People Tracking and Segmentation Using Efficient Shape Sequences Matching

Junqiu Wang, Yasushi Yagi, and Yasushi Makihara

The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, Japan
jerywangjq@gmail.com

Abstract. We design an effective shape prior embedded human silhouettes extraction algorithm. Human silhouette extraction is found challenging because of articulated structures, pose variations, and background clutters. Many segmentation algorithms, including the Min-Cut algorithm, meet difficulties in human silhouette extraction. We aim at improving the performance of the Min-Cut algorithm by embedding shape prior knowledge. Unfortunately, seeking shape priors automatically is not trivial especially for human silhouettes. In this work, we present a shape sequence matching method that searches for the best path in spatial-temporal domain. The path contains shape priors of human silhouettes that can improve the segmentation. Matching shape sequences in spatial-temporal domain is advantageous over finding shape priors by matching shape templates with a single likelihood frame because errors can be avoided by searching for the global optimization in the domain. However, the matching in spatial-temporal domain is computationally intensive, which makes many shape matching methods impractical. We propose a novel shape matching approach that has low computational complexity independent of the number of shape templates. In addition, we investigate on how to make use of shape priors in a more adequate way. Embedding shape priors into the Min-Cut algorithm based on distances from shape templates is lacking because Euclidean distances cannot represent shape knowledge in a fully appropriate way. We embed distance and orientation information of shape priors simultaneously into the Min-Cut algorithm. Experimental results demonstrate that our algorithm is efficient and practical. Compared with previous works, our silhouettes extraction system produces better segmentation results.

1 Introduction

Shape matching has been found useful in object recognition. In specific, shape matching based on silhouette information has been proved effective in human gait recognition. Gait recognition overcomes a few problems in an elegant manner that other people identification methods find difficult to handle. For example, a gait recognition system can identify a person from a distance. It is possible to recognize persons using silhouettes extracted from low-resolution images.

Reliable and accurate silhouettes are crucial for gait recognition. A gait recognition system tends to have poor performance when extracted silhouettes deviate from the real shapes in image sequences. Most gait recognition algorithms assume that silhouette information has been extracted precisely. However, silhouette extraction is a very challenging task especially when image sequences are captured by a moving camera, or the background contains clutters. In fact, silhouette extraction is not only important for gait recognition, but also can be used in human pose analysis and other applications. Human tracking and segmentation are challenging because of articulated structures, pose variations, and background clutters. Although some human tracking algorithm can provide foreground likelihood images [1], it is too difficult to calculate precise human silhouettes based on these likelihood images using simple image morphing techniques. As other segmentation methods, the Min-Cut algorithm also meets difficulties in human silhouette extraction. Markov Random Fields, which are the foundation of the Min-Cut algorithm, seldom present realistic shape priors. Therefore, the Min-Cut algorithm gives poor performance in human silhouette extraction, especially in cluttered backgrounds.

Shape priors play an important role in improving the performance of the Min-Cut algorithm. We develop a silhouette extraction algorithm based on the standard Min-Cut algorithm. Although shape priors have been incorporated in the Min-Cut algorithm in previous works [2], it is not trivial to compute shape priors automatically especially for human silhouettes. The likelihood images given by tracking algorithms are helpful in computing shape priors. Unfortunately, these likelihood images contain many errors. Matching a single likelihood image with a set of silhouettes templates is not reliable due to these errors. Matching shape sequences in spatial-temporal domain is advantageous over finding shape priors by matching shape templates with a single likelihood frame because errors can be avoided by searching for the global optimization in the domain. However, the matching in spatial-temporal domain is computationally intensive, which makes many shape matching methods impractical. We propose a novel shape matching approach that has low computational complexity independent of the number of shape templates.

Incorporating shape prior knowledge alleviates the problems in silhouette extraction. The Min-Cut algorithm allows for a straightforward incorporation of prior knowledge into its formulation. An important problem in employing shape priors is how to apply shape prior knowledge in an appropriate manner. Embedding shape priors into the Min-Cut algorithm based on distances [2] from shape templates is lacking because Euclidean distances cannot represent shape knowledge in a fully appropriate way. We embed distance and orientation information of shape priors simultaneously into the Min-Cut algorithm.

The rest of the paper is arranged as follows. Following the literature review, We describe a novel shape matching method and its application in optimal path searching in Section 3. We incorporate shape priors into the Min-Cut algorithm in Section 4. Both distance and orientation information of shape priors are embedded within the Min-Cut algorithm. Experimental results for image sequences are presented in Section 5. Section 6 concludes this work.

2 Previous Work

Human silhouette extraction is found challenging because of articulated structures, pose variations, and background clutters. Segmentation methods based solely on low-level information often provide poor performance in these difficult scenarios. Many segmentation algorithms meet difficulties in human silhouette extraction. The Min-Cut algorithm [3], which has achieved great success in interactive segmentation, faces problem in silhouette extraction.

The evident power of shape priors as an additional cue has been noticed by many researchers. Freedman and Zhang [2] define the coherence part of the Min-Cut algorithm by considering the shape distance transform results. In their work, shape priors are given manually, which is tedious for segmentation in video sequence. It is desirable to computer shape priors automatically for human silhouette extraction. Wang *et al.* [1] proposed a shape prior seeking algorithm by searching for the best path in spatio-temporal domain. The major drawback of their work is the heavy computational costs in shape matching, which makes their algorithm not practical for real applications. Although there is an effort in accelerating the shape matching process [4], the performance is still not efficient especially when the number of shape templates is large.

While pedestrian model representations have been employed for refining silhouettes in previous works [5, 6], they all assume that an foreground likelihood images can be obtained by background subtraction. In addition, these works do not address the shape matching problem, which is crucial for the applicability of silhouette extraction.

In visual tracking literature, temporal coherency was employed in particle filtering. Rathi *et al.* [7] formulated a particle filtering algorithm in a geometric active contour framework in which temporal coherency and curve topology are handled. In addition, shape and appearance information were considered in a unified metric framework by Toyama and Blake [8]. The use of exemplars alleviates the difficulty of constructing complex motion and shape models. Although these algorithms do improve the performance of tracking, few of them deal with silhouette extraction.

3 Computing Shape Priors

We adopt an adaptive mean-shift tracking approach [9, 10]. The adaptive tracker provides bounding boxes and Foreground Likelihood Images (FLI). We match FLI sequence with silhouette templates in the standard gait models. A Standard Gait Model (SGM) is constructed for the matching. Tanimoto distance [11] is taken as the similarity measure between FLIs and silhouette templates. Matching shape sequences in spatial-temporal domain [1] is advantageous over finding shape priors by matching shape templates with a single likelihood frame because errors can be avoided by searching for the global optimization in the domain. However, the matching in spatial-temporal domain is computationally intensive, which makes many shape matching methods impractical. We will introduce an efficient shape matching approach that has low computational complexity independent of the number of shape templates.

3.1 Matching Measure

FLIs generated by the tracker should be normalized to have the same size as the silhouette templates. An FLI in the nth frame is denoted by $\boldsymbol{f}(n)$. The center and the height of a human region's bounding box are denoted by (c_x, c_y) and h, respectively. Registration and scaling based on the bounding box are processed in the same way as the SGM, thus producing the normalized FLI $\boldsymbol{f}_N(n; c_x, c_y, h)$ in the nth frame.

Tanimoto distance [11] is exploited as the measure between the FLI $\boldsymbol{f}_N$ and SGM $\boldsymbol{g}$:

$$D_T(\boldsymbol{f}_N, \boldsymbol{g}) = 1 - \frac{\sum_{(x,y)} \min\{f_N(x,y), g(x,y)\}}{\sum_{(x,y)} \max\{f_N(x,y), g(x,y)\}}, \tag{1}$$

where $f_N(x,y)$ and $g(x,y)$ are the likelihood and silhouette values, respectively, at (x,y) . The Tanimoto distance between an FLI and SGM is 1 if they have identical shapes, and 0 if there is no overlap between them.

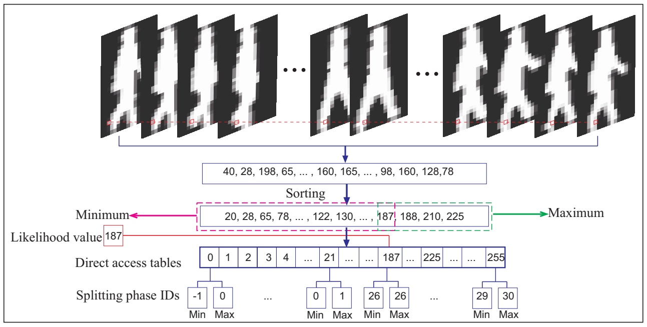

Fig. 1. The efficient Tanimoto distance method.

3.2 $O(1)$ Tanimoto Distance Computation

The matching between foreground likelihood images and silhouette templates using Tanimoto distance is computationally expensive because every pixel in every image has to be calculated individually. It takes more than 120 seconds when 30 silhouette templates are employed in the sequence matching. (The algorithm is run on a 1.6GHz laptop). This expensive Tanimoto distance computation makes the proposed approach impractical. The problem is exacerbated if we wish to include more shape variations by adding silhouette template images.

We present a novel distance calculation method whose matching complexity is $O(1)$. The calculation of the minimum and maximum in Eq. 1 is expensive if they are calculated individually. We noticed, however, that the silhouette template and likelihood images can be quantized in a limited range ($[0, 255]$ in this work) without any negative effect on the matching results. Because the silhouette template images given in the initialization have been aligned, we sort the values at each position in the template images. The phase information of these templates is kept in the sorted results. Based on the sorted results, a direct access table is built for each position of the silhouette template images. Each table has two splitting phase IDs corresponding to the calculation of the maximum and minimum in Eq. 1.

As shown in Figure 1, to calculate the minimum and maximum in Eq. 1, we do not need to compare the input likelihood value with all the values in the silhouette template images. To make the illustration clear, we use 30 phases here. In the initialization, we sort the values at each position in the template images. Then we build one direct access table for each value in the range [0, 255]. Each table contains two splitting phase IDs: the maximum phase ID p^S_{max} and the minimum phase ID p^S_{min}. For a given input likelihood value, the direct access table is found directly. Then the maximums and minimums can be assigned based on the splitting phase IDs corresponding to the input likelihood value. The value in a phase in the sorted results is assigned as the minimum, when its phase ID is equal to or smaller than the minimum splitting phase ID p^S_{min}, or as the maximum, when its phase ID is equal to or greater than the maximum splitting phase ID. For instance, if the input likelihood value is 0, all the values in the template images are assigned as maximums and the input likelihood value is always assigned as the minimum. The computational complexity of the matching process is $O(1)$. In other words, the matching is independent to the number of silhouette template images. This method is particularly important when many templates are necessary to cover large variations in shape. The template values are sorted only once during the initialization.

Tanimoto distance measures the overlapping regions of two input images. Its computation time is further reduced by reusing the calculated overlapping regions [12]. Tanimoto distance can be formulated as $D_T(\boldsymbol{f}_N, \boldsymbol{g}) = \frac{G+F-C}{C}$, where $F = \sum_{(x,y)} f_N(x, y)$, $G = \sum_{(x,y)} g(x, y)$, and $C(f_N, g) = \sum_{(x,y)} \min\{f_N(x, y), g(x, y)\}$. Based on this formulation, G (sum of gait template values) is calculated only once during the initialization. For each input foreground likelihood image sequence, F is also calculated only once. $C(f_N, g)$ is calculated based on the efficient method.

Using the method described above, it takes around 0.6 seconds to calculate distances between an input foreground likelihood sequence and all templates (including shifting and scaling) on the 1.6GHz laptop. Further computational cost reduction is expected when the number of shape templates becomes larger. In addition, the proposed distance calculation method can be used in other applications where silhouette template matching is necessary [12].

4 Embedding Shape Priors in Min-Cut Segmentation

Let $\mathcal{L} = \{1 \dots K\}$ be a set of labels. Let $G = (\mathcal{V}, \mathcal{E})$ be a graph with $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Segmentation is formulated in terms of energy minimization in the Min-Cut algorithm [13]. We embed shape priors into the algorithm.

4.1 Embedding Shape Priors

Shape priors can be embedded in the Min-Cut algorithm by adding an energy term [2] [14]:

$$E(A) = E_{smoothness}(A) + E_{data}(A) + E_{shape}(A). \tag{2}$$

The Min-Cut algorithm with shape priors includes shape fitness, smoothness and initial labeling. The energy function E_{shape} is penalized if the segmented contour deviates from the boundary of the silhouette. Shape priors are represented by a distance transform result.

We found that the method in [2] is deficient: the embedded shape priors need to be very accurate, otherwise the distance transform can misguide the segmentation. In contrast, we introduce orientation information in the shape priors to encourage smoothness of the segmentation. It has been found that the statistics of steered filters for human limbs are different from those of other natural scenes [15]. In this work, we learn a vocabulary that includes position and gradient orientations in human silhouettes. We calculate gradient orientations and normalized positions (in $[0, 1]$) in 400 segmented people silhouette images and detect edges in the silhouette templates using Canny edge detector. Then we calculate gradient orientations and normalized coordinates on the edges. We apply K-Means to form an initial vocabulary. An EM algorithm is adopted to get the final vocabulary. We compute the mean and covariance matrix for each word. The vocabulary has 10 words finally, allowing edges belonging to a same word formulate as an oriented template. Thus 10 oriented templates are gotten for every template image. Then Euclidean distance transform is applied to these oriented templates.

To improve the first term in Eq. 2, based on the distance transform results of the oriented templates, we calculate the minimum distance in corresponding to a pixel in the input image. If the minimum distance is greater than the threshold d^{DT}_{min}, the pixel is set as background. This method is effective in dealing with inaccurate shape priors. In contrast, the shape priors used in $[2, 1, 14]$ have to be very accurate, otherwise they can misguide the segmentation.

We also found that probabilities decrease too quickly near a contour. We decrease the distance values obtained from the distance transform by applying a local search. We then extract edges in the input images. The distance is kept as is if there are edges near the shape prior. Otherwise the distance values are multiplied by a constant factor c_{edge}. (The factor is set to 0.8 in this work). The cost function of shape priors is well described in the transformed image. The shape prior energy is written as $E_{shape} = \sum_{(pq) \in \mathcal{N} : A_p \neq A_q} \frac{\psi_{min}(p) + \psi_m in(q)}{2}$, where ψ_{min} is the minimum distance on the transformed image.

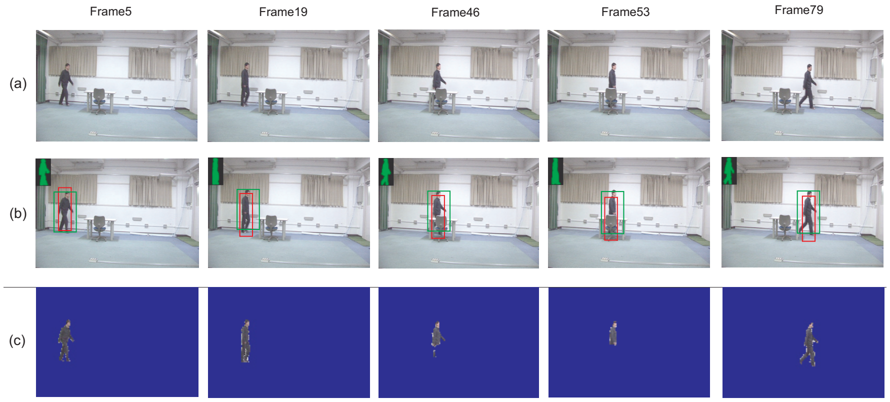

Fig. 2. Tracking and segmentation results for the indoor sequence. (a) Input images. (b) Initial bounding boxes (in red) generated by the tracker, optimal bounding boxes (in green) and gait models (phase) obtained using optimal path searching. (c) Segmentation results by embedding the shape priors in the Min-Cut algorithm.

5 Experimental Results

We tested the proposed algorithm on 8 sequences with tracking and segmentation ground truths. The size of all images in these sequences is 360 × 240 pixels. We show the results for two sequences in detail. Performance is evaluated with respect to refinement of bounding boxes and phase estimation and the improvement in segmentation. The Quantitative evaluation of other sequences is given in 5.3.

5.1 Refinement of Bounding Boxes and Phase Estimation

The results for the two sequences are shown in Figures 2 and 3, respectively. The initial bounding boxes produced by the tracker are not well aligned with the people regions and the initial foreground likelihoods are low for some parts.

Based on the optimal path searching results, the tracking bounding boxes are shifted to better positions. The bounding boxes are not accurately aligned with the person. The vertical centers in the initial bounding boxes deviate from their correct positions. The positions are adjusted downwards based on the optimal path searching results. The horizontal centers of the initial bounding boxes are relatively more accurate. They need to be shifted less frequently than the vertical centers.

The selected silhouette templates provided by the searching results are shown in Figure 2(b). The gait phases corresponding to the walking person are correct. The shape priors are incorporated in the Min-Cut algorithm giving the segmentation results shown in Figure 2(c).

Next we evaluate the smoothness of the walking phase transfer in Figure 4. The phases estimated with and without using shape sequence matching are com-

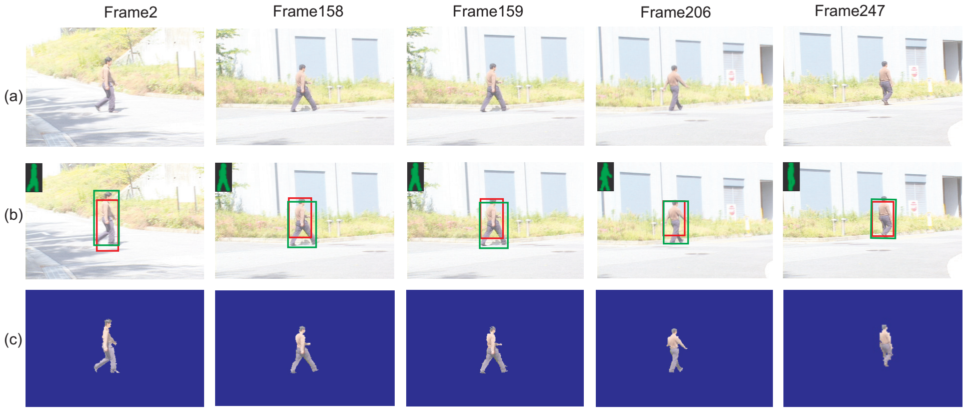

Fig. 3. Tracking and segmentation results for the outdoor sequence. (a) Input images. (b) Initial bounding boxes (in red) generated by the tracker, optimal bounding boxes (in green) and gait models (phase) calculated by the optimal path searching using shape sequence matching. (c) Segmentation results by embedding the shape priors in the Min-Cut algorithm.

pared. The estimation results without using shape sequence mathcing are obtained by matching an input likelihood image with all the silhouette templates. The phases estimated using shape sequence matching are much more accurate than those without shape sequence matching. This demonstrates the importance of shape sequence matching in optimal path searching. The phase estimation also verifies the necessity of searching in a spatiotemporal space instead of in a single frame. When the view changes, the phase estimation result is not as accurate as the side view. However, it is still much better than the estimation obtained from single image matching.

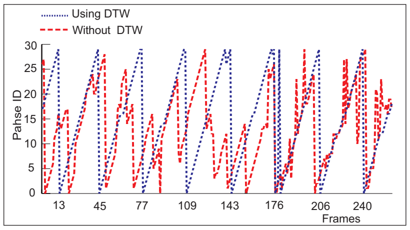

Fig. 4. Phase transition estimation for the sequence in Figure 3.

5.2 Segmentation Results

The segmentation ground truths of these sequences are obtained by labeling the images manually. Each pixel is labeled as background, foreground, or ambiguous. The ambiguous label is used to mark mixed pixels along the boundaries between foreground and background. We measure the error rate as a percentage of missegmented pixels, ignoring ambiguous pixels.

Test seqeunces	3	4	5
No Prior	0.18	0.30	0.32
Using Prior	0.092	0.24	0.23
Test seqeunces	6	7	8
No Prior	0.23	0.36	0.4
Using Prior	0.15	0.22	0.31

Table 1. Segmentation errors for eight of the test sequences.

5.3 Quantitative evaluation

Segmentation results using shape priors are shown for the indoor sequence (Figure 2) and the outdoor sequence (Figure 3). A quantitative evaluation of the segmentation results with and without shape priors is shown in Figure **??**. The segmentation results with shape priors embedded are compared with those without shape priors. The incorporation of shape priors improves the performance of the segmentation. Compared with the indoor sequence, the use of shape priors is more helpful in the outdoor sequence. Thus shape priors play a more important role in the challenging outdoor sequence.

Table 1 shows segmentation errors with respect to ground truth for six of our test sequences. Among them, sequence 3, 4, and 5 are indoor sequences, and 6, 7, 8 are outdoor sequences. The segmentation using shape priors has lower error rate in these sequences.

6 Conclusions

We find optimal paths for an input likelihood sequence by matching it with silhouette templates. The novel efficient shape matching method makes the proposed approach practical for real applications. The shape sequence matching provides shape priors for silhouette extraction. The proposed prior embedding method is effective. The segmentation performance is also improved based on shape constraints.

References

1. Wang, J., Makihara, Y., Yagi, Y.: Human tracking and segmentation supported by silhouette-based gait recognition. In: Proc. of IEEE Int. Conf. on Robotics and Automation. (2008)
2. Freedman, D., Zhang, T.: Interactive graph cut based segmentation with shape priors. In: Proc. of CVPR. (2004) 755–762
3. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: Proc. of ICCV. (2001) 105–112
4. Wang, J., Makihara, Y., Yagi, Y.: People tracking and segmentation using spatiotemporal shape constraints. In: Proc. of 1st ACM International Workshop on Vision Networks for Behavior Analysis, in conjunction with ACM Multimedia. (2008)
5. Baumberg, A., Hogg, D.: Learning flexible models from image sequences. In: Proc. of ECCV. (1994) 299–308
6. Lee, L., Dalley, G., Tieu, K.: Learning pedestrian models for silhouette refinement. In: Proc. of ICCV. (2003) 663–670
7. Rathi, Y., Vaswani, N., Tannenbaum, A., Yezzi, A.: Particle filtering for geometric active contours with application to tracking moving and deforming objects. In: Proc. of CVPR. (2005) 2–9
8. Toyama, K., Blake, A.: Probabilistic tracking in a metric space. In: Proc. of ICCV. (2001) 50–57
9. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(5) (2003) 564–577
10. Wang, J., Yagi, Y.: Integrating color and shape-texture features for adaptive realtime tracking. IEEE Trans. on Image Processing **17(2)** (2008) 235–240
11. Jr., K.S., Tanimoto, S.: Progressive refinement of raster images. IEEE Trans. on Computers **28**(11) (1979) 871–874
12. Marszalek, M., Schmid, C.: Accurate object localization with shape masks. In: Proc. of CVPR. (2007) 1–8
13. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans. Pattern Anal. Mach. Intell. **26**(9) (2004) 1124–1137
14. Bray, M., Kohli, P., Torr, P.H.S.: Posecut: simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In: Proc. of ECCV. (2006) 642–655
15. Sidenbladh, H., Black, M.J.: Learning the statistics of people in images and video. Int'l Journal of Computer Vision **54**(3) (2003) 183–209