People Tracking and Segmentation Using Spatiotemporal Shape Constraints

Junqiu Wang, Yasushi Makihara and Yasushi Yagi
The Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki
Osaka, 567-0047 Japan
jerywangjq@gmail.com; {makihara,yagi}@am.sanken.osaka-u.ac.jp

ABSTRACT

We present an efficient people tracking and segmentation algorithm for gait recognition. Even though most existing gait recognition algorithms assume that people have been tracked and that silhouettes are available for gait classification, tracking and segmentation are very difficult especially for articulated objects such as human beings. We improve the performance of tracking and segmentation based on spatiotemporal shape constraints. First of all, we track people using an adaptive mean-shift tracker which produces initial results consisting of bounding boxes and foreground likelihood images. The initial results, generally speaking, are not accurate enough to be applied in gait recognition directly. We refine the results by matching with silhouette templates sequences in a batch mode to find the optimal silhouette-based gait paths corresponding to the input. Since the process is computationally expensive, we propose a novel efficient distance computation method to accelerate the spatiotemporal silhouette matching. The spatiotemporal shape priors are embedded into the Min-Cut algorithm to segment people out. Experiments on indoor and outdoor sequences demonstrate the effectiveness of the proposed approach.

Categories and Subject Descriptors

I.5 [**Pattern recognition**]: Surveillance; I.5.4 [**Applications**]: Computer vision—*tracking,segmentation*

General Terms

Algorithms

Keywords

People tracking, people segmentation, Spatiotemporal shape priors

1. INTRODUCTION

People tracking and segmentation in natural environments are prerequisites for gait recognition, an effective approach to identify individuals at a distance from a camera. Spatiotemporal silhouette information has been widely used in gait recognition [16, 25, 27] because of its invariancy to illumination, background changes, and human clothing. Most existing gait recognition algorithms assume that people have been tracked and that silhouettes have been extracted successfully. However, people tracking and segmentation are very difficult due to occlusions, illumination changes, and the large variability in the shape and articulation of human body. This problem is further complicated by the quality of surveillance videos and the size of people within the frames.

This paper aims to improve people tracking and segmentation performance by incorporating spatiotemporal shape priors. Silhouettes are formed in image sequences when a person is performing certain activity or gesture. The shape deformation over time of such silhouettes depends on the activity performed. The deformation is under certain constraints that result from the physical body properties and the temporal continuities. These constraints can be used as priors for people tracking and segmentation.

We adopt a coarse-to-fine strategy to deal with tracking and segmentation. First, an adaptive mean-shift tracker [9, 26] is applied to provide preliminary tracking results including bounding boxes and foreground likelihood images. Since the results are not accurate enough for gait recognition, the Foreground Likelihood Images (FLI) are matched with Standard Gait Models (SGM) that have been built based on well segmentation silhouette sequences. The optimal path corresponding to the input is to be found in a 5-D space where position, scale, and gait phase are optimized. The spatiotemporal shape constraints are embedded into the Min-Cut algorithm [3] to improve people segmentation. The spatiotemporal shape priors are advantageous over single shape priors since monocular people shape information in one frame is inherently ambiguous due to Necker reversal [1]. In addition, matching one FLI with silhouettes can lead to problems because of the poor quality of the FLI. Such problems happen less frequently in the proposed matching method because we find the optimal path for all input FLIs. The ambiguity can be resolved by considering the spatiotemporal continuities.

Matching between sequences is computationally expensive especially when the optimal path has to be found in the 5-D space. We propose an efficient sequence matching method that accelerates the computation. The values at each position in the templates of the standard gait models are sorted during the initialization. Consequently, it is not necessary to compare the values of foreground likelihood images with the values of the templates directly. The computation is trans-

VNBA'08, October 31, 2008, Vancouver, British Columbia, Canada.
Copyright 2008 ACM 978-1-60558-313-6/08/10 ...$5.00.

formed into a binary search process. The computational complexity is reduced from $O(n)$ to $O(\log n)$. It is of great importance when large standard gait models are employed to represent wide varieties of activities. This method can be also used in other applications when silhouette matching is necessary(i.e., [15]).

The proposed approach matches shape sequences in a batch mode. Batch mode should be avoided in many cases for the sake of efficiency and effectiveness. However, the batch mode we apply is well fitted into gait recognition framework since shape information in a sequence is necessary to perform gait classification. While the computational cost is another concern of the matching in a batch mode, it is partially solved by the proposed efficient silhouette matching method.

Following a literature review, Section 2 briefly introduces the adaptive mean-shift tracker. Section 3 describes the optimal path searching using Dynamic Time Warping (DTW). Section 4 introduces the segmentation in which shape priors are embedded into the Min-Cut algorithm. Experimental results on real image sequences are demonstrated in Section 5. Section 6 concludes this work.

1.1 Related Work

Tracking and segmentation are, in essence, state estimation and image labeling based on observations and prior knowledge. Observations such as edges or photometric information are susceptible to noise and occlusions. It should help to introduce high-level knowledge such as shape or dynamics priors into tracking and segmentation. Filtering techniques have been used in people tracking [19] because shape or dynamics models can be added into a probabilistic framework. However, many such tracking algorithms require complex models defined for the object to be tracked. Toyama and Blake [23] proposed an exemplar-based probabilistic tracking algorithm. The use of exemplars alleviates the difficulty of constructing complex motion and shape models. However, their algorithm cannot deal with people segmentation.

Filtering based algorithms also suffer from the high dimensionality of human pose state space. It has been demonstrated that the space of possible human motions can be reduced into a lower dimensional space using dimensionality reduction algorithms [10]. Li *et al.* [14] proposed a coordinated mixture of factor analyzers for bidirectional mapping between the original body pose space and the low-dimensional space. Urtasun *et al.* [24] presented an impressive people tracking algorithm based Gaussian Process Dynamical Models (GPDM). Precise 3D motion data is necessary for the learning of GPDM. None of the above works handles people tracking and segmentation interactively.

Bray et al. [6] showed that segmentation and pose estimation can be integrated in a Bayesian framework simultaneously considering photometric and prior information. They used rough pose specific shape prior to improve segmentation results, which bears certain similarity to our work. However, the integration is carried out in a single frame. As we have mentioned, silhouette matching in one frame may generate ambiguous results. In contrast, our approach tries to find an optimal path by shape sequence matching that resolves the ambiguity in spatiotemporal context. Brox et. al [5] proposed another integration framework for segmentation and pose estimation where level sets are employed to do segmentation. They demonstrated the effectiveness of the integration on rigid objects but not articulated objects such as people. Rathi et al. [18] formulated a particle filtering algorithm in the geometric active contour framework in which temporal coherency and curve topology are handled. Although their approach can track moving and deformable objects under partial occlusions, it cannot deal with highly deformable objects [18].

Bilayer video segmentation for video-chat sequences has been intensively investigated based on Conditional Random Fields (CRF) [9, 12]. Segmentation for binocular stereo [12] and monocular video [9, 28] have achieved impressive results. People in video-chat sequences usually have few articulated actions, which alleviates the difficulty of the segmentation. In addition, these methods deal with static backgrounds with limited variations.

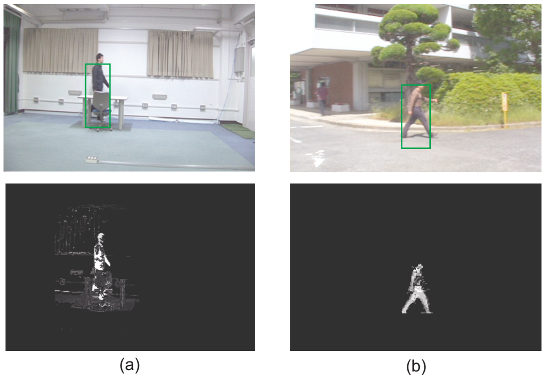


Figure 1: The tracking results of two frames from an indoor and an outdoor sequences. The images in the first row are the input images, and the bounding boxes computed by the tracker is overlaid on them; the images in the second row are the foreground likelihood images.

2. ADAPTIVE MEAN-SHIFT TRACKING

We adopt an adaptive mean-shift tracking approach [9, 13, 26]. The mean-shift algorithm and its variations have achieved success in object tracking due to its efficiency, simplicity, and robustness. The mean-shift algorithms find local maxima of a similarity measure between the color histograms (or kernel density estimations) of the model and the candidates in the image. Since fixed color features are not always discriminative enough, the basic mean-shift algorithm [8] has been extended to an adaptive tracker in which discriminative features are selected from multi-cue [7, 26].

The adaptive tracker provides bounding boxes and generates FLIs by back-projecting likelihood ratios into each pixel in the image [21].

Fig. 1 shows the tracking results of two frames in indoor and outdoor environments. It is clear that the bounding boxes computed by the tracker are not well aligned with the person in the images (Fig. 1(a),(b)). There are occlusions in Fig. 1(a). The foreground likelihood image contains many errors. Such errors are unavoidable in tracking due to the variations of the foreground or background. The optimal path searching described in next subsection improves the alignment using the imperfect bounding boxes and FLIs.

3. OPTIMAL PATH SEARCHING IN DTW

We match FLI sequence with silhouette templates in the standard gait models. The key is that the matching between the sequences solves the problems caused by imperfect tracking results. We can find the optimal path for the input FLI sequence because the sequence matching takes gait smoothness constraints into consideration. In contrast, the matching between one FLI and silhouette templates can be violated by the errors in tracking results.

A Standard Gait Model (SGM) is constructed for the matching. Tanimoto distance [22] is taken as the similarity measure between FLIs and silhouette templates. The computation of Tanimoto distances is very expensive. An efficient distance computation method is proposed to deal with this problem.

3.1 Standard Gait Model

We need high quality silhouette templates for the modeling of the SGM. We do temperature-based background subtraction in video sequences captured by an infrared-ray camera. The extracted silhouette templates are normalized by scaling and registration to produce SGM with the predefined size (The height and the width of SGM are denoted by h_g and w_g, respectively). The silhouettes are scaled so that each height is h_g and the aspect ratio is maintained. They are also registered to make the centers of these silhouette region corresponding to the SGM image center (c_{gx}, c_{gy}).

After the registration, the gait period N_{gait} is detected by maximizing autocorrelation of the normalized silhouette sequence for the temporal axis and the standard gait model $\vec{g}(\phi)$ is obtained as an averaged silhouette for each gait phase ϕ:

$$\vec{g}(\phi) = \frac{1}{N_P}\sum_{i=1}^{N_P}\vec{h}(iN_{gait}+\phi), \tag{1}$$

where $\vec{g}(\phi)$ is the SGM for phase ϕ, $\vec{h}(n)$ is the normalized silhouette image at nth frame, and N_P is the number of gait periods in the training sequence.

3.2 Matching Measure

FLIs generated by the tracker should be normalized to have the same size as the silhouette templates. An FLI at the nth frame is denoted by $\vec{f}(n)$. 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 and the normalized FLI $\vec{f_N}(n; c_x, c_y, h)$ at nth frame is produced.

Tanimoto distance [22] is exploited as the measure between the FLI $\vec{f_N}$ and the SGM $\vec{g}$:

$$D_T(\vec{f_N}, \vec{g}) = 1 - \frac{\sum_{(x,y)}\min\{f_N(x,y), g(x,y)\}}{\sum_{(x,y)}\max\{f_N(x,y), g(x,y)\}}, \tag{2}$$

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

3.3 Optimal State Estimation

The optimal SGM could be computed by minimizing the Tanimoto distance if initial tracking boxes are accurately aligned with the person. Unfortunately, the initial tracking bounding boxes always have certain deviations from the perfect alignment, which lead to false matching of the SGM. Therefore we have to search for the optimal SGM by translating and scaling the bounding boxes in FLIs. The translated and scaled bounding box candidates are defined as

$$\vec{f}_{NQ}(n;\vec{s}) = \vec{f}_N(n; (c_x^{init}+s_x\Delta c_x, c_y^{init}+s_y\Delta c_y, h^{init}+s_h\Delta h)), \tag{3}$$

where (c_x^{init}, c_y^{init}) and h^{init} are the center and the height of the tracking bounding box; Δc_x, Δc_y are quantization steps for translations in x and y directions; Δh is the quantization step for height scaling. The vector $\vec{s} = (s_x, s_y, s_h)$ represents translation and scaling coefficients. In this work, the steps are set to $\Delta c_x = \Delta c_y = \Delta h = 0.01h^{init}$ empirically.

The optimal state is estimated based on the searching. $\vec{x} = (\phi, \vec{s})$ denotes a state vector in the 4-D searching space. We define a cost function for silhouette matching for state $\vec{x}$ at nth frame. The optimal state is found by minimizing the following cost

$$\vec{x}^*_{sil} = \arg\min_{\vec{x}\in X} C_{sil}(n, \vec{x}), \tag{4}$$

where $C_{sil}(n,\vec{x}) = D_T(\vec{f_{NQ}}(n,\vec{s}), \vec{g}(\phi))$; X is the domain of $\vec{x}$. The parameters are set to $1 \le \phi \le N_{gait}, -5 \le s_x \le 5, -25 \le s_y \le 25, -5 \le s_h \le 5$ empirically.

This optimization described so far, however, does not consider gait smoothness constraints resulted from the physical body properties and the temporal continuities. The cost C_{sil} is minimized for each frame separately. We will introduce the global optimization using Dynamic Time Warping (DTW) in the next subsection.

3.4 Global Optimal Path Searching Using DTW

DTW is exploited in the global optimization to incorporate gait smoothness constraints. The initial tracking bounding boxes are preprocessed using a moving average filter. The filter size is set to 11 frames empirically.

After the preprocessing, the DTW is computed to find the optimal path in a 5-D space $(n, \vec{x})$. $C_{DTW}(n, \vec{x}(n))$ is a cumulative cost at state $\vec{x}(n)$ at nth frame when the optimal path from the first frame to nth frame is selected by the DTW algorithm. The DTW cost for the first frame is initialized as

$$C_{DTW}(1, \vec{x}(1)) = C_{sil}(1, \vec{x}(1)), \quad \forall\vec{x}(1). \tag{5}$$

Then, the DTW cost is calculated incrementally as

$$\begin{aligned} &C_{DTW}(n, \vec{x}(n)) = \\ &\quad C_{sil}(n, \vec{x}(n)) + C_{trans}(n, \vec{x}^*_p(n-1; \vec{x}(n)), \vec{x}(n)), \end{aligned} \tag{6}$$

where C_{trans} is a transition cost from the previous state $\vec{x}(n-1)$ to the current state $\vec{x}(n)$ at nth frame, which is defined as the sum of the previous DTW cost and smoothness constraint cost:

$$\begin{aligned} &C_{trans}(n, \vec{x}(n-1), \vec{x}(n)) = \\ &\quad C_{DTW}(n-1, \vec{x}(n-1)) + C_{smt}(\vec{x}(n-1), \vec{x}(n)), \end{aligned} \tag{7}$$

$$C_{smt}(\vec{x}(n-1), \vec{x}(n)) = \alpha|\min\{\delta\phi, N_{gait} - \delta\phi\}|, \tag{8}$$

$$\delta\phi = |\phi(n) - (\phi(n-1) + v_\phi)|, \tag{9}$$

where the v_ϕ is an averaged phase transition velocity and α is weight for smoothness constraint. The v_ϕ is set to 1 and α is set to 0.05 empirically.

$\vec{x}_p^*(n-1;\vec{x}(n))$ is the previous optimal state chosen from all the states which can transit to the current state ($\vec{x}(n)$). It is defined as

$$\vec{x}_p^*(n-1;\vec{x}(n)) = \arg \min_{\vec{x}(n-1)\in X(n-1;\vec{x}(n))} \{C_{trans}(n,\vec{x}(n-1),\vec{x}(n))\}, \quad (10)$$

where $X(n-1;\vec{x}(n))$ is a set of possible previous states $\vec{x}(n-1)$ and is defined as

$$\begin{aligned} |\min\{\delta\phi, N_{gait}-\delta\phi\}| &\leq \Delta\phi_{trans}, & (11)\\ |s_x(n)-s_x(n-1)| &\leq \Delta s_{x,trans}, & (12)\\ |s_y(n)-s_y(n-1)| &\leq \Delta s_{y,trans}, & (13)\\ |s_h(n)-s_h(n-1)| &\leq \Delta s_{h,trans}. & (14)\end{aligned}$$

Here, the transition is limited to adjacent states, that is, transition parameters $\Delta\phi_{trans}$, $\Delta s_{x,trans}$, $\Delta s_{y,trans}$, and $\Delta s_{h,trans}$ are set to be 1.

Once the DTW costs are calculated, the optimal path is found by back tracking from the last frame (Let it be Nth frame) as follows:

$$\vec{x}_{DTW}^*(N) = \arg\min_{\vec{x}(N)} C_{DTW}(N,\vec{x}(N)),$$

$$\vec{x}_{DTW}^*(n-1) = \vec{x}_p^*(n-1;\vec{x}_{DTW}^*(n)). \quad (15)$$

Based on the estimated path, the optimal silhouettes templates with the optimal bounding boxes are provided as shape priors for the Min-Cut segmentation.

3.5 Efficient Tanimoto Distance Computation

To find the optimal path, foreground likelihood sequences are matched with silhouette template. The matching includes Tanimoto distance computation and optimal path searching using DTW. The searching takes less than 0.2 seconds. Unfortunately, computing Tanimoto distance directly is very expensive, which takes more than 120 seconds when the algorithm runs on a 1.6GHz laptop. The expensive Tanimoto distance computation makes the proposed approach impractical.

We propose an efficient distance computation method which does not compute the minimum and maximum in Eq. 2 directly. Since the silhouette template images have been given in the initialization, we sort the values corresponding to each position in the silhouette template images. The phase information of these templates is kept in the sorted results. The minimum and maximum in Eq. 2 now can be computed based on nearest value searching. A binary searching is applied in the sorted results to find the value nearest to the input foreground likelihood value. Therefore the maximum and minimum can be assigned without further computation. The computational complexity for the searching is $O(\log n_t)$, which is much more practical than the complexity of the direct computation which is $O(n_t)$ (n_t is the number of the templates in the standard gait model). This method is particularly effective when more templates are necessary to cover large varieties of shapes. In our implementation, we first compare the input likelihood value with the largest and smallest values in the sorted results. If the input value is larger than the largest one or smaller than the smallest one, Tanimoto Distance computation is reduced to $O(1)$ operation. The proposed method is shown in Fig. 2. Note that 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 computed overlapping regions [15]. Tanimoto Distance can be formulated as

$$D_T(\vec{f_N},\vec{g}) = \frac{G+F-C}{C}, \quad (16)$$

where

$$F = \sum_{(x,y)} f_N(x,y), \quad (17)$$

$$G = \sum_{(x,y)} g(x,y), \quad (18)$$

$$C(f_N,g) = \sum_{(x,y)} \min\{f_N(x,y), g(x,y)\}. \quad (19)$$

Based on this formulation, G (sum of gait template values) is computed only once during the initialization. For each input foreground likelihood image sequence, F is also computed only once. $C(f_N,g)$ is computed based on the efficient method.

Using the approach described above, it takes around 0.8 seconds to compute distances between an input foreground likelihood sequence and all templates (including shifting, scaling) on the 1.6GHz laptop. The proposed distance computation method could be used in other applications when silhouette template matching is necessary [15].

4. MIN-CUT SEGMENTATION USING SHAPE PRIORS

The Min-Cut algorithm [3, 4] has achieved impressive results in interactive segmentation and 3D reconstruction. Automatic segmentation is extremely challenging based on color information alone. Markov Random Fields, which are the foundation of the Min-Cut algorithm, provide poor priors for specific shapes [6]. It is necessary to incorporate shape priors into the Min-Cut algorithm to achieve reasonable segmentation results.

4.1 Min-Cut Segmentation

Min-Cut algorithm is briefly revisited before the incorporation of shape priors. Let $\mathcal{L} = \{1\ldots K\}$ 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. The cost function is obtained in a context of MAP-MRF estimation. The purpose of the Min-Cut is to seek the labeling of image pixels ($\mathcal{P}$) by minimizing energy:

$$E(A) = E_{smooth}(A) + E_{data}(A), \quad (20)$$

where $A = (A_1,\ldots,A_{|P|})$ is a binary vector whose components specify label assignment; E_{smooth} measures the smoothness of neighboring pixels; and E_{data} measures the disagreement between labeling and the observed data. E_{smooth} and E_{data} are formulated respectively as

$$E_{smooth}(A) = \sum_{\{pq\}\in\mathcal{N}} V_{pq}(A_p,A_q)$$

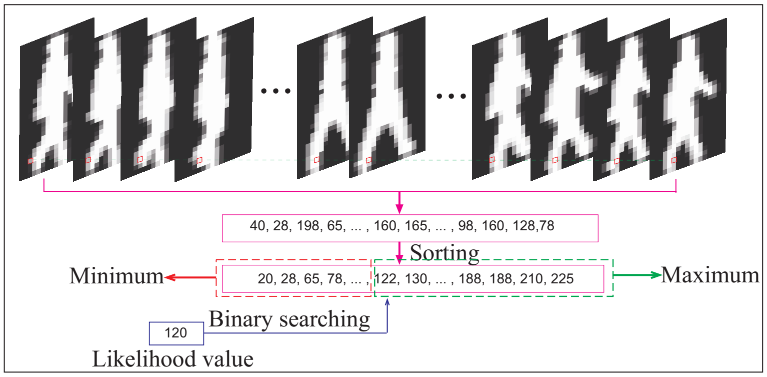


Figure 2: The efficient Tanimoto Distance Computation.

and

$$E_{data}(A) = \sum_{p \in P} D_p(A_p),$$

where $\mathcal{N}$ contains all unordered pairs of neighboring pixels; V_{pq} measures the smoothness of interacting pairs of pixels; D_p is determined by the fitness of p given the observed data. In this work, V_{pq} is formulated as $V_{pq} \propto \frac{e^{-(I(p)-I(q))^2/2}}{\|p-q\|}$ and D_p is computed as the probabilities of pixel p belonging to the foreground.

4.2 Min-Cut Segmentation Using Shape Priors

Shape priors add an energy term to the Min-Cut algorithm:

$$E(A) = E_{smooth}(A) + E_{data}(A) + E_{shape}(A). \tag{21}$$

The shape priors from the optimal path searching are silhouettes of people. The Min-Cut now includes the shape fitness, smoothness and data 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 [11]. Edges are detected in the silhouette image using Canny edge detector. Then Euclidean distance transform [2] is applied in the images. We found that the method in [11] is lacking since the probabilities decreases too quickly near the contour. We decrease the distance values obtained by distance transform by applying a local searching. We extract edges in the input images. The distance is kept without changes if there are edges near the shape prior. Otherwise the distance values are multiplied by a constant factor (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

$$E_{shape} = \sum_{(pq) \in \mathcal{N}: A_p \neq A_q} \frac{\psi(p) + \psi(q)}{2}, \tag{22}$$

where ψ is a value on the transformed image.

5. RESULTS

The proposed approach has been implemented and tested on indoor and outdoor sequences with ground truth. The first sequence is captured in an indoor environment by a stationary camera. The second sequence is captured in an outdoor environment by a moving camera. The tracking and segmentation are challenging due to the occlusions in the first sequence, the dynamic background in the second sequence and the appearance similarity between the foregrounds and the backgrounds in the two sequences. The images in these sequences have a size of 360 × 240 pixels. The heights of the persons in the sequences are less than 100 pixels.

5.1 Refinement of Bounding Boxes And Phase Estimation

The results of the indoor and the outdoor sequences are shown in Fig. 3 and Fig. 4 respectively. The initial bounding boxes produced by the tracker are shown in Fig. 3(b) and Fig. 4(b). Some of the bounding boxes are not well aligned with the people regions and that initial foreground likelihoods are low for some parts(Fig. 1). The person is occluded by a table in some frames.

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

The selected silhouette templates provided by the searching results are shown in Fig. 3(b). The gait phases corresponding to the walking person are correct. The shape priors are incorporated into the Min-Cut algorithm which gives the segmentation results in Fig. 3(c).

We evaluate the smoothness of walking phase transferring in Fig. 5. The phases estimated by using and without using DTW are compared in Fig. 5. The phases estimated using DTW are much more accurate than those without DTW. It demonstrate the importance of DTW for the optimal path searching. The phase estimation also verifies the necessity of searching in a spatiotemporal space instead on a single frame.

5.2 Segmentation Results

Segmentation performance of our algorithm is evaluated on the indoor and outdoor sequences. The ground truths of these sequences are got by labeling the images manually.

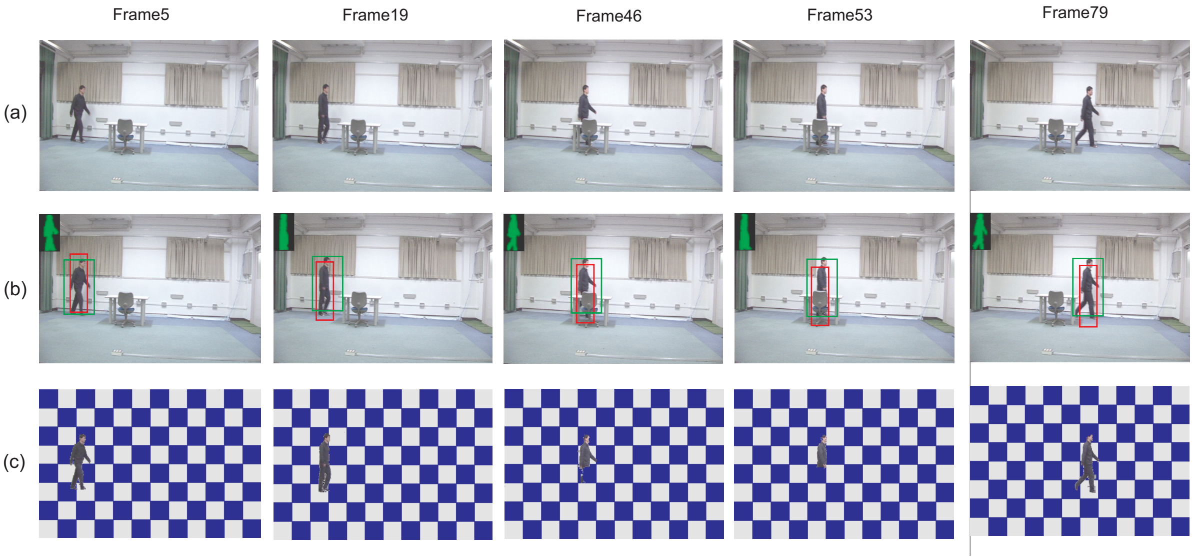


Figure 3: Tracking and segmentation results of the indoor sequence. (a) Input images; (b) Initial bounding boxes (in red) generated by the tracker, optimal bounding boxes (in green) and gait models (phase) obtained using the optimal path searching; (c) Segmentation results by embedding the shape priors into the Min-Cut algorithm.

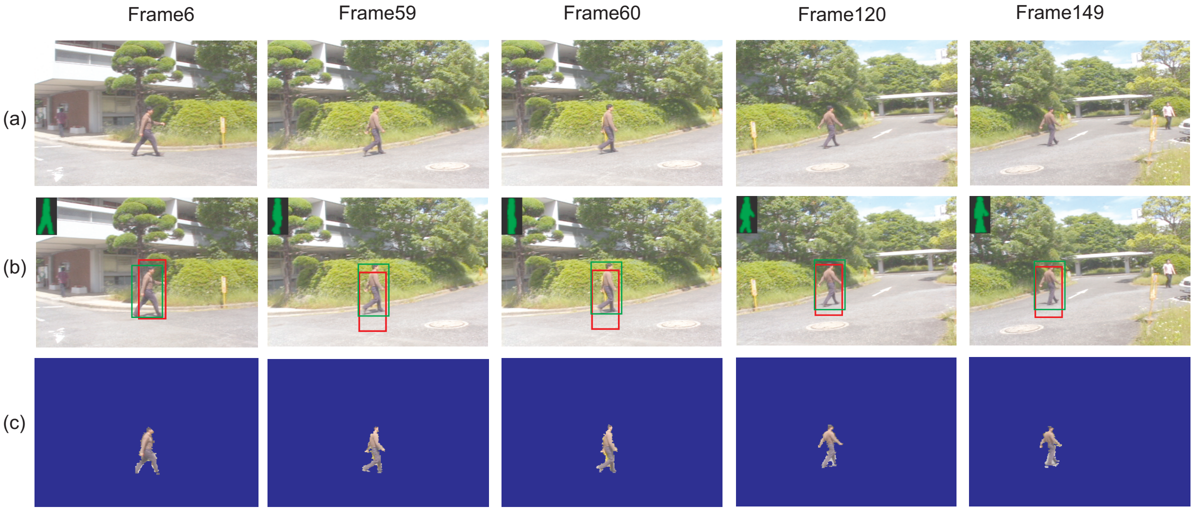
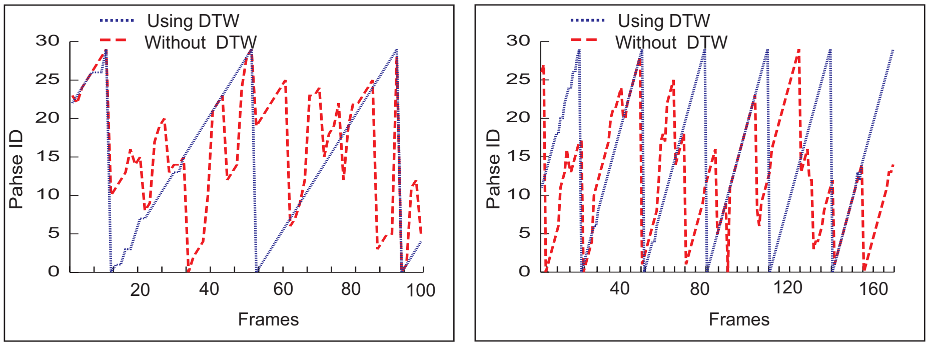


Figure 4: Tracking and segmentation results of the outdoor sequence. (a) Input images; (b) Initial bounding boxes (in red) generated by the tracker, optimal bounding boxes (in green) and gait models (phase) computed by the optimal path searching using DTW; (c) Segmentation results by embedding the shape priors into the Min-Cut algorithm.



(a) (b)

Figure 5: Phase transition estimation for the indoor sequence (a) and the outdoor sequence (b).

Each pixel is labeled as background, foreground, or ambiguous [28]. The ambiguous label is used to mark mixed pixels along the boundaries between foreground and background. We measure the error rate as percentage of mis-segmented pixels, ignoring ambiguous pixels.

Segmentation results using shape priors are shown for the indoor sequence (Fig. 3) and the outdoor sequence (Fig. 4). Quantitative evaluation of the segmentation results with and without shape priors is shown in Fig. 6. 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 to the indoor sequence, the use of shape priors is more helpful for the outdoor sequence. Thus shape priors play a more important role in the challenging outdoor sequence.

6. CONCLUSIONS

We apply spatiotemporal shape constraints in people tracking and segmentation. The optimal path searching results shift bounding boxes to correct positions. Moreover, they provide shape priors in the segmentation. The segmentation performance is also improved based on spatiotemporal shape constraints. The novel efficient silhouette template matching method makes the proposed approach practical for real surveillance applications.

More silhouette templates can be added into the standard gait model to cover the wide varieties of activities. Thanks to the low computational complexity of the proposed matching method, it will not bring much additional cost.

7. REFERENCES

[1] A. Agarwal and B. Triggs. ”Recovering 3D Human Pose from Monocular Images.” *IEEE Trans. Pattern Anal. Mach. Intell.* 28(1): 44-58 (2006)

[2] G. Boregefors. ”Distance Transformations in Digital Images”, *Computer Vision, Graphics and Image Processing*, Vol. 34(3), pp. 344-371, 1986.

[3] Y. Boykov and M-P. Jolly. ”Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in n-d images,” in *Proc. of Int'l Conf. on Computer Vision*, pp. 105-112, 2001.

[4] Y. Boykov, V. Kolmogorov. ”An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision.” *IEEE Trans. Pattern Anal. Mach. Intell.* 26(9): 1124-1137, 2004.

[5] T. Brox, B. Rosenhahn, J. Weickert. ”Three-dimensional shape knowledge for joint image segmentation and pose estimation,” in *Proc. 27th DAGM*, pp. 109-116, 2005.

[6] M. Bray, P. Kohli, P. H. S. Torr. ”PoseCut: Simultaneous Segmentation and 3D Pose Estimation of Humans Using Dynamic Graph-Cuts,” in *ECCV*, II, pp. 642-655, 2006.

[7] R. T. Collins and Y. Liu. ”On-line selection of discriminative tracking features,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 27(10), pp. 1631-1643, 2005.

[8] D. Comaniciu, V. Ramesh, and P. Meer. ”Kernel-based object tracking,” *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 25(5), pp. 564-577, 2003.

[9] A. Criminisi, G. Cross, A. Blake and V. Kolmogorov. ”Bilayer segmentation of live video,” in *Proc. of IEEE Conf. on Computer Vision and Patten Recognition*, pp. 53-60 , 2006.

[10] A. Elgammal and C-S. Lee. ”Inferring 3D body pose from silhouettes using activity manifold learning,” in *Proc. of IEEE Conf. on Computer Vision and Patten recognition*, pp. II-681-II-688, 2005.

[11] D. Freedman and T. Zhang. ”Interactive graph cut based segmentation with shape priors,” in *Proc. of IEEE Conf. on Computer Vision and Patten Recognition*, pp. 755-762, 2004.

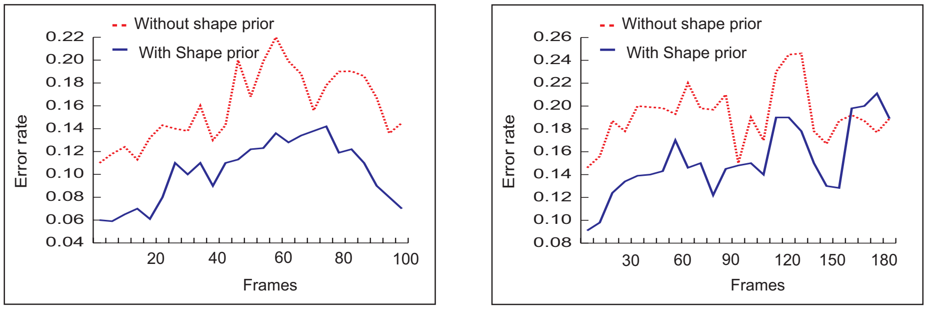
[12] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother. ”Bi-layer segmentation of binocular stereo video,” in *Proc. of IEEE Conf. on Computer Vision and Patten recognition*, pp. 407-414, 2005.

[13] D. Serby, E. Koller-Meier, L.J. Van Gool. ”Probabilistic object tracking using multiple features.” In *Proc. of Int'l Conf. on Pattern Recognition*, (2) pp. 184-187, 2004.

[14] R. Li, M-H Yang, S. Sclaroff, and T-P. Tian. ”Monocular tracking of 3D human motion with a coordinated mixture of factor analyzers”, in *Proc. of European Conf. on Computer Vision*, pp. 323-330, 2006.

[15] M. Marszalek, C. Schmid. ”Accurate Object Localization with Shape Masks.” in *Proc. of IEEE Conf. on Computer Vision and Patten Recognition*, pp. 1-8, 2007.

[16] Y. Makihara and R. Sagawa and Y. Mukaigawa and T. Echigo and Y. Yagi. ”Gait Recognition Using a



(a)

(b)

Figure 6: Segmentation performance evaluation of the indoor (a) and outdoor (b) sequences.

View Transformation Model in the Frequency Domain," in *Proc. of European Conf. on Computer Vision*, pp. 151-163, 2006.

[17] L. Rabiner and B. Juang. *Fundamentals of speech recognition*, Prentice Hall, 1993.

[18] Y. Rathi, N. Vaswani, A. Tannenbaum, A.J. Yezzi. "Particle Filtering for Geometric Active Contours with Application to Tracking Moving and Deforming Objects". In *Proc. of Computer Vision and Pattern Recognition* II, pp. 2-9, 2005.

[19] H. Sidenbladh, M. J. Black, D. J. Fleet. "Stochastic tracking of 3D human figures using 2D image motion," in *Proc. of European Conf. on Computer Vision*, pp. 323-330, 2000.

[20] C. Sminchisescu and A. Jepson. "Generative modeling for continuous non-linearly embedded visual inference," in *Proc. of Int'l Conf. on Machine Learning*, pp. 702-718, 2004.

[21] M. Swain and D. Ballard, "Color Indexing," *Int'l Journal of Computer Vision*, vol. 7, pp. 11-32, 1991.

[22] K. R. Sloan Jr. and S.L. Tanimoto, "Progressive Refinement of Raster Images", *IEEE Trans. on Computers*, Vol. 28(11), pp. 871-874, 1979.

[23] K. Toyama and A. Blake. "Probabilistic tracking in a metric space," in *Proc. of Int'l Conf. on Computer Vision*, pp. 50-57, 2001.

[24] R. Urtasun, D. J.Fleet and P. Fua. "3D people tracking with Gaussian Process Dynamical Models", in *Proc. of Conf. on Computer Vision and Pattern Recognition*, pp. 238-245, 2006.

[25] A. Veeraraghavan, A.K. Roy-Chowdhury and R. Chellappa. "Matching shape sequences in video with applications in human movement analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 27(12), pp. 1896-1909, 2005.

[26] J. Wang, Y. Yagi, "Integrating Color and Shape-texture Features for Adaptive Real-time Tracking", *IEEE Trans. on Image Processing*, vol.17, no.2, 2008.

[27] L. Wang, T. Tan, W. Hu, H. Ning. "Automatic gait recognition based on statistical shape analysis." *IEEE Trans. on Image Processing*, 12(9), pp. 1120-1131, 2003.

[28] P. Yin, A. Criminisi, J. Winn, and I. Essa. "Tree-based classifiers for bilayer video segmentation", in *Proc. of Conf. on Computer Vision and Pattern Recognition*, pp. 1-8, 2007.