A Dynamic Conditional Random Field Model for Object Segmentation in Image Sequences

Yang Wang and Qiang Ji
Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute, Troy, NY 12180, USA
yang.wang@ieee.org qji@ecse.rpi.edu

Abstract

This paper presents a dynamic conditional random field (DCRF) model to integrate contextual constraints for object segmentation in image sequences. Spatial and temporal dependencies within the segmentation process are unified by a dynamic probabilistic framework based on the conditional random field (CRF). An efficient approximate filtering algorithm is derived for the DCRF model to recursively estimate the segmentation field from the history of video frames. The segmentation method employs both intensity and motion cues, and it combines dynamic information and spatial interaction of the observed data. Experimental results show that the proposed approach effectively fuses contextual constraints in video sequences and improves the accuracy of object segmentation.

1. Introduction

Object segmentation in image sequences is very important to application areas such as human-computer interaction, content-based video coding, and multi-object tracking. To robustly differentiate independently moving objects composing the scene, the strategy to integrate temporal and spatial information from the video sequence is a key issue throughout the segmentation process.

Temporal correspondence among successive video frames can be constructed by motion estimation at the pixel level. Motion based segmentation approaches perform motion estimation and cluster pixels into regions of coherent motion [3] [20]. Moreover, layered approaches represent multiple objects in the scene with a collection of layers [10] [21] [25]. Typically parametric motion models and spatial supports for the layers are estimated through an EM algorithm [9] [24]. On the other hand, spatial regions from oversegmentation of the scene provide important hints of object boundaries [16]. Region merging methods divide the video scene into a set of regions with small intensity variation and extract objects by merging together regions of coherent motion [6] [17]. Temporal or spatial information alone may not be sufficient to discriminate objects since one area with similar intensity or motion sometimes contains multiple objects. From this point of view, joint spatio-temporal segmentation approaches have been proposed to deal with intensity and motion cues simultaneously [4] [12] [22]. Regions of homogeneous visual characteristics are extracted from the video volume by joint spatial and temporal grouping [5] [7].

Furthermore, contextual constraints are important to effectively fuse temporal and spatial information during the segmentation process. For instance, contiguous sites are likely to belong to the same object in the scene, and one site tends to remain in the same object at consecutive time instants. Markov random field (MRF) and hidden Markov model (HMM) have been extensively employed to formulate contextual constraints [1] [2] [11] [23]. However, conditional independence of observations is usually assumed in the previous work, which is too restrictive for visual scene modeling. Compared to generative models including MRF and HMM, the conditional random field (CRF) relaxes the strong independence assumption and captures dependencies between observations [14]. Originally proposed for one-dimensional text sequence labeling, the CRF has been applied to two-dimensional image labeling recently [8] [13].

Based on the CRF, a dynamic conditional random field (DCRF) model for object segmentation in image sequences is proposed in this paper. The DCRF extends the CRF model by incorporating temporal constraints among successive segmentation fields. Spatial and temporal dependencies during the segmentation process are unified in the dynamic probabilistic model. A computationally efficient approximate filtering algorithm is derived for the DCRF model to recursively estimate the segmentation field from the history of observed data. The method handles intensity and motion cues simultaneously and permits neighborhood interactions in both labels and observations. Experimental results show that the proposed approach effectively integrates spatio-temporal contextual constraints in video sequences and significantly improves the segmentation accuracy.

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 **IEEE**

The rest of the paper is arranged as follows: Section 2 proposes the DCRF model and derives its filtering algorithm. Section 3 formulates the observation model. Section 4 describes the implementation details. Section 5 discusses the experimental results. Then our technique is concluded in Section 6.

2. Dynamic conditional random field

For an image sequence, the segmentation label and observation of a point x within the kth frame are denoted by $s_k(x)$ and $z_k(x)$ respectively. Segmentation label $s_k(x)$ assigns the point x to one of L independently moving objects composing the scene. Here k, $L \in \mathbf{N}$, $x \in X$, and X is the spatial domain of the video scene. The label $s_k(x) = e_l$ if point x belongs to the lth object, where e_l is a L-dimensional unit vector with its lth component equal to one. The observation $z_k(x)$ consists of intensity and motion information at the site x. The entire label field and observed data over the video scene at time k are expressed compactly as s_k and z_k respectively. Spatial and temporal contextual constraints in the segmentation process can be imposed through a dynamic model of statistical dependencies among neighboring sites.

2.1. DCRF model

The definition of conditional random field (CRF) is given by Lafferty et al. in [14]. For two random fields r and r' over the video scene, (r, r') is a conditional random field if, when conditioned on r', the random field r obey the Markov property: $p(r(x) \mid r', r(y), y \neq x) = p(r(x) \mid r', r(y), y \in N_x)$, where set N_x denotes the neighboring sites of point x.

Given the observed data up to time k, the posterior probability distribution of the segmentation field s_k is modeled by a conditional random field to formulate spatial dependencies. The segmentation labels obey the Markov property when the observed data is given: $p(s_k(x) \mid z_{1:k}, s_k(y), y \neq x) = p(s_k(x) \mid z_{1:k}, s_k(y), y \in N_x)$, where $z_{1:k} = \{z_i, i = 1, 2, \ldots, k\}$ is the sequence of observed data up to time k. Hence s_k is a random field globally conditioned on the observed data. Using the Hammersley-Clifford theorem and considering only up to pairwise clique potentials, the posterior probability of the segmentation field is given by a Gibbs distribution with the following form.

$$p(s_k \mid z_{1:k}) \propto \exp\{-\sum_{x \in X}[V_x(s_k(x) \mid z_{1:k}) + \sum_{y \in N_x} V_{x,y}(s_k(x), s_k(y) \mid z_{1:k})]\}. \quad (1)$$

The one-pixel potential $V_x(s_k(x) \mid z_{1:k})$ reflects the information (or constraint) from the observed data for a single site. The two-pixel potential $V_{x,y}(s_k(x), s_k(y) \mid z_{1:k})$ imposes the spatial interaction (or pairwise constraint), and the interaction strength is dependent on the observed data.

To formulate temporal (or dynamic) dependencies of consecutive segmentation fields, the state transition probability of the segmentation field $p(s_{k+1} \mid s_k)$ is modeled using a Gibbs distribution defined on one-pixel and two-pixel cliques as well.

$$p(s_{k+1} \mid s_k) \propto \exp\{-\sum_{x \in X}[V_x(s_{k+1}(x) \mid s_k(M_x)) + \sum_{y \in N_x} V_{x,y}(s_{k+1}(x), s_{k+1}(y))]\}, \quad (2)$$

where M_x designates the set of sites in the kth image that impact on site x in the $(k+1)$th image. The one-pixel potential $V_x(s_{k+1}(x) \mid s_k(M_x))$ models the label state transition for a single site, and the two-pixel potential $V_{x,y}(s_{k+1}(x), s_{k+1}(y))$ imposes the spatial connectivity constraint to form contiguous regions. It should be noted that M_x is not equivalent to the neighborhood N_x. M_x and N_x may have different sizes. $x \notin N_x$ while $x \in M_x$. To distinguish them, N_x is called the spatial neighborhood, and M_x the temporal neighborhood (e.g. see Figure 1). The one-pixel potential is further expressed as

$$V_x(s_{k+1}(x) \mid s_k(M_x)) = \frac{1}{|M_x|} \sum_{y \in M_x} V_x(s_{k+1}(x) \mid s_k(y)), \quad (3)$$

where $|\cdot|$ denotes the size (number of points) of the set, and the potential $V_x(s_{k+1}(x) \mid s_k(y))$ imposes the temporal continuity constraint to encourage a point to have the same segmentation label as those of its temporal neighborhood.

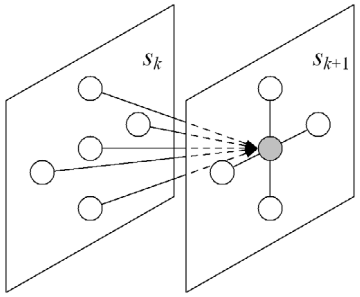

Figure 1. The 5-pixel temporal neighborhood and the 4-pixel spatial neighborhood.

The observation (or likelihood) model $p(z_k \mid s_k)$ is also formulated by a conditional random field to capture dependencies between observations. Here the likelihood is formulated as (see Section 3)

$$p(z_k \mid s_k) \propto \exp\{-\sum_{x \in X}[V_x(z_k(x) \mid s_k(x)) + \sum_{y \in N_x} V_{x,y}(z_k(x), z_k(y) \mid s_k(x), s_k(y))]\}. \quad (4)$$

The one-pixel potential $V_x(z_k(x) \mid s_k(x))$ reflects the local information for a single site. The two-pixel potential $V_{x,y}(z_k(x), z_k(y) \mid s_k(x), s_k(y))$ imposes the neighborhood interaction among observations. Thus the information of a point is influenced by the observed data from its spatial neighborhood.

By (1)-(4), spatial and temporal dependencies in the segmentation process are unified by a dynamic probabilistic framework based on the CRF model. Therefore it is called the dynamic conditional random field (DCRF). The DCRF extends the CRF for individual images [13] by incorporating temporal dependencies among segmentation fields, and it considers contextual constraints in more dimensions than the dynamic CRF for text sequences [19] since the model is built for three-dimensional image sequences.

2.2. DCRF filter

From a Bayesian perspective, the filtering algorithm is to recursively update the posterior distribution of the segmentation field. A first-order Markov assumption is made on segmentation fields to simplify the computation. Given the potentials of the distribution $p(s_k \mid z_{1:k})$, the posterior $p(s_{k+1} \mid z_{1:k+1})$ at time k+1 can be efficiently approximated by a conditional random field with the following potentials (see Appendix A).

$$V_x(s_{k+1}(x) \mid z_{1:k+1}) = \frac{1}{|M_x|} \sum_{y \in M_x} \sum_{s_k(y)} q_y(s_k(y) \mid z_{1:k}) V_x(s_{k+1}(x) \mid s_k(y)) + V_x(z_{k+1}(x) \mid s_{k+1}(x)), \quad (5a)$$

$$V_{x,y}(s_{k+1}(x), s_{k+1}(y) \mid z_{1:k+1}) = V_{x,y}(s_{k+1}(x), s_{k+1}(y)) + V_{x,y}(z_{k+1}(x), z_{k+1}(y) \mid s_{k+1}(x), s_{k+1}(y)), \quad (5b)$$

where $\{q_x(s_k(x) \mid z_{1:k})\}$ is given by mean field approximation of the segmentation field s_k (see Section 4). The one-pixel potential (5a) reflects the information for a single site. The first term in (5a) reflects the dynamic information (or temporal constraint) from its temporal neighborhood, which is controlled by previously observed data. The second term in (5a) reflects the information from the current observation. The two-pixel potential (5b) imposes the pairwise constraint from the spatial neighborhood. In (5b) the first term imposes the smoothness constraint independent of the observed data, and the second term imposes the spatial interaction dependent on neighboring observations.

3. Observation model

The observed data at time k is represented as $z_k = (g_k, m_k)$, where g_k and m_k are intensity and motion information for the kth video frame. For each point x, the pixel intensity $g_k(x)$ has three (R, G, and B) color components, and the motion vector $m_k(x)$ has two (horizontal and vertical) displacement components. The motion vector field is obtained by Bayesian MAP (maximum a posteriori) estimation with smoothness constraint [3]. When the segmentation field is given, it is reasonable to assume that intensity and motion cues are conditionally independent of each other since the color of an object will not affect its movement. Hence the observation model is factorized as

$$p(z_k \mid s_k) = p(g_k, m_k \mid s_k) = p(g_k \mid s_k) p(m_k \mid s_k). \quad (6)$$

Both the intensity likelihood $p(g_k \mid s_k)$ and the motion likelihood $p(m_k \mid s_k)$ are formulated as conditional random fields defined on one-pixel and two-pixel potential functions.

3.1. Likelihood model

For the intensity likelihood, the one-pixel potential is set as $V_x(g_k(x) \mid s_k(x)) = -\ln p(g_k(x) \mid s_k(x))$, so that the likelihood model becomes the product of local likelihood $p(g_k(x) \mid s_k(x))$ at each site if the two-pixel potential is ignored. Since an object may contain multiple colors, the probability density $p(g_k(x) \mid s_k(x))$ is modeled as a Gaussian mixture for each object.

$$p(g_k(x) \mid s_k(x) = e_l) = \sum_j w_{k,l,j} N(g_k(x); \theta^g_{k,l,j}), \quad (7)$$

where $1 \le l \le L$, $N(z; \theta)$ represents a Gaussian distribution with argument z and parameters (mean and covariance matrix) θ, and coefficients $\{w_{k,l,j}\}$ are the corresponding weights for the Gaussian mixture with $\sum_j w_{k,l,j} = 1$. Here the covariance matrixes are assumed to be diagonal to simplify the computation. The distribution parameters are estimated from the intensity information of the lth segmented object at time k-1 by an incremental EM algorithm termed adaptive mixtures [18]. The two-pixel potential for the intensity likelihood encourages color contrast between two objects.

$$V_{x,y}(g_k(x), g_k(y) \mid s_k(x), s_k(y)) \propto -\| g_k(x) - g_k(y) \|^2 (1 - \delta(s_k(x) - s_k(y))), \quad (8)$$

where $\delta(\cdot)$ is the Kronecker delta function, and $\|\cdot\|$ denotes the Euclidean distance. The pairwise constraint is imposed only when the two neighboring sites belong to different objects.

For the motion likelihood, similarly we have $V_x(m_k(x) \mid s_k(x)) = -\ln p(m_k(x) \mid s_k(x))$. The probability density $p(m_k(x) \mid s_k(x))$ is modeled as a Gaussian distribution for each object.

$$p(m_k(x) \mid s_k(x) = e_l) = N(m_k(x); \theta^m_{k,l}) . \quad (9)$$

The distribution parameters are estimated from the motion information of the *l*th segmented object at time *k*-1. For rapidly changing objects, it is easy to modify the parameter estimation to model the change of the probability density from time *k*-1 to time *k* (e.g. by Kalman filtering). The two-pixel potential for the motion likelihood encourages motion smoothness within an object.

$$V_{x,y}(m_k(x), m_k(y) \mid s_k(x), s_k(y)) \propto \| m_k(x) - m_k(y) \|^2 \, \delta(s_k(x) - s_k(y)) . \quad (10)$$

The pairwise constraint is imposed only if the two neighboring points belong to the same object.

Hence the potential functions for the observation model in (4) become

$$V_x(z_k(x) \mid s_k(x)) = -\ln p(g_k(x) \mid s_k(x)) - \ln p(m_k(x) \mid s_k(x)) , \quad (11a)$$

$$V_{x,y}(z_k(x), z_k(y) \mid s_k(x), s_k(y)) = -\beta_1 \| g_k(x) - g_k(y) \|^2 \, (1 - \delta(s_k(x) - s_k(y))) + \beta_2 \| m_k(x) - m_k(y) \|^2 \, \delta(s_k(x) - s_k(y)) , \quad (11b)$$

where positive β_1 and β_2 respectively weight the importance of intensity and motion cues in spatial interaction. The two-pixel potential function (11b) models the neighborhood interaction dependent on the observed data. Naturally, the potential imposes an adaptive contextual constraint that will adjust the interaction strength according to the similarity between neighboring observations. Moreover, when the two-pixel potential is set to zero, the observation model $p(z_k \mid s_k)$ will degenerate into the product of local likelihood, which is equivalent to the conditional independence assumption for the observed data (i.e. ignore interactions among observations when segmentation labels are given) in previous work [1] [15].

3.2. Notes on the likelihood formulation

Given the observations, there are two ways to estimate segmentation labels. In a generative framework, both the prior model of the segmentation field and the observation model are formulated to estimate the joint distribution of the observations and labels. Alternatively, in a discriminative framework the posterior distribution of the segmentation field is directly formulated. The conditional random field is originally proposed in the discriminative framework to avoid the formulation of the observation model [13]. Different from previous work, the observation model is formulated by the conditional random field in this paper. However, it should be noted that our formulation strictly complies with the definition of condition random field. There are two reasons to formulate the likelihood model in this dynamic conditional random field framework. First, the CRF formulation (4) relaxes the strong assumption of conditional independence and permits neighborhood interactions among observed data. Second, the derivation of the DCRF filter requires the formulation of the likelihood model in (15) for recursive estimation of the segmentation field.

4. Initialization and optimization

The smoothness constraints in (2) and (3) can be imposed by the following potentials.

$$V_{x,y}(s_{k+1}(x), s_{k+1}(y)) = \alpha_1 (1 - \delta(s_{k+1}(x) - s_{k+1}(y))) , y \in N_x, \quad (12a)$$

$$V_x(s_{k+1}(x) \mid s_k(y)) = \alpha_2 (1 - \delta(s_{k+1}(x) - s_k(y))) , y \in M_x, \quad (12b)$$

where positive α_1 and α_2 weight the importance of spatial connectivity and temporal continuity respectively. Thus two spatially or temporally neighboring pixels are more likely to belong to the same object than to different objects. To balance the influence of potential terms for the posterior distribution of the segmentation field, we assume that

$$\frac{1}{|X|} \sum_{x \in X} \sum_{y \in N_x} \alpha_1 = \frac{1}{|X|} \sum_{x \in X} \sum_{y \in M_x} \frac{\alpha_2}{|M_x|} = \lambda , \quad (13a)$$

$$|\Sigma_0^g|^{1/3} \beta_1 = |\Sigma_0^m|^{1/2} \beta_2 = \gamma , \quad (13b)$$

where Σ_0^g and Σ_0^m are the diagonal covariance matrixes of pixel colors and motion vectors for the initial frame. The parameters λ and γ are manually determined to reflect the influence of data-independent smoothness constraint and data-dependent spatial interaction respectively.

The initialization of the segmentation field s_0 is based on the procedure proposed in [21]. The initial frame is divided into small blocks and an affine transformation can be estimated for the motion of each block. To determine the number of objects *L*, a set of motion models is obtained by adaptively clustering the transformation parameters under a Euclidean distance measure. Each object is characterized by one motion model, and segmentation labels are assigned in a way that minimizes the motion distortion. In our work, the segmentation field is initialized by combining this procedure with smoothness constraint on the assignment of labels.

At each time instant, the segmentation field is optimized by mean field approximation [15]. The mean field algorithm suggests that when estimating the label mean at a single site, the influence from neighboring sites can be approximated by that of their means. The segmentation field at time *k* can be estimated as

$$p(s_k \mid z_{1:k}) \approx \prod_{x \in X} q_x(s_k(x) \mid z_{1:k}) , \quad (14a)$$

$$\hat{s}_k(x) = \arg\max_{e_l} q_x(s_k(x) = e_l \mid z_{1:k}) , \quad (14b)$$

where $\{q_x(s_k(x) \mid z_{1:k})\}$ is computed from an iterative procedure (see Appendix B). Initially $q_x(s_0(x) = e_l)$ is set as $\frac{1}{L}$ for all l and x.

5. Results and discussion

The proposed approach has been tested on video sequences captured under different environments and compared with other segmentation methods. The 24-pixel spatial neighborhood and 25-pixel temporal neighborhood are utilized in the algorithm. Our C program can process about two 320×240 frames per second on a Pentium 4 2.8G PC. Four spatio-temporal segmentation algorithms combining intensity and motion cues are studied in our experiments: A pixel based algorithm without using contextual constraints [12], an algorithm combining the first algorithm with spatio-temporal constraints imposed by the Markov random field in [22], the proposed algorithm without using dynamic information from previous frames (i.e. ignore the first term in (5a)), and the proposed algorithm. The same initialization and neighborhood are used in these algorithms (when applicable). The last three methods can be viewed as approaches to impose contextual constraints based on spatio-temporal Markov random field (S-T MRF), conditional random field (CRF), and dynamic conditional random field (DCRF) respectively.

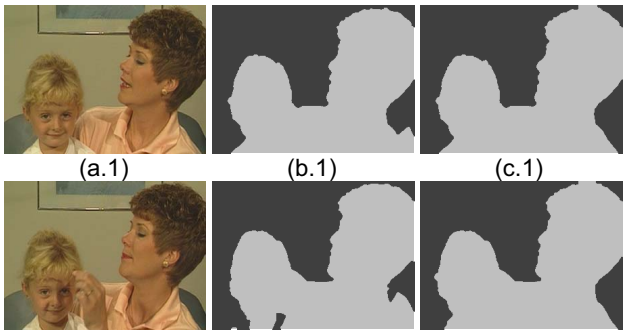

Figure 2. (a) Two frames of a sequence. (b) Segmentation results by CRF approach. (c) Segmentation results by the proposed method.

Figure 2 shows the segmentation results for the "mother and daughter" sequence by the conditional random field method and the proposed method. It can be seen in Figure 2b that as the mother raises her hand, the illumination variation makes the girl's shoulder partly misclassified as the background in the second image. Moreover, part of the sofa behind the mother is erroneously segmented by the CRF approach when instantaneous cues for different objects become similar. Using the temporal constraint imposed by previous frames, the erroneously segmented regions of the sofa and shoulder in Figure 2b are corrected in Figure 2c by the proposed dynamic approach.

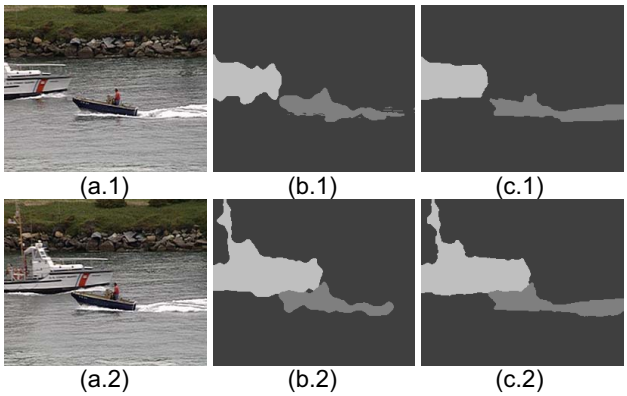

Figure 3. (a) Two frames of a sequence. (b.1) Segmentation result by pixel based approach for (a.1). (b.2) Segmentation result by spatio-temporal MRF approach for (a.2). (c) Segmentation results by the proposed method.

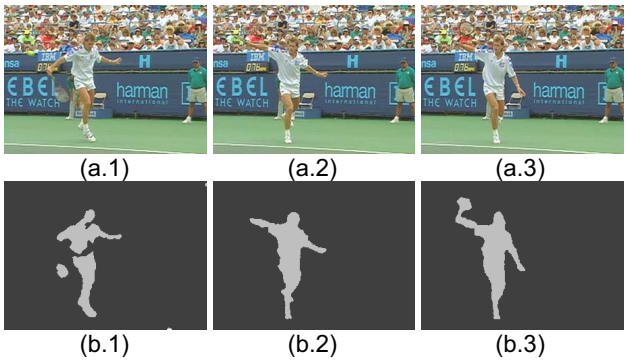

Figure 4. (a) The first, fifth, and ninth frames of a sequence. (b) Segmentation results by the proposed method.

Figure 3 shows the segmentation results for the "coastguard" sequence by the pixel based approach, the spatio-temporal Markov random field approach, and the proposed approach. It can be seen in Figure 3b that compared to the pixel based method, the MRF method generates smooth segmentation result. However, sometimes the Markov random field smoothes in a wrong way at object boundaries since it ignores the neighborhood interaction among observations. The DCRF approach can adjust the spatial interaction in terms of the similarity between neighboring observations, and it produces relatively accurate boundaries. Compared to the results in Figure 3b, the ship's bottom and the boat's trail are more accurately segmented in Figure 3c by the proposed method.

Figure 4 shows the segmentation results for the "tennis" sequence by the proposed method. Here motion or intensity information alone is not adequate to discriminate objects in the scene. The player's right leg

almost remains stationary in the second and third images with little motion information. Moreover, the similarity in skin color makes it difficult to distinguish the player's face and right arm from the audience in the scene. Initially, the player is separated into three parts in Figure 4b.1. Then the segmentation result is improved through the propagation of spatio-temporal information and contextual constraints from the video scene. After four frames the two half bodies are merged as a whole in Figure 4b.2. Eight frames later the tennis bat is correctly segmented as well in Figure 4b.3 by the proposed approach.

Table 1. Error rates of segmentation results.

	pixel	MRF	CRF	DCRF
error	6.3%	4.8%	3.1%	2.2%

The segmentation results are also evaluated quantitatively by comparing to the ground-truth images. The ground-truth images of frames in Figure 2-4 are manually labeled for comparison. Table 1 shows the average error rate (portion of misclassified points in the entire image) for the four algorithms. The MRF approach outperforms the pixel based approach by introducing contextual constraints. Compared to the MRF approach, the CRF approach takes advantage of data-dependent neighborhood interactions. The DCRF approach further improves the segmentation accuracy by utilizing dynamic information in image sequences. The substantial increase of the accuracy indicates that the DCRF approach effectively fuses spatial and temporal contextual constraints for object segmentation in video sequences.

6. Conclusion

There are two main contributions in this paper. First, based on the CRF, we have proposed a dynamic conditional random field (DCRF) model to unify spatial interaction and dynamic information in image sequences. Second, we have derived an efficient approximate DCRF filtering algorithm and applied it to object segmentation. The proposed approach effectively integrates contextual constraints in the segmentation process. Experimental results show that the method significantly improves the performance of object segmentation in video sequences. Our future study is to automatically determine the parameters and develop more accurate and efficient approximate filtering algorithms of the DCRF model.

Appendix A

At time k+1, observation z_{k+1} is used to update the posterior probability distribution of the segmentation field via Bayes' rule.

$$p(s_{k+1} \mid z_{1:k+1}) = \frac{p(s_{k+1} \mid z_{1:k}) p(z_{k+1} \mid s_{k+1})}{p(z_{k+1} \mid z_{1:k})}$$
$$\propto p(s_{k+1} \mid z_{1:k}) p(z_{k+1} \mid s_{k+1}) . \quad (15)$$

The conditional probability $p(s_{k+1} \mid z_{1:k})$ can be computed as

$$p(s_{k+1} \mid z_{1:k}) = \sum_{s_k} p(s_{k+1}, s_k \mid z_{1:k})$$
$$= \sum_{s_k} p(s_{k+1} \mid s_k) p(s_k \mid z_{1:k}) . \quad (16)$$

Accurate computation of (16) is intractable because all the possible assignments of the field s_k should be considered. By mean field approximation, the posterior $p(s_k \mid z_{1:k})$ can be factorized as (14a). Combining (2), (3), (14a), and (16), the probability $p(s_{k+1} \mid z_{1:k})$ becomes

$$p(s_{k+1} \mid z_{1:k}) \propto \exp[-\sum_{x \in X} \sum_{y \in N_x} V_{x,y}(s_{k+1}(x), s_{k+1}(y))] \cdot$$
$$\sum_{s_k} \{\exp[-\sum_{x \in X} \frac{1}{|M_x|} \sum_{y \in M_x} V_x(s_{k+1}(x) \mid s_k(y))] \cdot$$
$$\prod_{x \in X} q_x(s_k(x) \mid z_{1:k})\}$$
$$= \exp[-\sum_{x \in X} \sum_{y \in N_x} V_{x,y}(s_{k+1}(x), s_{k+1}(y))] \cdot$$
$$\prod_{x \in X} \{\sum_{s_k(x)} \exp[-\sum_{y \in M_x} \frac{1}{|M_y|} V_y(s_{k+1}(y) \mid s_k(x))] \cdot$$
$$q_x(s_k(x) \mid z_{1:k})\} . \quad (17)$$

Using Jensen's inequality, the term in (17) can be approximated by its lower bound,

$$\prod_{x \in X} \{\sum_{s_k(x)} \exp[-\sum_{y \in M_x} \frac{1}{|M_y|} V_y(s_{k+1}(y) \mid s_k(x))] \cdot$$
$$q_x(s_k(x) \mid z_{1:k})\}$$
$$\approx \prod_{x \in X} \exp\{-\sum_{s_k(x)} q_x(s_k(x) \mid z_{1:k}) \cdot$$
$$[\sum_{y \in M_x} \frac{1}{|M_y|} V_y(s_{k+1}(y) \mid s_k(x))]\}$$
$$= \exp\{-\sum_{x \in X} [\frac{1}{|M_x|} \sum_{y \in M_x} \sum_{s_k(y)} q_y(s_k(y) \mid z_{1:k}) \cdot$$
$$V_x(s_{k+1}(x) \mid s_k(y))]\} . \quad (18)$$

Combining (4), (15), (17), and (18), the posterior probability distribution of the segmentation field at time k+1 is updated as

$$p(s_{k+1} \mid z_{1:k+1}) \propto p(s_{k+1} \mid z_{1:k}) p(z_{k+1} \mid s_{k+1})$$
$$\propto \exp\{-\sum_{x \in X} [\frac{1}{|M_x|} \sum_{y \in M_x} \sum_{s_k(y)} q_y(s_k(y) \mid z_{1:k}) \cdot$$
$$V_x(s_{k+1}(x) \mid s_k(y)) + V_x(z_{k+1}(x) \mid s_{k+1}(x)) +$$

$$\sum_{y\in N_x} V_{x,y}(s_{k+1}(x), s_{k+1}(y)) + \sum_{y\in N_x} V_{x,y}(z_{k+1}(x), z_{k+1}(y) \mid s_{k+1}(x), s_{k+1}(y))]\} . \quad (19)$$

From (19), the posterior distribution at time k+1 can be approximated by a conditional random field with the one-pixel and two-pixel potentials in (5).

Appendix B

At time k, the mean field local energy for site x is defined as

$$U_x(s_k(x) \mid z_{1:k}) = V_x(s_k(x) \mid z_{1:k}) + \sum_{y\in N_x} V_{x,y}(s_k(x), \langle s_k(y)\rangle \mid z_{1:k}) , \quad (20)$$

where $\langle\cdot\rangle$ denotes the expectation or ensemble average. The conditional probability for site x is approximated by the mean field local probability,

$$q_x(s_k(x) \mid z_{1:k}) = Q_{x,k}^{-1} \exp[-U_x(s_k(x) \mid z_{1:k})] , \quad (21)$$

where $Q_{x,k}$ is called the mean field local partition,

$$Q_{x,k} = \sum_{s_k(x)} \exp[-U_x(s_k(x) \mid z_{1:k})] . \quad (22)$$

Hence the mean of a segmentation label is computed as

$$\langle s_k(x)\rangle \approx \sum_{s_k(x)} s_k(x) q_x(s_k(x) \mid z_{1:k}) = Q_{x,k}^{-1} \sum_{s_k(x)} s_k(x) \exp[-U_x(s_k(x) \mid z_{1:k})] . \quad (23)$$

It can be seen that in order to find the mean of a site, one has to know the means of its neighbors. Therefore the mean of the segmentation field can be iteratively estimated using (20), (22), and (23). Given the label means, the posterior distribution of the segmentation field is approximated by the product of mean field local probabilities.

$$p(s_k \mid z_{1:k}) \approx \prod_{x\in X} q_x(s_k(x) \mid z_{1:k}) = \prod_{x\in X} Q_{x,k}^{-1} \exp[-U_x(s_k(x) \mid z_{1:k})] . \quad (24)$$

References

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

[2] C. Bregler. "Learning and recognizing human dynamics in video sequences." *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 568-574, 1997.

[3] 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.

[4] D. DeMenthon and D. Doermann. "Video retrieval using spatio-temporal descriptors." *Proc. ACM Int'l Conf. Multimedia*, pp. 508-517, 2003.

[5] C. Fowlkes, S. Belongie, and J. Malik. "Efficient spatiotemporal grouping using the Nystrom method." *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 1, pp. 231-238, 2001.

[6] M. Gerlgon and P. Bouthemy. "A region-level graph labeling approach to motion-based segmentation." *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 514-519, 1997.

[7] H. Greenspan, J. Goldberger, and A. Mayer. "Probabilistic space-time video modeling via piecewise GMM." *IEEE Trans. Patt. Anal. Mach. Intel.*, vol. 26, pp. 384-396, 2004.

[8] X. He, R. S. Zemel, and M. A. Carreira-Perpinan. "Multiscale conditional random fields for image labeling." *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 695-702, 2004.

[9] 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*, vol. 1, pp. 692-706, 2002.

[10] 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.

[11] S. Kamijo, K. Ikeuchi, and M. Sakauchi. "Segmentations of spatio-temporal images by spatio-temporal Markov random field model." *Proc. EMMCVPR Workshop*, pp. 298-313, 2001.

[12] S. Khan and M. Shah. "Object based segmentation of video using color, motion, and spatial information." *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 746-751, 2001.

[13] S. Kumar and M. Hebert. "Discriminative fields for modeling spatial dependencies in natural images." *Advances in Neural Information Processing Systems*, pp. 1351-1358, 2004.

[14] J. Lafferty, A. McCallum, and F. Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." *Proc. Int'l Conf. Machine Learning*, pp. 282-289, 2001.

[15] S. Z. Li. *Markov Random Field Modeling in Image Analysis*, Springer-Verlag, 2001.

[16] 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.

[17] 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.

[18] C. E. Priebe. "Adaptive mixtures." *J. American Statistical Association*, pp. 796-806, 1994.

[19] C. Sutton, K. Rohanimanesh, and A. McCallum. "Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data." *Proc. Int'l Conf. Machine Learning*, 2004.

[20] R. Vidal and R. Hartley. "Motion segmentation with missing data using power factorization and GPCA." *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 310-316, 2004.

[21] J. Y. A. Wang and E. H. Adelson. "Representing moving images with layers." *IEEE Trans. Image Processing*, vol. 3, pp. 625-638, 1994.

[22] Y. Wang, T. Tan, and K.-F. Loe. "Video segmentation based on graphical models." *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 335-342, 2003.

[23] Y. Weiss. "Smoothness in layers: Motion segmentation using nonparametric mixture estimation." *Proc. Conf. Computer Vision and Pattern Recognition*, pp. 520-526, 1997.

[24] C. K. I. Williams and M. K. Titsias. "Learning about multiple objects in images: Factorial learning without factorial search." *Advances in Neural Information Processing Systems*, pp. 1415-1422, 2003.

[25] J. Xiao and M. Shah. "Motion layer extraction in the presence of occlusion using graph cut." *Proc. Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 972-979, 2004.