Video Segmentation Based on Graphical Models

Yang Wang   Tele Tan

Institute for Infocomm Research
21 Heng Mui Keng Terrace
Singapore 119613
{ywang, teletan}@i2r.a-star.edu.sg

Kia-Fock Loe

Department of Computer Science
National University of Singapore
Singapore 117543
loekf@comp.nus.edu.sg

Abstract

This paper proposes a unified framework for spatio-temporal segmentation of video sequences. A Bayesian network is presented to model the interactions among the motion vector field, the intensity segmentation field, and the video segmentation field. The notions of distance transformation and Markov random field are used to express spatio-temporal constraints. Given consecutive frames, an optimization method is proposed to maximize the conditional probability density of the three fields in an iterative way. Experimental results show that the approach is robust and generates spatio-temporally coherent segmentation results.

1. Introduction

Robust video segmentation is fundamental to such application areas as object-based video compression and multiple-object tracking. One of the key issues in the design of these vision systems is their ability to differentiate the objects composing the scene. However, the strategy to extract and couple temporal (or motion) information and spatial (or intensity) information with the segmentation process remains an open issue.

Motion information is one important element used for segmentation of video sequences. A moving object is characterized by the coherent motion over its support region. Layered approaches have been proposed to represent multiple moving objects in the scene with a collection of layers [10] [13] [19] [23]. The scene is segmented into a set of regions, such that pixel movements within each region are consistent with a motion model (a parametric transformation). Examples of motion models are the translational model (two parameters), the affine model (six parameters), and the perspective model (nine parameters). Furthermore, spatial constraints could be imposed on motion estimation in the form of a support region where the motion is assumed to follow a parametric transformation. Chang et al. [4] and Stiller [17] used the methods to simultaneously estimate the motion information and its support region. Moreover, spatial information provides important hints of object boundaries. Methods that combine an initial intensity segmentation with motion information have been proposed recently [12] [15] [20]. Given an oversegmentation of the current frame, objects are formed by merging together segments with spatio-temporal similarity. The region merging approaches have two disadvantages. Firstly, the intensity segmentation remains unchanged so that motion information has no influence upon the spatial information during the entire procedure. Secondly, even an oversegmentation sometimes cannot keep all the object edges, and the boundary information lost by the initial intensity segmentation cannot be recovered later. Since spatial information and temporal information should interact throughout the segmentation process, to utilize only motion information or fix intensity segmentation will degrade the performance of video segmentation. From this point of view, it is a relatively comprehensive idea to simultaneously adjust the motion vector field, the intensity segmentation field, and the spatio-temporal (or video) segmentation field.

On the other hand, graphical probabilistic models provide a natural tool for dealing with uncertainty and complexity, and they are playing an increasingly important role in the design and analysis of machine intelligent systems [7]. In particular, Markov random fields and Bayesian networks have attracted more and more attention as principled approaches for image and video processing [6] [16] [21].

In this paper, we present an approach 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 spatio-temporal segmentation field. The notions of distance transformation and Markov random field (MRF) are employed to express spatio-temporal constraints. A three-frame approach is adopted to deal with occlusions. The labeling criterion is the maximization of conditional probability density 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. Experiments show that our technique is robust and generates spatio-temporally consistent segmentation results. The rest of the paper is arranged as follows: Section 2 presents the formulation of our approach and compares the method with related work. Section 3 proposes the implementation details. Section 4 discusses the experiment results. Our technique is concluded in Section 5.

2. Method

2.1. Model representation

For a discrete image sequence, assume that the intensity distribution remains constant along a motion trajectory. Ignoring both illumination variations and occlusions, it may be stated as

$$y_k(\mathbf{x}) = y_{k-1}(\mathbf{x} - \mathbf{d}_k(\mathbf{x})),\ \forall \mathbf{x} \in \mathbf{X},\ k = 1, 2, \ldots, \qquad (1)$$

where $y_k(\mathbf{x})$ is the intensity of a single pixel within the kth video frame at spatial location $\mathbf{x}$, and $\mathbf{d}_k(\mathbf{x})$ is the displacement vector from frame k–1 to frame k. $\mathbf{X}$ is the spatial domain of each video frame. The entire motion vector field is expressed compactly as $\mathbf{d}_k$.

Since the video data is corrupted in the image acquisition process, an observation model is required for the image sequence. Assume that independent and identically distributed (i. i. d.) Gaussian noise corrupts each pixel, so that the observation model for the kth frame becomes

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

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 this work, video segmentation refers to grouping pixels that belong to independently moving objects in the frame. 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 to segment the video sequence. Given consecutive frames of the observed video sequence, g_{k-1}, g_k, and g_{k+1}, we wish to compute the maximum *a posteriori* (MAP) estimation of the displacement vector field $\mathbf{d}_k$, the intensity segmentation field s_k, and the spatio-temporal segmentation field z_k.

$$(\hat{\mathbf{d}}_k, \hat{s}_k, \hat{z}_k) = \arg \max_{(\mathbf{d}_k, s_k, z_k)} p(\mathbf{d}_k, s_k, z_k \mid g_k, g_{k-1}, g_{k+1}), \qquad (3)$$

where $p(\mathbf{d}_k, s_k, z_k \mid g_k, g_{k-1}, g_{k+1})$ is the posterior probability density function (pdf) given the three video frames. Using the Bayes' rule, the posterior probability density becomes

$$p(\mathbf{d}_k, s_k, z_k \mid g_k, g_{k-1}, g_{k+1}) = \frac{p(\mathbf{d}_k, s_k, z_k, g_k, g_{k-1}, g_{k+1})}{p(g_k, g_{k-1}, g_{k+1})}, \qquad (4)$$

where the denominator 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} can be modeled using the Bayesian network shown in Figure 1. Motion estimation establishes the pixel correspondence between the three consecutive frames. The intensity segmentation field provides a set of segments with relatively small intensity variation. In order to identify independently moving objects in the scene, these segments are encouraged to group into regions with the same parametric transformation. Meanwhile, if multiple motion models coexist within one segment, the segment may split into several spatio-temporally coherent regions.

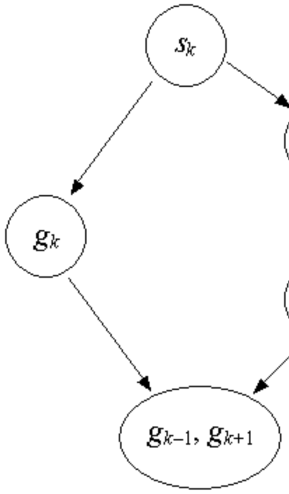

Figure 1. Bayesian belief network model for video segmentation.

The conditional independence relationships implied by the Bayesian network allow us to represent the joint more compactly [9]. Using the chain rule, the joint probability density can be factorized as

$$p(\mathbf{d}_k, s_k, z_k, g_k, g_{k-1}, g_{k+1}) = p(g_{k-1}, g_{k+1} \mid g_k, \mathbf{d}_k) \cdot p(g_k \mid s_k)\, p(s_k)\, p(\mathbf{d}_k \mid z_k)\, p(z_k \mid s_k). \qquad (5)$$

Then the MAP estimate becomes

$$(\hat{\mathbf{d}}_k, \hat{s}_k, \hat{z}_k) = \arg \max_{(\mathbf{d}_k, s_k, z_k)} p(g_{k-1}, g_{k+1} \mid g_k, \mathbf{d}_k) \cdot p(g_k \mid s_k)\, p(s_k)\, p(\mathbf{d}_k \mid z_k)\, p(z_k \mid s_k). \qquad (6)$$

2.2. Spatio-temporal constraints

The conditional probability density function $p(g_{k-1}, g_{k+1} \mid g_k, \mathbf{d}_k)$ qualifies how well the motion estimation fits the given frames. Assuming that the likelihood is completely specified by a random field that models the displaced frame difference (DFD) [18], the video observation model can be employed to compute $p(g_{k-1}, g_{k+1} \mid \mathbf{d}_k, g_k)$. We can define the backward DFD $e_k^b(\mathbf{x})$ and forward DFD $e_k^f(\mathbf{x})$ at site $\mathbf{x}$ as

$$e_k^b(\mathbf{x}) = g_k(\mathbf{x}) - g_{k-1}(\mathbf{x} - \mathbf{d}_k(\mathbf{x}))$$
$$= n_k(\mathbf{x}) - n_{k-1}(\mathbf{x} - \mathbf{d}_k(\mathbf{x})), \quad (7a)$$
$$e_k^f(\mathbf{x}) = g_k(\mathbf{x}) - g_{k+1}(\mathbf{x} + \mathbf{d}_k(\mathbf{x}))$$
$$= n_k(\mathbf{x}) - n_{k+1}(\mathbf{x} + \mathbf{d}_k(\mathbf{x})). \quad (7b)$$

The vector $(e_k^b(\mathbf{x}), e_k^f(\mathbf{x}))^T$ is denoted as $\mathbf{e}_k(\mathbf{x})$. Then the likelihood can be modeled using the bivariate normal distribution.

$$p(g_{k-1}, g_{k+1} \mid g_k, \mathbf{d}_k)$$
$$= \prod_{\mathbf{x}\in\mathbf{X}} p(\mathbf{e}_k(\mathbf{x}))$$
$$= \left(\frac{1}{2\pi\sqrt{|\mathbf{\Sigma}_\mathbf{e}|}}\right)^{|\mathbf{X}|} \exp\left\{\sum_{\mathbf{x}\in\mathbf{X}} -\frac{1}{2}[\mathbf{e}_k^T(\mathbf{x})\mathbf{\Sigma}_\mathbf{e}^{-1}\mathbf{e}_k(\mathbf{x})]\right\}$$
$$\propto \exp[-\sum_{\mathbf{x}\in\mathbf{X}} U_1(\mathbf{x} \mid \mathbf{d}_k(\mathbf{x}))], \quad (8a)$$
$$U_1(\mathbf{x} \mid \mathbf{d}_k(\mathbf{x})) = (e_k^b(\mathbf{x}))^2 - 2\rho \cdot e_k^b(\mathbf{x}) e_k^f(\mathbf{x}) + (e_k^f(\mathbf{x}))^2, \quad (8b)$$

where $\mathbf{\Sigma}_\mathbf{e}$ is the covariance matrix for each site $\mathbf{x}$, and ρ is the correlation coefficient of $e_k^b(\mathbf{x})$ and $e_k^f(\mathbf{x})$. With the i. i. d. Gaussian noise assumption, we have

$$\rho = \frac{\mathrm{Cov}[e_k^b(\mathbf{x}), e_k^f(\mathbf{x})]}{\sqrt{\mathrm{Var}[e_k^b(\mathbf{x})]\mathrm{Var}[e_k^f(\mathbf{x})]}} = \frac{\sigma_n^2}{2\sigma_n^2} = \frac{1}{2}. \quad (9)$$

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

$$p(g_k \mid s_k)$$
$$= \left(\frac{1}{\sqrt{2\pi}\sigma_\eta}\right)^{|\mathbf{X}|} \exp\left\{-\sum_{\mathbf{x}\in\mathbf{X}} \frac{1}{2\sigma_\eta^2}[g_k(\mathbf{x}) - \mu_{s_k(\mathbf{x})}]^2\right\}$$
$$\propto \exp[-\sum_{\mathbf{x}\in\mathbf{X}} U_2(\mathbf{x} \mid s_k(\mathbf{x}))], \quad (10a)$$
$$U_2(\mathbf{x} \mid s_k(\mathbf{x})) = [g_k(\mathbf{x}) - \mu_{s_k(\mathbf{x})}]^2, \quad (10b)$$

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

The pdf $p(s_k)$ represents the *a priori* probability of the intensity segmentation. We model the density $p(s_k)$ by a Markov random field [8]. That is, if $N_\mathbf{x}$ is a neighborhood of the pixel at $\mathbf{x}$, then the conditional distribution of a single variable at $\mathbf{x}$ is completely specified by the variables within its neighborhood $N_\mathbf{x}$. According to the Hammersley-Clifford theorem, the density is given by a Gibbs density that has the following form [18]:

$$p(s_k) \propto \exp[-\sum_{c\in C} V_c^s(s_k(\mathbf{x}) \mid \mathbf{x} \in c)], \quad (11)$$

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

Spatial connectivity can be imposed by the following two-pixel clique potential.

$$V_c^s(s_k(\mathbf{x}), s_k(\mathbf{y}))$$
$$= \frac{1}{\|\mathbf{x} - \mathbf{y}\|^2}[1 - \delta(s_k(\mathbf{x}) - s_k(\mathbf{y}))], \quad (12)$$

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

The term $p(\mathbf{d}_k \mid z_k)$ is the conditional pdf of the displacement field given the video segmentation field. To encourage the formation of continuous regions, it is modeled by a Gibbs distribution with the following potential function.

$$V_c^{\mathbf{d}|z}(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}) \mid z_k)$$
$$= V_c^{\mathbf{d}|z}(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}) \mid z_k(\mathbf{x}), z_k(\mathbf{y}))$$
$$= \frac{1}{\|\mathbf{x} - \mathbf{y}\|^2}\delta(z_k(\mathbf{x}) - z_k(\mathbf{y}))\,\|\mathbf{d}_k(\mathbf{x}) - \mathbf{d}_k(\mathbf{y})\|^2. \quad (13)$$

The piecewise smoothness constraint of the motion vectors is imposed only when the two pixels have the same video segmentation label.

The last term $p(z_k \mid s_k)$ represents the probability density of the spatio-temporal segmentation field when the intensity segmentation field is given. To employ the spatial information, distance transformation [1] is performed on the intensity segmentation field. Each pixel $\mathbf{x}$ in the distance transformed image has a value $DT_k(\mathbf{x})$ 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 density is modeled by a Gibbs distribution with the following potential function.

$$V_c^{z|s}(z_k(\mathbf{x}), z_k(\mathbf{y}) \mid s_k)$$
$$= V_c^{z|s}(z_k(\mathbf{x}), z_k(\mathbf{y}) \mid DT_k(\mathbf{x}), DT_k(\mathbf{y}))$$
$$= \frac{1}{\|\mathbf{x} - \mathbf{y}\|^2}[1 - \delta(z_k(\mathbf{x}) - z_k(\mathbf{y}))] \cdot$$
$$[1 + \alpha\theta(DT_k(\mathbf{x}), DT_k(\mathbf{y}))], \quad (14a)$$
$$\theta(DT_k(\mathbf{x}), DT_k(\mathbf{y})) = \begin{cases} 1, \text{if } DT_k(\mathbf{x}) < DT_k(\mathbf{y}). \\ 0, \text{otherwize.} \end{cases} \quad (14b)$$

The first term on the right side of (14a) encourages the spatial connectivity, while the second term gives a penalty on the pixel closer to the boundary of the intensity segmentation field if the two pixels are not of the same video segmentation class. The parameter α controls the strength of the constraint imposed by the intensity segmentation field.

Combining the above models, the Bayesian MAP estimation criterion becomes

$$\begin{aligned}
&(\hat{\mathbf{d}}_k, \hat{s}_k, \hat{z}_k) \\
&= \arg\min_{(\mathbf{d}_k, s_k, z_k)} \Big[\sum_{\mathbf{x}\in\mathbf{X}} U_1(\mathbf{x} \,|\, \mathbf{d}_k(\mathbf{x})) + \\
&\quad \lambda_1 \sum_{\mathbf{x}\in\mathbf{X}} U_2(\mathbf{x} \,|\, s_k(\mathbf{x})) + \\
&\quad \lambda_2 \sum_{\{\mathbf{x},\mathbf{y}\}\in C} V_c^s(s_k(\mathbf{x}), s_k(\mathbf{y})) + \\
&\quad \lambda_3 \sum_{\{\mathbf{x},\mathbf{y}\}\in C} V_c^{\mathbf{d}|z}(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}) \,|\, z_k(\mathbf{x}), z_k(\mathbf{y})) + \\
&\quad \lambda_4 \sum_{\{\mathbf{x},\mathbf{y}\}\in C} V_c^{z|s}(z_k(\mathbf{x}), z_k(\mathbf{y}) \,|\, DT_k(\mathbf{x}), DT_k(\mathbf{y})) \Big], \quad (15)
\end{aligned}$$

where the parameters λ_1, λ_2, λ_3, and λ_4 control the contribution of the terms.

2.3. Notes on the Bayesian network

In our model, the motion vector field establishes the correspondence between the current frame and its two neighboring frames. The video segmentation is influenced by both spatial information and temporal information. It should be noted that the direction of the links in the Bayesian network model does not mean that the action between the cause and consequence is one-way.

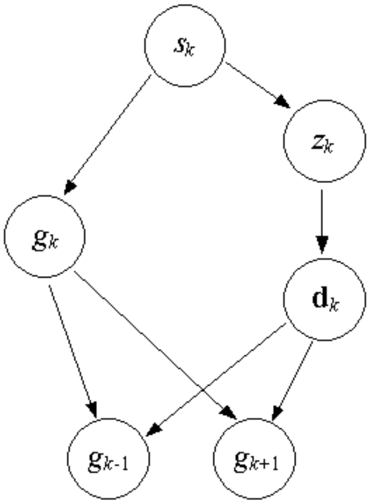

Figure 2. Simplified Bayesian belief network model for video segmentation.

In a video sequence, the current frame could be thought as the cause of the next frame. For most of sequences (including the two test sequences in Section 4), both the original sequence and the one in reverse order are understandable from the viewpoint of segmentation. 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}.

When frame g_{k+1} and frame g_{k-1} are separated (as shown in Figure 2), the model seems more clear at the first glance. However, from the chain rule we know that in this case,

$$\begin{aligned}
&p(g_{k-1}, g_{k+1} \,|\, g_k, \mathbf{d}_k) \\
&= p(g_{k-1} \,|\, g_k, \mathbf{d}_k)\, p(g_{k+1} \,|\, g_k, \mathbf{d}_k) \\
&= \prod_{\mathbf{x}\in\mathbf{X}} p(e_k^b(\mathbf{x}))\, p(e_k^f(\mathbf{x})) \\
&\propto \exp\{-\sum_{\mathbf{x}\in\mathbf{X}} [(e_k^b(\mathbf{x}))^2 + (e_k^f(\mathbf{x}))^2]\}\,. \qquad (16)
\end{aligned}$$

Comparing with (8), the correlation coefficient of $e_k^b(\mathbf{x})$ and $e_k^f(\mathbf{x})$ is zero in (16). The Bayesian belief network in Figure 2 neglects the interaction between the forward DFD and the backward DFD. Therefore, the Bayesian network model in Figure 2 is just a simplification of the original model.

2.4. Related work

Our method is mostly related to the work of Chang et al. [4] and Patras et al. [15], and it could be viewed as the generalization of the former one. 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 information of 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 keeps the disadvantages of region merging approaches. The boundary information neglected by the initial intensity segmentation field could no longer be recovered by the motion vector field, and the temporal information could not act on the spatial information.

In order to overcome the above problems, our algorithm simultaneously estimates the three fields to form spatio-temporal coherent results. Described by the Bayesian network model in Figure 1, the interaction between spatial information and temporal information is bi-directional. Boundaries of the video segmentation field are supplied by both the intensity segmentation field and the motion vector field.

3. Implementation

3.1. Optimization

Obviously, there is no simple method of directly minimizing (15) with respect to all unknowns. We perform the minimization by iterating over the following two steps.

Firstly, 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 video segmentation field z_k and the three successive frames are given. The joint estimation can be factorized as

$$
\begin{aligned}
(\hat{\mathbf{d}}_k, \hat{s}_k) &= \arg\max_{(\mathbf{d}_k, s_k)} p(\mathbf{d}_k, s_k \mid g_k, g_{k-1}, g_{k+1}, z_k) \\
&= (\arg\max_{\mathbf{d}_k} p(\mathbf{d}_k \mid g_k, g_{k-1}, g_{k+1}, z_k), \\
&\quad \arg\max_{s_k} p(s_k \mid g_k, z_k)). \qquad (17)
\end{aligned}
$$

Using the chain rule, the MAP estimate becomes

$$
\begin{aligned}
\hat{\mathbf{d}}_k &= \arg\max_{\mathbf{d}_k} p(\mathbf{d}_k \mid g_k, g_{k-1}, g_{k+1}, z_k) \\
&= \arg\max_{\mathbf{d}_k} p(g_{k-1}, g_{k+1} \mid g_k, \mathbf{d}_k)\, p(\mathbf{d}_k \mid z_k), \qquad (18a) \\
\hat{s}_k &= \arg\max_{s_k} p(s_k \mid g_k, z_k) \\
&= \arg\max_{s_k} p(g_k \mid s_k)\, p(z_k \mid s_k) p(s_k). \qquad (18b)
\end{aligned}
$$

Secondly, update the spatio-temporal segmentation field z_k, assuming the motion field $\mathbf{d}_k$ and the intensity segmentation field s_k are given.

$$
\begin{aligned}
\hat{z}_k &= \arg\max_{z_k} p(z_k \mid g_k, g_{k-1}, g_{k+1}, \mathbf{d}_k, s_k) \\
&= \arg\max_{z_k} p(z_k \mid \mathbf{d}_k, s_k) \\
&= \arg\max_{z_k} p(\mathbf{d}_k \mid z_k)\, p(z_k \mid s_k). \qquad (19)
\end{aligned}
$$

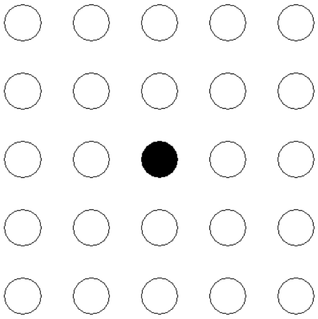

Figure 3. The fifth order neighborhood system.

In our work, the 24-point neighborhood system (the fifth order neighbor system, see Figure 3) is used, and potentials are defined only on two-point cliques. Using the terms in (15), the Bayesian MAP estimates in (18) and (19) can be obtained by minimizing the following objective functions.

$$
\begin{aligned}
F_{\mathbf{d}}(\mathbf{d}_k) &= \sum_{\mathbf{x}\in\mathbf{X}} [U_1(\mathbf{x} \mid \mathbf{d}_k(\mathbf{x})) + \\
&\quad \frac{1}{2}\lambda_3 \sum_{\mathbf{y}\in N_{\mathbf{x}}} V_c^{\mathbf{d}|z}(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}) \mid z_k(\mathbf{x}), z_k(\mathbf{y}))], \qquad (20a) \\
F_s(s_k) &= \sum_{\mathbf{x}\in\mathbf{X}} [\lambda_1 U_2(\mathbf{x} \mid s_k(\mathbf{x})) + \\
&\quad \frac{1}{2}\lambda_2 \sum_{\mathbf{y}\in N_{\mathbf{x}}} V_c^{s}(s_k(\mathbf{x}), s_k(\mathbf{y})) + \\
&\quad \frac{1}{2}\lambda_4 \sum_{\mathbf{y}\in N_{\mathbf{x}}} V_c^{z|s}(z_k(\mathbf{x}), z_k(\mathbf{y}) \mid DT_k(\mathbf{x}), DT_k(\mathbf{y}))], (20b) \\
F_z(z_k) &= \sum_{\mathbf{x}\in\mathbf{X}} [\frac{1}{2}\lambda_3 \sum_{\mathbf{y}\in N_{\mathbf{x}}} V_c^{\mathbf{d}|z}(\mathbf{d}_k(\mathbf{x}), \mathbf{d}_k(\mathbf{y}) \mid z_k(\mathbf{x}), z_k(\mathbf{y})) + \\
&\quad \frac{1}{2}\lambda_4 \sum_{\mathbf{y}\in N_{\mathbf{x}}} V_c^{z|s}(z_k(\mathbf{x}), z_k(\mathbf{y}) \mid DT_k(\mathbf{x}), DT_k(\mathbf{y}))], (20c)
\end{aligned}
$$

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

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 sub-optimal estimate of each objective function [3]. The scheme employs the greedy strategy in the iterative local minimization. Given the observed data and the other labels, the algorithm sequentially updates the label by locally minimizing the objective function at each site.

3.2. Initialization and parameter determination

The intensity segmentation field is initialized using a generalized K-means clustering algorithm to include the spatial constraint. Each cluster is characterized by a constant intensity, and the spatial constraints are performed by the two-point clique potential in (12). The initialization algorithm is actually a simplification of the adaptive clustering algorithm proposed by Papps [14]. The initial motion vector field is obtained by using Bayesian MAP estimation with a global smoothness constraint [18]. Given the initial motion estimates, Wang and Adelson [22] have proposed a procedure for initialization of the video segmentation field. The current frame is divided into small blocks and a set of affine parameters is computed for each block. By adaptively clustering the affine parameters under a distance measure, a set of motion models is known. Then regions within the image are assigned in a way that minimizes the motion distortion. In our work, the video segmentation field is initialized by combining this procedure with the spatial constraint on the assignment of regions. Given the initial estimates of the three fields, we employ the idea for parameter selection proposed by Chang et al. [4]. The parameters (λ_1, λ_2, λ_3, and λ_4) are determined by equalizing the contributions of the terms in (15). Details can be found in the references.

4. Results and discussion

The results tested on the "flower garden" sequence and the "table tennis" sequence (see Figure 4 and 5) are shown in our experiments. We assume that there are four layers in the video segmentation field. The value of α in (14a) is set as 4.

The motion vector field, intensity segmentation field, and the spatio-temporal 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 Figure 4c and 5c, we can see that motion occlusions are successfully overcome. The results of the intensity segmentation are depicted in Figure 4d and 5d, where an area with constant intensity represents an intensity segment. Figure 4e and 5e are the corresponding distance transformed images. Darker gray levels are used to represent the pixels with smaller distance values. In Figure 4f-i and 5f-i, we represent the video segmentation results obtained by our approach. In the "flower garden" sequence, the edge information is preserved well in intensity segmentation field (see Figure 4d). The algorithm is capable of distinguishing the different objects in the scene by successfully grouping the small regions that are spatio-temporally coherent. While in the "table tennis" sequence, the boundary information lost in Figure 5d (boundary information may be lost even in an oversegmentation, e.g., the left arm in Figure 5i) 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 Figure 4i and the body in Figure 5h). The segmentation algorithm is robust even at the largely homogeneous areas (e.g., the sky in Figure 4h and table in Figure 5f), where there is little motion information.

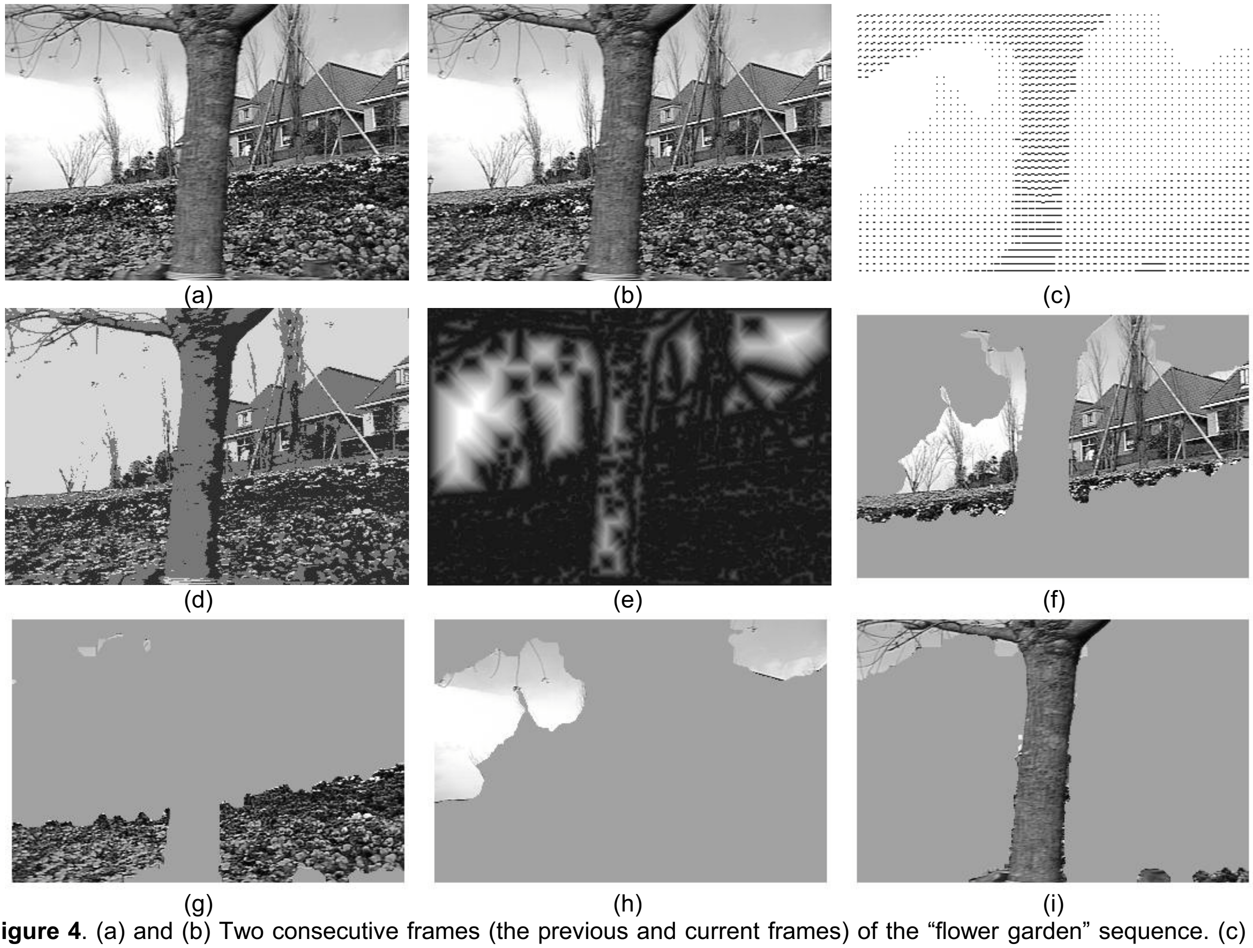

Figure 4. (a) and (b) Two consecutive frames (the previous and current frames) of the "flower garden" sequence. (c) The motion vector field. (d) The intensity segmentation field. (e) The distance transformed image. (f)-(i) The video segmentation results.

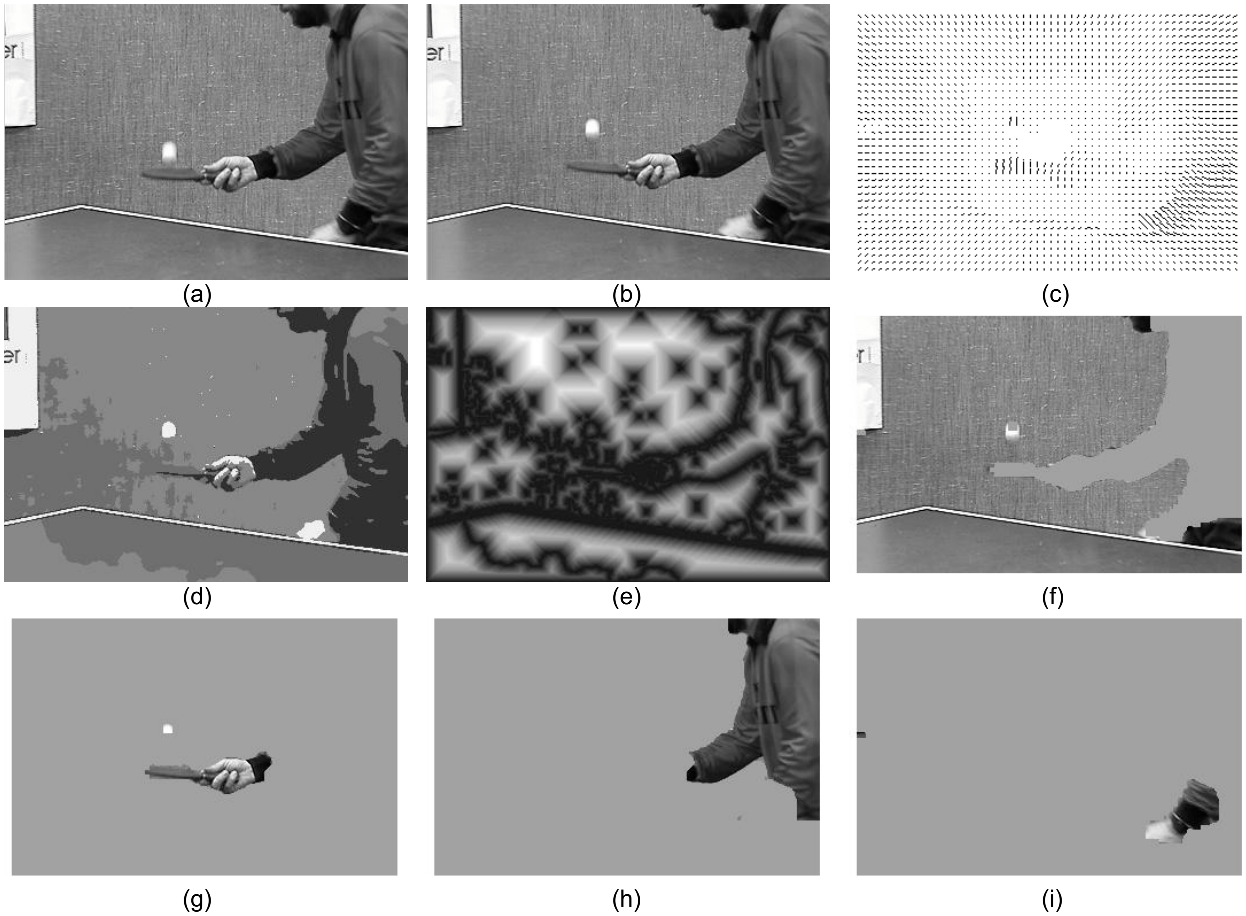

Figure 5. (a) and (b) Two consecutive frames (the previous and current frames) of the “table tennis” sequence. (c) The motion vector field. (d) The intensity segmentation field. (e) The distance transformed image. (f)-(i) The video segmentation results.

It should be noted that eq. (14b) does not destroy the symmetry of the two-pixel clique potential in MRF. (14b) is associated with the objective function (20) and the optimization algorithm in Section 3.1. The optimization algorithm updates the label by locally minimizing the objective function at each site. A two-point potential is accounted on both sites. (14b) is equivalent to the following $\theta'(DT_k(\mathbf{x}), DT_k(\mathbf{y}))$ for the objective function.

$$\theta'(DT_k(\mathbf{x}), DT_k(\mathbf{y})) = \frac{1}{2}[1-\delta(DT_k(\mathbf{x})-DT_k(\mathbf{y}))]. \quad (21)$$

(21) is symmetric and it complies with the definition of MRF. (21) and (14b) are equivalent for the object function because the total penalty for the entire field (or the objective function) is the same. The difference between them occurs in the local minimization of the optimization process. We prefer the form of (14b) since in our experiments, we found that the convergence can be fastened by giving all the penalty to the site near the boundary (see (14b)) instead of evenly allocating the penalty for both sites (see (21)).

The intensity segmentation constraint helps generate accurate boundaries in spatio-temporally coherent areas. Since one area of similar intensity may belong to different objects, the intensity segmentation constraint becomes weak when the motion information in an intensity segment 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. Our approach may not consistently produce accurate segmentation edges in the entire field. However, the 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.

5. Conclusion

In this paper, we have proposed a unified framework for segmentation of video sequences. The spatial and temporal consistency is expressed in terms of interactions between the motion field, the intensity segmentation field, and the video segmentation field. The solution is obtained by the MAP criteria and an optimization strategy that iteratively maximizes the conditional probability density of the three fields is proposed. There are two main contributions within the paper. The first is building a belief network based framework that combines both the spatial and temporal information in the video segmentation process. The second is to formulate the spatio-temporal constraints by utilizing distance transformation, Markov random fields, and multivariate normal distribution. 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 spatio-temporal 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 spatio-temporal information could be adopted, but that requires load reduction through efficient optimization schemes [5] [11]. This could be our future study. Moreover, adaptive methods for automatic determination of the number of layers and selection of the parameters would be beneficial [2].

6. References

[1] C. Arcelli and G. S. di Baja, "Ridge points in Euclidean distance maps," *Patt. Recognit. Lett.*, vol. 13, pp. 237-243, 1992.

[2] S. Ayer and H. S. Sawhney, "Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding," *Proc. International Conf. Computer Vision*, pp. 777-784, 1995.

[3] J. Besag, "On the statistical analysis of dirty pictures," *J. Roy. Stat. Soc. B*, vol. 48, pp. 259-302, 1986.

[4] M. M. Chang, A. M. Tekalp, and M. I. Sezan, "Simultaneous motion estimation and segmentation," *IEEE Trans. Image Processing*, vol. 6, pp. 1326-1333, 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] S. L. Dockstader and A. M. Tekalp, "Multiple camera tracking of interacting and occluded human motion," *Proc. IEEE*, vol. 89, pp. 1441-1455, 2001.

[7] P. A. Flach, "On the state of the art in machine learning: a personal review," *Artificial Intelligence*, vol. 131, pp. 199-222, 2001.

[8] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 6, pp. 721-741, 1984.

[9] F. V. Jensen, *Bayesian Networks and Decision Graphs*, Springer-Verlag, 2001.

[10] A. D. Jepson, D. J. Fleet, and M. J. Black, "A layered motion representation with occlusion and compact spatial support," *Proc. European Conf. Computer Vision*, pp. 692-706, 2002.

[11] S. Z. Li, *Markov Random Field Modeling in Computer Vision*, Springer-Verlag, 1995.

[12] F. Moscheni, S. Bhattacharjee, and M. Kunt, "Spatiotemporal segmentation based on region merging," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 20, pp. 897-915, 1998.

[13] N. Jojic and B. J. Frey, "Learning flexible sprites in video layers," *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 199-206, 2001.

[14] T. N. Papps, "An adaptive clustering algorithm for image segmentation," *IEEE Trans. Image Processing*, vol. 4, pp. 901-914, 1992.

[15] I. Patras, E. A. Hendriks, and R. L. Lagendijk, "Video segmentation by MAP labeling of watershed segments," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 23, pp. 326-332, 2001.

[16] C. S. Regazzoni and A. N. Venetsanopoulos, "Group-membership reinforcement for straight edges based on Bayesian networks," *IEEE Trans. Image Processing*, vol. 7, pp. 1321-1339, 1998.

[17] C. Stiller, "Object-based estimation of dense motion fields," *IEEE Trans. Image Processing*, vol. 6, pp. 234-250, 1997.

[18] A. M. Tekalp, *Digital Video Processing*, Prentice Hall, 1995.

[19] P. H. S. Torr, R. Szeliski, and P. Anandan, "An integrated Bayesian approach to layer extraction from image sequences," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 23, pp. 297-303, 2001.

[20] Y. Tsaig and A. Averbuch, "Automatic segmentation of moving objects in video sequences: a region labeling approach," *IEEE Trans. Circuit Sys. Video Technol.*, vol. 12, pp. 597-612, 2002.

[21] N. Vasconcelos and A. Lippman, "Empirical Bayesian motion segmentation," *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 23, pp. 217-221, 2001.

[22] J. Y. A. Wang and E. H. Adelson, "Representing moving images with layers," *IEEE Trans. Image Processing*, vol. 3, pp. 625-637, 1994.

[23] C. K. I. Williams and M. K. Titsias, "Learning about multiple objects in images: factorial learning without factorial search," *Proc. Advances in Neural Information Processing Systems* (to be appeared), 2002.