Challenges in Segmentation of Human Forms in Outdoor Video

Zongyi Liu and Sudeep Sarkar
Computer Science and Engineering
University of South Florida; Tampa; FL 33647
{zliu4, sarkar}@csee.usf.edu

Abstract

Most gait and activity recognition algorithms rely on the use of silhouettes as the low-level representation. However, the detection of good silhouettes is still an open problem, particularly for sequences that are taken outdoors. Illumination conditions, compression artifacts, and low number of pixels on the subject, contribute to the difficulty. Presently, these issues are either ignored by using indoor data or addressed, on a case by case basis, by employing, essentially, a “bag of tricks” based approach. We argue for a more formal approach, based on generic shape and motion models to handle a variety of these problems, under the umbrella of one formalism. We present an HMM-based Eigen Stance model, built based on manually created silhouettes from 71 individuals. The population HMM helps map a frame in any given sequence to a stance and the appearance based Eigen-Stance model is used to reconstruct the computed silhouette in that frame. We quantify the performance in terms of signal based criteria of missed detection and false positive prediction rate. We also show results on three different databases.

1. Introduction

Extracting moving objects from a video sequence has been studied for a long time in computer vision. Lately, the detection and segmentation of humans in video have gained renewed interest with increased research on gait and activity recognition. Given their possible application to security and surveillance, this interest will only grow. Surprisingly, the segmentation problem is usually under-emphasized in these problems, either by using indoor sequences or using well controlled backgrounds. However, for the robust deployment of these ideas to outdoors or even indoors with challenging illumination conditions, it is essential that segmentation of human forms in video be approached in a more systematic manner.

Not surprisingly, the common segmentation strategy relies on background subtraction, followed by an amalgamation of fixes to correct for errors due to shadows, missing background, and spurious background. This strategy is decades old. In the early of the 80s, Jain et al. [7, 8] used inter-frame differencing to estimate moving objects. More recent works, using background subtraction based segmentation, range from modeling pixels as mixture Gaussian distribution [19, 5, 3, 4] to the more involved Wallflower method [18]. Another common approach to motion segmentation is based on grouping optic flow to find coherent motion [20, 17]. Such algorithms are good at detecting motions where the illumination variation is not significant and the background is well controlled. A completely different approach to motion segmentation relies on processing the spatio-temporal volume [15, 1], instead of frame by frame processing. Perhaps the most versatile system for people detection across activities is the W4 approach [6], however, it too relies on silhouettes detected by background subtraction. The subsequent part labeling and person to person segmentation steps are strongly dependent on the accuracy of the silhouette boundary. There are also model based approaches that are aimed at directly interpreting the given segmented forms. A recent example is the Bayesian framework presented in [21], which integrates human shape, human height, camera model, and image cues including human head candidates, foreground-background separation in one framework.

Typical errors encountered in silhouette segmentation are due to: (i) shadows, (ii) inability to segment parts because they fall just below the threshold and are classified as background, (iii) moving objects in the background, such as fluttering leaves and moving persons, and (iv) compression artifacts near the boundaries of the person, which are present in, medium cost, consumer grade cameras.
fill in the holes and to prevent small pieces. This smoothed distance is then used to classify pixels into foreground or background using Expectation Maximization (EM) with a Gaussian mixture model. The third column are generated by the modified version of the baseline strategy. Specifically, we replace the smoothing step with a proximity based grouping process that assembles disconnected components. The fourth column shows the silhouettes detected also by background subtraction, but using slightly different thresholding scheme [2]. The fifth column shows the silhouettes that were cleaned using average silhouettes models [10]. We see that some silhouettes look “thinner” than others but the effect of shadows and missing body portions still exist in all of them.

Various algorithms have been proposed to handle the shadows in segmentation. Most of them are based on pixel-based processing of photometric attributes, which has problems in removing strong shadows. Of course, such methods cannot handle missing body parts or extraneous background moving objects merged with the foreground. Instead of a unstructured solution, constructed out of a set of essentially “pixel-punching” tricks, we advocate the use of generic human action models to reconstruct and refine silhouettes that have been extracted by simple, bottom-up, low-level strategies. Such an approach will be able to handle a wide range of segmentation errors in a single, structured, framework. The process involves the collection of a dataset of many subjects performing different activities, construction of generic movement and shape models, and reconstruction or refinement of silhouettes, detected by background subtraction or other simple means, using these models. We demonstrate these ideas in this paper by presenting results on segmenting walking human forms.

2. Dataset

To model walking human forms we use the recently formulated HumanID Gait Challenge dataset [9]. The full dataset consists of walking sequences of 122 subjects, collected outdoors from roughly fronto-parallel viewpoints. For each person, we have sequences exercising different combinations of up to 5 different covariates: (i) 2 different shoe types, (ii) 2 different surface-types (grass and surface), (iii) 2 different carrying conditions (with or without a briefcase), (iv) 2 different viewpoints (30’ difference), and (v) 2 different times (about 6 months apart, May and Nov).

Manual silhouettes were created for a subset of Gait Challenge dataset. More details about the process and quality checks can be found in [13, 11], here we highlight some salient aspects. About 70 subjects from one of the two collection periods (May collection) were chosen for manual silhouette specification. For each subject, we chose sequences taken (i) on grass, with shoe type A, right camera view, (ii) on grass, with shoe type B, right camera view, (iii) on concrete, with shoe type A, right camera view, and (iv) on grass, with shoe type A, carrying briefcase, right camera view. We also created the silhouettes for repeat subject taken on grass but collected in November. Manual silhouettes for one condition were used to construct the generic model and the others were used to test the performance of the reconstruction process, thus separating the train and test sets.

We manually specified the silhouette in each frame over one walking cycle, of approximately 30 to 40 image frames. This cycle was chosen to begin at the right heel strike phase of the walking cycle through to the next right heel strike. We attempted to pick this gait cycle from the same 3D location in each sequence, whenever possible. We did not just mark a pixel as being from the background or subject, but provided more detailed specifications in terms of body parts too. We explicitly labeled the head, torso, left arm, right arm, left upper leg, left lower leg, right upper leg, and right lower leg using different colors. The middle row in Fig. 2 shows some examples of part-level ground truth silhouettes corresponding to the images in the top row. The bottom row shows the height scaled and centered silhouettes of the

![Figure 1. Typical silhouette errors.](image-url)
kind used by gait recognition algorithms. Quality control checks looked for miscolored parts and backgrounds, randomly colored isolated pixels, errors on the boundary of the body, and missed body parts. Some of the difficulties encountered during the creating process include low-image quality due to varying overall intensity, occlusion of feet in the grass sequences, similarity of dark skin tones of some subjects with the background, frequent occlusion of the right arm, and the presence of dark or baggy clothing, which made it hard to delineate various body parts. However, despite these difficulties we were able to create pretty consistent silhouettes across the subjects.

3. Generic Gait Model

Strictly speaking, gait refers to the dynamics of human walking. However, as it is used in computer vision, gait refers to both the shape and the dynamics of the movement. The underlying body shape and movement stance configuration contribute to overall shape in each time frame. So, a generic gait model should be able to capture both shape and dynamics. We use a population based Hidden Markov Model (HMM) to model the dynamics and an Eigen Stance gait shape model, associated with each HMM state, as the shape model.

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

3.1. Exemplar Construction

The observation in our model is the distance between a given silhouette $f$ and an exemplar set $E$, which is computed by clustering silhouettes of similar shape in a set of sequences. The distance measurement we used here between two horizontal aligned frames $f_i$ and $f_j$ is defined as:

$$D(i, j) = 1 - \frac{f_i^T f_j}{f_i^T f_i + f_j^T f_j - f_i^T f_j}$$  \tag{1}$$

which is also known as the Tanimoto distance. Based on this distance, we form 20 stance exemplar sets using K-means technique with the constraints that (i) there is at least one frame in each sequence for every exemplar, and (ii) a frame can only be re-assigned to the exemplar that its immediate neighbors. By setting these two constraints we essentially enforce a left-right cyclic Bakis model for the HMM transitions, which is consistent with the cyclical nature of gait. The K-means process is initialized by equally partitioning each sequence into 20 sections; frames from $i$-th section from all sequences forms the initial estimate for the $i$-stance exemplar set, $E_i$. Fig. 3(a) shows a sampling of the averages of the final exemplar sets, found by K-means until no frames are re-assigned, which is about 10 iterations.

3.2. Population Hidden Markov Model

A Hidden Markov Model (HMM) is specified by the possible states, $q_t \in \{1, \cdots, N_s\}$ and the triple $\lambda = (A, B, \pi)$, representing the state transition matrix, observation model, and priors, respectively. We have defined the states by forming 20 exemplars as described above. Note that we use a full gait cycle in order to pick up the asymmetry between the
two strides. And, we adopt the uniform distribution for the priors $\pi$ because a given gait sequence could start from any state.

We choose the observation model to be exponential in terms of the Tanimoto distance, $D$, between any given silhouette, $f_i$, to the mean of the state exemplars, $E_j$.

$$b_j(f_i) = \frac{1}{\mu_j} e^{-\frac{D(f_i, E_j)}{\epsilon_j}} \quad (2)$$

The observation model is thus parameterized by the mean $\mu_j$. The HMM structure is somewhat similar to that used in [16] for recognition, but in our case it is designed to model gait dynamics over a population. Differences are also there in the observation model and the state definitions; our model takes into account the gait asymmetry between the two strides in a cycle.

The entries of the state transition matrix are also initialized from the exemplars and then refined using Levinson’s method for training with multiple observation sequences based on the iterative Baum-Welch algorithm [14] until the likelihood of the observation sequences are maximized.

### 3.3. Eigen-stance Model

The goal of our eigen-stance model is to remove noise from silhouettes, while retaining the shape variations. We model the variation in each stance (state) as a multivariate Gaussian distribution, which is represented compactly using the principal component analysis (PCA). The number of eigenvectors, $N_e$, is chosen so that at least 80% of the variation is modeled. Considering the strong impact on gait for surface variation, we built different eigen-space along with the HMM model. Fig. 3(b) shows some sample of the first eigen-stances.

### 3.4. Reconstruction

Before a silhouette is reconstructed, it has to be matched to one of the states in our gait-stance model. We use the dynamic programming based Viterbi algorithm [14] to perform the silhouette decoding, which returns the most likely state assignment for the silhouettes in a given sequence. To reduce the combinatorics of this assignment process, we partition the input sequence into subsequences of roughly one gait cycle length, which is estimated from the periodic variation in the number of foreground pixels in the bottom half of the silhouettes. Note that the starting state of these subsequences need not match the starting HMM state; the cyclical nature of the HMM model can handle this.

After a frame $f_i$ is estimated to be at state $k$, it is projected into the corresponding eigen-space $\Phi(k) = \{\mu_k, e_{k,1}, \ldots, e_{k,N_e}\}$, and then reconstructed as $f^*_i$:

$$f^*_i = \mu_k + \sum_{j=1}^{N_e} (e_{k,j}^T (f_i - \mu_k)) e_{k,j} \quad (3)$$

Since the eigen-space is built on the “cleanest” manual silhouettes, the reconstruction is expected to remove most of the noise in the input silhouettes. On the other hand, the output of the reconstruction has a continuous value ranging from 0 to 1 which needed to be thresholded into binary silhouette. To reconstruct the human form by removing noise, while retaining gait and shape features for biometrics, we use a two-level thresholding strategy:

$$F^*_i(j) = \begin{cases} 
\text{Foreground} & \text{if } f^*_i(j) > T_{\text{high}} \text{ or } \mu_k(j) = 1 \\
\text{Background} & \text{if } f^*_i(j) < T_{\text{low}} \\
\text{f_i(j)} & \text{otherwise}
\end{cases} \quad (4)$$

For the experiments in this paper, $T_{\text{low}} = 0.2$ and $T_{\text{high}} = 0.8$. We later show results justifying these choices.

### 4. Evaluating the Reconstruction Quality

How does the strategy work? Can it remove unwanted background blobs, shadows, and fill in “holes”? As input, we use the raw silhouettes generated by the modified base-
line algorithm (see the third column of Fig. 1), which contain fewer noise but the artifacts still exist. Fig. 4 shows some examples of the quality of reconstruction (bottom row) for poor quality input silhouettes (top row). The cases include examples of shadow removal, hole filling, and removal of spurious background blobs from another person in the background. Fig 5 compares silhouettes before and after reconstruction with the corresponding manual silhouettes over one gait cycle. It shows that our model works well for all gait phases. Fig 6 shows results of reconstruction of silhouettes of persons carrying briefcase. Notice how the briefcases are erased in the reconstructed version. This ability would be particularly useful to identify objects that are being carried by a subject.

4.1. Pixel Level Evaluation

To measure the silhouette quality, we compute the detection rate ($P_D$) and false positive prediction rate ($PPV$), with respect to the manual silhouettes. The PPV is the ratio of incorrectly classified foreground pixel number to the total classified foreground pixel number. Note that this is not the same as false alarm rate. The computation of the false alarm rate requires the definition of the background class, which is unbounded in this case. The performance of the model is quantified by the percentage improvement in detection ($\Delta P_D = 100 \frac{P_D(\text{After}) - P_D(\text{Before})}{P_D(\text{Before})}$) and false positive prediction ($\Delta PPV = 100 \frac{PPV(\text{After}) - PPV(\text{Before})}{PPV(\text{Before})}$) with reconstruction. Fig. 7 reports the improvement for 71 subjects walking on grass field, concrete field, and carrying briefcase, where each plot is the average of the frames of a sequence for which we have manually created silhouettes. For reference, we also show the absolute $P_D$ and $PPV$ before and after reconstruction in Fig. 8 and 9, respectively. We find that the detection rate slightly dropped by about 1%. Nonetheless, the false positive prediction rate has dramatically dropped around 20%-30%, suggesting that noise has been substantially removed. As for the outlier points in the scatter plot, we found that the sequences with dramatic $\Delta PPV$ drops were for sequences taken in the evening when the shadows are longer, which were removed during reconstruction.

4.2. Parameter Variations

There are two sets of parameters that appear to be critical to the reconstruction process. The first is the choice of the number of eigenvectors to keep in the Eigen-stance model. Fig. 10 plots the variation in median pixel detection rate over all persons with variation in the eigenspace energy preserved. We see that the change in the detection rate stabilizes at around 80%, which is about 30 eigenvectors. And, the second choice involves the double thresholds in Eq. 4. Fig. 11 shows the variation in the pixel-level detection and false positive prediction rates for three different choices of the double thresholds. We see that the choice of $T_{low} = 0.2$ and $T_{high} = 0.8$ offers a good compromise between high detection rate and low false positive prediction rate.

4.3. Viewpoint Variation Robustness

The generic gait model is a view-based model. It is reasonable to ask how effective is the reconstruction for sequences taken from a viewpoint different from the model viewpoint. The Gait Challenge dataset includes frames of the same gait event viewed from two different angles, with the verging angle of roughly 30°. The manual silhouettes, which were used to construct the Eigen-Stance model, only exist for the sequences viewed from the right camera. Fig. 12 shows the reconstruction results over 3 sequences collected from the left camera. As we have seen before, the model is able to successfully remove most shadows and other background noise artifacts.
Figure 8. Scatter plot of pixel level detection rate and false positive prediction rate before reconstruction.

Figure 10. The plot of the median, pixel level, detection rate for silhouettes on grass and concrete with variation in energy preserved in the Eigen-stance model.

Figure 9. Scatter plot of pixel level detection rate and false positive prediction rate after reconstruction.

Figure 11. Scatter plot of pixel level detection rate and false positive prediction rate for different values of the double thresholds ($T_{low}, T_{high}$).
4.4. Generalizability to Different Datasets

Although, our test and train sequences were separate, our results, presented so far, has been on one dataset. We also test the model with other gait database to evaluate its generalizability. The first tested dataset is the Georgia Tech. outdoor dataset available at http://www.cc.gatech.edu/cpl/projects/hid/. It consists of subjects walking on concrete surface. We use the modified baseline algorithm to generate silhouettes, which are then reconstructed by the developed model. Fig. 13 shows the original and reconstructed silhouettes over one gait cycle for one subject. We found that the qualities of the original silhouettes are poor due to strong illumination and low contrast. However, the qualities of the reconstructed silhouettes are significantly improved.

The second tested dataset is the CMU Mobo indoor dataset. It consists of subjects walking on a treadmill under different conditions, which are specified in http://www.hid.ri.cmu.edu/HidEval/. We used the sequences taken from side view with slow walking holding a ball. Fig 14 shows the result of reconstruction over these silhouettes generated by CMU HumanID group. We see that most body holes are “filled”, and ball portions are erased. Moreover, the model rebuilds silhouettes severely corrupted by the noise, e.g. see the third column.

5 Conclusions

We demonstrated how a generic model based silhouette reconstruction technique can effectively enhance segmentation of walking human forms in video, by overcoming a variety of errors arising due to shadows, missing pixels, spurious background blobs, and even objects that are being carried. The dynamic aspects were modeled by an HMM...
model, whereas the shape aspects were modeled by eigen-
stance shape models.

We also tested its robustness with the data collected un-
der different conditions on which it is trained, specifically,
data collected with viewpoint difference about 30° and sub-
jects carrying briefcase or a ball. In addition, we also pro-
cessed the silhouettes of Georgia Tech. outdoor dataset and
CMU Mobo indoor dataset. All these experiments suggest
that our model is very effective and robust in removing
noises and building “better” silhouettes of walking humans.

The impact of the computed silhouettes on the gait
recognition problem can be found in [12].

Future directions would include, learning probabilistic
models to simultaneously label parts of a silhouette and de-
dect it. The present manual silhouette database supports
such as study for walking human forms. It explicitly la-
ables the head, torso, left arm, right arm, left upper leg, left
lower leg, right upper leg, and right lower leg using dif-
ferent colors. Other directions could involve the extension
to other human movements, both periodic and non-periodic
ones. Databases for such activities are starting to crop up,
but mostly for indoor scenarios. In addition, views other
than frontal parallel ones could be added to the model to
arrive at a view-independent reconstruction models.

6. Acknowledgment

This research was supported by funds from DARPA
(AFOSR-F49620-00-1-00388), NSF (EIA-0130768), and
the USF College of Engineering REU program. Laura
Malave, Adebola Osuntogun, Preksha Sudhakar, and Chris-
tine Bexley helped to meticulously put together the manual
silhouette database. Thanks also the CMU and MIT gait
research groups for making their silhouettes available.

References

structures exploiting spatio-temporal image slices. Proc. of the IEEE International Conference on Image Processing,
identification from body shape and gait. In International
Conference on Automatic Face and Gesture Recognition,
model for background subtraction. 6th European Conference
on Uncertainty in Artificial Intelligence, page 175181, Au-
gust 1997.
adaptive tracking to classify and monitor activities in a site.
IEEE Conference on Computer Vision and Pattern Recogni-
surveillance of people and their activities. IEEE Trans. Pat-
tive difference pictures from image sequences of real world
[8] S. N. Jayaramamurthy and R. Jain. An approach to the seg-
mmentation of textured dynamic scences. Computer Vision,
K. Bowyer. Baseline results for the challenge problem of
Human ID using gait analysis. In International Conference
on Automatic Face and Gesture Recognition, pages 137–
142, 2002.
for silhouette refinement. In International Conference on
Toward understanding the limits of gait recognition. In SPIE
Proceedings of Defense and Security Symposium: Biometric
[12] Z. Liu, L. Malave, and S. Sarkar. Studies on silhouette qual-
ity and gait recognition. In Computer Vision and Pattern
Recognition, 2004.
resource and limits. Master’s thesis, University of South
[14] L. Rabiner and B. H. Juang. Fundamental of Speech Recog-
sation based computational model for robust segmentation
of moving objects. Computer Vision and Image Understand-
A hidden markov model based framework for recognition
of humans from gait sequences. In IEEE International Confer-
Wallflower: Principles and practice of background mainta-
nence. Proc. Internal Conference on Computer Vision, pages
[20] Y. Yacob and M. J. Black. Parameterized modeling and
recognition of activities. Computer Vision Image Under-
[21] T. Zhao and R. Nevatia. Bayesian human segmentation in
crowded situations. Computer Vision and Pattern Recogni-