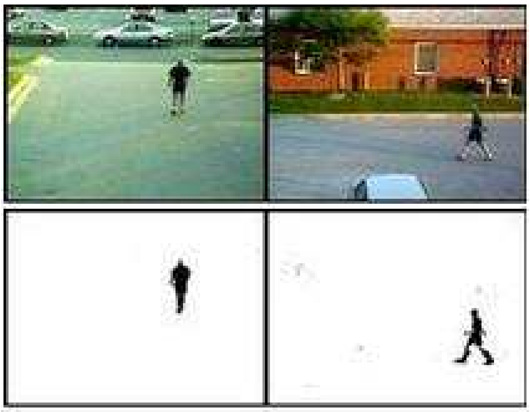
Table 1: Overview of the databases used in the evaluation.

Following Phillips et.al. [12] we distinguish between *gallery* and *probe* images. The gallery contains images used

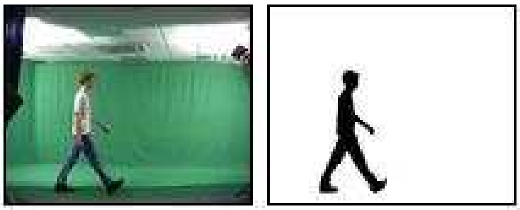
(a) The six CMU Mobo database viewpoints.



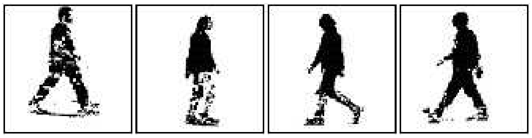
(b) Sample silhouettes from the CMU Mobo database.



(c) Samples from the U.Maryland database.



(d) Samples from the U.Southampton database.



(e) Sample silhouettes from the MIT gait database.

Figure 3: Gait databases used for algorithm evaluation. **(a and b) The CMU MoBo database** [5] contains six simultaneous motion sequences of 25 subjects (23 male, 2 female) walking on a treadmill. The 3CCD progressive scan images have a resolution of 640x480. Each subject is recorded performing four different types of walking: slow walk, fast walk, inclined walk, and slow walk holding a ball (to inhibit arm swing). Each sequence is 11 seconds long, recorded at 30 frames per second. More than 8000 images are captured per subject. **(c) The U.Maryland database** [2] contains two datasets of people walking outside. Our evaluation concentrates on the second, larger dataset with 55 individuals (46 male, 9 female). The subjects are walking a T-shaped pattern in a parking lot and are recorded with two orthogonally positioned surveillance cameras (Philips G3 EnviroDome). A total of four different body poses (frontal, right, left and back) are visible during each sequence. For each pose typically 9-11 steps are recorded. This database is challenging due to the recording conditions (outside, surveillance camera) and number of subjects. **(d) The University of Southampton database** [10] comprises 28 subjects walking indoors on a track. The subjects are imaged with a camera view perpendicular to the walking direction. Each subject appears in four sequences, recorded in direct succession. Each sequence consists of a complete stride from heel strike to heel strike. The subjects are recorded against a uniform green background, so the application of chromakey extraction results in extremely clean silhouettes. **(e) The MIT database** shows 25 subjects (14 male, 11 female) walking twice indoors on a path perpendicular to a single camera (Sony Handycam). 13 out of the 25 subjects were recorded in at least two and up to four sessions over the span of a three month period. Silhouette images are already cropped and subsampled to size 128x128.

during training of the algorithm, while the probe set contains test images. All results reported here are based on non-overlapping gallery and probe sets. We use the *closed universe* model for evaluating performance, meaning that every subject in the probe set is also present in the gallery.

Table 2 summarizes the database collection conditions that are varied within each of our experiments. Experiment 1 considers tests of the algorithm when the gallery and probe image sets have the same gait, and are taken on the same day. In Experiment 2, we train on a slow walk gait, and then test on a fast walk, and a walk carrying a ball. Experiment 3 considers subjects with the same gait, but viewed on different days. All results are presented as cumulative match scores which plot the probability of correct identification against relative rank K. For example, a value of 85% at a rank of 10% means that the correct subject label is included within the top 10% of subjects (ranked by match score) 85% of the time.

	Exp 1			Exp 2	Exp 3
Database	MIT	UMD	USH	CMU	MIT
Variable day					x
Variable gait				x	
Variable view				x	
Variable session	x	x	x		

Table 2: Overview of experimental conditions.

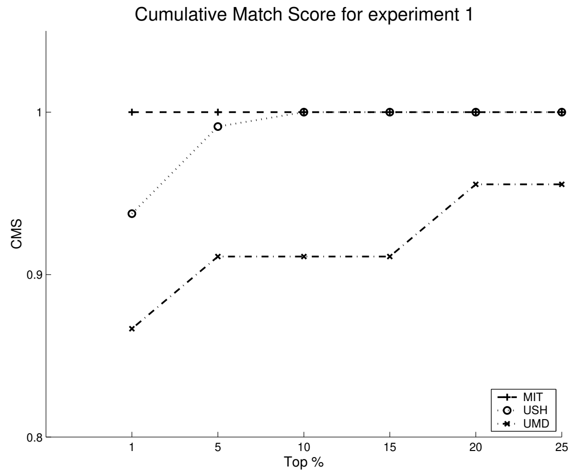


Figure 4: Cumulative match score: within-gait condition.

Training conditions	Testing conditions	top 1	top 5%	top 10%
profile slow	profile fast	76%	92%	92%
profile slow	profile ball	92%	96%	96%
frontal slow	frontal fast	100%	100%	100%
frontal slow	frontal ball	92%	100%	100%

Table 3: Match scores for the across-gait condition.

3.1. Within gait condition

This set of tests examines the ability of our algorithm to recognize individuals across multiple recordings of the same gait. The different sequences in the databases were recorded during the same session. Figure 4 shows the cumulative match scores for ranks up to 25% for the UMD, USH and MIT datasets. The algorithm shows excellent performance on the USH and MIT datasets. It scales well from the two small indoor datasets (USH, MIT) to the large outdoor dataset (UMD).

3.2. Across gaits condition

We evaluate our algorithm for three different gaits of the CMU dataset: slow walk, fast walk and slow walk holding a ball. Table 3 shows the results for slow walk (gallery) vs. fast walk (probe) and slow walk (gallery) vs. ball (probe) for two different view angles (profile, frontal). Again the algorithm shows excellent performance.

3.3. Across days condition

The across days condition represents the hardest test of this evaluation. This is due in part to same-subject differences caused by changing clothing (bulky vs thin) and hairstyles, both of which alter 2D silhouette shape. However, a major factor is differences in lighting and the contrast between clothing and background across two measurement sessions, which leads to significant differences in silhouette accuracy for the same individual across different days. As a result, classification rates are much lower (Figure 5). We hypothesize that more robust methods for silhouette extraction would yield an improvement in scores.

4. Conclusion

We have presented a simple method for human identification from body shape and gait. The method is based on matching 2D silhouettes extracted from key frames across a gait cycle sequence. These key frames are compared to training frames using normalized correlation, and subject classification is performed by nearest neighbor matching among correlation scores. The approach implicitly captures biometric shape cues such as body height, width, and body-part proportions, as well as gait cues such as stride length and amount of arm swing.

We have evaluated the method on four databases with varying viewing angles, background conditions (indoors and outdoors), walk styles and pixels on target. Overall,

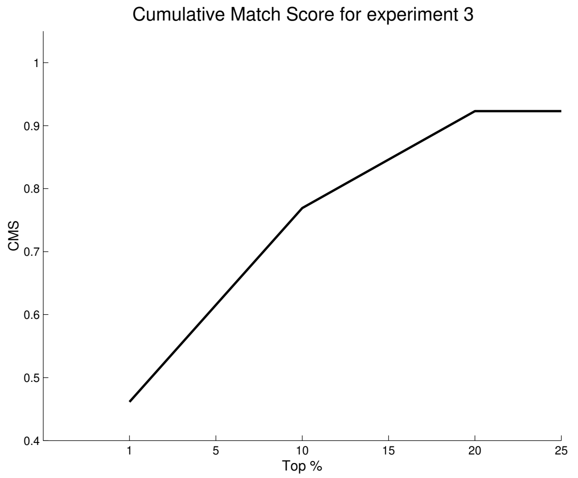


Figure 5: Cumulative match score: across-days condition.

the method performs well when used within a single viewpoint, even recognizing people when the testing gait type (fast walk, walking with ball) differs from the training gait (slow walk). The method can handle noisy silhouettes, such as those extracted from typical surveillance video data, and it can be used on sequences as short as a single stride.

Because it is based on 2D template matching, the approach is obviously limited to classifying test sequences taken from roughly the same viewing angle as the training sequences. In operational settings with cooperative subjects, the viewpoint can be controlled and this is not a problem. Even with subjects who are unaware that they are being watched, cameras can be placed at "choke points" where walking direction is limited, or multiple cameras can be used to ensure that a range of viewing directions is available. The obvious way to generalize the algorithm itself is to store training sequences taken from multiple viewpoints, and classify both the subject AND the viewpoint. However, the inability to generalize to situations where a person must be recognized from a totally new viewpoint is a fundamental limitation that we feel should be addressed by other approaches based on recovery of 3D shape, or discovery of relative phase between different moving body parts. These approaches are the subject of our current research.

References

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

[2] T. Chalidabhongse, V. Kruger, and R. Chellappa. The UMD database for human identification at a distance. Technical report, University of Maryland, 2001.

[3] E.Ayyappa. Normal human locomotion, part 1: Basic concepts and terminology. In *Journal of Prosthetics and Orthotics*, volume 9(1), pages 10–17. The American Academy of Orthotists and Prosthetists, 1997.

[4] D.M. Gavrila. The visual analysis of human movement: A survey. *CVIU*, 73(1):82–98, January 1999.

[5] R. Gross and J. Shi. The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, 2001.

[6] R. Gross, J. Shi, and J. Cohn. Quo vadis face recognition? In *Third Workshop on Empirical Evaluation Methods in Computer Vision*, December 2001.

[7] J.J.Little and J.E.Boyd. Recognizing people by their gait: The shape of motion. In *Videre (online journal)*, volume 1(2), Winter 1998.

[8] M.Nixon, J.Carter, D.Cunado, P.Huang, and S.Stevenage. Automatic gait recognition. In A.Jain, R.Bolle, and S.Pankanti, editors, *Biometrics: Personal Identification in Networked Society*, pages 231–249. Kluwer Academic Publishers, 1999.

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

[10] M. Nixon, J. Carter, J. Shutler, and M. Grant. Experimental plan for automatic gait recognition. Technical report, University of Southampton, 2001.

[11] S.A. Niyogi and E.H. Adelson. Analyzing and recognizing walking figures in xyt. In *IEEE Proceedings Computer Vision and Pattern Recognition*, pages 469–474, 1994.

[12] P.J.Phillips, H.Moon, S.Rizvi, and P.Rauss. The feret evaluation methodology for face recognition algorithms. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 22(10), pages 1090–1104, 2000.

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

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