2008 IEEE International Conference on
Robotics and Automation
Pasadena, CA, USA, May 19-23, 2008

Human Tracking and Segmentation Supported by Silhouette-based Gait Recognition

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

***Abstract*— Gait recognition has recently gained attention as an effective approach to identify individuals at a distance from a camera. Most existing gait recognition algorithms assume that people have been tracked and silhouettes have been segmented successfully. Tacking and segmentation are, however, very difficult especially for articulated objects such as human beings. Therefore, we present an integrated algorithm for tracking and segmentation supported by gait recognition. After the tracking module produces initial results consisting of bounding boxes and foreground likelihood images, the gait recognition module searches for the optimal silhouette-based gait models corresponding to the results. Then, the segmentation module tries to segment people out using the provided gait silhouette sequence as shape priors. Experiments on real video sequences show the effectiveness of the proposed approach.**

I. INTRODUCTION

Human gait is an effective biometric for person identification. The major advantage of gait recognition is the ability to identify persons at a distance from a camera, which is a desirable property in surveillance and other applications. Shape-based methods are popular in gait recognition because they are invariant to human clothing, illumination, and background changes. In shape-based methods, silhouette information can be extracted by background subtraction or tracking and then segmentation. Background subtraction is not very effective due to background clutter and foreground variations. In addition, background subtraction cannot deal with video sequences captured by a moving camera. Tacking and segmentation are very difficult especially for articulated objects such as human beings. Despite of these difficulties, most existing gait recognition algorithms assume that tracking and segmentation have been done and there are perfect spatial-temporal silhouettes available for gait classification.

In this work, we aim to improve tracking and segmentation performance by incorporating shape priors. Silhouettes are formed in image sequences when a human 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 human tracking and segmentation.

*The first two authors have same contributions to this work.

In the proposed approach, there are three modules: the tracking module, the gait recognition module and the segmentation module. The tracking module provides preliminary input to the gait recognition module. Bounding boxes and Foreground Likelihood Images (FLI) generated by the tracking module are not perfect. Errors always exist in tracking results due to the variations of foreground or background. The recognition module finds the optimal path for the people in a 5-D space where position, scale, and gait phase are optimized. The searching results contain shape priors computed using a Standard Gait Model (SGM). Since it takes gait smoothness constraints into consideration, gait pathes found by the recognition module is optimal. These shape priors are embedded into the Min-Cut algorithm to improve segmentation.

This paper is organized as follows. The remainder of Section I discusses related work. Section II introduces the tracking module. Section III describes gait recognition using Dynamic Time Warping (DTW). Section IV presents the segmentation module in which shape priors are embedded into the Min-Cut algorithm. Experimental results on real image sequences are demonstrated in Section V. Section VI concludes this work.

A. Related Work

Shape and dynamics prior models play an important role in people tracking. Filtering techniques have been widely used in tracking human [13] because shape or dynamics models can be added into a probabilistic framework. However, many such tracking algorithms require that complex models have been defined for the object to be tracked. Toyama and Blake [16] 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 the problem of human segmentation.

Filtering based algorithms also suffer from the curse of dimensionality due to 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 [7]. Li et al. [10] proposed a coordinated mixture of factor analyzers for bidirectional mapping between the original body pose

978-1-4244-1647-9/08/$25.00 ©2008 IEEE.

Fig. 1. The tracking results of one frame from an outdoor sequence. The image on the left is the input image, and the bounding box computed by the tracking module is overlaid on it; the image on the right is the foreground likelihood image.

space and the low-dimensional space. Urtasun et al. [17] use Gaussian Process Dynamical Models (GPDM) for people tracking. Precise 3D motion data is necessary for the learning of GPDM. None of the above works handles people tracking and segmentation interactively.

Shape-based methods are popular in gait recognition because of the invariance of shape features. Gait recognition was formulated as a shape sequence matching problem [3] or spatial-temporal frequency domain analysis [11]. Nevertheless, these approaches do not provide an effective human segmentation algorithm crucial for the success of gait recognition.

Bilayer video segmentation for video-chat sequences has been widely studied based on Conditional Random Fields (CRF) [6][9]. Segmentation for binocular stereo [9] and monocular video [6][19] have achieved impressive results. People in video-chat sequences usually have few articulated actions, which alleviates the difficulty of the segmentation.

II. Tracking

Our tracking module is formulated based on the mean-shift algorithm. The mean-shift algorithm [5] has achieved success in object tracking because of its simplicity and robustness. It finds 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 mean-shift algorithm has been extended to an adaptive tracker in which discriminative features are selected from multi-cue [4][18].

The tracking module compute bounding boxes and generates FLIs by back-projecting likelihood ratios into each pixel in the image [18].

Fig. 1 shows the tracking results of one frame. It is clear that the bounding box computed by the tracking module are not well aligned with the person in the image (Fig. 1(a)). Moreover, the foreground likelihood image contains many errors (Fig. 1(b)). Such errors are unavoidable in tracking due to the variations of the foreground or background. The gait recognition module can be helpful in computing better alignment based on imperfect bounding boxes and FLIs.

III. Gait Recognition by DTW

The gait recognition module matches FLI sequence and standard gait models in a 5-D space. The key idea is that the matching between these sequences can help solve the problem brought by imperfect tracking results. The gait recognition module can find an 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 gait models could be violated by the errors in tracking results.

Fig. 2. The Standard Gait Model.

The first step of gait recognition is to construct a Standard Gait Model (SGM). Then an measure is defined for the matching between an FLI and SGM. Finally, the global optimization is achieved based on DTW by considering gait smoothness.

A. SGM Construction

The SGM involves one gait-cycle silhouette image used to estimate gait phase transition from FLI sequence. Whereas high quality silhouettes are required for the modeling of the SGM, temperature-based background subtraction is applied to video sequences captured by an infrared-ray camera. The extracted silhouettes 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 SGM is obtained as an averaged silhouette for each gait phase ϕ :

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

where $\boldsymbol{g}(\phi)$ is the SGM for phase ϕ, $\boldsymbol{h}(n)$ is the normalized silhouette image at nth frame, and N_P is the number of gait periods in the training sequence. The constructed SGM is shown in Fig. 2. Note that components of vector $\boldsymbol{g}$ and $\boldsymbol{h}$ are silhouette values of all positions in each image and the dimension of the vector is $(h_g \times w_g)$.

B. Matching Measure

FLIs generated by the tracking module should be normalized to have the same size as the GSMs'. An FLI at the nth frame is denoted as $\boldsymbol{f}(n)$. 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 SGMs and the normalized FLI $\boldsymbol{f}_N(n; c_x, c_y, h)$ at nth frame is produced.

Tanimoto distance [15] is exploited as the measure between the FLI $\boldsymbol{f}_N$ and the SGM $\boldsymbol{g}$:

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

where $f_N(x,y)$ and $g(x,y)$ are likelihood and silhouette values at (x,y) respectively.

The optimal GSM could be computed by minimizing the above 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 GSMs. Therefore we have to search for the optimal GSM by translating and scaling the bounding boxes in FLIs. The translated and scaled bounding box candidates are defined as

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

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

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

where $C_{sil}(n,\boldsymbol{x}) = D_T(\boldsymbol{f}_{NQ}(n,\boldsymbol{s}), \boldsymbol{g}(\phi))$; X is the domain of $\boldsymbol{x}$, 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.

C. Global Optimization by 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 $(n,\boldsymbol{x})$. $C_{DTW}(n,\boldsymbol{x}(n))$ is a cumulative cost at state $\boldsymbol{x}(n)$ 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

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

Then, the DTW cost is calculated incrementally as

$$C_{DTW}(n,\boldsymbol{x}(n)) = C_{sil}(n,\boldsymbol{x}(n)) + C_{trans}(n,\boldsymbol{x}_p^*(n-1;\boldsymbol{x}(n)),\boldsymbol{x}(n)), \quad (6)$$

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

$$C_{trans}(n,\boldsymbol{x}(n-1),\boldsymbol{x}(n)) = C_{DTW}(n-1,\boldsymbol{x}(n-1)) + C_{smt}(\boldsymbol{x}(n-1),\boldsymbol{x}(n)), \quad (7)$$

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

$$\delta\phi = |\phi(n) - (\phi(n-1) + v_\phi)|, \quad (9)$$

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.

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

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

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

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

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:

$$\begin{aligned} \boldsymbol{x}_{DTW}^*(N) &= \arg\min_{\boldsymbol{x}(N)} C_{DTW}(N,\boldsymbol{x}(N)),\\ \boldsymbol{x}_{DTW}^*(n-1) &= \boldsymbol{x}_p^*(n-1;\boldsymbol{x}_{DTW}^*(n)). \quad (15) \end{aligned}$$

Finally, based on the estimated path, the optimal gait silhouette with the optimal bounding box is provided as shape priors to the segmentation module.

IV. Min-Cut Segmentation using Shape Priors

Boykov and Jolly [2] proposed the Min-Cut algorithm for interactive segmentation. Since then, the Min-Cut algorithm has achieved excellent results in segmentation and 3D reconstruction. However, fully automatic segmentation based on color distributions alone is extremely challenging. Markov Random Fields, which are the foundation of the Min-Cut algorithm, provide poor prior for specific shape. It is necessary to embed shape priors into the Min-Cut algorithm to achieve reasonable segmentation results.

The gait recognition module provides an optimal silhouette corresponding to a certain pose in an image. The silhouette from the SGM is adopted as the shape prior for the segmentation. The problem is how to embed shape priors into the Min-Cut algorithm.

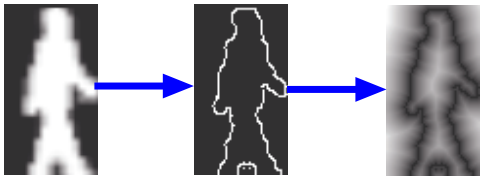


Fig. 3. Generation of shape priors for the segmentation. (a) The silhouette provided by the gait recognition module; (b) the Edge detection result; and (c) the Euclidean distance transform result.

A. Min-Cut Segmentation

Before the embedding of shape priors, the Min-Cut algorithm is briefly revisited. 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:

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

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

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

and

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

where $\mathcal{N}$ contains all unordered pairs of neighboring pixels; $V_{p,q}$ measures the smoothness of interacting pairs of pixels; D_p is determined by the fitness of p given the observed data. In this work, $V_{p,q}$ 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.

B. Min-Cut Segmentation Using Shape Priors

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

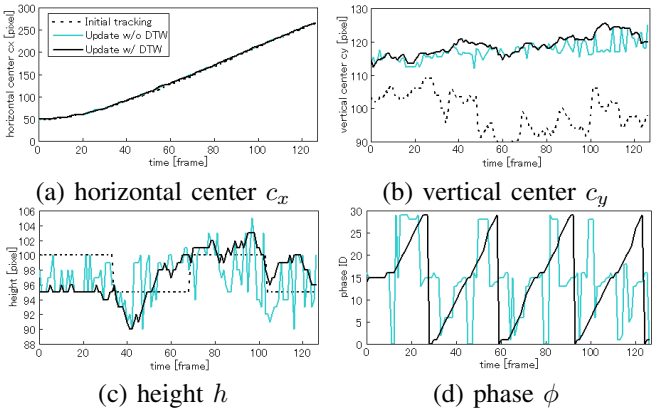
$$E(A) = E_{smooth}(A) + E_{data}(A) + E_{shape}(A). \quad (16)$$

The shape priors from the gait recognition module 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 [8]. In Fig. 3, edges are detected in the silhouette image using Canny edge detector. We found that the method in [8] is lacking since the probabilities decreases too quickly near the contour. In our implementation, edge images are enlarged by scale 1.05 (The scale is set empirically). Then we compute the Euclidean distance transformation [1] of the edge image based on the enlarged edge image. The cost function of shape priors is well described in the transformed image where costs depend on the distance from the edges. The shape prior energy is written as

$$E_{shape} = \sum_{(p,q) \in \mathcal{N}: A_p \neq A_q} \frac{\psi(p) + \psi(q)}{2}, \quad (17)$$

where ψ is a value on the transformed image.



(a) horizontal center c_x (b) vertical center c_y

(c) height h (d) phase ϕ

Fig. 4. Bounding box refinement and phase transition estimation of the indoor sequence.

V. Experiment

In the proposed algorithm, the gait recognition module finds optimal pathes in the 5-D space. Bounding boxes provided by the tracking module are refined by the gait recognition module. The gait models computed based on the searching results can be incorporated into the Min-Cut algorithm to improve the performance of people segmentation.

In order to evaluate its performance, the proposed algorithm was tested on real video sequences with ground truth. Two sequences were used in the experiments: the first one is an indoor sequence of 128 frames captured by a stationary camera; the second is an outdoor sequence of 121 frames captured by a moving camera. The images in these sequences have a size of 360 $\times$ 240 pixels.

A. Refinement of bounding boxes and Phase transition

The results of the indoor sequence are shown in Fig. 5. The initial bounding boxes and FLIs are shown in Fig. 5(a) and (b). We compute segmentation results by simply thresholding the FLIs and the segmentation results are shown in Fig. 5(c). We can find that the some initial bounding boxes are not well aligned with the human regions and that initial foreground likelihoods are low for some parts of head and leg. As a result, the initial segmentation (Fig. 5(c)) does not cover the whole human regions.

The updated tracking bounding boxes and selected GSMs computed by the gait recognition model without and with DTW are shown in Fig. 5(d) and (e) respectively. In addition, estimated state transitions without and with DTW are shown in Fig. 4 accompanied with transitions by initial tracking results.

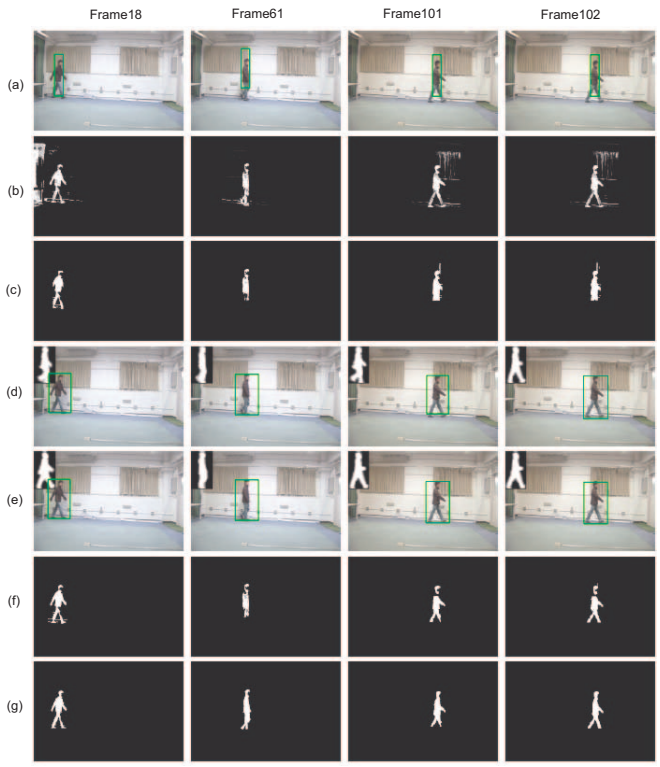


Fig. 5. Tracking, recognition, and segmentation results of the indoor sequence. (a) Input images and initial tracking bounding boxes; (b) Initial FLIs; (c) Initial segmentation results by thresholding FLIs; (d) Optimal bounding box and selected gait model (phase) *without* DTW; (e) Optimal bounding box and selected gait model (phase) *using* DTW; (f) Final segmentation results *without* shape priors; (g) Final segmentation results by *embedding* shape priors. respectively.

Since the horizontal centers of the bounding boxes (Fig. 4(a)) are relatively accurate, there are little differences among the initial tracking, and the updated results without (Fig. 4(d)) and with DTW (Fig. 4(e)). As to the vertical center (Fig. 4(b)), the initial tracking results are shifted upward incorrectly in the latter frames as shown in Fig. 5(a). The recognition module adjusts the positions downward correctly. The state transition is not smooth and the positions change rapidly when the gait recognition does not use DTW. In contrast, the states computed using DTW transit smoothly and the estimated positions are much more accurate.

In Fig. 4(d), when DTW is not in use, phase jumps can be observed and false gait phases are selected in some frames (18th and 101st frames in Fig. 5(d)). The phase transition with DTW is very smooth and estimated phase is correct (Fig. 5(e)).

In Fig. 6, the bounding boxes of the outdoor sequence are refined by the recognition module. The use of DTW improves the smoothness of state transition. It is demonstrated that the gait recognition module is also effective in challenging outdoor sequences.

B. 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. 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 percentage of mis-segmented pixels, ignoring ambiguous pixels.

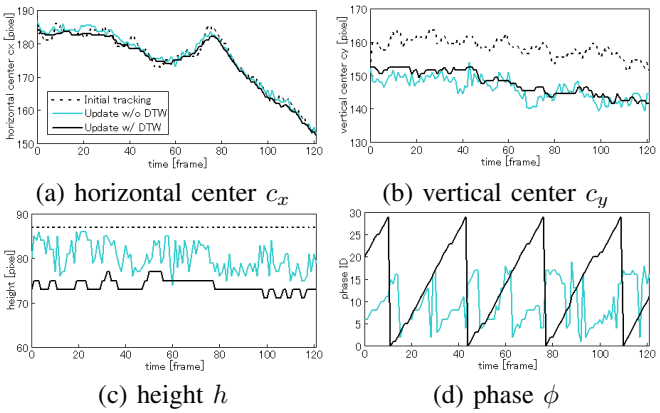


Fig. 6. Bounding box refinement and phase transition estimation of the outdoor sequence.

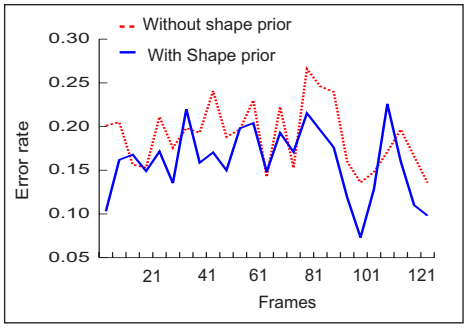


Fig. 8. Segmentation performance evaluation of the indoor sequence.

Segmentation performance for the indoor sequence is qualitatively compared for the results computed without and with shape priors in (Fig. 5). Though the results without shape prior (Fig. 5(f)) are better than the initial result by thresholding the FLIs (Fig. 5(c)), some parts are not segmented correctly. The segmentation results in Fig. 5(g) are much better than those in Fig. 5(c).

Segmentation performances for the indoor and outdoor sequences evaluated in Fig. 8 and Fig. 9. Compared to the performance of the indoor sequence, the use of shape priors for the outdoor sequence is more helpful. Thus shape priors play an important role in the challenging outdoor sequence.

VI. Conclusions and Future Work

We described a tracking and segmentation algorithm supported by gait recognition. Based on the preliminary results provided by the tracking module, the gait recognition module fits SGMs by computing Tanimoto distance. The recognition module then finds the optimal state including gait phase and tracking bounding boxes using Dynamic Time Warping. We incorporate shape and dynamic priors implicitly into the gait recognition module. The Min-Cut based segmentation

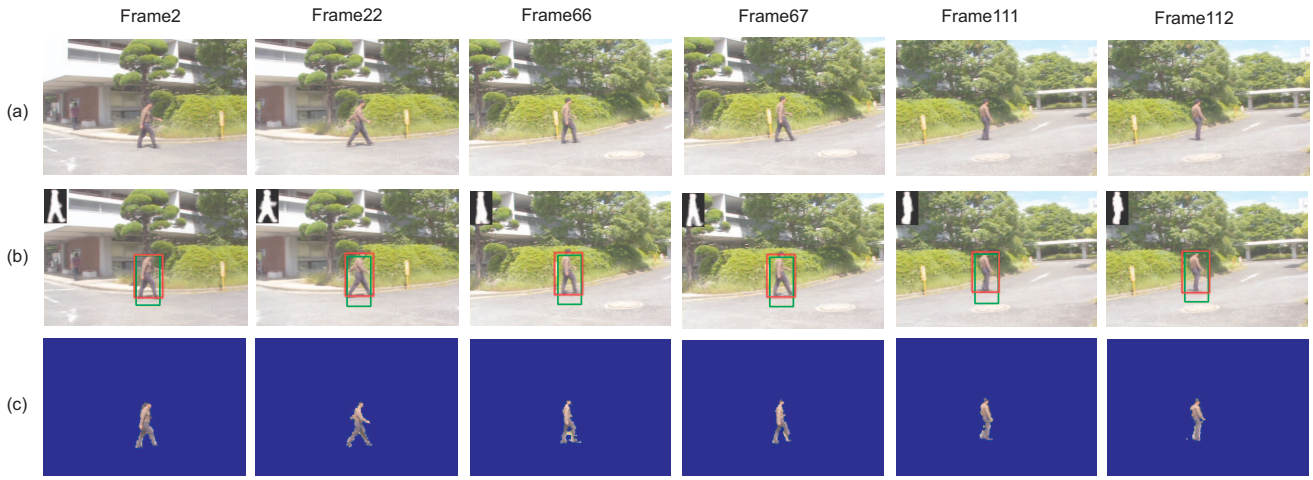


Fig. 7. Tracking, recognition, and segmentation results of the outdoor sequence. (a) Input images; (b) Initial bounding boxes (in green) generated by the tracking module, optimal bounding boxes (in red) and gait models (phase) computed by the gait recognition module using DTW; (c) Segmentation results by embedding the shape priors into the Min-Cut algorithm.

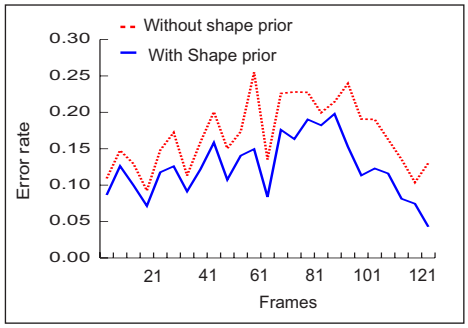


Fig. 9. Segmentation performance evaluation of the outdoor sequence.

module improves the results using the standard gait silhouette sequence as shape priors. Experiments with real gait scene demonstrates that the proposed method improves initial tracking and segmentation results.

We are working on constructing a more generic SGM. Though the SGM was constructed by a specific person's sequence in this paper, it should be constructed by multiple persons' sequences to handle gait type variation. Moreover, it is also necessary to construct multi-view SGMs to cope with changes of view points and walking directions.

REFERENCES

[1] A. Boregefors. "Distance transformations in digital images", *Computer Vision, Graphics and Image Processing*, Vol. 34(3), pp. 344-371, 1986.

[2] 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.

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

[4] R. T. Collins and Y. Liu. "On-line selection of discriminative tracking features," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 27(10), pp. 1631-1643, 2005.

[5] D. Comaniciu, V. Ramesh, and P. Meer. "Kernel-based object tracking," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 25(5), pp. 564-577, 2003.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Y. Makihara and R. Sagawa and Y. Mukaigawa and T. Echigo and Y. Yagi. "Gait Recognition Using a View Transformation Model in the Frequency Domain," in *Proc. of European Conf. on Computer Vision*, pp. 151-163, 2006.

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

[13] 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.

[14] 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.

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

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

[17] 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.

[18] J. Wang and Y. Yagi. "Integrating shape and color features for adaptive real-time object tracking", *2006 IEEE Int'l Conf. on Robotics and Biomimetics*, 2006.

[19] 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.