Learning Pedestrian Models for Silhouette Refinement

L. Lee G. Dalley K. Tieu
Massachusettes Institute of Technology Artificial Intelligence Laboratory
{llee, dalleyg, tieu}@ai.mit.edu

Abstract

We present a model-based method for accurate extraction of pedestrian silhouettes from video sequences. Our approach is based on two assumptions, 1) there is a common appearance to all pedestrians, and 2) each individual looks like him/herself over a short amount of time. These assumptions allow us to learn pedestrian models that encompass both a pedestrian population appearance and the individual appearance variations. Using our models, we are able to produce pedestrian silhouettes that have fewer noise pixels and missing parts. We apply our silhouette extraction approach to the NIST gait data set and show that under the gait recognition task, our model-based sulhouettes result in much higher recognition rates than silhouettes directly extracted from background subtraction, or any non-model-based smoothing schemes.

1. Introduction

The ability to accurately segment pedestrians from a video stream is important for applications such as gait recognition, person height/girth estimation [4], articulated body tracking, pedestrian activity description [3], and 3D reconstruction of people from silhouettes [7]. We use a model-based approach to pedestrian segmentation that incorporates information from background subtraction, pedestrian shape models, and an individual shape model sampled at discrete phases of the walking cycle. Our approach eliminates noise introduced by background subtraction, and fills in missing parts of the pedestrian silhouette, which often result from camera noise or lack of color/intensity difference between the pedestrian and the background. In addition, our pedestrian models are learned from a noisy background subtraction process, hence making the entire process completely automatic.

Traditional approaches to pedestrian segmentation from video generally involve using a background subtraction algorithm to arrive at foreground silhouettes, then post-processing to refine the silhouettes. Because background subtraction inherently detects pixel value changes in the video/image, spurious foreground pixels are formed by noise in the video, and pedestrian silhouettes will have holes and missing parts if there is not enough contrast between the pedestrian and the background scene. The general solutions to these difficulties are to apply large numbers of morphological operations to fill in holes and remove noise in the silhouettes, or to apply a smoothing process at the background subtraction stage [9]. In either case, these operations tend to systematically distort the silhouettes and remove fine details which may be important for identification.

We investigate the particular case of pedestrians walking in a plane roughly parallel to the camera image plane. Under this scenario, it is easy to see that there are commonalities between all pedestrian shapes. Moreover, the cyclic nature of the walking action ensures that the silhouette appearance of each individual pedestrian is repeated at each stride at regular intervals. These observations make it possible to improve the estimation of silhouette appearance over time and over a population of pedestrians. We take advantage of these characteristics to learn two types of pedestrian models, one that represents all pedestrians, and one that represents each individual walking video sequence. Using the pedestrian population model and individual sequence models, we are able to remove noise from each frame of a silhouette sequence and fill in missing parts of each silhouette. To show that these silhouettes are an improvement over the traditional methods of silhouette smoothing, we apply our approach to the NIST gait data set[1] to produce a set of silhouttes and use these silhouettes in a set of baseline gait recognition tests introduced in [9]. Our results show that recognition results are improved using our model-based silhouettes.

While we are only concerned with extracting pedestrian silhouettes in this paper, the method we propose is generally applicable to any moving object that demonstrates the cyclic nature and common overall appearance that are observed in pedestrians. For example, models for joggers, or trotting dogs or horses may be built using the same technique.

2. Previous Work

There are many works related to the problem of pedestrian detection/tracking/segmentation. We indicate several of relevance to our approach. We categorize these works into two

[1] See http://www.gaitchallenge.org.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

types: pedestrian detection, and pedestrian shape representation.

Oren *et al.* [8] trained a set of wavelet template representations of the frontal view of pedestrians. These representations capture the shape gradient difference between the pedestrian and the surrounding background. The authors applied their pedestrian representation to images to detect roughly frontal (or back) views of pedestrians. Gavrila [2] used a set of edge models of pedestrian shapes to detect pedestrians from video sequences taken with a moving camera. While pedestrian detection is the goal of both algorithms, additional steps are needed to extract the silhouette of pedestrians.

Haritaoglu *et al.* [3] used background subtraction to detect, segment, and track pedestrians, but they did not eliminate the errors introduced by background subtraction. Baumberg and Hogg [1] represented the pedestrian shape by a chain of edge points. However, a clean segmentation of the pedestrian is assumed, and point selection requires human intervention.

Kale *et al.* [5] use a five state HMM for gait identification and reduce their observation distributions to a single Gaussian per state using noisy silhouettes. Zhou and Chellappa [11] use a time series continuous state space model to recognize people walking toward the camera.

While there are pedestrian model representations presented in these papers, they do not address problems inherent to background subtraction that make accurate extraction of pedestrian silhouettes difficult. Without the full silhouette, questions such as "what color of clothing is the pedestrian wearing" can be hard to answer. Our approach seeks to overcome these difficulties by learning a probability distribution of pedestrian foreground models at different phases of a walking cycle over time and then using these models to provide better shape definitions and to recover from errors in the background subtraction process.

3. The Need for Model-based Segmentation

If pedestrians always appeared in colors that are drastically different from the surrounding background, and there were no cast shadows, then pedestrian segmentation from any image would be a simple task. However, in any realistic video monitoring situation people may have colors on the body that are close to the background, and shadows will appear. For example, Figure 1 shows the intensity of one color channel of one pixel location in a video sequence. We manually found the frames for which the pedestrian is the foreground at that location. Clearly, there are some frames for which the foreground process is indistinguishable from the background process. The pedestrian in this case is wearing a black shirt and walking past a black background. As a result there are large holes in the torso of the silhouette. These are difficulties that no local background subtraction algorithm can solve, because background subtraction only detects changes in pixel intensities. A model-based pedestrian representation imparts expectations on the structures of pedestrians, and the confidence level associated with the expectations will allow us to ignore the noise in the video data and fill in the expected structure where data is missing.

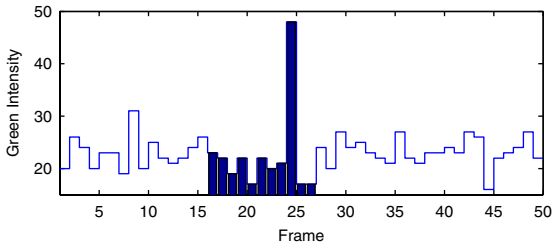


Figure 1: The intensity of a pixel through time. Within the foreground segment, indicated by the dark bars, there are times when the intensity is indistinguishable from the background values.

4. Learning Pedestrian Models

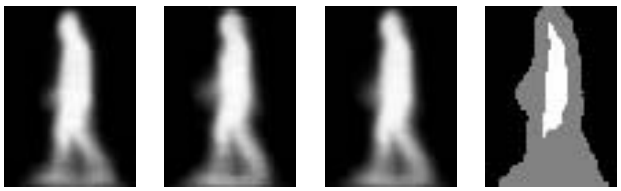
We consider the case where the pedestrian is walking in a plane that is roughly parallel to the image plane and always in the same direction. Under this scenario, the cyclic nature of pedestrian silhouette appearance is readily apparent. The same phase of a walking cycle will appear repeatedly in a sequence. Hence we can obtain a better estimate of a silhouette by using all silhouettes that correspond to that same phase. To further simplify the problem, we assume that the walking direction is known. Hence we only need to represent the silhouettes in one direction while the silhouette appearance from the opposite direction is a mirror image of the standard direction.

The above observations lead to a straightforward method for obtaining a pedestrian model within a silhouette sequence using a number of discrete phase representations:

1. Detect the period of the silhouette sequence using periodic features, such as the silhouette aspect ratio.
2. Align all silhouettes by the phase of the walking cycle assuming a constant walking period.
3. Average all silhouettes assigned to the same phase.

Assuming that there is no systematic error in the backgrounding process or in the environment, and that pedestrians walk at roughly constant speed, this method will generate a good representation of silhouettes over different phases of a walking cycle that captures the shape of the pedestrian in the walking sequence. However, both of these assumptions are occasionally violated. If the pedestrian has consistent patches of clothing that match the background environment, the raw silhouette sequence will have many

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE



(a) (b) (c) (d)

Figure 2: The pedestrian population models: (a) male, (b) female, (c) average model, and, (d) mask for pedestrian shape–black for turning off a pixel, white for turning on a pixel, and gray for unchanged pixel).

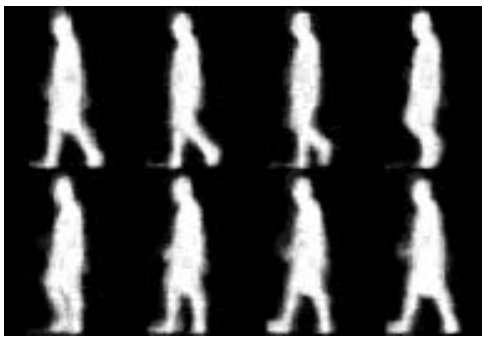


Figure 3: Sample model of 8 phases of the walking cycle for one of our sequences after HMM training.

frames with large holes in the body. If the walking speed of the pedestrian changes in a sequence, assigning a silhouette to its correct phase may be difficult.

4.1. Pedestrian Population Model

To address the issue of systematic noise in a gait video sequence, we devise a separate model that represents the appearance of all pedestrians, which we name the pedestrian population model. We assume that while systematic errors in background subtraction may occur for one walking sequence, they are unlikely to occur at the same location for a population of pedestrians. Hence, a silhouette model constructed using a sampling of silhouette sequences from a general population of pedestrians will not suffer from systematic background errors. However, because different individuals have different stride lengths, aligning and averaging silhouettes from different pedestrians by phase results in blurred legs, especially for the phase with the widest stance. As a consequence, we choose to represent the silhouette of all pedestrians with the mean silhouette of a training set that is representative of the population.

There are some postural differences between the silhouette appearances of male and female pedestrians (see Figure 2a and 2b), thus the training sequences need to contain an equal number of male and females. Figure 2c shows the average pedestrian model computed from 5 males and 5 females randomly chosen from our gait data set. This population model is generated using 100 random silhouette frames from each of the 10 training subjects. The amount of data used represents 1% of all the data frames, and 8% of the total number of frames of the training subjects.

4.2. Pedestrian Sequence Model

To overcome the constraint on constant walking period, we construct a hidden Markov model (HMM) of the silhouette appearances where each state represents the silhouette at different stages of walk for each pedestrian silhouette sequence. The transitions between the states in an HMM contain information about the relative amount of time a pedestrian stays at each state and thus covertly constrains the period, but this is not a hard constraint and does allow for adaptation to changing walking speed in a walking sequence. In addition, because an HMM is trained on each sequence, the states of the HMM will represent the silhouette appearance of each sequence much better than a pedestrian model constructed using any generic silhouette sequence.

An HMM is a probabilistic model of a random process with discrete states. In a first order Markov model, the state of the system at time $t+1$ can be predicted knowing only the state at time t, *e.g.* $P(s_{t+1}|s_t, s_{t-1}, ..., s_0) = P(s_{t+1}|s_t)$. In our case, the states are 8 phases of the walking cycle, represented as images in Fig. 3. A Markov model is hidden when we are unable to directly observe the states. Instead we observe some output of the system, characterized by a probability distribution, $P(y_t|s_t)$. For the pedestrians, we see images (our observations) of a person instead of having a perfect noise-free "phase detector" (our states).

An HMM is characterized by the probability of starting in some state, $P(s_0 = i)$, the transition probabilities, $P(s_{t+1} = i|s_t = j)$, and the observation probabilities, $P(y|s)$. These probabilities are estimated using standard techniques [10], with the following caveats. First, we model walking as a set of cyclical transitions between N discrete states, where we have selected $N = 8$. Second, we assume that a person will start being filmed at a random time with respect to the phase, so $P(s_0 = i) = \frac{1}{N}$ for all i. Third, we set the transition probabilities to be:

$$P(s_{t+1} = i|s_t = j) = \begin{cases} 1 - \frac{1}{f/N} & \text{if } i = j \\ \frac{1}{f/N} & \text{if } i = j + 1 \bmod N \\ 0 & \text{otherwise} \end{cases}$$

where f is the average number of frames in a walking cycle for the given sequence ($f > N$). $\frac{f}{N}$ is the average number of frames per phase transition, so $\frac{1}{f/N}$ is the probability of transitioning out of a state after one frame. Finally, our observations are binary silhouette images. We model the

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

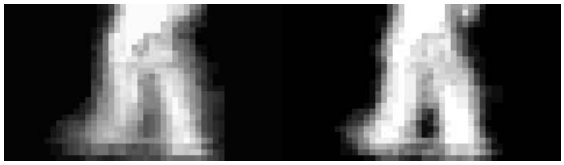


Figure 4: Closeup of the legs for the sixth state of a sequence. Left: original estimate based on averaged frames. Right: refined estimate after HMM training.

probability of each individual pixel being turned on as an independent Bernoulli random variable. This model can be represented as an image where the intensity of a model pixel is the probability that that pixel will be on in an observed binary silhouette image. As previously mentioned, Fig. 3 is a rendering of this model for a particular HMM we trained.

To train the HMM, we must supply initial estimates of these probabilities. In our case, the transition and initial state probabilities are fixed as described. For the observation probabilities, we start by assuming a near-constant walking speed and assigning the widest stance to be state 0. We then estimate the state of the frames:

$$s_t = \left\lfloor s_0 + t\frac{1}{f/N} \mod N \right\rfloor$$

where s_0 is the state for frame 0 and t is a non-negative integer. For our initial observation probabilities estimates, we then average all of the frames assigned to each state. For the frames in the NIST gait data we are using, the assumption of near-constant walking speed is valid, and these initial estimates work well. A more robust method would be necessary if there were significant changes or drift in the walking speed.

Once we have the initial estimate of the observation probabilities, we train an HMM on the sequence silhouettes to refine the probabilities. The HMM is able to adapt to smaller fluctuations in walking speed and make the observation model sharper, as seen in Figure 4 .

5. Raw Silhouette Extraction

We have assumed to this point that the raw pedestrian silhouettes used as input for our model-based pedestrian silhouette extraction method had been obtained and that the tracking of the silhouette is accurate. Below we describe the process by which we obtain such a set of silhouettes.

5.1. The Gait Data

The data set we are using is the standard NIST gait data set; the details of the data collection method are described in [9]. Subjects were asked to walk along a smoothly curving path under differing environmental and imaging conditions.

The difficulties posed by automatically extracting good silhouettes from this data set include: shadows on the ground, grass covering feet, moving objects (including people, palm trees, fluttering construction tape, etc.) in the background, subjects wearing clothing that is largely indistinguishable from the background. All of these make the tracking and background subtraction problem difficult. However, the predefined pedestrian path allows us to apply global constraints to simplify the tracking problem. Because all frames of a gait video sequence are available at processing time, we are able to use a batch background subtraction algorithm to extract the foreground.

5.2. Tracking and Background Subtraction

Because of the moving objects in the scene and the amount of harsh shadows, tracking the pedestrian accurately becomes a challenging problem if we make no assumption about the gait data. To simplify the tracking problem, we instead use frame differencing (i.e. we subtract color values at each pixel between successive frames) to initially locate the pedestrian in the image. Frame differencing has the advantage that it is robust to gradual lighting change, large shadows, and even waving trees, thus allowing us to localized the pedestrian accurately. However, it does suffer from missing pixels from the upper portion of the body at times, because the torso generates less motion than the legs. Hence we have to choose a large bounding box to outline the pedestrian.

In addition to using the frame difference image to localize the pedestrian, we also impose a constraint that the paths of the silhouette centroid must be smooth to a 2nd degree polynomial. We use a repeated robust estimation process to generate a path and a set of bounding boxes containing the pedestrian.

Given the bounding boxes for the tracked pedestrian in each frame, silhouette extraction using background subtraction is relatively straightforward. A background is modeled as an array of Gaussian distributions in RGB color space, one for each pixel location. For each image frame, the Mahalanobis distance of each pixel location is computed using its corresponding pixel value and background Gaussian. This distance image is thresholded to yield a binary silhouette. The parameters for each Gaussian are estimated using all pixel values occurring at the pixel location except when the location is within a pedestrian bounding box.

6. Model-based Silhouette Refinement

Given the raw pedestrian silhouettes generated in the process described in Section 5, and the pedestrian models described in Section 4, we can post-process the raw silhouettes by scale normalizing the silhouettes and then using the silhouette models to remove noise and fill in holes at each

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

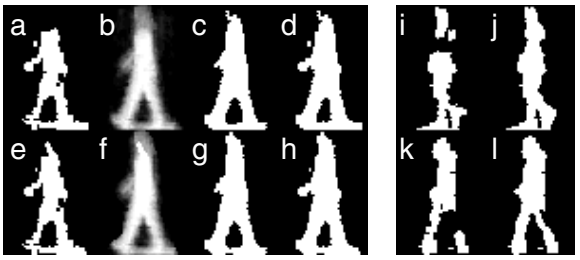


Figure 5: Silhouette filling examples. (a) raw silhouette (b) HMM model for the state most likely to have generated the silhouette (c) mask made by thresholding b (d) logical-OR of a and c. (e)-(h) are the same as a-d, except e is the population-filled silhouette and the HMM in f was trained on the population-filled sequence. (i) another raw silhouette from a different person (j) i after population- and HMM-filling. (k)-(l) same as i and j for a third silhouette.

frame. Our pedestrian silhouette model involves two levels of representations: the pedestrian population representation and the pedestrian sequence representation, each requiring a different treatment.

The pedestrian population model, generated by averaging a set of training silhouettes equally representing men and women, is used to refine the raw silhouettes. We can interpret the average as the maximum likelihood estimate of the parameters of a population silhouette generative process. Each pixel location L is an independent Bernoulli process with parameter $\theta_L = p(L = 1)$. Given a sequence of silhouettes from a pedestrian, we want to choose a binary value for each pixel location in every frame. We can obtain the posterior distribution of θ_L given the sequence and a prior based on the population parameters. In principle, we could threshold the maximum *a posteriori* value of θ_L. However because the population model prior is only valid for static binary shapes, we can only confidently threshold at pixel locations for which the shape is static across time (i.e., low variance Bernoulli processes). Empirically, we found that restricting the prior to be valid only in the range $\theta_L \geq 0.9$ and $\theta_L \leq 0.05$ worked well . All other pixel locations in the pedestrian silhouette sequence are left unchanged. This set of thresholds gives us the mask shown in Figure 2d. Note that the pixels that are consistently turned on are the ones interior to the pedestrian torso and head region, and the ones that are turned off are far from the edge of the silhouettes, whereas the unchanged pixels are the edge of the silhouettes and the legs.

The pedestrian sequence model, a cyclic silhouette model representing discrete phases of a walking cycle, is used to produce silhouettes that preserve the fine details of individual pedestrian. This model is trained on each sequence and hence is able to preserve the detailed shape of the silhouette in the sequence. We begin by training an HMM on the sequence, as described in Section 4.2. Using that HMM, we determine the most likely state assignments for each of the silhouettes using the Viterbi algorithm [10]. To do the filling, we turn on any pixel in a silhouette that has a likelihood of greater than 0.5 in the HMM.

In Fig. 5, we see an example of the two filling methods: (a) has no filling, (d) is HMM-filled, (e) is population-filled, and (h) is both population and HMM-filled. In this example, the population-filling recovers part of the head and removes a few spurious pixels. The HMM-filling is able to fill in more of the head and parts of the lower torso. In (i) and (j), we see an example of filling in the entire upper torso and part of the hair for a different person. The legs are filled in for a frame of a third person in (k) and (l).

7. Evaluation Methods

To evaluate the quality of our model-based silhouettes, we apply these silhouettes in a gait recognition task. We use two gait recognition algorithms—an existing algorithm described in [6] briefly summarized below, and a distance metric based on the silhouette HMM states.

Our gait dynamics feature vector consists of smoothed versions of moment features in image regions containing the walking person. For each silhouette of a gait sequence, we find the centroid and divide the silhouette into 7 parts roughly corresponding to head/shoulder, arms/torso (front and back), thighs(left/right), and calves/feet(left/right) (see Figure 6(a)). For each of the regions, we fit an ellipse to describe the centroid, the aspect ratio and the angle of the portion of foreground object visible in that region(Figure 6(b)). We assume that all of these features–the centroid, aspect ratio, and angle of each region–are sampled from a Gaussian distribution and compute the mean and standard deviation for each of these parameters across each walking sequence. The feature vector of mean and standard deviation of each region is used in a nearest neighbor classifier to retrieve the identity whose walking dynamics feature vector is closest to the query feature vector.

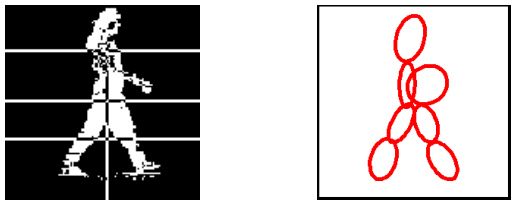
In addition to the region-based features, we also use the states of our HMM silhouette model as a gait representation. We use the Euclidean distances between the 8 HMM state observation models as comparison between two gait silhouette sequences.

8. Results

Given a set of gait data, we perform the following steps,

1. Track the pedestrian and extract a set of raw silhouettes using the algorithm described in Section 5.
2. Build the following pedestrian models:

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE



(a) Partition of a silhouette (b) Ellipse fit to each region

Figure 6: Computing the feature vector for gait recognition

- A population model that represents the appearance of all pedestrians. This model is constructed using 100 random raw silhouettes from each of 5 male and 5 female subjects (Section 4).
- A sequence-based HMM that models the silhouette at discrete phases of a walking cycle.

Note that these two models can be constructed independently of each other, or with the HMM following the pedestrian population model.

3. For each sequence, refine the silhouettes using the pedestrian population model and/or the state models of the HMM.
4. Generate a set of region-based gait features for recognition, or use the HMM states directly for recognition.

We applied the above steps to the HID gait challenge data set, which resulted in a suite of silhouettes and gait features. These silhouettes and gait features were then used in a set of gait recognition tasks.

8.1. Silhouette Comparisons

For each gait silhouette sequence, the following types of silhouettes are used in our experiments[2].

S_r the raw silhouettes, which are the results of the background subtraction process described in Section 5;

S_{d3} S_r dilated with a neighborhood size of 3;

S_{d6} S_r dilated with a neighborhood size of 6;

S_p S_r cleaned and filled using the population model;

S_{Hr} S_r filled using an HMM trained using S_r;

S_{Hp} S_p filled using an HMM trained using S_p;

S_N a set of silhouettes provided with the NIST gait data.

The silhouettes that are provided with the NIST gait data are semi-automatically generated in the following process:

1. Manually track the pedestrian in the video sequences.
2. Compute the Mahalanobis distance between the image containing the pedestrian and a background model.
3. Smooth the Mahalanobis distance image with a 9×9 filter.
4. Threshold the smoothed image to obtain the silhouette.

The smoothing process in step 3 has a side effect of smearing out the fine features of the silhouette and possibly removing some features that may be important to the identification of individuals.

Excluding the raw silhouettes, the set of silhouettes that we have chosen fall into two classes, those that reduce noise by a non-model-based process, such as smoothing or morphological operation, which are S_{d3}, S_{d6}, and S_N, and those that reduce noise by a model-based method, as in S_p, S_{Hr}, and S_{Hp}. We will show through gait recognition experiments that the silhouettes generated using a model-based method are consistently better.

We generate for each set of silhouettes the region-based gait features described in Section 7. In addition, the two types of HMM, generated using S_r and S_p, are also used for gait recognition.

8.2. The Recognition Task

The NIST gait challenge data is comprised of gait video of individuals taken under different conditions. A standard set of tests, described in [9], examines the gait recognition rate across different conditions.

The NIST gait data set contains pedestrians walking on different surfaces (concrete and grass), with camera view change (left and right views), and shoe type change. The data set is divided into a gallery set and a number of probe sets. The gallery set contains sequences of pedestrian walking on grass wearing one particular type of shoes and viewed from one of two cameras. The probe sets differs from the gallery in the following ways:

Probe Set	Difference
A	view
B	shoe
C	shoe, view
D	surface
E	surface, shoe
F	surface, view
G	surface, shoe, view

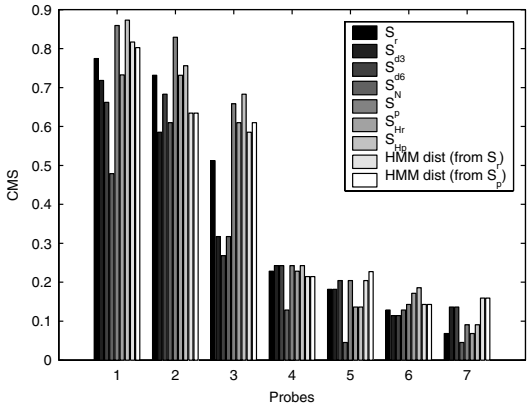
There are seven corresponding recognition experiments A through G, each testing a probe set against the gallery set. The task of a recognition algorithm is to rank the sequences in the gallery by their distances to the probe sequences. The recognition performance is evaluated using a cumulative match score (CMS), which measures the percentage of probes correctly identified at each ranking.

8.3. Recognition Results

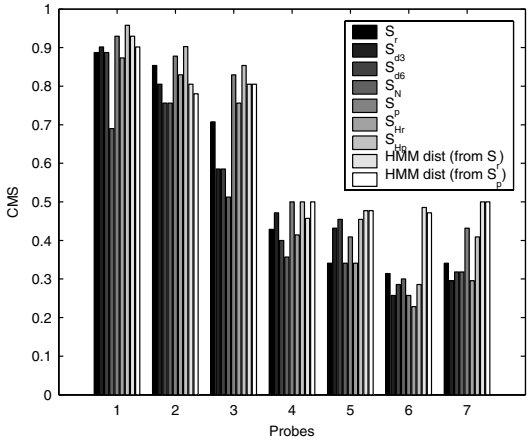
As in [9], we report the gait recognition rate using the cumulative match score at ranks 1 and 5, as shown in Fig-

[2]The silhouettes produced for this paper are available for download at `http://www.ai.mit.edu/people/llee/HID/NIST_sil.htm`.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE



CMS at rank=1



CMS at rank=5

Figure 7: Comparison of recognition rate using different silhouettes using CMS at rank 1 and 5.

ure 7. While we observe that the experiments D, E, F, and G, which all involve surface change, presented the more challenging recognition problems, it is less clear how much the silhouette type affects the recognition performance.

We are interested in the question of how well each silhouette type and gait feature type perform in all recognition experiments. To present a clearer picture, we average the CMS for each silhouette type across all recognition experiments A through G, across experiments A through C (the same surface condition), and across experiments D through G (the change-of-surface condition). The surface condition demands further investigation because it is the most challenging test. The averaged CMS are shown in Figure 8. The general trends presented in the recognition results are:

- Recognition rates using the region based features on the NIST silhouette set, S_N, are consistently worse.
- Using the region based features, raw silhouettes performed better than their dilated cousins in the same surface condition, but are comparable or slightly worse in the change of surface experiments and the average of all experiments.
- Using the region based features, the silhouettes S_p, and S_{Hp}, ranked by increased recognition performance, resulted in better recognition rate than the raw silhouettes or the non-model based smoothed silhouettes.
- Using the region based features, the silhouette set S_{Hr}, filled using an HMM trained on raw silhouette sequence, performed only marginally better than using the raw silhouettes.
- Using the distance of the HMM states, the recognition performances are comparable between the HMMs trained using the raw silhouettes and the HMMs trained on preprocessed silhouettes, S_p. They also performed much better than all other features in the change of surface condition.

8.4. Discussion

Our gait recognition experiments above show that incorporating a pedestrian model component, be it using HMM states for recognition or the region features on silhouettes filled with a pedestrian population model, resulted in better recognition rates than the non-model based silhouettes and the raw silhouettes. Based on the gait recognition performance using region based features on the various silhouette types, we rank, in increasing recognition rate, the quality of the silhouettes as follows: S_{Hr}, S_p, S_{Hp}. Simply using a silhouette model based on one sequence is not adequate because there may be persistent silhouette errors through a large number of frames. These systematic errors in the raw silhouette tend to be caused by lack of contrast between the foreground object and the background environment. The pedestrian population model is able to recover from this type of error because the persistent errors for one sequence are unlikely to persist through a population of pedestrians. Using the HMM silhouette model is an improvement over using just the pedestrian population model because it is able to improve the estimate of the individual shape over time and capture the appearance of the legs at discrete walking phase. The recognition rates using the state observation models of the HMM trained on raw silhouettes and the HMM trained on S_p were among the best three algorithms/silhouette data. This indicates that for recognition purposes, HMM silhouette models are robust to some systematic silhouette errors.

9. Conclusions and Future Work

We have proposed a method to automatically construct models of pedestrian silhouettes in a walking cycle. Our model contains two components, a pedestrian population based model, and an individual gait silhouette sequence

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE

model that is comprised of discrete phase states of walking cycles. The population model is used to recover from systematic noise of a particular gait sequence. The sequence model is used to correct for sporatic noise that occur from time to time within a video sequence. This model construction process can be applied to any moving object that exhibits cyclic properties and/or overall shape commonalities that allows one to improve the estimate of shape over time. Our silhouette models can be used in two ways: to fill in silhouettes for any algorithm that needs accurate silhouette sequences, and to be used directly for gait recognition. In both cases, we have shown that using a model based silhouette extraction is superior to using a non-model based silhouette smoothing algorithm, such as morphological operations, or a smoothing process in the background subtraction process. We are investigating the ability of these pedestrian models to remove persistent artifacts in silhouettes, such as shadows.

Acknowlegement

This research is supported in part by DARPA under contract N00014-00-1-0907.

References

[1] A. Baumberg and D. Hogg. Learning flexible models from image sequences. In *Proc of ECCV*, pages 299–308, 1994.

[2] D. Gavrila. Pedestrian detection from a moving vehicle. In *Proc of ECCV*, 2000.

[3] I. Haritaoglu, Davie Harwood, and L. Davis. W4: Who? when? where? what? In *Proc of Face and Gesture*, 1998.

[4] A. Johnson and A. Bobick. Gait recognition using static, activity-specific parameters. In *Proc of CVPR*, 2001.

[5] A. Kale, A.N. Rajagopalan, N. Cuntoor, and V. Kruger. Gait based recognition of humans using continuous HMMs. In *Proc of Face and Gesture*, 2002.

[6] L. Lee and W.E.L. Grimson. Gait aanalysis for recognition and classification. In *Proc of Face and Gesture*, 2002.

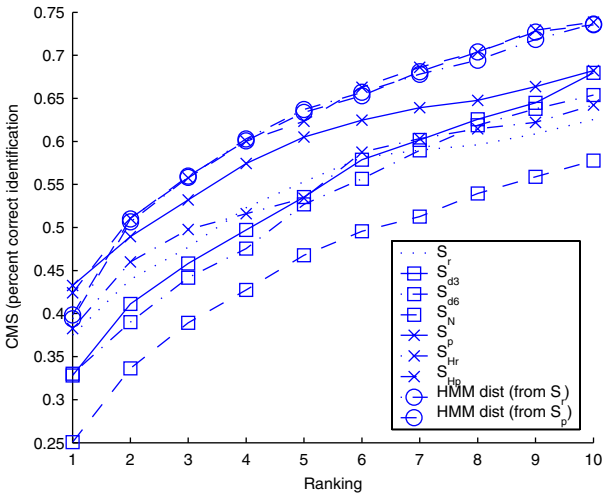
[7] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan. Image-based visual hulls. In *SIGGRAPH Computer Graphics Proceedings*, 2000.

[8] M. Oren, C. Papageorgio, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *Proc of CVPR*, 1997.

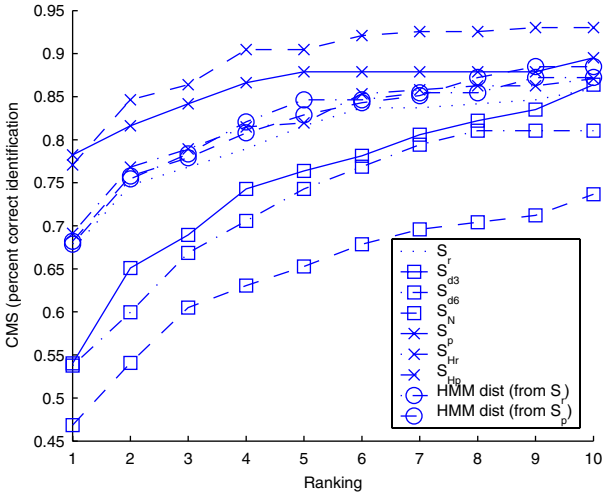
[9] J. Phillips, S. Sarkar, I. Robledo, P. Grother, and K. Bowyer. Baseline results for the challenge problem of human id using gait analysis. In *Proc of Face and Gesture*, 2002.

[10] L. Rabiner. A tutorial on hidden markov model for speech recognition. *Proc of the IEEE*, 77(2), Feb 1989.

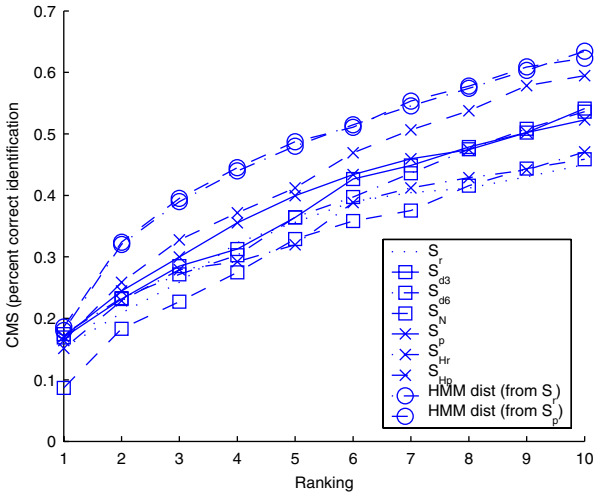
[11] S. Zhou and R. Chellappa. Probabilistic human recognition from video. In *Proc of ECCV*, pages 681–697, 2002.



(a) Average CMS for all probes



(b) Average CMS for probes A, B, C



(c) Average CMS for probes D, E, F, G

Figure 8: Comparison of recognition rate using different silhouettes using average CMS over all probes, grass probes, and concrete probes.

Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set
0-7695-1950-4/03 $17.00 © 2003 IEEE