[24] S. Velasco and J. Angulo, "Morphological processing of hyperspectral images using kriging-based supervised ordering," in *Proc. 17th IEEE Int. Conf. Image Process.*, 2010, pp. 1409–1412.
[25] S. Roman, *Lattices and Ordered Sets*. Berlin, Germany: Springer-Verlag, 2008.
[26] H. Heijmans, "Theoretical aspects of gray-level morphology," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 6, pp. 568–582, Jun. 1991.
[27] I. T. Jolliffe, *Principal Component Analysis*. Berlin, Germany: Springer-Verlag, 1986.
[28] B. Scholkopf, A. Smola, and K. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comp.*, vol. 10, no. 5, pp. 1299–1319, 1998.
[29] J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.
[30] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comp.*, vol. 15, pp. 1373–1396, 2002.
[31] N. Kambhatla and T. Leen, "Dimension reduction by local principal component analysis," *Neural Comp.*, vol. 9, no. 7, pp. 1493–1516, 1997.
[32] Z. Ghahramani and G. E. Hinton, "The EM algorithm for mixtures of factor analyzers," Dept. of Comput. Sci., Univ. of Toronto, Tech. Rep., 1997.
[33] G. Matheron, "Le Krigeage Universel," Ecole des Mines de Paris, 1969.
[34] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press, 2006.
[35] E. H. Isaaks and R. M. Srivastava, *An Introduction to Applied Geostatistics*. Oxford, U.K.: Oxford Univ. Press, 1989.
[36] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[37] V. Vapnik and A. Lerner, "Pattern recognition using generalized portrait method," *Autom. Remote Control*, vol. 24, pp. 774–780, 1963.
[38] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[39] A. Kiely and M. Klimesh, "Exploiting calibration-induced artifacts in lossless compression of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 8, pp. 2672–2678, Aug. 2009.
[40] F. Melgani and L. Bruzzone, "Classification of hyperspectral remote sensing images with support vector machines," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 8, pp. 1778–1790, Aug. 2004.

Segmentation and Tracking Multiple Objects Under Occlusion From Multiview Video

Qian Zhang, *Student Member, IEEE*, and King Ngi Ngan, *Fellow, IEEE*

***Abstract*—In this paper, we present a multiview approach to segment the foreground objects consisting of a group of people into individual human objects and track them across the video sequence. Depth and occlusion information recovered from multiple views of the scene is integrated into the object detection, segmentation, and tracking processes. *Adaptive background penalty with occlusion reasoning* is proposed to separate the foreground regions from the background in the initial frame. Multiple cues are employed to segment individual human objects from the group. To propagate the segmentation through video, each object region is independently tracked by motion compensation and uncertainty refinement, and the motion occlusion is tackled as layer transition. The experimental results implemented on both our sequences and other's sequence have demonstrated the algorithm's efficiency in terms of subjective performance. Objective comparison with a state-of-the-art algorithm validates the superior performance of our method quantitatively.**

***Index Terms*—Graph cut, layer transition, multiview video, object segmentation, object tracking.**

I. Introduction

Object detection, segmentation, and tracking are the key topics in computer vision and have facilitated many important applications, such as visual surveillance, human behavior analysis, and object recognition. Segmenting and tracking multiple human objects correctly and consistently when overlapping with each other under occlusion in a complex scene is a more challenging task than when the targets are separated due to the nonrigid motion of deformable objects and the dynamic change of object attributes, such as color distribution, shape, and visibility.

Recently, segmenting and tracking multiple simultaneous objects under occlusion have been addressed in the literature [1]–[9]. A number of video segmentation and tracking approaches and systems have been proposed [1]–[3] for handling objects occlusion in a single view. However, in the cluttered scene and wide-range surveillance, segmenting and tracking crowded people scene with high density using monocular camera is insufficient due to the limited visibility and substantial occlusion. To solve this challenging problem, stereo/multiple cameras are reasonable alternatives to collect more information from different perspectives of the same scene to improve the segmentation and tracking efficiency. This improvement can be achieved by gathering evidence from multiple views [4], combining probabilistic occupancy map [5] or camera collaboration [6]. A relatively new area uses the ground plane homography [7]–[9], which projects feature in each view onto the common view for data fusion.

Manuscript received July 16, 2010; revised November 30, 2010 and March 28, 2011; accepted May 25, 2011. Date of publication June 09, 2011; date of current version October 19, 2011. The work described in this paper was supported by the Research Grants Council of Hong Kong Special Administrative Region, China, under Project CUHK415707. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. A. G. Bors.

The authors are with the Department of Electrical Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong (e-mail: qzhang@ee.cuhk.edu.hk; knngan@ee.cuhk.edu.hk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2011.2159228

Similar to the objective in [7], the proposed algorithm seeks to segment foreground objects corresponding to a group of people into individual objects and track them from the multiview video. The multiple cameras collaboratively work to compute depth and occlusion information of a specific view from the other views of the scene, which are integrated into the object detection, segmentation, and tracking processes. Instead of only tracking feature points of the target, we update the entire objects across videos. Additionally, we obtain more accurate object silhouette, rather than the approximate localization. Furthermore, without knowing the objects' priors and the background information beforehand, we automatically segment the human object from the initial frame, even when the object is occluded by and overlaps with other objects.

With respect to our previous work [10], we proposed three novelties in this paper. First, *adaptive background penalty with occlusion reasoning* is developed to separate the foreground regions from the background. Second, depth, occlusion, motion, and color information are deployed to achieve the segmentation of individual human objects from a group of people. Third, labeled motion occlusion is modeled as layer transition for tracking the multiple overlapping objects.

In the reminder of this paper, Section II describes the segmentation of a group of human objects in the initial frame with overlapping object regions. In Section III, the algorithm extends to the video domain to track the trajectory of individual object. Experimental results tested on a number of sequences are presented with ensuing discussion in Section IV. Section V contains the conclusion.

II. Object Segmentation in the Initial Frame

Identifying the tracked objects and separating them from background is the first step in video tracking. Given the background image, algorithms in [7]–[9] utilize the background subtraction to extract the tracked objects. Here, an automatic algorithm is presented to first extract the foreground as multiple overlapping human objects and then segment them into individual object.

A. Related Work

In our most recent work [10], we have proposed an automatic algorithm to segment multiple objects from a multiview video based on the assumption that the interested objects are visually separated without overlapping regions. Unsupervised object extraction is performed using a saliency model in the initial frame, where a saliency map is calculated by combining motion and depth cues estimated offline. Pixel-based motion estimation is implemented by the proposed method in [11]. The disparity and occlusion maps are estimated using the refinement stage of algorithm in [12] for narrow-baseline stereo estimation, which is a discontinuity-preserving regularization method directly coupling the disparity and occlusion labeling. The depth map and combined occlusion map of a specific view are generated using the disparity map and occlusion map with respect to (w.r.t.) its neighboring views, respectively. We define the energy functions in (1)–(3) and employ the bilabel graph cut for energy minimization to segment individual objects as

$$E(f) = \sum_{(p\in P)} E_p(f_p) + \lambda \sum_{(p,q\in N)} E_{p,q}(f_p, f_q) \tag{1}$$

$$E_p(f_p) = -\log g(c_p|f_p,\theta_c) - \log h(d_p|f_p,\theta_d) \tag{2}$$

$$E_{p,q}(f_p,f_q) = \frac{\exp(-\mathrm{diff}(\mathrm{c_p},\mathrm{c_q})) + \exp(-\beta_{\mathrm{mr}}\cdot(\mathrm{mr_p}-\mathrm{mr_q})^2)}{\mathrm{dist}(\mathrm{p},\mathrm{q})} \tag{3}$$

where f is the labeling field, P is the set of pixels, and N is the neighborhood. Equation (1) is the general formulation of energy function. $E_p(f_p)$ is the basic likelihood energy combining the likelihoods of color and depth cues, and $E_{p,q}(f_p, f_q)$ is the prior energy by *foreground contrast enhancement* [10] incorporating color and motion residual contrast. λ is a parameter to control the weight of $E_p(f_p)$ and $E_{p,q}(f_p, f_q)$, which is user specified to get the optimal results. A Gaussian mixture model (GMM) $g(.)$ [13] is used to model the color distribution θ_c and the histogram model $h(.)$ to model the depth distributions from Fig. 2(b). One GMM, which is a full-covariance Gaussian mixture, is built for each of object and the background using an expectation–maximization (EM) method. The number of components is five in each foreground GMM and ten in the background GMM. Histograms are normalized to sum to 1 over the gray-scale range. c_p and d_p are the color and depth of pixel p, respectively; $\mathrm{mr_p}$ and $\mathrm{mr_q}$ [10] are the motion residual of p and q, respectively. $\mathrm{dist}(p,q)$ and $\mathrm{diff}(\mathrm{c_p},\mathrm{c_q}) = (\beta_{\mathrm{L}}\cdot(\mathrm{L_p}-\mathrm{L_q})^2 + \beta_{\mathrm{a}}\cdot(\mathrm{a_p}-\mathrm{a_q})^2 + \beta_{\mathrm{b}}\cdot(\mathrm{b_p}-\mathrm{b_q})^2)/3$ are the coordinate distance and average color difference in the Lab space between p and q, respectively. β is a constant to control the extent of smoothness and $\beta_{\mathrm{mr}} = (2\langle\|(\mathrm{mr_p}-\mathrm{mr_q})^2\|\rangle)^{-1}$, where $\langle\cdot\rangle$ is the expectation operator over the image. β_L, β_a, and β_b are similarly defined for the Lab color channels, respectively.

In the complex scene with cluttered background, the modified energy function involved the background penalty with occlusion reasoning based on an important observation that the focused objects commonly appear in all the cameras, and the inter-view occlusions are mostly the background regions occurring around either the image boundary or the object boundary (referred as inter-object occlusion). Thus, we impose background penalty factor $\alpha_{bp} = 3.5$ to enforce the likelihood to be the background for the occluded pixels in the combined occlusion map CO_t^v in view v at time t, i.e.,

$$E_p^*(f_p) = \alpha_{bp}\cdot E_p(f_p),\ (f_p = 0, CO_t^v(p) = 128) \tag{4}$$

where $f_p = 0$ and $CO_t^v(p) = 128$ if p is defined as the occluded background. $E_p^*(f_p)$ is the likelihood energy with occlusion penalty.

B. Adaptive Background Penalty With Occlusion Reasoning

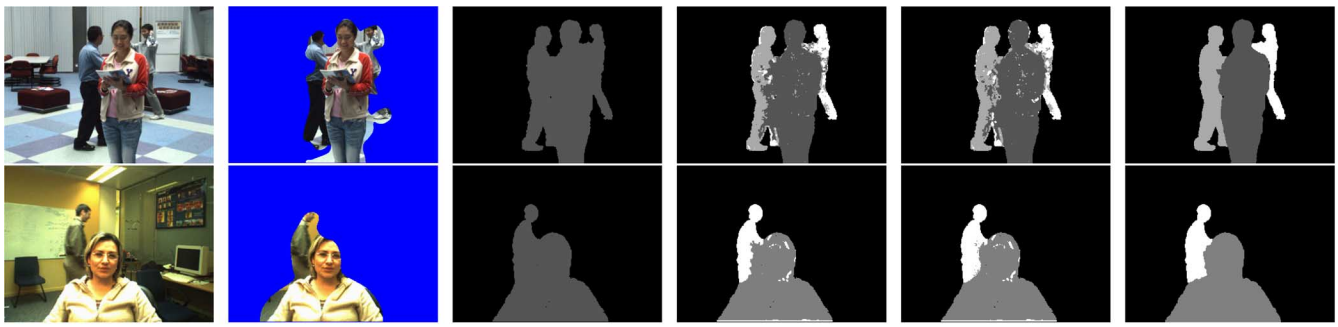
As stated in Section II-A, to segment the spatially separated objects, we propose the background penalty with occlusion reasoning in (4) to enforce the likelihood to be the background for the occluded pixels by introducing constant penalty factor α_{bp}. When the segmentation starts from the initial frame with overlapping objects, not all parts of the objects in the target view in Fig. 2(a) are also visible in other reference views, as shown in Fig. 1(a) and (b). The inter-view occlusions displayed in Fig. 1(c) contain not only the inter-object occlusion but also the intra-object occlusion at the interior of the object. Since the inter-object occlusion is mainly located in the background regions, it deserves a larger value of α_{bp} to enforce its likelihood to be the background. On the other hand, the intra-object occlusion should be assigned a small value of α_{bp} since it is more likely to be the foreground. We achieve this in (5) to adaptively change the value of α_{bp} using the following motion statistics:

$$\alpha_{bp} = \frac{\log h(m_p|f_p = 0,\theta_m)}{\log h(m_p|f_p = 1,\theta_m) + \eta} \tag{5}$$

where $f_p = 0$ for the static background and $f_p = 1$ for the moving object. m_p is the motion vector of p, and θ_m is the motion foreground and background distribution modeled in Fig. 2(b) using the histogram. η is a small value to avoid the division by zero. Equation (5) indicates that if the motion log likelihood of the occluded pixel that is deemed to be the background is larger than that of the object, it is more likely to be the inter-object occlusion, and a large value of α_{bp} is introduced. Otherwise, a smaller value of α_{bp} is used, enforcing a higher probability of the intra-object occlusion. Fig. 1(d) illustrates the segmentation result

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

Fig. 1. Adaptive background penalty with occlusion reasoning. (a) Left reference view of Fig. 2(a). (b) Right reference view of Fig. 2(a). (c) Combined occlusion (CO) map of Fig. 2(a). $\mathrm{CO}(p) = 0$ is not occluded, and $\mathrm{CO}(p) = 128$ is occluded. (d) Result with constant α_{bp}. (e) Result with adaptive α_{bp}.



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

Fig. 2. Segmentation of individual object. (a) Target view of initial frame. (b) Extracted initial objects from saliency map. (c) Foreground regions. (d) Initial labeling by depth clustering. (e) Improved classification using depth ordering. (f) Objects segmentation results. (Top row) *Three-People* sequence. (Bottom row) *IU* sequence.

with constant α_{bp}, where the inter-view occlusions are equally penalized to be the background using the same factor, whereas the improved result using adaptive α_{bp} is evident in Fig. 1(e), where the likelihood to be the background is changed with the value of α_{bp}.

C. Segmentation of Individual Objects

In Fig. 2(c), multiple overlapping objects are first separated from the background as foreground regions. Segmentation of individual objects is equivalent to a k-class pixel labeling problem. By assuming that the human objects are in the different depth layers, a coarse labeling field shown in Fig. 2(d) can be obtained by the k-means clustering of the depth map, where the number of human hypotheses is automatically determined as the number of continuous bins of the depth histogram. The coarse labeling is further improved according to the *depth ordering* [14] stated as follows: If we know that layer L_1 is behind layer L_2, the occlusion region must belong to L_1. Thus, the occlusions in foreground regions around the intersection of different objects are labeled as the indexes of the object in the bottom layer, as shown in Fig. 2(e). Finally, graph cut with α-expansion [15] is employed for the multiple label energy minimization to simultaneously segment the multiple human objects, as illustrated in Fig. 2(f). In the energy function, $E_{p,q}(f_p, f_q)$ is similarly defined as in (3) and $E_p(f_p)$ combines the color and motion cues as

$$E_p(f_p) = -\log g(c_p|f_p, \theta_c) - \log h(m_p|f_p, \theta_m) \tag{6}$$

where θ_c and θ_m are learned in Fig. 2(e). In Fig. 2(f), objects with lower depth values are assigned smaller object indexes, which will be tracked first.

III. Object Tracking in the Video

The goal of our algorithm is to track the entire deformable object and update its connected regions in the spatiotemporal video frames under complex nonrigid motion.

A. Motion Compensation and Uncertainty Refinement

In the earlier approach [10] for video tracking, coarse prediction of the consecutive frames is projected using pixel-based motion compensation exploiting the temporal consistency. To refine the predictions, we define the *activity* measure of a pixel as the motion variance within its second-order neighborhood and shape an uncertain band along the object boundary centered at the most active pixel. The pixels in the uncertain band are segmented by minimizing the energy functions in (1)–(3) using a graph cut for boundary refinement. The performance of the tracking strategy has been demonstrated on a couple of real and complex videos containing spatially separated objects.

In the region of overlapping human objects, the uncertain band of each object is shaped based on the interaction with other objects, as shown in Fig. 3(e). Due to the dynamic movements of the human objects, the motion field between adjacent frames cannot be accurately estimated, particularly between the intersections of different object layers, that is caused by the motion occlusion without correspondence. These motion compensation errors highlighted in Fig. 3(d) degrades the segmentation results, which are accumulated into the following frames. The object mask shown in Fig. 4(d) is the tracking results of the seventh frame by only the motion compensation and uncertainty refinement, where the new uncovered regions across different object layers are lost or mislabeled resulting from the prediction and segmentation errors accumulating frame after frame.

B. Motion Occlusion With Layer Transition

Tracking the human objects by only the motion compensation and uncertainty refinement with overlapping area introduces errors because of the motion occlusion even in the newly exposed regions. To handle this problem, we model the motion occlusion as layer transition since the emergence of occlusion is always accompanied by the label transition between different object layers. The motion occlusion of the two successive frames is detected using the algorithm in [16]. We now discuss two distinct classes of layer transitions for the occluded pixels

Fig. 3. Motion compensation and uncertain band. (a) Zoom-in of previous frame. (b) Zoom-in of current frame. (c) Segmentation results of the previous frame. (d) Prediction after motion compensation. (e) Uncertain band. (Left) Object 1. (Right) Object 2.

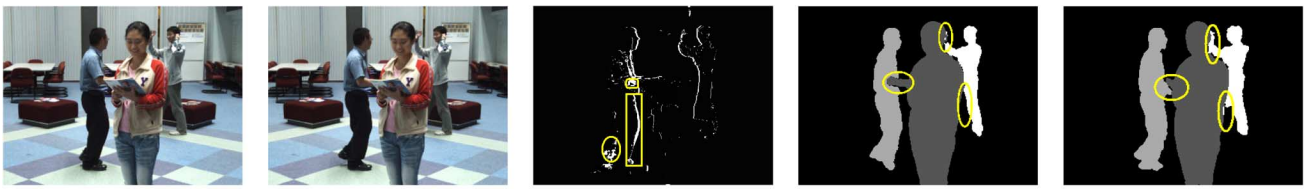


Fig. 4. Objects tracking without and with motion occlusion analysis. (a) Sixth frame. (b) Seventh frame. (c) Motion occlusion: Occlusion in the ellipse is the background to be covered, in the round rectangle is the exposure of uncovered object, and in rectangle is the exposure of uncovered background. (d) Mask of (b) by motion compensation and uncertainty refinement. (e) Mask of (b) by motion occlusion with layer transition.

respectively corresponding to the *background to be covered* and the *uncovered new regions*, as shown in Fig. 4(c).

1) Background to be Covered: We know that if the pixel in the previous frame is labeled as the background layer ($f_p^{t-1} = 0$), it will only transit to a certain foreground object in the current frame.

The determination of the object index is a Bayesian maximum *a posteriori* (MAP) problem, i.e.,

$$f_p^t = \underset{f_p \in F_{\text{fore}} = \{1,2,\ldots,N\}}{\arg\max} P(f_p|x_p) \tag{7}$$

where f_p^t is the label of pixel p in the current frame at time t. F_{fore} is the foreground label set, and N is the number of objects. According to the Bayesian rule, the posterior probability $P(f_p|x_p)$ that an observation of pixel x_p belonging to an object can be decomposed into joint likelihood function $P(x_p|f_p)$ and prior $P(f_p)$ as

$$P(f_p|x_p) \propto P(x_p|f_p)P(f_p). \tag{8}$$

By assuming the uniform distribution of prior $P(f_p)$, the MAP problem is reduced to a maximum likelihood (ML) problem to maximize joint likelihood function $P(x_p|f_p)$, which is evaluated using the color cue modeled by the GMM, combined with the depth and motion cues modeled using the histogram given as follows:

$$P(x_p|f_p) = \log g(c_p|f_p, \theta_c) + \log h(d_p|f_p, \theta_d) + \log h(m_p|f_p, \theta_m) \tag{9}$$

where θ_c, θ_d, and θ_m are learned from the results of previous frame.

2) Uncovered New Regions: Furthermore, we know that if the pixel in the previous frame is labeled as the foreground object ($f_p^{t-1} \in F_{\text{fore}}$), it will only transit to the intersected same layer or the back layer.

Similarly, finding the corresponding layer is an ML problem, i.e.,

$$f_p^t = \underset{f_p \in F_{\text{feasible}} = \left\{\left(f_p^t \geq f_p^{t-1} \cup f_p^t = 0\right) \cap \text{Ins}\left(f_p^{t-1}\right)\right\}}{\arg\max} P(x_p|f_p) \tag{10}$$

where f_p^{t-1} is the label of p in the previous frame at $\{t-1\}$. $\text{Ins}(f_p^{t-1})$ is the set of layer that f_p^{t-1} intersects with, and F_{feasible} is the feasible label set for a foreground object, which is located in the same or back layer in $\text{Ins}(f_p^{t-1})$. The transition between the same layer corresponds to the new uncovered part of object. The transition from the front layer to the back layer indicates the exposure of the occluded part.

Feature Selection: For the uncovered regions appearing on the scene, the new exposed parts may not be consistent with its associated object, or even they are very similar to the other object. For example, the arm belonging to object 2 highlighted as the round rectangle in Fig. 4(c) is more similar in appearance to object 1 than object 2. Under such condition, the color component in the joint likelihood function in (9) will mislead the label decision, which makes the color evidence invalid. To avoid this from happening, we select the appropriate features in the evaluation of the joint likelihood function. We traverse all the possible labels and find the label that corresponds to the maximum color likelihood in

$$f_p^t = \underset{f_p \in F_{\text{fore}} \cup 0}{\arg\max} \log g(c_p|f_p, \theta_c). \tag{11}$$

Based on the statement of the new uncovered regions described in (10), the label that will be transited to should exist in F_{feasible}. If $f_p^t \notin F_{\text{feasible}}$, this bias indicates that the new uncovered parts have nonhomogenous appearance with the associated object, and we use the motion and depth terms to define $P(x_p|f_p)$ in (12a). Otherwise, we combine the color and depth cues to calculate $P(x_p|f_p)$ in (12b), which is distinctive enough to make a good decision, i.e.,

$$P(x_p|f_p) = \begin{cases} \log h(d_p|f_p, \theta_d) \\ \quad + \log h(m_p|f_p, \theta_m) : f_p^t \notin F_{\text{feasible}} & \text{(12a)} \\ \log g(c_p|f_p, \theta_c) \\ \quad + \log h(d_p|f_p, \theta_d) : \text{otherwise}. & \text{(12b)} \end{cases}$$

IV. Experimental Results

We implemented the proposed algorithm on three stereo/multiview video sequences with resolution of 320 × 240. The *Three-People* and *Passing* sequences are three-view videos captured at 25 fps by our multiview acquisition system. Three synchronized cameras are individually mounted on the tripod and placed along an arc spanning a

Fig. 5. Object tracking results in the video. (a) (Left) Object mask and (right) superimposed mask of every 11th frame in the *Three-People* sequence. (b) (Left) Object mask and (right) superimposed mask of every 11th frame in *IU* sequence.

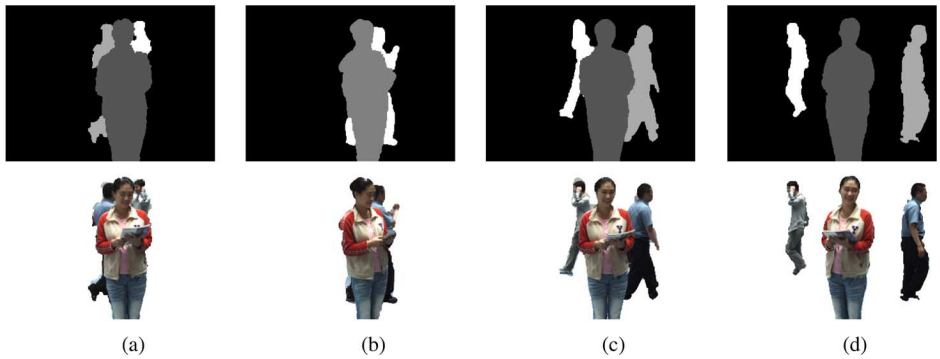


Fig. 6. (Top) Object mask and (bottom) superimposed mask for complete occlusion and reappearance. (a) Nearly complete occlusion. (b) Complete occlusion. (c) Reappearance. (d) Separation.

small angle with nearly equal distance between the neighbors. We also chose the *IU* binocular video captured by a stereo camera from the i2iDatabase [17] to show the robustness of our algorithm on other's database. The algorithm proposed in [18] is adopted for camera calibration, and the camera parameters are used for stereo estimation and depth reconstruction from disparity.

A. Qualitative Evaluation

Segmentation results in the initial frame of the *Three-People* and *IU* sequences are presented in Fig. 2(f), which shows overlapping human objects. Accurate pixel labeling of individual object can be obtained using the proposed segmentation approach in the complex scene with cluttered background, object interocclusion, and color mixing between different layers. From the comparison of Fig. 4(d) and (e), the tracking errors in Fig. 4(d) without motion occlusion analysis have been successfully handled in Fig. 4(e) using the proposed tracking strategy toward modeling the motion occlusion as layer transition, which can achieve more precise representation of individual object by eliminating the lost regions or mislabeling of the new uncovered parts. Object tracking results on every 11th frames on the *Three-People* and *IU* sequences are shown in Fig. 5. The selected frames show the performance of video tracking, which typically contain tracking problems such as objects' partial occlusion, separation, and appearance of a new part. Two-dimensional regions of each object are consistently and correctly tracked across the video sequence with various dynamics in the scene. The satisfactory segmentation and tracking results in Fig. 5 demonstrate the efficiency and robustness of our algorithm in the subjective performance.

With multiple overlapping objects in the scene, some objects may suffer from complete occlusion. We implemented our algorithm on the *Passing* sequence, where one of the objects is completely occluded by the other objects in certain frames and then reappear in the scene. To retrack the object as it emerges, the features' (color, depth, and motion) distributions of each object in the initial frame whether they are visible or partially occluded are recorded. During the object's complete occlusion, its tracker disappears, and the number of objects is reduced in Fig. 6(b). The distributions of the disappeared object in the initial frame and those of the other objects in the previous frame are used to calculate the joint likelihood function to detect the layer transition of motion occlusion and capture the reappearance of the disappeared object. The successful capture of the completely occluded object in Fig. 6(b) as it emerges in Fig. 6(c) until separation in Fig. 6(d) shows the algorithm's capability to handle the complete occlusion following reappearance.

B. Quantitative Evaluation

To further evaluate the performance of our method, we carried out the quantitative comparison with a bilayer segmentation algorithm layered graph cut (LGC) reported in [19]. We compared the segmentation results of our method w.r.t. the ground truth in the *IU* sequence, which can be freely downloaded from [17]. We define the absolute-mean-error rate (AMER) of every fifth frame (left view) as the number of misclassified pixels over the total number of pixels in the image, which is the same measurement adopted in [19]. Additionally, we use the relative-mean-error rate (RMER), which is calculated by the number of misclassified pixels w.r.t. the number of foreground pixels to evaluate the relative segmentation quality. We first compare our results with the ground-truth segmentation based on the depth, which manually labels

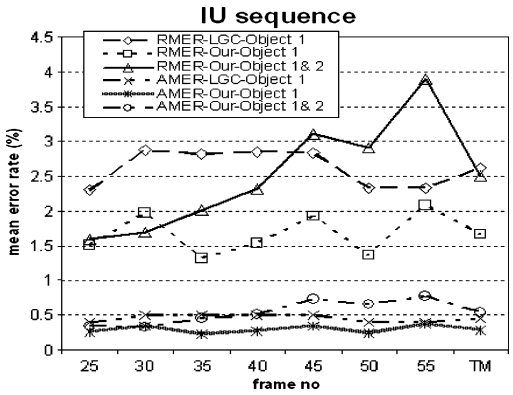


Fig. 7. Quantitative comparison.

the background, object 1 (front lady), and unknown regions. The comparison results in Fig. 7 show that the segmentation quality of our proposed algorithm outperforms that of the LGC in every compared frame, as well as the temporal mean (TM) across the video, by producing lower mean error rates in both numerical measurements. Since our algorithm is designed to handle the object occlusion, it is capable of simultaneously segmenting multiple objects, as shown in Fig. 5(b). Furthermore, we calculate the AMER and the RMER w.r.t. the ground truth based on the motion, where the masks of both objects 1 (front lady) and 2 (back man) are provided. It should be noticed that the RMER of object 1 may be larger than that of object 2, as in frame 30 in Fig. 7, because the increased number of errors is much smaller than the increased number of foreground pixels. Generally, the increased number of the foreground objects degrades the segmentation accuracy because of object occlusion and added complexity. However, in Fig. 7, the TMs of the AMER and the RMER of both objects in the sequence obtained by our method approach those of object 1 by the LGC and are a little higher than those of object 1 by the proposed method. This attests to the advantage obtained by the tracking strategy with occlusion analysis.

C. Computational Efficiency

We implemented all of the algorithms in C++ on a personal computer with Intel Core 2 Duo 1.86-GHz CPU and 1-GB random access memory, without code optimization. The running speed of online segmentation and tracking relies on the number and size of the tracked objects. In the *Three-People* and *Passing* sequences containing three people with whole bodies, our algorithm achieves a processing speed of an average of 0.5 fps, which increases to 0.625 fps in the *IU* sequence including two human objects occupying a smaller area. However, currently, the whole system cannot be realized in real time since the offline operations are quite time consuming. Motion estimation between two successive frames requires 1 min, with a search range of ± 16 pixels in the horizontal and vertical directions. Two-view disparity and occlusion estimation costs 3.5 min, where the maximum disparity is around 50 pixels. To speed up the offline operations, we are going to investigate fast algorithms for motion estimation with occlusion reasoning. Integrating the 3-D depth camera into the system is an alternative way to generate the depth on the fly. Additionally, object extraction from background subtraction by time recursive filtering will accelerate the online processing speed. Furthermore, implementing the program on the computer with advanced configurations and faster processors by code optimization is also expected to increase the computational efficiency.

V. Conclusion

In this paper, we have presented a novel multiview approach, which aims to segment a group of people into individual human object and track them across the video sequence with high accuracy. Based on our previous work on foreground extraction, adaptive background penalty with occlusion reasoning is proposed to extract foreground objects from the background under object occlusion. Segmentation of individual objects is realized using the depth, occlusion, color, and motion cues. To track each object region in the videos, appearance-based tracking approach is employed by the motion compensation and uncertainty refinement, where the motion occlusion is handled as layer transition. Good subjective quality on both others' and our sequences involving various tracking problems validated the algorithm's efficiency and robustness. Quantitative comparison demonstrated the superiority of our algorithm over a state-of-the-art segmentation work.

References

[1] C. Toklu, A. Murat Tekalp, and A. Tanju. Erdem, "Semi-automatic video object segmentation in the presence of occlusion," *IEEE Trans. Circuits Syst. Video Technol.*, no. 4, pp. 624–629, 10, Jun. 2000.

[2] A. M. Elgammal and L. S. Davis, "Probabilistic framework for segmenting people under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001, vol. 2, pp. 145–152.

[3] T. Zhao and R. Nevatia, "Tracking multiple humans in complex situations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1208–1221, Sep. 2004.

[4] A. Mittal and L. S. Davis, "M2Tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo," in *Proc. Eur. Conf. Comput. Vis.*, 2002, pp. 18–36.

[5] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multicamera people tracking with a probabilistic occupancy map," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 267–282, Feb. 2008.

[6] W. Qu, D. Schonfeld, and M. Mohamed, "Distributed Bayesian multiple target tracking in crowded environments using multiple collaborative cameras," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, p. 21, Jan. 2007.

[7] Z. Lin, L. S. Davis, D. Doermann, and D. DeMenthon, "Simultaneous appearance modeling and segmentation for matching people under occlusion," in *Proc. Asian Conf. Computer Vis.*, 2007, pp. 404–413.

[8] K. Kim and L. S. Davis, "Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 98–109.

[9] S. M. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 505–519, Mar. 2009.

[10] Q. Zhang and K. N. Ngan, "Multi-view video segmentation using graph cut and spatiotemporal projections," *J. Vis. Commun. Image Represent.*, vol. 21, no. 5/6, pp. 453–461, Jul./Aug. 2010.

[11] W. Yang, K. N. Ngan, J. Lim, and K. Sohn, "Edge-preserving regularization of disparity and motion fields," in *Proc. 4th EURASIP Conf. Video/Image Process. Multimedia Commun.*, 2003, pp. 71–76.

[12] Q. Zhang and K. N. Ngan, "Dense stereo matching from separated views of wide-baseline images," in *Proc. Adv. Concepts Intell. Vis. Syst.*, 2010, vol. 6474/2010, Lecture Notes in Computer Science, pp. 255–266.

[13] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[14] A. S. Ogale, C. Fermuller, and Y. Aloimonos, "Motion segmentation using occlusions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 6, pp. 988–992, Jun. 2005.

[15] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[16] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001, pp. 508–515.

[17] [Online]. Available: http://research.microsoft.com/en-us/projects/i2i/data.aspx

[18] C. Cui, W. Yang, and K. N. Ngan, "External calibration of multi-camera system based on pair-wise estimation," in *Proc. PSIVT Adv. Image Video Technol.*, 2007, vol. 4872, pp. 497–509.

[19] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother, "Probabilistic fusion of stereo with color and contrast for bi-layer segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1480–1492, Sep. 2006.