

Spatiotemporal Video Segmentation Based on Graphical Models

Yang Wang, *Member, IEEE*, Kia-Fock Loe, Tele Tan, and Jian-Kang Wu

Abstract—This paper proposes a probabilistic framework for spatiotemporal segmentation of video sequences. Motion information, boundary information from intensity segmentation, and spatial connectivity of segmentation are unified in the video segmentation process by means of graphical models. A Bayesian network is presented to model interactions among the motion vector field, the intensity segmentation field, and the video segmentation field. The notion of the Markov random field is used to encourage the formation of continuous regions. Given consecutive frames, the conditional joint probability density of the three fields is maximized in an iterative way. To effectively utilize boundary information from the intensity segmentation, distance transformation is employed in local objective functions. Experimental results show that the method is robust and generates spatiotemporally coherent segmentation results. Moreover, the proposed video segmentation approach can be viewed as the compromise of previous motion based approaches and region merging approaches.

Index Terms—Bayesian network, graphical model, motion segmentation, Markov random field (MRF), region merging, spatiotemporal segmentation.

I. INTRODUCTION

ROBUST video segmentation is very important to application areas such as human-computer interaction, object-based video compression, and multiobject tracking. To differentiate independently moving objects composing the scene, one of the key issues in the design of these vision systems is the strategy to extract and couple temporal (or motion) information and spatial (or intensity) information in the segmentation process.

Motion information is one fundamental element used for segmentation of video sequences. A moving object is characterized by coherent motion over its support region. The scene can be segmented into a set of regions, such that pixel movements within each region are consistent with a motion model (or a parametric transformation) [25]. Examples of motion models are the translational model (two parameters), the affine model (six parameters), and the perspective model (eight parameters).

Manuscript received July 9, 2003; revised June 11, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Philippe Salembier.

Y. Wang was with the Institute for Infocomm Research, Singapore 119613. He is now with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180 USA (e-mail: yang.wang@ieee.org).

K.-F. Loe is with the Department of Computer Science, National University of Singapore, Singapore 117543 (e-mail: loekf@comp.nus.edu.sg).

T. Tan was with the Institute for Infocomm Research, Singapore 119613. He is now with the Department of Computing, Curtin University of Technology, Western Australia 6102, Australia (e-mail: teletan@cs.curtin.edu.au).

J.-K. Wu is with the Institute for Infocomm Research, Singapore 119613 (e-mail: jiankang@i2r.a-star.edu.sg).

Digital Object Identifier 10.1109/TIP.2005.849330

Furthermore, spatial constraints could be imposed on the segmented region where the motion is assumed to be smooth or follow a parametric transformation. In the work of [4], [20], and [24], the motion information and segmentation are simultaneously estimated. Moreover, layered approaches have been proposed to represent multiple moving objects in the scene with a collection of layers [12], [13], [22]. Typically, the expectation maximization (EM) algorithm is employed to learn the multiple layers in the image sequence.

On the other hand, intensity segmentation provides important hints of object boundaries. Methods that combine intensity segmentation with motion information have been proposed [16], [18], [23]. A set of regions with small intensity variation is given by intensity (over)segmentation of the current frame. Usually, a region adjacency graph or a partition tree can be used to represent the regions in the scene [10], [19]. Objects are then formed by merging together regions with coherent motion. The region merging approaches have two disadvantages. First, the intensity segmentation remains unchanged so that motion information has no influence upon the spatial information during the entire process. Second, even an oversegmentation sometimes cannot keep all the object edges, and the boundary information lost in the initial intensity segmentation cannot be recovered later. Since spatial information and temporal information should interact throughout the segmentation process [6], to utilize only motion information or fix intensity segmentation will degrade the performance of video segmentation. From this point of view, it is relatively comprehensive to simultaneously estimate the motion vector field, the intensity segmentation field, and the object (or video) segmentation field.

Fortunately, graphical models provide a natural tool for handling uncertainty and complexity through a general formalism for compact representation of joint probability distribution [14]. In particular, Bayesian networks and Markov random fields are playing an increasingly important role in the design and analysis of machine intelligent systems [8] including image and video processing [7], [15].

In this paper, we present a probabilistic framework in which spatial information and temporal information act on each other during the video segmentation process. A Bayesian network is proposed to model the interactions among the motion vector field, the intensity segmentation field, and the video segmentation field. The notion of the Markov random field (MRF) is employed to boost spatial connectivity of segmented regions. A three-frame approach is adopted to deal with occlusions. The segmentation criterion is the maximum *a posteriori* (MAP) estimate of the three fields given consecutive video frames. To perform the optimization, we propose a procedure that minimizes the corresponding objective functions in an iterative

way. Distance transformation is employed in local optimization to effectively couple the boundary information from intensity segmentation. Experiments show that our technique is robust and generates spatiotemporally consistent segmentation results. Theoretically, the proposed video segmentation approach can be viewed as the compromise of motion based approaches and region merging approaches.

Our method is mostly related to the work of Chang *et al.* [4] and Patras *et al.* [18]. Both approaches simultaneously estimate the motion vector field and the video segmentation field using a MAP-MRF algorithm. The method proposed by Chang *et al.* adopts a two-frame approach and does not use the constraint from the intensity segmentation field during the video segmentation process. Although the algorithm has successfully identified multiple moving objects in the scene, the object boundaries are inaccurate in their experimental results. The method of Patras *et al.* employs an initial intensity segmentation and adopts a three-frame approach to deal with occlusions. However, the method retains the disadvantages of region merging approaches. The temporal information could not act on the spatial information, and the boundary information neglected by the initial intensity segmentation field could no longer be recovered by the motion vector field.

In order to overcome these problems, our algorithm simultaneously estimates the three fields to form spatiotemporally coherent results. The interrelationships among the three fields and successive video frames are described by a Bayesian network model, in which spatial information and temporal information interact on each other. In our approach, regions in the intensity segmentation can either be merged or split according to the motion information. Hence, boundary information lost in the intensity segmentation field can be recovered by the motion vector field.

The rest of the paper is arranged as follows. Section II presents the formulation of our approach. Section III proposes the optimization scheme. Section IV discusses the experimental results. Then, our technique is concluded in Section V.

II. METHOD

A. Model Representation

For an image sequence, it is assumed that the intensity of a pixel remains constant along its motion trajectory. Ignoring both illumination variations and object occlusions, it may be stated as

$$y_k(\mathbf{x}) = y_{k-1}(\mathbf{x} - \mathbf{d}_k(\mathbf{x})) \quad (1)$$

where $y_k(\mathbf{x})$ is the pixel intensity within the k th video frame at site \mathbf{x} , with $k \in \mathbf{N}$, $\mathbf{x} \in \mathbf{X}$, and \mathbf{X} is the spatial domain of each video frame. $\mathbf{d}_k(\mathbf{x})$ is the motion vector from frame $k-1$ to frame k . The entire motion vector field is expressed compactly as \mathbf{d}_k .

Since the video data are corrupted in the image acquisition process, an observation model is required for the sequence. Assume that independent and identically distributed (i.i.d.) Gaussian noise corrupts each point; thus, the observation model for the k th frame becomes

$$g_k(\mathbf{x}) = y_k(\mathbf{x}) + n_k(\mathbf{x}) \quad (2)$$

where $g_k(\mathbf{x})$ is the observed image intensity at site \mathbf{x} , and $n_k(\mathbf{x})$ is the independent zero-mean additive noise with variance σ_n^2 .

In our paper, video segmentation refers to grouping pixels that belong to independently moving objects in the scene. To deal with occlusions, we assume that each site \mathbf{x} in the current frame g_k cannot be occluded in both the previous frame g_{k-1} and the next frame g_{k+1} . Thus, a three-frame method is adopted for object segmentation. Given consecutive frames of the observed video sequence, g_{k-1} , g_k , and g_{k+1} , we wish to estimate the joint conditional probability distribution of the motion vector field \mathbf{d}_k , the intensity segmentation field s_k , and the object (or video) segmentation field z_k . Using the Bayes' rule, we know

$$p(\mathbf{d}_k, s_k, z_k | g_k, g_{k-1}, g_{k+1}) = \frac{p(\mathbf{d}_k, s_k, z_k, g_k, g_{k-1}, g_{k+1})}{p(g_k, g_{k-1}, g_{k+1})} \quad (3)$$

where $p(\mathbf{d}_k, s_k, z_k | g_k, g_{k-1}, g_{k+1})$ is the posterior probability density function (pdf) of the three fields, and the denominator on the right side is constant with respect to the unknowns.

The interrelationships among $\mathbf{d}_k, s_k, z_k, g_k, g_{k-1}, g_{k+1}$ are modeled in the following aspects. First, motion estimation establishes the pixel correspondence among the three consecutive frames. Given the current frame and motion vector field, pixels in the previous frame and the next frame should follow the constant intensity assumption in (1). Second, the intensity segmentation field provides a set of regions with relatively small intensity variation in the current frame. Third, in order to identify independently moving objects in the scene, these regions are encouraged to group into segments with coherent motion. Fourth, if multiple motion models coexist within one region, the region may split into several segments. These four interrelationships are modeled, respectively, by the Bayesian networks in Fig. 1(a)–(d). Combining these four relationships, our video segmentation model can be represented by the Bayesian network in Fig. 1(e). Thus, according to the motion vector field, regions in the intensity segmentation field can either merge or split to form spatiotemporally coherent segments. Moreover, spatial connectivity should be encouraged during the video segmentation process.

The conditional independence relationships implied by the Bayesian network allow us to compactly represent the joint distribution. Using the chain rule [11], the joint probability density can be factorized as the product of the conditional distribution of each element in the Bayesian network given its parents

$$p(\mathbf{d}_k, s_k, z_k, g_k, g_{k-1}, g_{k+1}) = p(g_{k-1}, g_{k+1} | g_k, \mathbf{d}_k) \times p(g_k | s_k) p(s_k) p(\mathbf{d}_k | z_k) p(z_k | s_k). \quad (4)$$

Hence, the MAP estimate of the three fields becomes

$$\begin{aligned} (\hat{\mathbf{d}}_k, \hat{s}_k, \hat{z}_k) &= \arg \max_{(\mathbf{d}_k, s_k, z_k)} p(\mathbf{d}_k, s_k, z_k | g_k, g_{k-1}, g_{k+1}) \\ &= \arg \max_{(\mathbf{d}_k, s_k, z_k)} p(\mathbf{d}_k, s_k, z_k, g_k, g_{k-1}, g_{k+1}) \\ &= \arg \max_{(\mathbf{d}_k, s_k, z_k)} p(g_{k-1}, g_{k+1} | g_k, \mathbf{d}_k) \\ &\quad \times p(g_k | s_k) p(s_k) p(\mathbf{d}_k | z_k) p(z_k | s_k). \end{aligned} \quad (5)$$

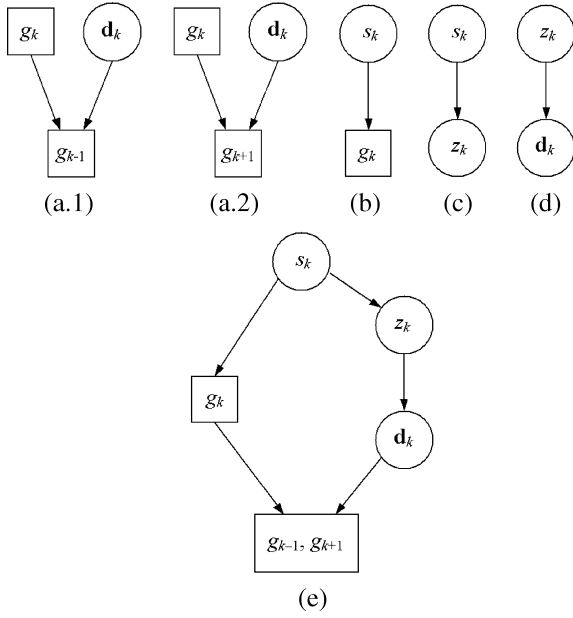


Fig. 1. Bayesian network model for video segmentation.

B. Spatiotemporal Constraints

The conditional probability density $p(g_{k-1}, g_{k+1} | g_k, \mathbf{d}_k)$ shows how well the motion estimation fits the consecutive frames. Assuming that the probability is completely specified by the random field of the displaced frame difference (DFD) [21], the video observation model can be employed to compute $p(g_{k-1}, g_{k+1} | \mathbf{d}_k, g_k)$. We can define the backward DFD $e_k^b(\mathbf{x})$ and forward DFD $e_k^f(\mathbf{x})$ at site \mathbf{x} as

$$\begin{aligned} e_k^b(\mathbf{x}) &= g_k(\mathbf{x}) - g_{k-1}(\mathbf{x} - \mathbf{d}_k(\mathbf{x})) \\ &= n_k(\mathbf{x}) - n_{k-1}(\mathbf{x} - \mathbf{d}_k(\mathbf{x})) \end{aligned} \quad (6a)$$

$$\begin{aligned} e_k^f(\mathbf{x}) &= g_k(\mathbf{x}) - g_{k+1}(\mathbf{x} + \mathbf{d}_k(\mathbf{x})) \\ &= n_k(\mathbf{x}) - n_{k+1}(\mathbf{x} + \mathbf{d}_k(\mathbf{x})). \end{aligned} \quad (6b)$$

The vector $(e_k^b(\mathbf{x}), e_k^f(\mathbf{x}))^T$ is denoted as $\mathbf{e}_k(\mathbf{x})$. With the i.i.d. Gaussian noise assumption, we know that $\mathbf{e}_k(\mathbf{x})$ is of zero mean bivariate normal distribution. The correlation coefficient of $e_k^b(\mathbf{x})$ and $e_k^f(\mathbf{x})$ is

$$\rho = \frac{\text{Cov}[e_k^b(\mathbf{x}), e_k^f(\mathbf{x})]}{\sqrt{\text{Var}[e_k^b(\mathbf{x})] \text{Var}[e_k^f(\mathbf{x})]}} = \frac{\sigma_n^2}{2\sigma_n^2} = \frac{1}{2}. \quad (7)$$

Assuming conditional independence among spatially distinct observations, the probability density can be factorized as

$$\begin{aligned} & p(g_{k-1}, g_{k+1} | g_k, \mathbf{d}_k) \\ & \approx \prod_{\mathbf{x} \in \mathbf{X}} p(g_{k-1}(\mathbf{x} - \mathbf{d}_k(\mathbf{x})), g_{k+1}(\mathbf{x} + \mathbf{d}_k(\mathbf{x})) | g_k(\mathbf{x})) \\ & \approx \prod_{\mathbf{x} \in \mathbf{X}} p(e_k^b(\mathbf{x}), e_k^f(\mathbf{x})) \\ & = \left(\frac{1}{2\pi\sqrt{|\Sigma_e|}} \right)^{|\mathbf{X}|} \exp \left[- \sum_{\mathbf{x} \in \mathbf{X}} \frac{1}{2} \mathbf{e}_k^T(\mathbf{x}) \Sigma_e^{-1} \mathbf{e}_k(\mathbf{x}) \right] \\ & \propto \exp \left[- \frac{1}{3\sigma_n^2} \sum_{\mathbf{x} \in \mathbf{X}} U_{\mathbf{x}}^{g|d}(\mathbf{d}_k(\mathbf{x})) \right] \end{aligned} \quad (8a)$$

$$\begin{aligned} & U_{\mathbf{x}}^{g|d}(\mathbf{d}_k(\mathbf{x})) \\ & = (e_k^b(\mathbf{x}))^2 - 2\rho e_k^b(\mathbf{x})e_k^f(\mathbf{x}) + (e_k^f(\mathbf{x}))^2 \end{aligned} \quad (8b)$$

where Σ_e is the covariance matrix for each site \mathbf{x} , and the correlation coefficient ρ has been computed in (7).

The term $p(g_k | s_k)$ shows how well the intensity segmentation fits the scene. Assuming Gaussian distribution for each segmented region in the current frame, the conditional probability density could be factorized as

$$\begin{aligned} & p(g_k | s_k) \\ & = \prod_{\mathbf{x} \in \mathbf{X}} p(g_k(\mathbf{x}) | s_k(\mathbf{x})) \\ & = \left(\frac{1}{\sqrt{2\pi}\sigma_\eta} \right)^{|\mathbf{X}|} \exp \left[- \sum_{\mathbf{x} \in \mathbf{X}} \frac{1}{2\sigma_\eta^2} (g_k(\mathbf{x}) - \mu_{s_k(\mathbf{x})})^2 \right] \\ & \propto \exp \left[- \frac{1}{2\sigma_\eta^2} \sum_{\mathbf{x} \in \mathbf{X}} U_{\mathbf{x}}^{g|s}(s_k(\mathbf{x})) \right] \end{aligned} \quad (9a)$$

$$U_{\mathbf{x}}^{g|s}(s_k(\mathbf{x})) = (g_k(\mathbf{x}) - \mu_{s_k(\mathbf{x})})^2 \quad (9b)$$

where $s_k(\mathbf{x}) = l$ assigns site \mathbf{x} to region l , μ_l is the intensity mean of region l , and σ_η^2 is the variance for each region.

The pdf $p(s_k)$ represents the prior probability of the intensity segmentation. To encourage the formation of continuous regions, we model the density $p(s_k)$ by a Markov random field [9]. That is, if $N_{\mathbf{x}}$ is the neighborhood of a pixel at \mathbf{x} , then the conditional distribution of a single variable at site \mathbf{x} depends only on the variables within its neighborhood $N_{\mathbf{x}}$. According to the Hammersley-Clifford theorem, the density is given by a Gibbs distribution with the following form:

$$p(s_k) \propto \exp \left[- \sum_{c \in C} V_c^s(s_k(\mathbf{x}) | \mathbf{x} \in c) \right] \quad (10)$$

where C is the set of all cliques c and V_c^s is the clique potential function. A clique is a set of pixels that are neighbors of each other, and the potential function V_c^s depends only on the points within clique c .

Spatial constraint can be imposed by the following two-pixel clique potential:

$$\begin{aligned} & V_c^s(s_k(\mathbf{x}), s_k(\mathbf{y})) \\ & \propto U_{\mathbf{x}, \mathbf{y}}^s(s_k(\mathbf{x}), s_k(\mathbf{y})) \\ & = \frac{1}{\|\mathbf{x} - \mathbf{y}\|^2} [1 - \delta(s_k(\mathbf{x}) - s_k(\mathbf{y}))] \end{aligned} \quad (11)$$

where

$$\delta(x) = \begin{cases} 1, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases}$$

is the Kronecker delta function and $\|\cdot\|$ denotes the Euclidean distance. Thus, two neighboring pixels are more likely to belong to the same class than to different classes. The constraint becomes stronger with the decrease of the distance between the neighboring sites.

The term $p(\mathbf{d}_k | z_k)$ is the conditional probability density of the motion vector field given the video segmentation field.

To boost spatial connectivity, it is modeled by a Gibbs distribution with the following potential function

$$\begin{aligned} V_c^{\mathbf{d}|z}(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}) | z_k) \\ \propto U_{\mathbf{x},\mathbf{y}}^{\mathbf{d}|z}(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}), z_k(\mathbf{x}), z_k(\mathbf{y})) \\ = \frac{1}{\|\mathbf{x} - \mathbf{y}\|^2} \delta(z_k(\mathbf{x}) - z_k(\mathbf{y})) \|\mathbf{d}_k(\mathbf{x}) - \mathbf{d}_k(\mathbf{y})\|^2. \end{aligned} \quad (12)$$

The pairwise smoothness constraint of the motion vectors is imposed only when the two neighboring points share the same video segmentation label. It encourages one region to split into several segments when different motion models coexist. Hence, $U_{\mathbf{x},\mathbf{y}}^{\mathbf{d}|z}$ can be viewed as the region splitting force.

The last term $p(z_k | s_k)$ represents the posterior probability density of the video segmentation field when the intensity segmentation field is given. The density is modeled by a Gibbs distribution with the following potential function:

$$\begin{aligned} V_c^{z|s}(z_k(\mathbf{x}), z_k(\mathbf{y}) | s_k) \\ \propto U_{\mathbf{x},\mathbf{y}}^{z|s}(z_k(\mathbf{x}), z_k(\mathbf{y}), s_k(\mathbf{x}), s_k(\mathbf{y})) \\ = \frac{1}{\|\mathbf{x} - \mathbf{y}\|^2} [1 - \delta(z_k(\mathbf{x}) - z_k(\mathbf{y}))] \\ + \frac{\alpha}{\|\mathbf{x} - \mathbf{y}\|^2} \delta(s_k(\mathbf{x}) - s_k(\mathbf{y})) [1 - \delta(z_k(\mathbf{x}) - z_k(\mathbf{y}))]. \end{aligned} \quad (13)$$

The first term on the right side encourages the spatial connectivity of video segmentation, while the second term encourages two neighboring pixels to share the same video segmentation label when they are within one region of the intensity segmentation field. Therefore, $U_{\mathbf{x},\mathbf{y}}^{z|s}$ encourages intensity segmentation regions to group altogether and can be viewed as the region merging force. The parameter α controls the strength of the constraint imposed by the intensity segmentation.

The interactions in the Bayesian network are modeled by the above spatiotemporal constraints. Combining these pdf terms, the MAP estimation criterion becomes

$$\begin{aligned} (\hat{\mathbf{d}}_k, \hat{s}_k, \hat{z}_k) \\ = \arg \min_{(\mathbf{d}_k, s_k, z_k)} \left[\sum_{\mathbf{x} \in \mathbf{X}} U_{\mathbf{x}}^{g|d}(\mathbf{d}_k(\mathbf{x})) \right. \\ + \lambda_1 \sum_{\mathbf{x} \in \mathbf{X}} U_{\mathbf{x}}^{g|s}(s_k(\mathbf{x})) \\ + \lambda_2 \sum_{\{\mathbf{x}, \mathbf{y}\} \in C} U_{\mathbf{x},\mathbf{y}}^s(s_k(\mathbf{x}), s_k(\mathbf{y})) \\ + \lambda_3 \sum_{\{\mathbf{x}, \mathbf{y}\} \in C} U_{\mathbf{x},\mathbf{y}}^{\mathbf{d}|z}(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}), z_k(\mathbf{x}), z_k(\mathbf{y})) \\ \left. + \lambda_4 \sum_{\{\mathbf{x}, \mathbf{y}\} \in C} U_{\mathbf{x},\mathbf{y}}^{z|s}(z_k(\mathbf{x}), z_k(\mathbf{y}), s_k(\mathbf{x}), s_k(\mathbf{y})) \right] \end{aligned} \quad (14)$$

where the parameters $\lambda_1, \lambda_2, \lambda_3,$ and λ_4 control the contribution of individual terms.

C. Notes on the Bayesian Network Model

In our model, the video segmentation is influenced by both spatial information and temporal information. It should be noted

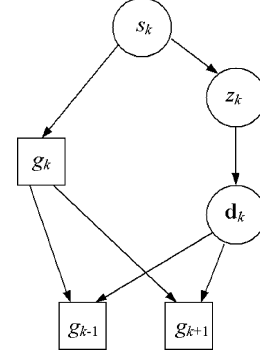


Fig. 2. Simplified Bayesian network model for video segmentation.

that the direction of the links in the Bayesian network model does not mean that the influence between the cause and consequence is only one way.

The current video frame could be thought as the cause of the next frame. For an image sequence, both the original sequence and the one in the reverse sequence order are understandable from the viewpoint of segmentation (in the reversed sequence, object appearances and occlusion relationships are the same as those in the original sequence, while motion models are reversed for all the objects in the scene). Thus, the current frame could also be viewed as the cause of the previous frame (in the reversed sequence). In our model, g_k is the cause of both the next frame g_{k+1} and the previous frame g_{k-1} .

The motion vector field establishes the correspondence between the current frame and its two neighboring frames. When frame g_{k+1} and frame g_{k-1} are separated (as shown in Fig. 2), the interrelationship seems clearer at the first glance. However, from the structure of the Bayesian network, we know that, in this case

$$\begin{aligned} p(g_{k-1}, g_{k+1} | g_k, \mathbf{d}_k) \\ = p(g_{k-1} | g_k, \mathbf{d}_k) p(g_{k+1} | g_k, \mathbf{d}_k) \\ = \prod_{\mathbf{x} \in \mathbf{X}} p(e_k^b(\mathbf{x})) p(e_k^f(\mathbf{x})) \\ \propto \exp \left[-\frac{1}{4\sigma_n^2} \sum_{\mathbf{x} \in \mathbf{X}} (e_k^b(\mathbf{x}))^2 + (e_k^f(\mathbf{x}))^2 \right]. \end{aligned} \quad (15)$$

Compared with (8), the correlation coefficient of $e_k^b(\mathbf{x})$ and $e_k^f(\mathbf{x})$ is zero in (15). The Bayesian network in Fig. 2 neglects the interaction between the forward DFD and the backward DFD. Therefore, the Bayesian network model in Fig. 2 is just a simplification of the original model.

In (13), when the parameter α becomes zero, the constraint from the intensity segmentation disappears so that our method degenerates into a motion based approach. Meanwhile, when α becomes infinity, boundaries in the video segmentation field must come from the intensity segmentation field, and our technique turns into a region merging approach. Therefore, the proposed method can be viewed as the compromise of previous motion based approaches and region merging approaches.

III. MAP ESTIMATION

A. Iterative Estimation

Obviously, there is no simple method of directly minimizing (14) with respect to all unknowns. We propose an optimization strategy iterating over the following two steps.

First, we update \mathbf{d}_k and s_k given the estimate of the video segmentation field z_k . From the structure of the proposed Bayesian network, we can see that \mathbf{d}_k and s_k are conditionally independent when the video segmentation field z_k and the three successive frames are given. The joint estimation can be factorized as

$$\begin{aligned} (\hat{\mathbf{d}}_k, \hat{s}_k) &= \arg \max_{(\mathbf{d}_k, s_k)} p(\mathbf{d}_k, s_k | g_k, g_{k-1}, g_{k+1}, \hat{z}_k) \\ &= \left(\arg \max_{\mathbf{d}_k} p(\mathbf{d}_k | g_k, g_{k-1}, g_{k+1}, \hat{z}_k) \right. \\ &\quad \left. \arg \max_{s_k} p(s_k | g_k, \hat{z}_k) \right). \end{aligned} \quad (16)$$

Using the chain rule, the MAP estimate becomes

$$\begin{aligned} \hat{\mathbf{d}}_k &= \arg \max_{\mathbf{d}_k} p(\mathbf{d}_k | g_k, g_{k-1}, g_{k+1}, \hat{z}_k) \\ &= \arg \max_{\mathbf{d}_k} p(g_{k-1}, g_{k+1} | g_k, \mathbf{d}_k) p(\mathbf{d}_k | \hat{z}_k) \end{aligned} \quad (17a)$$

$$\begin{aligned} \hat{s}_k &= \arg \max_{s_k} p(s_k | g_k, \hat{z}_k) \\ &= \arg \max_{s_k} p(g_k | s_k) p(s_k) p(\hat{z}_k | s_k). \end{aligned} \quad (17b)$$

Second, we update z_k given the estimate of the motion field \mathbf{d}_k and the intensity segmentation field s_k

$$\begin{aligned} \hat{z}_k &= \arg \max_{z_k} p(z_k | g_k, g_{k-1}, g_{k+1}, \hat{\mathbf{d}}_k, \hat{s}_k) \\ &= \arg \max_{z_k} p(z_k | \hat{\mathbf{d}}_k, \hat{s}_k) \\ &= \arg \max_{z_k} p(\hat{\mathbf{d}}_k | z_k) p(z_k | \hat{s}_k). \end{aligned} \quad (18)$$

In our work, the 24-point neighborhood system (the fifth-order neighbor system, see Fig. 3) is used, and potentials are defined only on two-point cliques. Using the terms in (14), the Bayesian MAP estimates in (17) and (18) can be obtained by minimizing the following objective functions:

$$\begin{aligned} F^{\mathbf{d}}(\mathbf{d}_k) &= \sum_{\mathbf{x} \in \mathbf{X}} \left[U_{\mathbf{x}}^g |^{\mathbf{d}}(\mathbf{d}_k(\mathbf{x})) \right. \\ &\quad \left. + \frac{1}{2} \lambda_3 \sum_{\mathbf{y} \in N_{\mathbf{x}}} U_{\mathbf{x}, \mathbf{y}}^{\mathbf{d}} |^z(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}), \hat{z}_k(\mathbf{x}), \hat{z}_k(\mathbf{y})) \right] \end{aligned} \quad (19a)$$

$$\begin{aligned} F^s(s_k) &= \sum_{\mathbf{x} \in \mathbf{X}} \left[\lambda_1 U_{\mathbf{x}}^g |^s(s_k(\mathbf{x})) \right. \\ &\quad \left. + \frac{1}{2} \lambda_2 \sum_{\mathbf{y} \in N_{\mathbf{x}}} U_{\mathbf{x}, \mathbf{y}}^s(s_k(\mathbf{x}), s_k(\mathbf{y})) \right. \\ &\quad \left. + \frac{1}{2} \lambda_4 \sum_{\mathbf{y} \in N_{\mathbf{x}}} U_{\mathbf{x}, \mathbf{y}}^z |^s(\hat{z}_k(\mathbf{x}), \hat{z}_k(\mathbf{y}), s_k(\mathbf{x}), s_k(\mathbf{y})) \right] \end{aligned} \quad (19b)$$

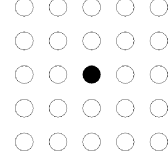


Fig. 3. Fifth-order neighborhood system.

$$\begin{aligned} F^z(z_k) &= \sum_{\mathbf{x} \in \mathbf{X}} \left[\frac{1}{2} \lambda_3 \sum_{\mathbf{y} \in N_{\mathbf{x}}} U_{\mathbf{x}, \mathbf{y}}^{\mathbf{d}} |^z(\hat{\mathbf{d}}_k(\mathbf{x}), \hat{\mathbf{d}}_k(\mathbf{y}), z_k(\mathbf{x}), z_k(\mathbf{y})) \right. \\ &\quad \left. + \frac{1}{2} \lambda_4 \sum_{\mathbf{y} \in N_{\mathbf{x}}} U_{\mathbf{x}, \mathbf{y}}^z |^s(z_k(\mathbf{x}), z_k(\mathbf{y}), \hat{s}_k(\mathbf{x}), \hat{s}_k(\mathbf{y})) \right] \end{aligned} \quad (19c)$$

where $N_{\mathbf{x}}$ is the neighborhood of the pixel at \mathbf{x} .

B. Local Optimization

In general, the objective functions are nonconvex and do not have a unique minimum. The iterated conditional modes (ICM) algorithm is used to arrive at a suboptimal estimate of each objective function [2]. The ICM algorithm employs the greedy strategy in iterative minimization. Given the observed data and other estimated labels, the segmentation label is sequentially updated by locally minimizing the objective function at each site.

To effectively employ boundary hints from the spatial information in local optimization, distance transformation [3] is performed on the intensity segmentation field. Each pixel \mathbf{x} in the distance transformed image has a value $d_{\mathbf{x}}(s_k)$ representing the distance between the pixel and the nearest boundary pixel in s_k . Here, a boundary pixel \mathbf{x} has at least one point \mathbf{y} within its neighborhood where $s_k(\mathbf{y})$ is not the same as $s_k(\mathbf{x})$. The term $U_{\mathbf{x}, \mathbf{y}}^z |^s$ in (19c) is replaced by

$$\begin{aligned} U_{\mathbf{x}, \mathbf{y}}^z |^s(z_k(\mathbf{x}), z_k(\mathbf{y}), \hat{s}_k(\mathbf{x}), \hat{s}_k(\mathbf{y})) &= \frac{1}{\|\mathbf{x} - \mathbf{y}\|^2} [1 - \delta(z_k(\mathbf{x}) - z_k(\mathbf{y}))] \\ &\quad + \frac{\alpha \theta(d_{\mathbf{x}}(\hat{s}_k) - d_{\mathbf{y}}(\hat{s}_k))}{\|\mathbf{x} - \mathbf{y}\|^2} \\ &\quad \times \delta(\hat{s}_k(\mathbf{x}) - \hat{s}_k(\mathbf{y})) [1 - \delta(z_k(\mathbf{x}) - z_k(\mathbf{y}))] \end{aligned} \quad (20)$$

where

$$\theta(x) = \begin{cases} 2, & \text{if } x < 0 \\ 1, & \text{if } x = 0 \\ 0, & \text{otherwise.} \end{cases}$$

The term θ helps to give a penalty on the pixel closer to the boundary in the intensity segmentation field if the two neighboring pixels within an intensity segmentation region do not share the same video segmentation label. It should be noted that $U_{\mathbf{x}, \mathbf{y}}^z |^s$ does not destroy the symmetry of the two-pixel clique potential in MRF. $U_{\mathbf{x}, \mathbf{y}}^z |^s$ is associated with the objective function (19c) and the optimization algorithm. The optimization algorithm updates the label by locally minimizing the objective function at each site. A two-point potential is accounted on both sites. $U_{\mathbf{x}, \mathbf{y}}^z |^s$ is equivalent to $U_{\mathbf{y}, \mathbf{x}}^z |^s$ for the objective function because the total penalty for the entire field is the same. $U_{\mathbf{x}, \mathbf{y}}^z |^s$ is

symmetric and it complies with the definition of MRF. The difference between them occurs in the local minimization of the optimization process. We prefer the form of (20) since, in our experiments, the boundary information is more accurately estimated by giving the entire penalty to the site near the boundary instead of evenly allocating the penalty for both sites in local optimization (see Section IV).

Similarly, in (19b), $U_{\mathbf{x},\mathbf{y}}^z|s$ could be replaced by

$$\begin{aligned} & U_{\mathbf{x},\mathbf{y}}^{uz}|s(\hat{z}_k(\mathbf{x}), \hat{z}_k(\mathbf{y}), s_k(\mathbf{x}), s_k(\mathbf{y})) \\ &= \frac{\alpha\theta(d_{\mathbf{x}}(\hat{z}_k) - d_{\mathbf{y}}(\hat{z}_k))}{\|\mathbf{x} - \mathbf{y}\|^2} \\ & \quad \times \delta(s_k(\mathbf{x}) - s_k(\mathbf{y}))[1 - \delta(\hat{z}_k(\mathbf{x}) - \hat{z}_k(\mathbf{y}))]. \end{aligned} \quad (21)$$

Compared to (13), (21) ignores the first term in (13) since it is constant when the video segmentation field is given.

Thus, we obtain the actual local objective functions that are sequentially optimized at each site

$$\begin{aligned} F_{\mathbf{x}}^d(\mathbf{d}_k) &= U_{\mathbf{x}}^g|d(\mathbf{d}_k(\mathbf{x})) \\ & \quad + \frac{1}{2}\lambda_3 \sum_{\mathbf{y} \in N_{\mathbf{x}}} U_{\mathbf{x},\mathbf{y}}^d|z(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}), \hat{z}_k(\mathbf{x}), \hat{z}_k(\mathbf{y})) \end{aligned} \quad (22a)$$

$$\begin{aligned} F_{\mathbf{x}}^s(s_k) &= \lambda_1 U_{\mathbf{x}}^g|s(s_k(\mathbf{x})) \\ & \quad + \frac{1}{2}\lambda_2 \sum_{\mathbf{y} \in N_{\mathbf{x}}} U_{\mathbf{x},\mathbf{y}}^s(s_k(\mathbf{x}), s_k(\mathbf{y})) \\ & \quad + \frac{1}{2}\lambda_4 \sum_{\mathbf{y} \in N_{\mathbf{x}}} U_{\mathbf{x},\mathbf{y}}^{uz}|s(\hat{z}_k(\mathbf{x}), \hat{z}_k(\mathbf{y}), s_k(\mathbf{x}), s_k(\mathbf{y})) \end{aligned} \quad (22b)$$

$$\begin{aligned} F_{\mathbf{x}}^z(z_k) &= \frac{1}{2}\lambda_3 \sum_{\mathbf{y} \in N_{\mathbf{x}}} U_{\mathbf{x},\mathbf{y}}^d|z(\hat{\mathbf{d}}_k(\mathbf{x}), \hat{\mathbf{d}}_k(\mathbf{y}), z_k(\mathbf{x}), z_k(\mathbf{y})) \\ & \quad + \frac{1}{2}\lambda_4 \sum_{\mathbf{y} \in N_{\mathbf{x}}} U_{\mathbf{x},\mathbf{y}}^t|s(z_k(\mathbf{x}), z_k(\mathbf{y}), \hat{s}_k(\mathbf{x}), \hat{s}_k(\mathbf{y})). \end{aligned} \quad (22c)$$

C. Initialization and Parameters

The intensity segmentation field is initialized using a generalized K-means clustering algorithm to include the spatial constraint. Each intensity segment is characterized by a constant intensity, and the spatial constraint is imposed by two-point clique potential, which actually is a simplification of the adaptive clustering algorithm proposed by Pappas [17]. The motion vector field is initialized by the MAP estimation with pairwise smoothness constraint [21]. Given the initial motion estimation, Wang and Adelson [25] have proposed a procedure for initialization of the video segmentation field. The current frame is divided into small blocks and an affine transformation can be estimated for the motion of each block. A set of motion models is estimated by adaptively clustering the affine parameters. Then, video segmentation labels are assigned in a way that minimizes the motion distortion. In our work, the video segmentation field is initialized by combining this procedure with pairwise spatial constraint on the assignment of regions. For the parameter selection, the idea proposed by Chang *et*

al. is employed [4]. After the initialization of the three fields, the parameters $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are determined by equalizing the contributions of the potentials in objective functions. First, λ_3 is computed by balancing the two potentials in (19a). Then, λ_4 can be calculated by balancing the two potentials in (19c). Finally, λ_1 and λ_2 are determined by balancing the three potentials in (19b). Details can be found in the references.

The parameter α in (13) controls the constraint imposed by the intensity segmentation field. The more comprehensive information of object boundaries is kept in the intensity segmentation, the higher penalty (larger α) should be paid when the object edge in the video segmentation field does not come from the intensity segmentation field. The parameter α is manually determined for individual sequence. In our experiments, the parameter is empirically set as $0.5 \leq \alpha \leq 2$ to achieve robust video segmentation. The neighborhood size also influences the strength of spatiotemporal constraints. The segmentation results will be too noisy or over smoothed if the neighborhood size is excessively small or large. Compared to the 8-pixel (3×3) neighborhood and the 48-pixel (7×7) neighborhood, the 24-pixel (5×5) neighborhood obtains better video segmentation results in our practice. During the optimization process, the Euclidean distance is approximated by the Chamfer distance to simplify the computation of distance transformation [3].

IV. EXPERIMENTS AND DISCUSSION

The results tested on the ‘‘flower garden’’ sequence and the ‘‘table tennis’’ sequence are shown in Figs. 4 and 5. We assume that there are four objects in the video segmentation field. The motion vector field, intensity segmentation field, and the video segmentation field are recovered using the proposed technique for both sequences. The spatial connectivity is clearly exhibited in the estimation results. From the motion vector fields shown in Figs. 4(b) and 5(b), we can see that motion occlusions are successfully overcome. The results of the four-level intensity segmentation are depicted in Figs. 4(c) and 5(c), where an area with constant intensity represents an intensity segment. Figs. 4(d) and 5(d) are the corresponding distance transformed images. Darker gray levels are used to represent the pixels with smaller distance values. In Figs. 4(e)–(h) and 5(e)–(h), we represent the video segmentation results obtained by our approach. In the ‘‘flower garden’’ sequence, the edge information is preserved well in the intensity segmentation field [see Fig. 4(c)]. The algorithm is capable of distinguishing the different objects in the scene by successfully grouping the small regions that are spatiotemporally coherent. While in the ‘‘table tennis’’ sequence, the boundary information lost in the intensity segmentation field [boundary information may be lost even in an oversegmentation, e.g., the boundary between the body and the left arm is lost in Fig. 5(c)] is recovered according to the information from the motion vector field. However, boundaries are detected more accurately when both spatial and temporal features are matched [e.g., the tree in Fig. 4(h) and the body in Fig. 5(f)]. The segmentation algorithm is robust even at the largely homogeneous areas [e.g., the sky in Fig. 4(e) and table in Fig. 5(e)], where there is little motion information.

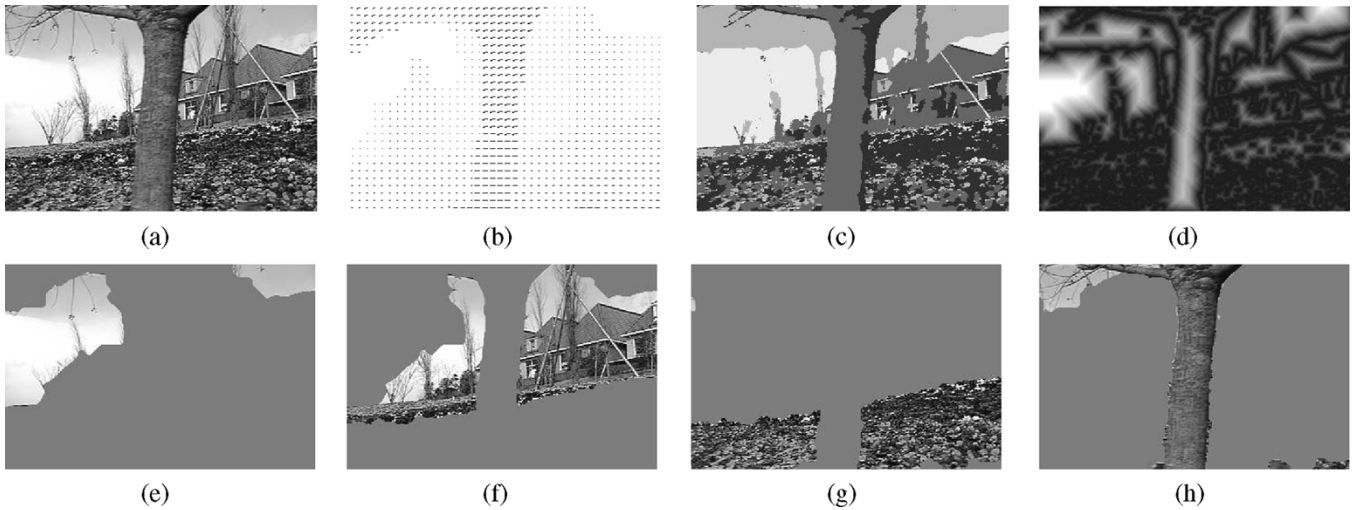


Fig. 4. (a) One frame of the “flower garden” sequence. (b) Motion vector field. (c) Four-level intensity segmentation field. (d) Distance transformed image. (e)–(h) Video segmentation results.

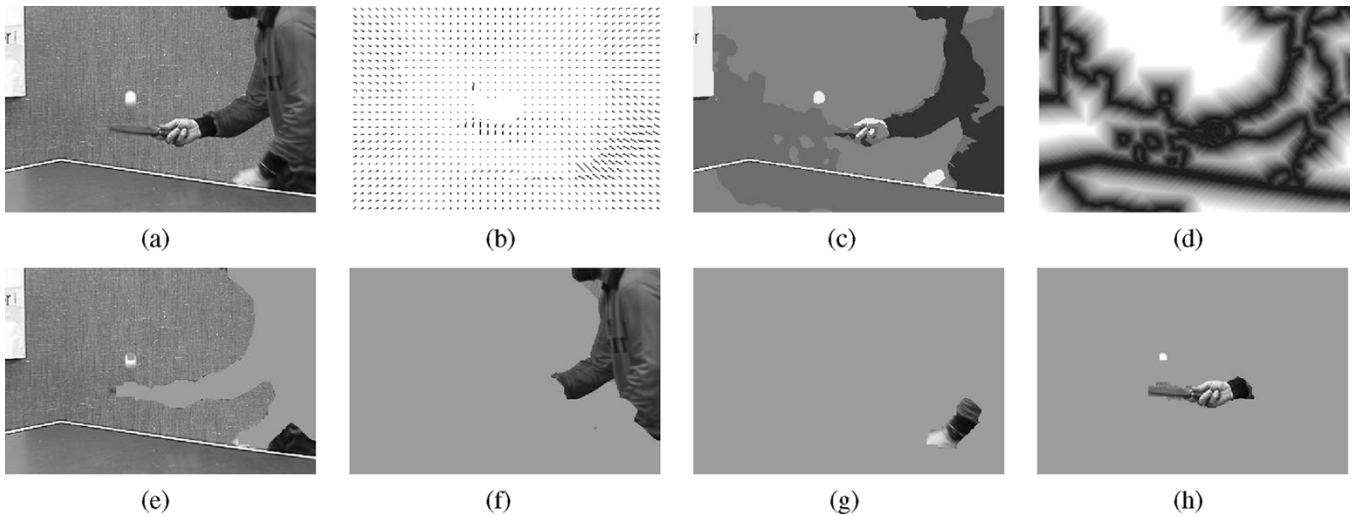


Fig. 5. (a) One frame of the “table tennis” sequence. (b) Motion vector field. (c) Four-level intensity segmentation field. (d) Distance transformed image. (e)–(h) Video segmentation results.

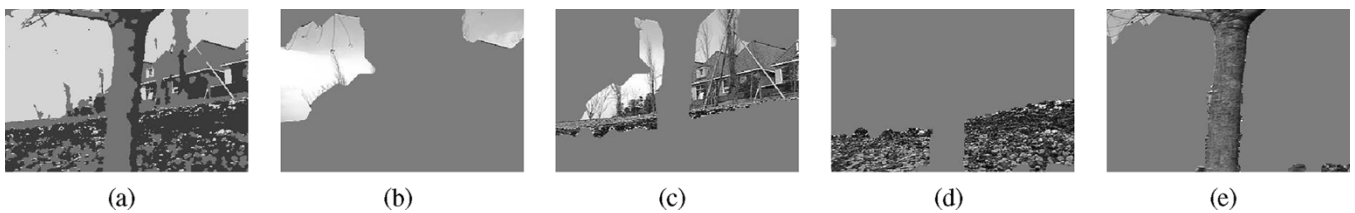


Fig. 6. (a) Three-level intensity segmentation field for the frame of the “flower garden” sequence. (b)–(e) Video segmentation results.

Figs. 6 and 7 show the video segmentation results with three-level and six-level intensity segmentation for the “flower garden” sequence and the “table tennis” sequence, respectively. Comparing with the video segmentation results shown in Figs. 4 and 5, it can be seen that our method is robust to achieve spatiotemporally coherent results without strong requirement of intensity segmentation. Fig. 8 shows part of the video segmentation results using (13) in local objective functions instead of (20) for the two sequences. Comparing with the segmented results in Figs. 4 and 5, it can be known that the utilization

of distance transformation in local optimization substantially improves the boundary accuracy of video segmentation.

Figs. 9–10 show the video segmentation results for the two sequences by simultaneous motion estimation and segmentation [4], and Fig. 11 shows the corresponding Bayesian network model for the motion based method. The method adopts a two-frame approach and does not utilize the constraint from the intensity segmentation field. Compared with Figs. 4–5, both the motion based approach and our approach have successfully identified multiple moving objects composing the scene, but ob-

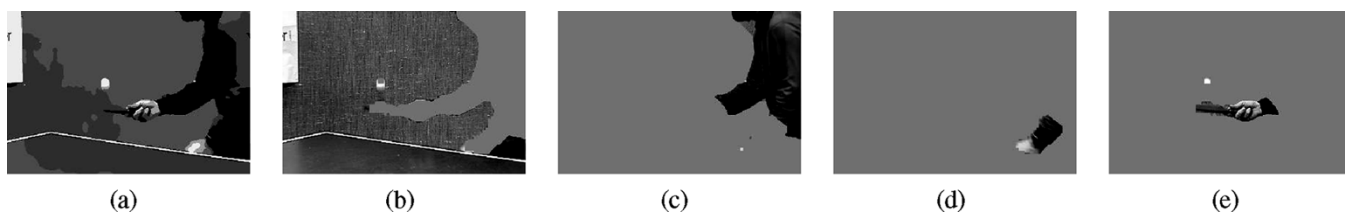


Fig. 7. (a) Six-level intensity segmentation field for the frame of the “table tennis” sequence. (b)–(e) Video segmentation results.

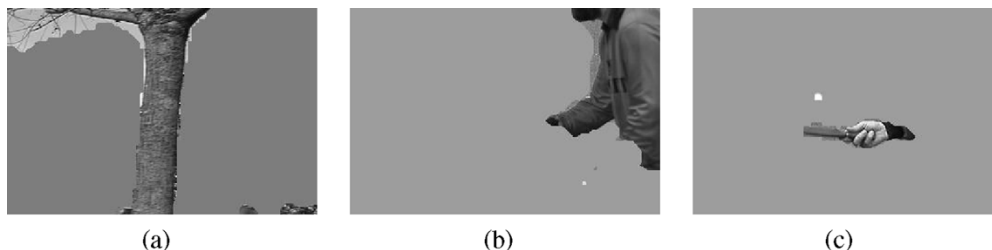


Fig. 8. Video segmentation results without using distance transformation in local optimization for (a) the “flower garden” sequence and (b), (c) the “table tennis” sequence.



Fig. 9. (a)–(d) Video segmentation results by simultaneous motion estimation and segmentation for the “flower garden” sequence.



Fig. 10. (a)–(d) Video segmentation results by simultaneous motion estimation and segmentation for the “table tennis” sequence.

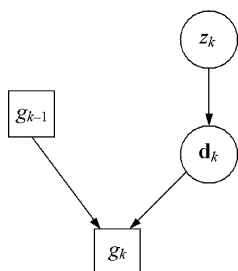


Fig. 11. Bayesian network model for simultaneous motion estimation and segmentation.

ject boundaries are estimated more accurately by the proposed method.

The video segmentation results are also quantitatively evaluated by comparing to manually segmented ground-truth images (see Fig. 12). Table I shows error rates of the results by the

proposed method with three-level, four-level, and six-level intensity segmentation for the two sequences. Table II shows error rates of the results by the proposed method with and without using distance transformation in local optimization. Moreover, Table III compares the error rates of the results by the proposed method and the motion based method. For Tables II–III, the four-level intensity segmentation is used in the proposed method. The quantitative evaluation also indicates that the proposed method effectively improves the video segmentation accuracy.

To test the robustness of the algorithm, Figs. 13–14 show the video segmentation results by the proposed method for another frame of the “flower garden” sequence and the “table tennis” sequence, respectively. Figs. 15–16 show the video segmentation results by the proposed method for the “coastguard” sequence and the “sign” sequence, respectively. In Figs. 13–16, it is assumed that there are three objects in the scene. The motion

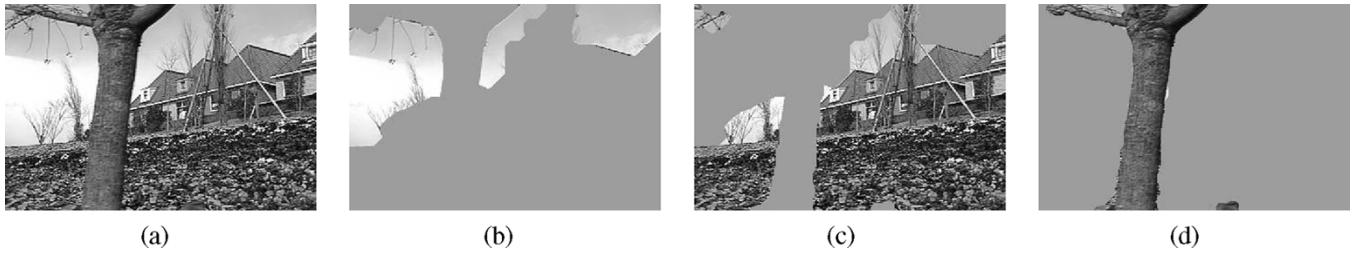


Fig. 13. (a) Another frame of the “flower garden” sequence. (b)–(d) Video segmentation results.



Fig. 12. (a) Ground-truth video segmentation for the frame of the “flower garden” sequence. (b) Ground-truth video segmentation for the frame of the “table tennis” sequence.

TABLE I
ERROR RATES OF THE VIDEO SEGMENTATION RESULTS BY THE PROPOSED METHOD WITH THREE-LEVEL, FOUR-LEVEL, AND SIX-LEVEL INTENSITY SEGMENTATION

	flower garden	table tennis
three-level	12.0%	4.9%
four-level	13.6%	4.6%
six-level	13.0%	4.8%

TABLE II
ERROR RATES OF THE VIDEO SEGMENTATION RESULTS BY THE PROPOSED METHOD WITH AND WITHOUT USING DISTANCE TRANSFORMATION IN LOCAL OPTIMIZATION

	flower garden	table tennis
without distance transformation	14.9%	5.8%
with distance transformation	13.6%	4.6%

TABLE III
ERROR RATES OF THE VIDEO SEGMENTATION RESULTS BY THE PROPOSED METHOD AND THE MOTION BASED METHOD

	flower garden	table tennis
motion based	15.7%	6.2%
proposed	13.6%	4.6%

vector field and the intensity segmentation field for the “sign” sequence are also shown in Fig. 16. The experimental results exhibit satisfactory spatiotemporal coherence.

The intensity segmentation constraint helps generate accurate boundaries in spatiotemporally coherent areas. Since sometimes one area of similar intensity may belong to different objects, the intensity segmentation constraint is weakened when the motion information within one intensity segmentation region is incoherent. This is why boundaries lost in the intensity

segmentation can be recovered by the motion information in our work. As a compromise, the boundary is not anticipated to be accurate in the incoherent area because the intensity segmentation constraint is weak there. Moreover, the incoherence of spatiotemporal information may be caused by the boundary information loss in the intensity segmentation field or the estimation error in the motion vector field. In the worst case, our algorithm will fail in the area where the boundary is lost in intensity segmentation, and the motion is erroneously estimated at the same time [e.g., the segmentation error for the part of the right hand in Fig. 16(e)]. Hence, our approach may not consistently produce accurate edges in the entire video segmentation field. However, the proposed approach has an advantage in application areas where it is important to discover areas with different motions (such as in human-machine interaction and video indexing). Therefore, the new approach is complementary to region merging methods in this aspect.

V. CONCLUSION

In this paper, we have proposed a unified framework for video segmentation based on graphical models. The spatiotemporal consistency of segmentation is expressed in terms of interactions among the motion field, the intensity segmentation field, and the video segmentation field. The solution is obtained by the MAP estimate, and an optimization procedure that iteratively maximizes the conditional probability density of the three fields is proposed. There are three main contributions within the paper. The first is building a Bayesian network based framework that combines both the spatial and temporal information in the video segmentation process. The second is formulating the spatiotemporal constraints by utilizing Markov random fields, distance transformation, and multivariate normal distribution. The third is the theoretical compromise of previous motion based approaches and region merging approaches. The approach deals with video segmentation from a relatively comprehensive and general viewpoint and, thus, can be universally applied. Our method exhibits good robustness and spatiotemporal coherence.

To simplify the computation, we do not consider the localization properties in the sequences. More advanced segmentation techniques that account for both local information and spatiotemporal information could be adopted, but that requires load reduction through efficient optimization schemes [5], [15]. This could be our future study. Moreover, adaptive methods for automatic determination of the number of objects and strength of the spatiotemporal constraints would be beneficial [1].

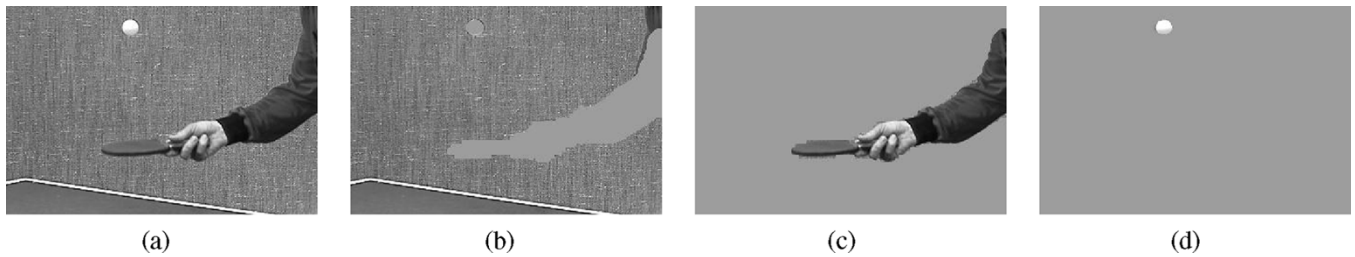


Fig. 14. (a) Another frame of the “table tennis” sequence. (b)–(d) Video segmentation results.



Fig. 15. (a) One frame of the “coastguard” sequence. (b)–(d) Video segmentation results.

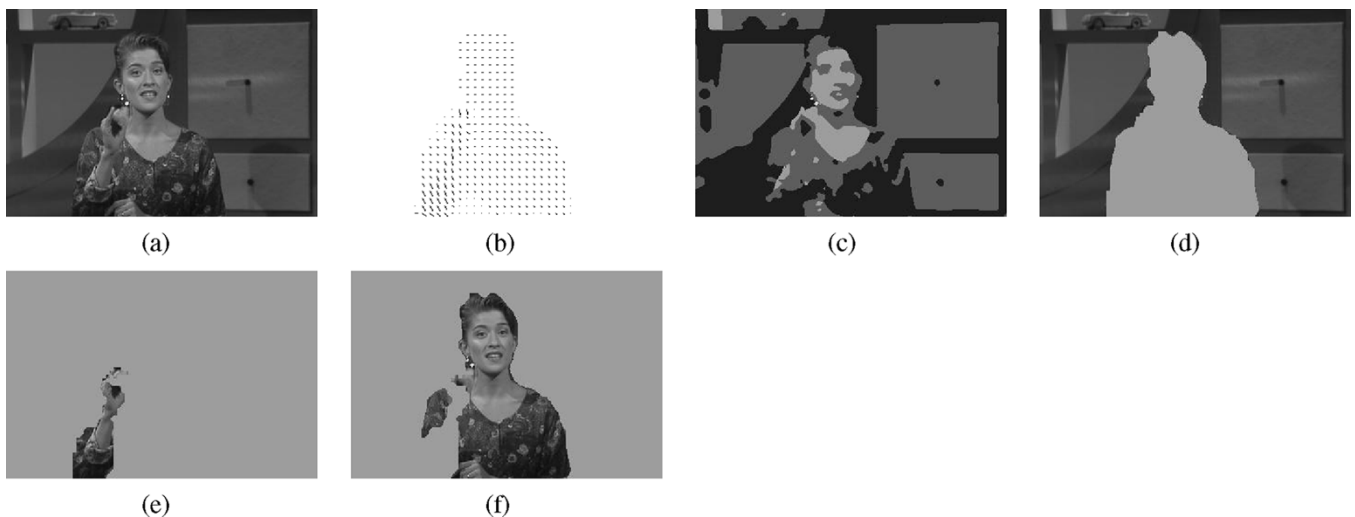


Fig. 16. (a) One frame of the “sign” sequence. (b) Motion vector field. (c) Four-level intensity segmentation field. (d)–(f) Video segmentation results.

REFERENCES

- [1] S. Ayer and H. S. Sawhney, “Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding,” in *Proc. Int. Conf. Computer Vision*, 1995, pp. 777–784.
- [2] J. Besag, “On the statistical analysis of dirty pictures,” *J. Roy. Stat. Soc. B*, vol. 48, pp. 259–302, 1986.
- [3] G. Borgefors, “Distance transformation in digital images,” *Comput. Vis., Graph., Image Process.*, vol. 34, pp. 344–371, 1986.
- [4] M. M. Chang, A. M. Tekalp, and M. I. Sezan, “Simultaneous motion estimation and segmentation,” *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1326–1333, Aug. 1997.
- [5] P. B. Chou and C. M. Brown, “The theory and practice of Bayesian image labeling,” *Int. J. Comput. Vis.*, vol. 4, pp. 185–210, 1990.
- [6] D. DeMenthon and D. Doermann, “Video retrieval using spatio-temporal descriptors,” in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 508–517.
- [7] S. L. Dockstader and A. M. Tekalp, “Multiple camera tracking of interacting and occluded human motion,” *Proc. IEEE*, vol. 89, no. 6, pp. 1441–1455, Jun. 2001.
- [8] P. A. Flach, “On the state of the art in machine learning: A personal review,” *Artif. Intell.*, vol. 131, pp. 199–222, 2001.
- [9] S. Geman and D. Geman, “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 4, pp. 721–741, Apr. 1984.
- [10] M. Gerlgon and P. Bouthemy, “A region-level graph labeling approach to motion-based segmentation,” in *Proc. Conf. Computer Vision and Pattern Recognition*, 1997, pp. 514–519.
- [11] F. V. Jensen, *Bayesian Networks and Decision Graphs*. New York: Springer-Verlag, 2001.
- [12] A. D. Jepson, D. J. Fleet, and M. J. Black, “A layered motion representation with occlusion and compact spatial support,” in *Proc. Eur. Conf. Computer Vision*, 2002, pp. 692–706.
- [13] N. Jovic and B. J. Frey, “Learning flexible sprites in video layers,” in *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 199–206.
- [14] M. I. Jordan, Ed., *Learning in Graphical Models*. Cambridge, MA: MIT Press, 1999.
- [15] S. Z. Li, *Markov Random Field Modeling in Image Analysis*. New York: Springer-Verlag, 2001.
- [16] F. Moscheni, S. Bhattarjee, and M. Kunt, “Spatiotemporal segmentation based on region merging,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 5, pp. 897–915, May 1998.

- [17] T. N. Pappas, "An adaptive clustering algorithm for image segmentation," *IEEE Trans. Image Process.*, vol. 4, no. 5, pp. 901–914, May 1992.
- [18] I. Patras, E. A. Hendriks, and R. L. Lagendijk, "Video segmentation by MAP labeling of watershed segments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 326–332, Feb. 2001.
- [19] P. Salembier and F. Marques, "Region-based representations of image and video: Segmentation tools for multimedia services," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 6, pp. 1147–1169, Aug. 1999.
- [20] C. Stiller, "Object-based estimation of dense motion fields," *IEEE Trans. Image Process.*, vol. 6, no. 2, pp. 234–250, Feb. 1997.
- [21] A. M. Tekalp, *Digital Video Processing*. Upper Saddle River, NJ: Prentice-Hall, 1995.
- [22] P. H. S. Torr, R. Szeliski, and P. Anandan, "An integrated Bayesian approach to layer extraction from image sequences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 297–303, Feb. 2001.
- [23] Y. Tsai and A. Averbuch, "Automatic segmentation of moving objects in video sequences: A region labeling approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 3, pp. 597–612, Apr. 2002.
- [24] N. Vasconcelos and A. Lippman, "Empirical Bayesian motion segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 2, pp. 217–221, Feb. 2001.
- [25] J. Y. A. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Trans. Image Process.*, vol. 3, no. 4, pp. 625–638, Apr. 1994.



Yang Wang (M'03) was born in China in 1976. He received the B.Eng. degree in electronic engineering and the M.Sc. degree in biomedical engineering from Shanghai Jiao Tong University, China, in 1998 and 2001, respectively, and the Ph.D. degree from the Department of Computer Science, National University of Singapore, attached to Institute for Infocomm Research, in 2004.

He is now a Research Scholar with the Intelligent Systems Laboratory, Rensselaer Polytechnic Institute, Troy, NY. He has published about ten

international journal and conference papers. His current research interests are in the area of artificial intelligence and computer vision.

Dr. Wang was awarded the National Excellence Scholarship of China and the President's Graduate Fellowship of Singapore.



Kia-Fock Loe received the Ph.D. degree from the University of Tokyo, Tokyo, Japan.

He is an Associate Professor with the Department of Computer Science, National University of Singapore. His current research interests are in pattern recognition, computer vision, neural networks, machine learning, and uncertainty reasoning.



Tele Tan is Senior Lecturer with the Division of Engineering, Science, and Computing, Curtin University of Technology, Australia, where he is affiliated with the Department of Computing, the Department of Electrical and Computer Engineering, and the Applied Physics Department. His research interests are in human motion analysis, security and surveillance, multimodal system considerations, and technology commercialization. He helped contribute to the original commercialization plan of a biometrics start-up company, XiD Technologies (<http://www.xidtech.com>), in late 2002. The biometrics software developed by the company was nominated for the 2004 World Technology Awards (software category) that was held in conjunction with the World Technology Summit 2004. He was made Technical Advisor to Miltrade Technologies in 2003 and was appointed International Reader with the Australian Research Council (ARC) in mid 2004.



Jian-Kang Wu received the B.Sc. degree from the University of Science and Technology of China and the Ph.D. degree from Tokyo University, Tokyo, Japan.

He is currently the Principal Scientist and Department Manager of New Initiatives at the Institute for Infocomm Research (I²R), Singapore, which is formally known as Kent Ridge Digital Labs (KRDL), Institute of Systems Science (ISS), National University of Singapore. Prior to joining ISS in 1992, he was a Full Professor with the Uni-

versity of Science and Technology of China. He was also with universities in the USA, the U.K., Germany, France, and Japan. He pioneered several researches in the area of visual information processing. This includes adaptive image coding in late 1970s, object-oriented GIS in the early 1980s, face recognition systems in 1992, content-based multimedia indexing and retrieval in the early 1990s, and neuroinformatics and physioinformatics most recently. He initiated and led three large intentional collaboration projects in the 1990s. He is the author of 18 patents, 60+ journal publications, and five books. He developed novel technologies for the authentication of the electronic document and its printed version, and founded Trustcopy Pte, Ltd., a high-tech spinoff from the institute in 2000. He served as Chairman and CEO for the first two years, developed products for online bills of lading, letters of credit, and brand protection labels for high-value wines and pharmaceuticals, which are used by lead players in the market. He secured the investment of \$7M for the company from 3i, the number one investor in Europe.

Dr. Wu received nine distinguished awards from the Ministry of Education and the Ministry of Science of China and the Chinese Academy of Science.