A Survey of Silhouette-Based Gait Recognition Methods

Craig Martek, cmm6857@rit.edu

Abstract

Gait recognition is a relatively new area being studied as a possible biometric. This paper presents a survey of some recent gait recognition methods based on the silhouette images acquired from a sequence. Individual methods are discussed and some of the comparable results produced from these are discussed.

1 Introduction

A person's gait is influenced by a variety of factors, including age, height, and weight, just to name a few. In fact, gaits have been shown to be unique enough for someone to recognize a friend by only looking at lights attached to their joints as they walk [1]. Accordingly, gaits have been increasingly studied as a way of uniquely identifying a person.

Two main groups of gait recognition approaches exist: those that employ a model of the human body and those that rely on motion present in a frame. Model-based approaches tend to be difficult to implement, as they usually require some mapping from two to three dimensions. Motion-based approaches often operate on the silhouettes of a subject, which can be easily acquired through preprocessing methods.

In [2], motion-based recognition methods are further divided into two categories: those that represent motion as a sequence of poses and those that map the distribution of the motion through space and time. This paper will review both of these types of motion-based recognition and compare some of the results achieved.

Figure 1: Examples of Silhouettes from the NLPR Database

2 Related Methods

Model-based approaches generally attempt to fit the input to some model of the human body. These models usually include locations and limitations of the body segments and joints. One such approach observes the rotation angle of the hip across time, as this is a roughly periodic feature in gait [3]. Another method first segments the person into body parts and then models hip rotation and gait period [4].

Model-based approaches have some advantages over those that are motion-based. As many are based on silhouette images, occlusion and segmentation errors can both have a damaging impact on recognition rates. Shadows are also a frequent problem, as they can be detected as motion and tend to be connected to the human. However, the problem of locating and segmenting body parts in a two-dimensional image can be difficult and computationally expensive, while silhouettes are relatively simple to obtain.

3 Preprocessing

Many silhouette-based recognition methods have been built around the assumption of a fixed camera observing a mostly static scene. This enables silhouettes to be obtained through simple background modeling and subtraction methods.

Several methods of background segmentation are used for gait recognition. One of the simplest methods described in [5] first calculates the mean and covariance values for each pixel across an entire sequence. Mahalanobis distance is then used to determine which pixels in each frame are distant enough to be considered foreground. Another approach presented in [6] and

utilized in [7] is the Least Median of Squares method. In this, if I is a sequence of N frames, the background is computed with

$$b_{xy} = \min_p med_t(I_{xy}^t - p)^2 \tag{1}$$

where t is the index of the frame and p is the background brightness value for that particular pixel. Foreground pixels can then be determined by thresholding the difference between the frame and the background measure.

4 Spatiotemporal Approaches

Spatiotemporal gait recognition methods are those that observe the distribution across the XY and time dimensions. Commonly this will create large feature spaces that need to be mapped to lower-dimensionality ones for use in training and classification.

4.1 Baseline Algorithm

A baseline algorithm is proposed in [5] for the sake of measuring significant contributions to the problem of gait recognition. A dataset was also defined that consists of video sequences of 74 subjects walking outside under varying parameters, including different shoe types, surfaces, and camera angles. As these parameters are varied either individually or together, the baseline algorithm demonstrated a decreasing level of recognition.

The recognition algorithm in [5] is quite simple yet effective. For each sequence, foreground pixels are segmented as described in Section 3. After smaller regions are discarded, a bounding box is drawn around the remaining foreground silhouette. This resulting box is then scaled to a common size. Classification of gait sequences is based on a similarity measure between the training sequences, or the "gallery," and the test sequences, the "probe." The similarity between two sequences is defined as the median value of the maximum correlations between the gallery sequence and each subsequence of the probe. This similarity measure is defined as

$$Sim(S_P, S_G) = Med_k(\max_l Corr(S_{Pk}, S_G)(l)) \tag{2}$$

where Corr(l) is simply the sum of the frame similarity measures between each frame of the probe subsequence P_k and a gallery subsequence S_G starting

at frame l. This frame similarity measure is just the ratio of the number of pixels that the two frames have in common to the number of pixels they have combined.

Some interesting experiments were performed with this baseline algorithm. By slowly removing parts of the silhouettes from both top to bottom and bottom to top, it was determined that the area from approximately the knees downward is responsible for almost all of the successful recognition rate achieved by [8], an optimized version of [5] that performs noticeably better with variations in both surface and shoe type and slightly better in most other variations.

4.2 PCA

Due to the high-dimensionality of the feature space generated by many of these methods, principal component analysis is frequently used to determine where the primary variation is in the feature space and possibly eliminate some features that do not produce significant variation. After these spaces have been mapped to an eigenspace, classifiers such as k-nearest neighbor will be much more computationally tractable.

4.2.1 Distance Signals

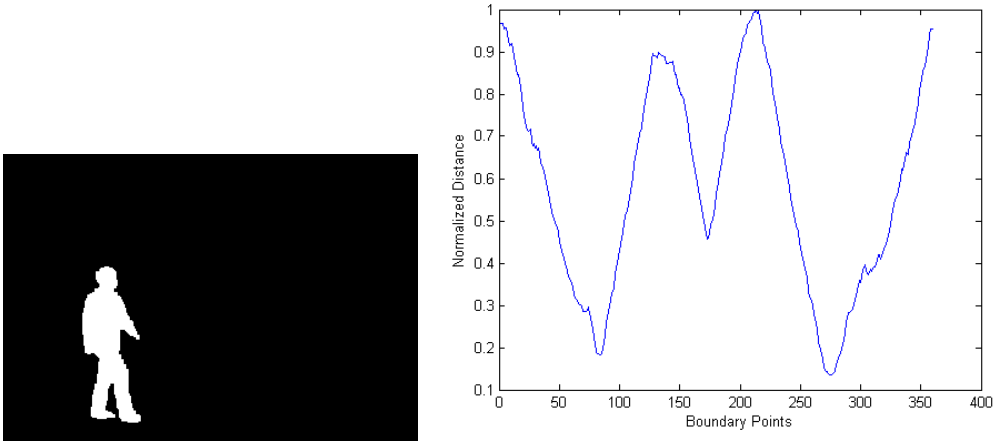
One spatiotemporal approach presented in [7] bases its recognition on the distance between each silhouette boundary pixel and the centroid of the person blob. Looking at the outline of a silhouette allows the system to better respond to noisy images. A signal is produced by starting at the topmost point on the contour directly above the centroid and moving counterclockwise, calculating the distance from each contour point to the centroid. The signals for each sequence are then normalized with respect to length

$$S_i(j) = D_i\left(\left[\frac{j * N_D}{N_S}\right]\right) \tag{3}$$

and magnitude

$$S_i = \frac{S_i}{\max S_i} \tag{4}$$

where D_i is the ith signal in the sequence, S_i is the normalized signal, and N_{D_i} and N_{S_i} are the number of points in each. This effectively represents



(a) (b)

Figure 2: Normalized distance signal (b) for the silhouette (a)

each gait sequence as a set of normalized 1-dimensional signals, such as the one seen in Figure 2.

The training phase of this algorithm determines the eigenvalues and eigenvectors across the set of training sequences. In the experiments, this becomes a 15-dimensional eigenspace into which all training sequences are first mapped. A similarity measure, either normalized Euclidean distance or spatial-temporal correlation, is then used with 1-nearest neighbor to classify the test sequence.

Normalized Euclidean distance is simply the norm of the difference between the average projections for the sequence. This can be written as

$$d^2 = \left|\left| \frac{C_1}{||C_1||} - \frac{C_2}{||C_2||} \right|\right|^2 \tag{5}$$

Spatial-temporal correlation, or STC, is a distance measure that accounts for different temporal alignments between sequences. This measure is defined by

$$d^2 = \min_{ab} \sum_{t=1}^{T} ||P_1(t) - P_2'(at+b)||^2 \tag{6}$$

in which $P_1(t)$ and $P_2(t)$ are the eigenspace projections for each sequence,

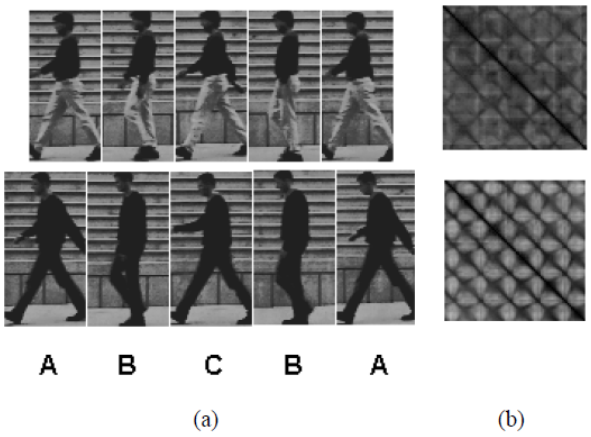


Figure 3: Self similarity plots (b) for the sequences shown in (a).

and $P_2'(at+b)$ is a temporal transformation of P_2. It seems that this distance measure would produce better identification results, but STC performed worse than the normalized Euclidean difference in each of the experiments.

4.2.2 EigenGait

Another spatiotemporal method is described in [2] and [9] in which a plot of image self-similarity is used to classify sequences. For each frame, a bounding box is drawn around the person. Image self-similarity between two frames from the same sequence is defined as the sum of all the pixels in the difference image of the two frames. A self-similarity plot can therefore be constructed by plotting this value for all combinations of sequence frames. This plot can be seen to contain information about the period and magnitude of a person's gait. Two of these plots can be seen in Figure 3. The main diagonal will always be darkest, but other dark diagonals can also be seen where the pose is either the same or opposite.

The self-similarity plot is used as an input to the classification system. The plots need to be normalized with respect to phase, so that cycles will start at approximately the same location; and frequency, because the same person may walk at a different pace between two iterations of recording.

4.2.3 Hough Transform

A third method that employs PCA to reduce the feature space is described in [10]. The Hough transform is employed, which is a method of locating certain features through voting. For each sequence, the period of the gait cycle is determined in order to acquire a single cycle. The Hough transform is then computed for the edge images of each of these silhouettes, as indicated in Figure 4. These template images essentially contain information about where straight lines exist in the sequence of silhouettes. PCA is then used to decrease the amount of features, as the templates are still the same size as the silhouettes.

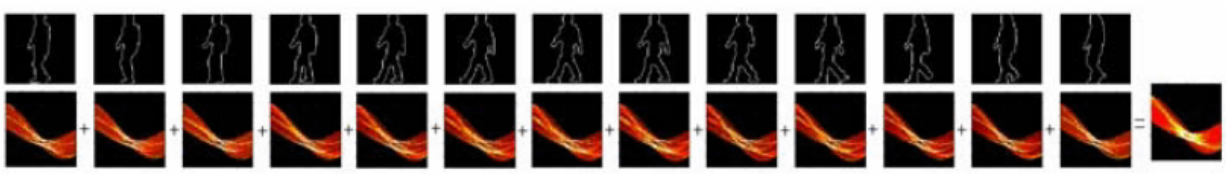


Figure 4: Hough transforms for a sequence of silhouettes and the template computed for the entire sequence.

4.3 Other Methods

Some approaches attempt to use distinct body parts as a method of classifying people by gait. These are the closest to model-based approaches, yet they take a more simplistic view of the construction of a human being. For example, [11] breaks each silhouette into seven regions based on location with respect to the centroid, specifically the head, front and back of torso, and thighs and feet for each leg.

To generate features out of this division, an ellipse is fit to each of the segmented parts and the centroid, aspect ratio, and orientation of the ellipse is calculated. Across the whole sequence, the mean and standard deviation of each of these features and the mean of the height of the silhouette is determined. Another vector containing magnitude and phase information with respect to a single step is also calculated using a Fourier transform whose input is the previous feature vector. This produces two different feature vectors for each sequence that can be used for classification.

Since this produces a large feature vector, the entirety of which may not be necessary for good classification, a method known as analysis of variance, or ANOVA, is used to determine features that do not discriminate well between classes. This method produces a value known as the p-value that determines the probability that the variation occurs as a result of chance. The features are ranked using this value to determine those with the most actual variation. The classifier used is a simple 1-nearest neighbor method, using Mahalanobis distance between the pruned input and feature vectors to make the classification.

5 Discrete Approaches

In contrast to the spatiotemporal methods described above, discrete approaches aim to represent human gait with by considering variations over time with respect to a set of static configurations of the body [7].

5.1 PCA

Similar to the spatiotemporal approaches, many discrete approaches need to map a large feature space to a smaller one for classification purposes. A method is proposed in [12] that uses optical flow templates as a feature for classification. Each sequence of templates is first collapsed to an eigenspace and then to a smaller space with canonical space transformation. Recognition in this particular method is done by first projecting the input sequence into the canonical space as for training and then finding its nearest neighbor in the space.

5.2 Hidden Markov Models

Some other discrete methods aim to represent gait cycles with a hidden Markov model. These are systems in which the state of the system is not actually observable, but each state will generate output based on probability.

One such system described in [13] attempts to describe each person's gait in five separate stances with a hidden Markov model. For such a system, silhouettes are compacted to width vectors containing the difference between the right and leftmost pixels across the height of the person. The five stances

are determined by applying k-means with $k = 5$ to the subject's width vectors. The width vectors are then encoded with respect to the stances using

$$d^2 = \left|\left|OW^j(k) - S_l^j\right|\right| \tag{7}$$

where $OW^j(k)$ is the width vector for the kth frame of the jth person and S_l^j is the width vector for the lth stance of the jth person.

Another method presented in [14] uses Hu image moments as an input to the hidden Markov model. These are moments are defined in [15] that do not change with scale or rotation of the image. In this algorithm, the rough symmetry of a gait cycle is exploited to determine how many frames are in the cycle, and therefore how many states are in the model. An HMM is constructed and trained on two sequences for each subject. Classification is then defined as

$$M = \underset{j=1,2,\ldots,C}{\arg\max}\, P(M_j|S) \tag{8}$$

where S is the vector quantization of the feature vectors obtained through the Hu moments and C is the number of HMMs. This returns the HMM that is most likely to have produced the particular sequence S.

A third more recent approach is demonstrated in [16] that aims to eliminate some of the issues with broken or occluded silhouettes, using a frame difference energy image. The silhouette images in a sequence are first clustered and then averaged to obtain a gait energy image. These energy images are then cleared of noise by removing any pixels that fall below a certain threshold. A separate set of difference images are calculated by finding the difference between each silhouette and the one before it. The frame difference energy image is then the sum of each distance image with its respective denoised image. An example of this is demonstrated in Figure 5.

The frieze features of the FDEIs are used as inputs to the HMM. These features are simply

$$F(y,t) = \sum_x B(x,y,t) \tag{9}$$

where B is the silhouette image and t is the time. The frieze feature vectors are then clustered and the center of each cluster is calculated as an exemplar. The HMM contains one state for each cluster, and each state transitions to either the next state or the same state on a 50/50 chance.

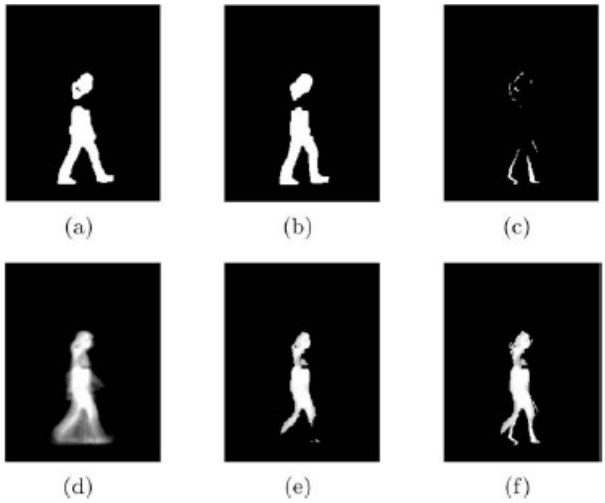


Figure 5: Construction of the FDEI. (a) and (b) are two sequential silhouette images, (c) is the difference image between (a) and (b), (d) is the gait energy image, (e) is the denoised image, and (f) is the constructed frame difference energy image

The frieze features of the FDEI are shown to perform better than the frieze features of the initial silhouette images.

6 Published Results

Comparing results between different methods of gait recognition is currently difficult, as there are no major standard databases for testing. However, some approaches have started to utilize the data set from [5] to compare the performance of their algorithm. Also, [7] implemented some similar algorithms and compared their results on the lateral view NLPR database they developed, which contains four sequences for each of twenty subjects. These results are indicated in Table 1. The results from [10] are appended to this table, although no data is available on computational time.

The baseline algorithm developed in [5] showed the difficulty of varying the different conditions (surface, shoe type, view). Experiments in which the view, shoe type, or both the shoe type and angle of view were changed between the training and testing phases boasted significant identification rates. Varying other parameters severely decreased the effectiveness of the

Table 1: Comparison of Results on NLPR Database (0°) [7]

Methods	Top 1 (%)	Top 5 (%)	Top 10 (%)	Computational cost (min/seq)
BenAbdelkader 2001 [2]	72.50	88.75	96.25	Medium (8.446)
Collins 2002 [17]	71.25	78.75	87.50	High (17.807)
Lee 2002 [11]	87.50	98.75	100	Low (2.2365)
Phillips 2002 [5]	78.75	91.25	98.75	Highest (200)
Wang 2003 (no validation) [7]	75.00	97.50	100	Lowest (2.054)
Wang 2003 (w/ validation) [7]	82.50	100	100	
Liu 2009 [10]	97.5	100	100	no data

identifier.

Running the challenge experiments from [5] on the method in [7] produced roughly comparable results, although the latter enjoys a much smaller computational cost. Similarly, the two methods had about the same level of performance on the NLPR database introduced by [7]. Identification rates in which the correct choice is in the top 1, 5, and 10 from those selected are shown in Table 1 for some of the algorithms discussed earlier in this paper.

7 Discussion of Results

Many of these methods are limited in overall effectiveness. For instance, methods like that proposed in [11] currently only work on sequences of people walking parallel to the viewing plane. For all other methods, varying the viewing angle will almost certainly degrade the results achieved. All silhouette-based methods can be adversely affected by artifacts in the generation of the actual silhouette, whether natural in the form of bulky clothing or artificial through background modeling problems. Also, current methods are reliant upon a stationary camera over a mostly invariant scene.

On a single angle in the NLPR database, the methods in [11], [7], and [10] exhibited the best performance. Of note, due to the significantly larger size of the feature space in [10], computational time is likely much higher than in the other two methods. In the experiments defined in [5], the method in [7] generally performed slightly worse than did the baseline algorithm. It can be seen in this that performance of these algorithms depends heavily on the construction of the dataset.

Most of the gait databases currently in use do not deal with occlusion. However, background segmentation errors are quite frequent. The frame difference energy image introduced in [16] seems to compensate for some of these issues and would presumably improve upon other methods that operate off of the raw silhouettes.

8 Conclusion

As interest in person identification increases, unique and efficient methods are constantly being developed. Gait recognition is particularly interesting because of the possibility of recognizing a person from a significant distance. This is a relatively new area of research, but a wide variety of promising methods are demonstrated in this paper as well as with model-based approaches. While not currently effective enough to be used as a sole basis for identification, many of these methods could certainly be used for verification of existing results.

References

[1] C. Barclay, J. Cutting, and L. Kozlowski, "Temporal and spatial factors in gait perception that influence gender recognition," *Perception and Psychophysics*, vol. 23, no. 2, pp. 145–152, 1978.

[2] C. BenAbdelkader, R. Cutler, H. Nanda, and L. Davis, "Eigengait: Motion-based recognition of people using image self-similarity," in *Proc. Int'l Conf. Audio- and Video-Based Biometric Person Authentication*, 2001, pp. 284–294.

[3] D. Cunado, M. Nixon, and J. Carter, "Automatic extraction and description of human gait models for recognition purposes," *Computer Vision and Image Understanding*, vol. 90, no. 1, pp. 1–41, 2003.

[4] D. Wagg and M. Nixon, "On automated model-based extraction and analysis of gait," in *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, 2004, pp. 11–16.

[5] P. J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer, "Baseline results for challenge problem of human ID using gait analysis," in

Proc. Int'l Conf. Automatic Face and Gesture Recognition, 2002, pp. 137–142.

[6] Y. Yang and M. Levine, "The background primal sketch: An approach for tracking moving objects," *Machine Vision and Applications*, vol. 5, pp. 17–34, 1992.

[7] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1505–1517, 2003.

[8] P. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer, "The gait identification challenge problem: Data sets and baseline algorithm," in *Proc. Int'l Conf. Pattern Recognition*, 2002, pp. 385–388.

[9] C. BenAbdelkader, R. Cutler, and L. Davis, "Motion-based recognition of people in EigenGait space," in *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, 2002, pp. 267–274.

[10] L. Liu, W. Jia, and Y. Zhu, "Gait recognition using hough transform and principal component analysis," *Emerging Intelligent Computing Technology and Applications*, vol. 5754, pp. 363–370, 2009.

[11] L. Lee and W. Grimson, "Gait analysis for recognition and classification," in *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, 2002, pp. 148–155.

[12] P. Huang, C. Harris, and M. Nixon, "Human gait recognition in canonical space using temporal templates," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 146, no. 2, pp. 93–100, 1999.

[13] A. Kale, A. Rajagopalan, N. Cuntoor, and V. Krüger, "Gait-based recognition of humans using continuous HMMs," in *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, 2002.

[14] Q. He and C. Debrunner, "Individual recognition from periodic activity using hidden markov models," in *Proc. IEEE Workshop Human Motion*, 2000, pp. 47–52.

[15] M. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions Information Theory*, vol. 8, no. 2, pp. 179–187, 1962.

[16] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian, "Frame difference energy image for gait recognition with incomplete silhouettes," *Pattern Recognition Letters*, vol. 30, no. 11, pp. 977–984, 2009.

[17] R. Collins, R. Gross, and J. Shi, "Silhouette-based human identification from body shape and gait," in *Proc. Int'l Conf. Automatic Face and Gesture Recognition*, 2002, pp. 366–371.