SHAPE-ORIENTED SEGMENTATION WITH GRAPH MATCHING CORROBORATION FOR SILHOUETTE TRACKING

Qingxiang Zhu, Hongkai Xiong

Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, 200240, China

Xiaoqian Jiang

Division of Biomedical Informatics
University of California, San Diego
La Jolla 92093, CA, USA

ABSTRACT

This paper addresses the problem of advanced silhouette tracking with no prior information, and proposes shape-oriented segmentation together with graph matching corroboration. In terms of unified energy minimization, the shape-oriented graph cut in segmentation exploits the shape information by penalizing the feature points in alignment with shape-oriented map of adjacent frames. While reducing the temporal inconsistencies and improve the accuracy of segmentation, the energy model of graph matching is further designed to compensate the validity of segmentation. To be concrete, it is involved with structural matching cost and unmatched penalty cost to deal with occlusion during tracking. The effectiveness of the proposed scheme is shown with experiments on challenging real-world image sequences.

Index Terms— Silhouette tracking, shape-oriented graph cut, segmentation, graph matching

1. INTRODUCTION

Visual tracking is a challenging and important problem in computer vision. According to a recent review by Yilmaz et al.[1], there are three types of tracking forms: point tracking, silhouette tracking, and kernel tracking. The difference between silhouette tracking and other kinds of tracking is that silhouette tracking not only localizes the position of the objects, but also segments the objects from the background. Moreover, silhouette and contour rather than primitive geometric shapes could properly represent the time-lapse non-rigid objects for animation, surveillance, human-computer interaction, medical diagnosis, and further event recognition. Hence, silhouette tracking has been an active research field for decades and draws attention of this paper.

We can classify silhouette tracking as contour evolution and segmentation-based methods. Contour evolution approaches evolve an initial contour to a new position in the current frame by either using state space searching or minimizing certain contour energy functional. Isard and Blake [8] defined a state space including the spline shape parameters and the affine motion parameters. Particle filter was employed to update the state in the current frame based on the edge observation along the normal lines at the control points on the contour. However, the parametric representation set the limits of this approach for which certain model assumptions must be satisfied. Segmentation based approaches track silhouettes by segmenting every frame into foreground and background regions. Lu et al. [9] addressed the problem of localizing a target's position and segmenting a target as an online binary classification problem using dynamic foreground/background appearance models. The appearance models were formulated as bags of image patches, which were maintained using temporal adaptive importance resampling procedure based on simple nonparametric statistics of the appearance patch bags.

Malcolm et al. used graph cuts for multi-objects tracking through clustering in [2], by introducing a distance penalty and location prediction. However, it may lead to an inaccurate result when there is an occlusion or large scale change since it does not take the shape of the objects into account. In [3], a combined scheme on silhouette tracking against drastic scale change and occlusion was composed of particle filter tracking, 3D graph cut based segmentation. When predicting the position by particle filter, the result of segmentation still depends on graph cut and the mask which may not be always accurate. A combined shape and feature based video analysis for non-rigid object tracking was proposed in [4], which is tightly coupled with an adaptive background generation method to compensate the weakness of block matching. It generates a set of features called shape control points (SCPs) by detecting edges in the neighboring four directions. However, it estimates the objects' boundary using the second-order derivative which is prone to local noises. Conditional random field (CRF) has ever been adopted for silhouette tracking [5], where different visual cues are fused by means of a graphical model and the temporal shape continuity is neglected.

In terms of unified energy minimization, this paper proposes shape-oriented graph cut segmentation together with graph matching corroboration as shown in Fig.1. To reduce

X. Jiang is supported partially by 1K99LM011392-01 and U54HL108460

the temporal inconsistencies, the shape-oriented graph cut in segmentation exploits the shape information by penalizing the feature points in alignment with shape-oriented map of adjacent frames. For the validity of segmentation, the energy model of graph matching is further designed to involve with structural matching cost and unmatched penalty cost. It compensates for the segmentation results if needed and cope with occlusion during tracking. The effectiveness of the proposed approach is demonstrated with experiments on challenging real-world image sequences.

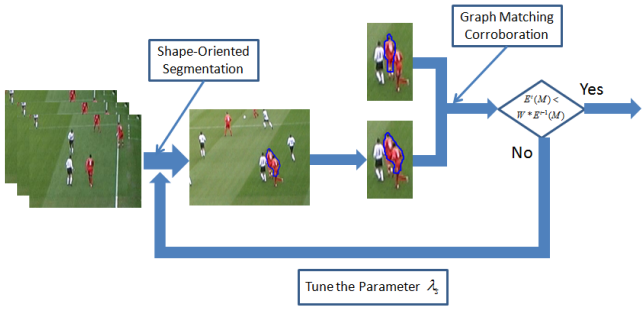


Fig. 1. The diagram of the proposed scheme with shape-oriented segmentation and graph matching corroboration.

The remainder of the paper is organized as follows. Section 2 details our proposed method for silhouette tracking, including shape-oriented segmentation and graph matching corroboration. Section 3 gives some experiments on real life sequences and Section 4 is the conclusion of the paper.

2. THE PROPOSED SCHEME FOR SILHOUETTE TRACKING

Our proposed method for silhouette tracking consists of two parts, shape-oriented segmentation and graph matching corroboration. We first segment the object from background via shape-oriented graph cut method, which can reduce the temporal inconsistencies and improve the accuracy and correctness of segmentation. This part is described in Section 2.1. Graph matching technique is then used to check the validity of segmentation and compensates for the results if needed, which will be elaborated in Section 2.2. The framework of our proposed method is shown in Fig.1, and algorithm 1 depicts the overview algorithm of our silhouette tracking method.

2.1. Shape-Oriented Graph Cut Segmentation

We denote the image sequences as $\{\Omega^1, \Omega^2, ..., \Omega^n\}$. To distinguish foreground from background, we also introduce a labeling function f: for $\forall x \in \Omega^t$, $f(x) = 0$ if x belongs to the background, otherwise $f(x) = 1$.

Algorithm 1 Overview of our proposed silhouette tracking algorithm

INPUT: Image Sequence$\{\Omega^0, \Omega^1, ..., \Omega^N\}$ and the initial object shape C^0 in frame Ω^0
OUTPUT: The object silhouette $C^1, C^2, ..., C^N$ in frame $\{\Omega^1, \Omega^2, ..., \Omega^N\}$
Initialize the maximum iterations $ITERATIONS$
for $n = 1$ to N **do**
  Set $E^n(\mathbf{M})$ infinite.
  $iter\_number = 0$
  while $E^n(\mathbf{M}) > W \times E^{n-1}(\mathbf{M})$ and $iter\_number < ITERATIONS$ **do**
    Segment object in frame Ω^n using Shape-Oriented Graph Cut and get segmentation result $\hat{C}^n$; (See Section 2.1)
    Match $\hat{C}^n$ with C^{n-1} using Graph Matching technique and minimize the energy function $E^n(\mathbf{M})$; (See Section 2.2)
    $iter\_number++$
  end while
  $C^n = \hat{C}^n$
end for

The energy function for graph cut segmentation consists of three parts: the data term ε_D, the regularization term ε_R, and the shape-oriented term ε_S. The data term ε_D measures the likelihood $P_i(x)$ of a pixel belonging to background$(i = 0)$ or foreground$(i = 1)$

$$\varepsilon_D(f) = \sum_{x \in \Omega^t} \sum_{i=0}^{1} P_i(x) \delta(f(x) - i) \quad (1)$$

where $\delta(l)$ is the characteristic function.

The regularization term ε_R penalizes the situation when two neighboring pixels belong to different classes.

$$\varepsilon_R(f) = \sum_{x \in \Omega^t} \sum_{y \in N^l(x)} F(I_x, I_y)[1 - \delta(f(x) - f(y))] \quad (2)$$

where the neighbor of the pixels N^l is defined by:

$$N^l(x) = \{y \in \Omega^t \quad such \quad that \quad 0 < |y - x| \leq l\} \quad (3)$$

, and $F : R \times R \rightarrow R^+$ is a decreasing function that penalizes the spatial discontinuities of the segmentation according to the image data.

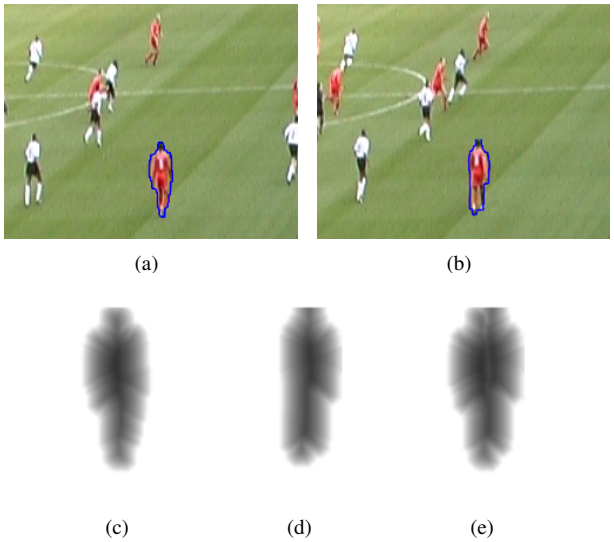
With only the data term and the regularization term, the segmentations obtained at each frame would still suffer from temporal inconsistencies. Therefore, a third term is introduced to improve the graph cut algorithm by penalizing pixels in alignment with their distance from the expected location, which is called shape-oriented term.

Denote the segmentation obtained at frame Ω^{t-1} and Ω^{t-2} as $C^{t-1}(x)$ and $C^{t-2}(x)$, respectively. The shape-oriented map $S^t(x)$ at frame Ω^t is the function of $C^{t-1}(x)$ and $C^{t-2}(x)$.

Algorithm 2 illustrates how the shape-oriented map is generated, and a corresponding example is shown in Fig.2.

Algorithm 2 Generation Of Shape-Oriented Map

INPUT: Results of segmentation $C^{t-1}(x)$ and $C^{t-2}(x)$ at frame Ω^{t-1} and Ω^{t-2}
OUTPUT: Shape-Oriented map $S^t(x)$ at frame Ω^t
Initialize $factor$ and $count$
$C_D^k(x) = C_E^k(x) = C^k(x)$ $(k = t-1, t-2)$
for Erode(C_E^k) $\neq$ NULL **do**
  $count$++
  C_E^k = **Erode** (C_E^k)
  $\hat{S}^k = \min(\hat{S}^k, C_E^k \times factor^{count})$
  C_D^k = **Dilate** (C_D^k)
  $\hat{S}^k = \min(\hat{S}^k, C_D^k \times factor^{-count})$
end for
$S^t(x) = \min(\hat{S}^{t-1}, \hat{S}^{t-2})$
Calculate the center of $C^{t-1}(x)$ and $C^{t-2}(x)$, denote them as m^{t-1} and m^{t-2}
Predict the center of $S^t(x)$, $\hat{m}^t = 2m^{t-1} - m^{t-2}$
Shift the center of $S^t(x)$ to $\hat{m}^t$



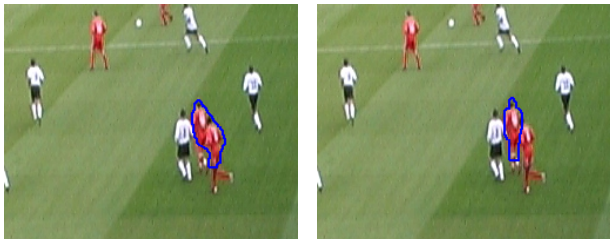
(a) (b) (c) (d) (e)

Fig. 2. (a) and (b) are the segmentation results of $C^{t-1}(x)$ and $C^{t-2}(x)$ at frame Ω^{t-2} and Ω^{t-1}. (c) and (d) are the results of $\hat{S}^{t-1}$ and $\hat{S}^{t-2}$, and (e) shows the result of shape-oriented map S^t.

Thus, the shape-oriented term is formulated as

$$\varepsilon_S(f) = \sum_{x \in \Omega^t} [S^t(x)\delta(f(x)-1) + \frac{1}{S^t(x)}\delta(f(x))] \quad (4)$$

where $S^t(x)$ is the shape-oriented map at frame Ω^t. When x belongs to the foreground $(f(x) = 1)$, more cost will be required if x is far away from its expected position, and vice versa.



(a) $\lambda_S = 2$ (b) $\lambda_S = 10$

Fig. 3. Segmentation results using different λ_S.

Fig.3 shows how shape-oriented term affects the segmentation results. When the value of λ_S is small, the segmentation result largely depends on the data term and regularization term, which may not correctly segment the objects from background when occlusion exists, as shown in Fig.3(a). However, if we set λ_S a large value, it is likely to obtain the similar shape of the objects as the preceding frame, as shown in Fig.3(b).

The shape-oriented term is neither rough nor precise, because the rough information may not include the importance of shape or silhouette of the objects, while precise information may violate the situation of objects deformation.

Combining the Eq.(1), (2), and (4), the total energy function for segmentation can be written as

$$\varepsilon(f) = \varepsilon_D(f) + \lambda_R\varepsilon_R(f) + \lambda_S\varepsilon_S(f) \quad (5)$$

There are at least two advantages introducing shape-oriented method into graph cut method. First, it alleviates the problem of temporal inconsistencies by considering the shape of objects in the two former frames. Secondly, it can help to improve the accuracy of segmentation because the shape-oriented term makes full use of the shape information besides the intensity information.

2.2. Graph Matching Corroboration

We now describe the energy function of our graph matching model. Here, we first clarify some important definitions will be used in graph matching algorithm. Denote P and Q as the sets of feature points extracted from two images, where $P = \{p_1, p_2, ..., p_n\}$ and $Q = \{q_1, q_2, ..., q_m\}$. Here $p_i (i = 1, 2, ..., n)$ are the feature points from the first image, and $q_j (j = 1, 2, ..., m)$ are the feature points from the second one.

Denote G as the set of possible feature point correspondences, where $G \subseteq P \times Q$. We also introduce a binary valued matching vector $\mathbf{M} \in \{0, 1\}^G$ to indicate whether the feature points from set P and Q are matched or not. $\forall a = (p_i, q_j) \in G$ $(i = 1, 2, ..., n; j = 1, 2, ..., m)$, $m_a = 1$ if feature points p_i and q_j are matched, $m_a = 0$ otherwise.

In graph matching problems, we also constrain that at most one active match per feature point is allowed. This re-

quirement is known as the uniqueness constraint, and it is commonly used in correspondence problems. Therefore, the matching vector should satisfier

$$\sum_{a \in G(k)} m_a \leq 1 \tag{6}$$

where $k \in P \cup Q$ is a feature point from either of the image , and $G(k)$ is the set of correspondences involving feature point k.

The energy function of our graph matching model consists of three parts: local matching cost, structural matching cost, and unmatched penalty cost. And it is formulated as following:

$$E(\mathbf{M}) = \lambda_{local} E^{local}(\mathbf{M}) + \lambda_{struc} E^{struc}(\mathbf{M}) + E^{penal}(\mathbf{M}) \tag{7}$$

where λ_{local} and λ_{struc} are scalar weights. We describe each of the energy term below.

Local matching cost E^{local} is defined in Eq. (8).

$$E^{local}(\mathbf{M}) = \sum_{a \in G} \alpha_a m_a \tag{8}$$

Considering a matching pair of the feature points $a = (p_i, q_j) \in G$ $(i = 1, 2, ..., n; j = 1, 2, ..., m)$, α_a is defined as the distance between the local descriptor of the feature nodes p_i and q_j. Therefore, E^{local} can achieve a rather small value when two matched feature points have quite similar local information.

Before describing E^{struc}, we clarify the definition of neighbor. We denote N_p as K nearest neighbor feature points of p, where K is a positive integer controlling the size of the neighborhood. And for two assignments $a = (p_i, q_j) \in G$ and $b = (p_{ii}, q_{jj}) \in G$. We say $(a, b) \in N$ if p_i and p_{ii} are the neighbors of each other, and so are q_i and q_{ii}. Thus, the set N is defined as following:

$$N = \{((p_i, q_j), (p_{ii}, q_{jj})) \mid (p_i \in N_{p_{ii}}) \wedge (p_{ii} \in N_{p_i}) \wedge (q_j \in N_{q_{jj}}) \wedge (q_{jj} \in N_{q_j})\} \tag{9}$$

The term E^{struc} measures the geometric agreement between two neighboring matching pairs a and b by evaluating how well the segment $p_i\bar{p}_{ii}$ matches the segment $q_j\bar{q}_{jj}$ both in length and direction.

$$E^{struc}(\mathbf{M}) = \sum_{(a,b) \in N} \beta_{ab} m_a m_b \tag{10}$$

where β_{ab} measures the structural similarity between matching correspondences a and b.

$$\beta_{ab} = e^{dis(a,b)} + e^{arg(a,b)} - 2 \tag{11}$$

$$dis(a, b) = \frac{|\|p_i - p_{ii}\| - \|q_j - q_{jj}\||}{\|p_i - p_{ii}\| + \|q_j - q_{jj}\|} \tag{12}$$

$$arg(a, b) = \arccos\left(\frac{p_i - p_{ii}}{\|p_i - p_{ii}\|} \frac{q_j - q_{jj}}{\|q_j - q_{jj}\|}\right) \tag{13}$$

Unmatched penalty cost E^{penal} is defined to punish feature points that are unmatched.

$$E^{penal}(\mathbf{M}) = 1 - \frac{1}{\min\{|P|, |Q|\}} \sum_{a \in G} m_a \tag{14}$$

where $|P|$ and $|Q|$ are the number of feature points in graph P and Q respectively, and $\sum_{a \in G}$ is the total number of matching pairs between the two graphs.

We can rewrite the energy function Eq. (7) into the form of Eq. (15) which can be solved by the Dual Decomposition algorithm [6].

$$\min E(\mathbf{M}) = \sum_{a \in G} \bar{\alpha_a} m_a + \sum_{(a,b) \in N} \beta_{ab} m_a m_b \tag{15}$$

In the proposed method, we conduct the graph matching between $\hat{C}^t$ and C^{t-1}, where $\hat{C}^t$ is the result obtained by shape-oriented segmentation at frame Ω^t, and C^{t-1} is the segmentation result of frame Ω^{t-1}. The point sets P and Q are uniformly sampled along the contour of $\hat{C}^t$ and C^{t-1}, and we use the Shape Context descriptors [7] as the local feature descriptor.

To ensure that the segmentation result is reasonable, the minimized graph matching energy at frame Ω^t, denote as $E^t(\mathbf{M})$, should be smaller than the minimized energy at the previous frame multiply a factor.

$$E^t(\mathbf{M}) < W \times E^{t-1}(\mathbf{M}) \tag{16}$$

where W is the constant factor.

To be concrete, we will increase the value of λ_S to give more attention to the shape of the objects and do the shape-oriented segmentation again until it satisfies the criterion in Eq. (16) or reaches the maximum number of iterations.

3. EXPERIMENTS

The proposed scheme is validated on several challenging image sequences, where the tracker is developed using Matlab combined with a C++ implementation of graph cut and Dual Decomposition [6] algorithms.

Within implementation, we set $P_i(x) = (I_x - \mu_i)^2$ in the data term of Eq. (1). The loss function $F(I_x, I_y)$ in the regularization term of Eq. (2) is defined as $F(I_x, I_y) = \exp(\frac{-(I_x - I_y)^2}{2\sigma^2}) \frac{1}{\|x-y\|}$, and the neighborhood size $N^l = N^5$. As to the parameters configuration, we set $\lambda_R = 5$ and the value $\lambda_S = 2$ of Eq. (5) in all experiments. We also set constant factor $W = 1.2$ in Eq. (16), and the maximum number of iterations is 5.

We first compare the segmentation results based on a football match sequence with resolution 240×200 as shown in

Fig. 4. The top row shows the segmentation result using [2], which takes into account the distance penalty and location prediction, while the bottom row is the result by using our proposed graph cut method in Sec. 2.1. Taking the advantage of shape-oriented term, our graph cut method achieves a more accurate segmentation result of the football player, especially in Frame #153 after overlapping occurs.

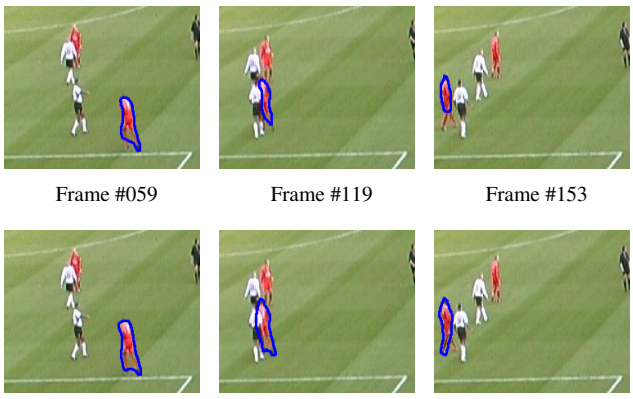


Fig. 4. The top row shows the segmentation results by using [2], and the bottom row by using our proposed segmentation method in Sec. 2.1.

Fig. 5 shows the result on another football match sequence with resolution 240×200. In Fig. 5, the top row is the result by [2], the middle row by using shape-oriented segmentation in Sec. 2.1 only, and the bottom row is the result by our proposed scheme. When occlusion happens in frame #053, the proposed scheme maintains a better performance because of the shape-oriented term and the graph matching corroboration.

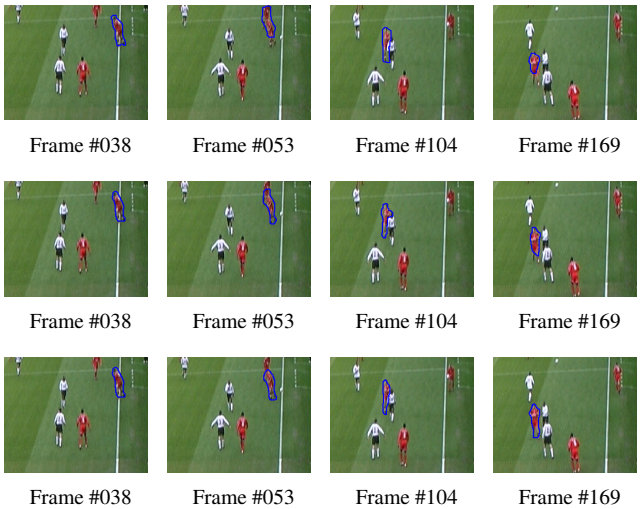


Fig. 5. The top row shows the result of a football match sequence by [2], the middle row by using shape-oriented segmentation only, and the bottom row by our proposed scheme.

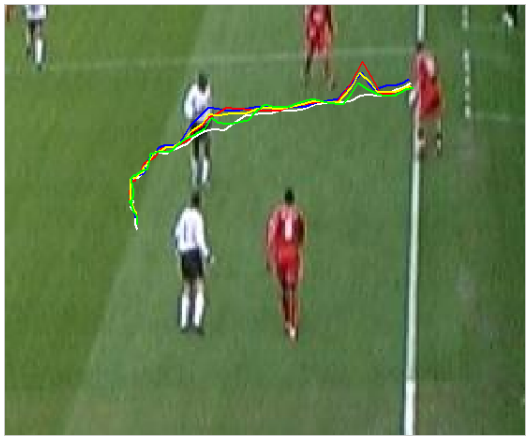


Fig. 6. The trace of the player in football match sequence, where white line is the manually labeled results, blue by [2], red by [3], yellow by shape-oriented method and green by our proposed scheme.

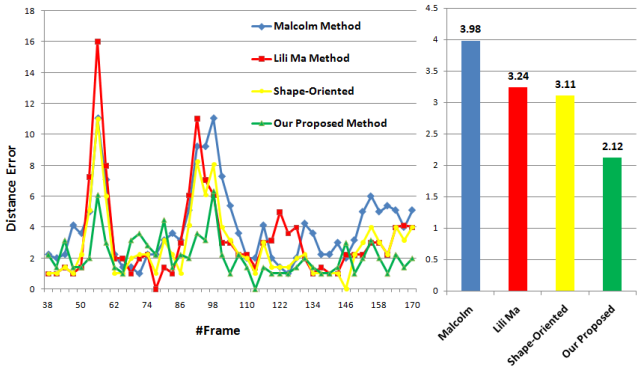


Fig. 7. The left figure shows the distance error of each frame by using different methods, and the right chart is the average distance error of the four methods.

Fig. 6 shows the trace of the football player by using different methods. White line is our manually labeled results, blue by [2], red by [3], yellow by shape-oriented method and green by our proposed scheme. And the statistical result can be seen in Fig. 6. The distance error is defined as the Euclidean distance between the center of the segmentation result and the ground truth. Fig. 7 computes the average distance error of the four different methods.

We also validate our proposed method on a pedestrian sequence from PETS 2010 with resolution 768×576, which is a more challenging video clip. Fig. 8 shows the result, where the top row by [2] and the bottom row by our proposed scheme. Fig. 9 is the trace of the man by using different methods, and Fig. 10 plots the distance error of each frame. We can see that both [2] and [3] fail to track the people in serious occlusion, while our proposed scheme can be immune to the occlusion.

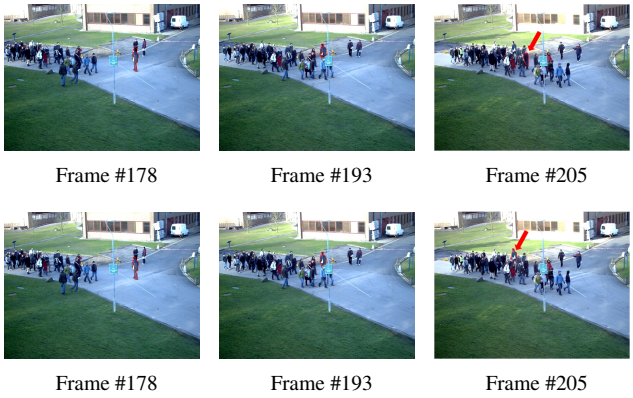


Fig. 8. The top row is the result of a pedestrian sequence by [2] and the bottom row by our proposed scheme.

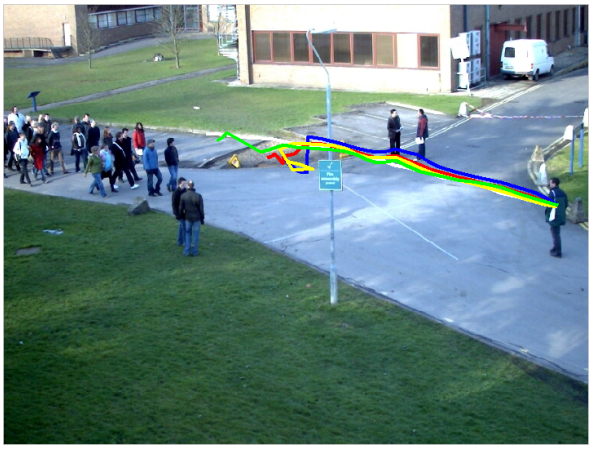


Fig. 9. The trace of the man in pedestrian sequence, where white line is the manually labeled results, blue by [2], red by [3], yellow by shape-oriented method and green by our proposed scheme.

4. CONCLUSION

In this paper, we proposed a novel shape-oriented segmentation and graph matching corroboration scheme for more accurate silhouette tracking. The shape-oriented graph cut segmentation takes into account the shape-oriented map of the objects in adjacent frames. The compensated energy model of graph matching is devised to validate whether previous segmentation results are causal, and accommodate occlusion during tracking. The shape-oriented segmentation and graph matching corroboration are combined to achieve robust detection and tracking. Experiments in real life sequences demonstrate the accuracy of our framework.

There is still some future work to do. A more efficient algorithm may be needed for checking the accuracy of segmentation since the graph matching algorithm still takes much of the time. Besides that, the object is initialized in the first frame of the video, and an detection mechanism may be helpful such that moving objects can be automatically tracked in the video.

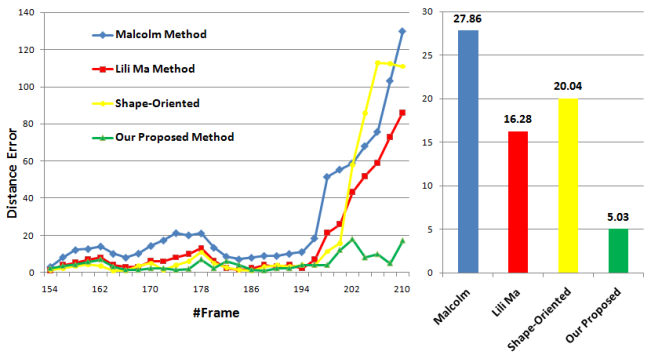


Fig. 10. The left figure shows the distance error of each frame by using different methods, and the right chart is the average distance error of the four methods.

5. REFERENCES

[1] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: a survey," *ACM Computing Surveys*, vol. 38, no. 4, pp. 87-100, Dec. 2006.

[2] J. Malcolm, Y. Rathi, and A. Tannenbaum, "Multi-object tracking through clutter using graph cuts," in *Proc. IEEE Int. Conf. on Computer Vision*, Oct. 2007.

[3] L. Ma, J. Liu, J. Wang, et al., "A improved silhouette tracking approach integrating particle filter with graph cuts," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, pp. 1142-1145, Mar. 2010.

[4] T.Kim, S. Lee, and J. Paik, "Combined shape and featurebased video analysis and its application to non-rigid object tracking," *IET Image Processing*, vol. 5, no. 1, pp. 87-100, Feb. 2011.

[5] G. Boudoukh, I. Leichter, E. Rivlin, "Visual tracking of object silhouettes," in *Proc. IEEE Int. Conf. Image processing*, pp. 3625-3628, Nov. 2009.

[6] L. Torresani, V. Kolmogorov, and C. Rother, "Feature correspondence via graph matching: Models and global optimization," in *Proc. European Int. Conf. on Computer Vision*, Oct. 2008.

[7] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, pp. 509-522, Apr. 2002.

[8] M.Isard, A.Blake, "Condensation-conditional density propagation for visual tracking," *International journal of computer vision*, vol. 29, pp. 5-28, 1998.

[9] L.Lu, G.D.Hager, "A nonparametric treatment for location/segmentation based visual tracking," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2007.