Segmentation of Video Sequences using Spatial-temporal Conditional Random Fields

Lei Zhang and Qiang Ji
Rensselaer Polytechnic Institute
110 8th St., Troy, NY 12180
{zhangl2@rpi.edu,qji@ecse.rpi.edu}

Abstract

Segmentation of video sequences requires the segmentations of consecutive frames to be consistent with each other. We propose to use a three dimensional Conditional Random Fields (CRF) to address this problem. A triple of consecutive image frames are treated as a small 3D volume to be segmented. Our spatial-temporal CRF model combines both local discriminative features and the conditional homogeneity of labeling variables in both the spatial and the temporal domain. After training the model parameters with a small set of training data, the optimal labeling is obtained through a probabilistic inference by Sum-product loopy belief propagation. We achieve accurate segmentation results on the standard video sequences, which demonstrates the promising capability of the proposed approach.

1. Introduction and the Related Work

Object segmentation of video sequences has many applications in security surveillance, video tracking, video compression, etc. Given a sequence of image frames, the goal is to segment the objects of interest (ROI) in each frame. Object segmentation of image sequences is different from segmentation of each frame independently. Assuming the motion of the object of interest is smooth, the segmentations of consecutive frames shall be consistent with each other. In another sense, the temporal homogeneity of segmentation is one constraint for video sequence segmentation.

Different approaches have been proposed to address the problem of video sequence segmentation. Since the objects of interest are often the moving objects in image frames, motion information (e.g. the optical flow) has been successfully used for video sequence segmentation [4] [3] [20]. However, accurate calculation of optical flow is not a trivial problem, especially at the boundary of objects [13]. On the other hand, pixel colors also provide useful information and have been used for segmentation. It is possible to use color features to find accurate segmentation along the object boundary. But color segmentation does not exploit the relationship between two consecutive image frames in a video sequence. Segmentation based on only color cannot guarantee the segmentations of two consecutive frames to be consistent with each other. In addition, there are obvious relationships among the labeling of spatially adjacent pixels because the natural objects tend to have a smooth change of features in the spatial domain. It is natural that adjacent pixels with similar features tend to belong to the same object. Due to these reasons, it is necessary to exploit both the temporal and spatial relationships to facilitate the segmentation of video sequences.

Researchers have applied Markov Random Fields (MRF) for segmenting video sequences. MRF is an undirected graphical model. It provides a general method for modeling the relationship of neighboring labeling random variables. Standard MRF models the apriori homogeneity assumption of labels in certain neighborhood systems. According to the Hammersley-Clifford theorem [8], a MRF model can be easily defined by a set of potential functions specified in cliques. For a detailed introduction to MRF model, we refer to [16] [23]. MRF models have been widely used in solving segmentation problems [6] [21] [19] [1] [11] [33] [30]. Although MRF models have been successfully applied in computer vision, they have some limitations. Standard MRF segmentation models often assume the conditional independent likelihood of the observed data, given the label of a site (i.e. the single node in the MRF model). This assumption is too restrictive because there are often complex relationships among the data. Moreover, standard MRF segmentation models enforce the homogeneity of the labeling variables (i.e. the Markovian property) everywhere. However, the homogeneity of labeling variables may not be true for every place, especially in the regions with strong discontinuity. Additional process such as the line process [16] can deal with the discontinuity of labeling variables. However, this makes the segmentation process more complex.

Conditional Random Fields (CRF) [15] is another class of undirected graphical model. It makes the Markovian assumption of the labeling variables conditioned on the observation. It relaxes the normal conditional independence assumption of the likelihood model in MRF models. CRF model enforces the homogeneity of labeling variables conditioned on the observation. It therefore can automatically handle the discontinuity of labeling variables. Different from the MRF model, the CRF model does not model the probability density of the data. Therefore it is basically a discriminative model that aims to distinguish different sites. Due to the weak assumptions of CRF model and its discriminative nature, CRF model allows arbitrary relationship among data and may require less resources to train its parameters [15] [14]. For the problem of labeling sequence data, CRF has also overcome the problem of label bias problem [15]. These advantages make CRF model more and more popular in the computer vision world.

Several previous works have demonstrated the successfulness of CRF models in some computer vision fields. Lafferty *et al.* [15] [26] [18] [12] apply CRF models for labeling sequence data. They have shown better performance of CRF models than the Hidden Markov Model (HMM) and Maximum Entropy Markov models (MEMMs), especially for language and text processing problem. Quattoni and Kumar *et al.* [24] [14] [17] [32] present different CRF models for solving object recognition problems. They also extend the CRF models to incorporate the hidden layers and nonlinear kernels. These extensions make the CRF models more powerful. For example, the hidden layers can model the parts in part-based models [24]. He *et al.* [9] [10] [2] have used CRF models to deal with image segmentation on individual images and demonstrated that the CRF models generally outperform the MRF models.

CRF models have also been applied in video segmentation [5] [29] [31]. We notice that all of them have retained certain generative models, which somewhat contradicts the discriminative nature of the CRF models. All of them model the likelihood of color and the likelihood of motion using generative models. In general, it requires more training data to learn the complex generative models that tend to explain the distribution of the data. On the contrary, the discriminative CRF model focuses on discriminating the data instead of explaining their distribution. It may require less data for training [15] [14]. Since the discriminative characteristic of CRF model is one of its advantages, we are interested in how to use pure discriminative CRF models to solve video sequence segmentation problem. In Section 4, we will show that we actually use very few data to train our spatial-temporal CRF model. This demonstrates the advantage of CRF models.

The CRF model is also used to solve the tracking problem in video sequences [25]. This approach first oversegments the image frames into superpixels. A process of constrained Delaunay triangulation (CDT) is needed to partition the image into a set of triangles. A CRF model is then constructed on these triangles. The superpixel based approach may have problems when the initial oversegmentation fails to accurately find the object boundary.

In this paper, we propose a pixel based spatial-temporal CRF model to segment video sequences. We treat video sequences as a group of small volume data. We extend two dimensional CRF segmentation model to a three dimensional (3D) CRF model, including both the spatial domain and the temporal domain. Each small volume is then segmented by the 3D CRF model. In the 3D CRF model, the pairwise neighborhood contains not only the adjacent pixels in the spatial domain, but also the adjacent pixels in the temporal domain. The homogeneity constraint of random labeling variables is enforced in both domains. In this way, we build a pure discriminative segmentation model that enforces both spatial consistency and temporal consistency of the labeling variables.

2. Overview of the Approach

Our goal is to segment the foreground objects from the backgrounds in a video sequence. Firstly, we define our notations. Let $\mathbf{x} := \{x_u, u \in V\}$ be the observed image, where V denotes all the sites (i.e. the pixels) in the image. Let $\mathbf{y} = \{y_u \in \mathcal{Y}, u \in V\}$ be the corresponding set of labeling random variables, where $\mathcal{Y}$ is the set of possible labels at a single site. We assume $\mathcal{Y} = \{-1, +1\}$, where $+1$ represents the foreground and -1 represents the background. Let t denote the temporal index. Given a triple of consecutive images x^{t-1}, x^t, x^{t+1}, the goal is to find the optimal labeling $\mathbf{y}^t$ for the image frame x^t.

We extend the Conditional Random Fields (CRF) [15] to the spatial-temporal domain to model the sequence segmentation problem. CRF model only assumes that the labeling variables $\mathbf{y}$ follows the Markovian property, conditioned on the observation. CRF model relaxes the normal assumption in a MRF model by allowing arbitrary relationship among the observed data. Moreover, the interaction potential of a CRF model may depend on all the observation. It allows to model data-adaptive potential functions that can deal with discontinuity in the observed data. For example, if there is a strong edge, it may not be necessary to impose the homogeneous constraint on neighboring labels.

For video sequence segmentation, the labeling random variables shall be homogenous not only in the spatial neighborhood, but also in the temporal neighborhood. In order to incorporate the homogeneous constraint in the temporal neighborhood, we add additional interaction potentials in the temporal domain into the two dimensional CRF segmentation model. The neighborhood of a site u includes not only those neighbors in the spatial domain, but also those neighbors in the temporal domain. Figure 1 il-

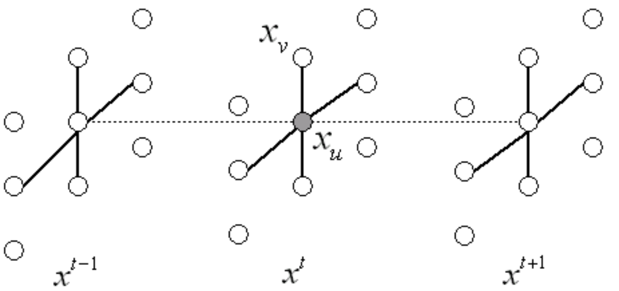

Figure 1. The spatial neighborhood and the temporal neighborhood in the spatial-temporal CRF model. The shaded circle represents the current site. The solid lines are the pairwise cliques in the spatial domain. The dotted lines are the pairwise cliques in the temporal domain.

lustrates the spatial-temporal neighborhood in our model. For simplicity, only pairwise cliques in the spatial domain and in the temporal domain are considered. We call this extended model as Spatial-Temporal Conditional Random Fields (STCRF). In the spatial domain, we use the standard 4-neighborhood system. In the temporal domain, we currently only add two links between the pixels at the same row and column locations in consecutive frames.

3. Spatial-temporal CRF Segmentation Model

The Spatial-temporal CRF model directly model the posteriori probability distribution of the labeling variables $\mathbf{y}^t$, given three consecutive image frames x^{t-1}, x^t, x^{t+1}. The posteriori distribution of the labeling variables is defined as

$$\begin{aligned} P(y|x) &= \frac{1}{Z} exp\{\sum_{u \in V} [logP(y_u, x_u) + \sum_{v \in \mathcal{N}_\mathrm{u}} y_u y_v \mu^T g_{uv}(x) \\ &\quad + \sum_{v \in \mathcal{M}_\mathrm{u}} y_u y_v \gamma^T g_{uv}(x)]\} \end{aligned} \quad (1)$$

where v is a site in the spatial-temporal neighborhood of the site u. $\mathcal{N}_u$ is the spatial neighborhood of the site u and $\mathcal{M}_u$ is the temporal neighborhood. μ and γ are the parameter vectors.

There are three terms in Eq.(1). The first term is the unary potential, which tries to label the site u according to its local features. For this purpose, we use a discriminative classifier based on a three-layer perceptron. Let $net(x_u)$ denotes the output of the perceptron when the features x_u are the input. The output of the three-layer perceptron is converted to a probabilistic interpretation using a logistic function, i.e.,

$$P(y_u, x_u) = \frac{1}{1 + exp(-y_u \frac{net(x_u)}{\tau})} \quad (2)$$

where τ is a constant that can adjust the curve of the logistic function. In our experiments, τ is fixed as 0.34. The three-layer perceptron is trained with a set of training data (see more detail in Section 3.1).

The second term in Eq.(1) is the pairwise interaction potential in the spatial domain. The third term in Eq.(1) is the pairwise interaction potential in the temporal domain. We separate these two kinds of pairwise potentials because the homogeneity constraints in the spatial domain and that in the temporal domain may be differently emphasized. $g_{uv}(\cdot)$ represents the feature vector for a pair of sites u and v. An additional bias term (fixed as 1) is also added into the feature vector $g_{uv}(\cdot)$. In this paper, we only use color features in the CIELAB color space to define the feature vector $g_{uv}(\cdot)$. However, our model can use arbitrary feature vectors as the observed data x_u and x_v. The feature vector $g_{uv}(x)$ is defined as

$$g_{uv}(x) = [1, \ |x_u - x_v|]^T \quad (3)$$

where T is the transpose of a vector. The operator $|\cdot|$ represents the absolute value of each component in the vector.

The variable Z in Eq. (1) is the normalization term (i.e. the partition function). It can be calculated by summing out all the possible configurations of the labeling variables $\mathbf{y}$, i.e.,

$$\begin{aligned} Z &= \sum_y exp\{\sum_{u \in V} [logP(y_u, x_u) + \sum_{v \in \mathcal{N}_\mathrm{u}} y_u y_v \mu^T g_{uv}(x) \\ &\quad + \sum_{v \in \mathcal{M}_\mathrm{u}} y_u y_v \gamma^T g_{uv}(x)]\} \end{aligned} \quad (4)$$

3.1. Parameter Estimation

The three-layer perceptron classifier and the parameters of the pairwise potentials (i.e., $\theta = [\mu, \gamma]^T$) are automatically learned from the training data. Different from the parameter estimation for two dimensional CRF model, each training data here is a triple of consecutive image frames. With a little abuse of notations, we denote $x^{(i)}$ as the ith triple of the training images. $y^{(i)}$ are the corresponding ground truth labeling for the ith triple images. Assume we have $x^{(1)}, x^{(2)}, ..., x^{(m)}$ such triple images and their ground truth labeling $y^{(1)}, y^{(2)}, ..., y^{(m)}$, where m is the number of triple images, the aim of parameter estimation is to automatically learn the parameters θ and the three-layer preceptron classifier from these data.

First, we train the three-layer perceptron classifier. The structure of our three-layer perceptron includes 3 input nodes, 8 hidden nodes and 1 output node. In the training step, the target output of the three-layer perceptron classifier is +1 (foreground pixel) or -1 (background pixel). The CIELAB color features of each pixel are the input of the three-layer perceptron. Given the input and the desired output, the three-layer perceptron classifier is trained using the standard BFGS quasi-Newton backpropagation method [7].

Next, we fix the three-layer perceptron classifier and use the Maximum Likelihood Estimation (MLE) method to learn the parameter θ for the pairwise potentials. Assuming all the training data are independently sampled, the log-likelihood of the parameters is calculated as

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{m}\{\sum_{u\in V}[logP(y_u^{(i)},x_u^{(i)}) + \sum_{v\in\mathcal{N}_{\mathrm{u}}} y_u^{(i)}y_v^{(i)}\mu^T g_{uv}(x^{(i)}) \\ &\quad + \sum_{v\in\mathcal{M}_{\mathrm{u}}} y_u^{(i)}y_v^{(i)}\gamma^T g_{uv}(x^{(i)})] - z^{(i)}\} \end{aligned} \tag{5}$$

where $z^{(i)}$ is the logarithm of the partition function, i.e.

$$z^{(i)} = logZ^{(i)}$$

The optimal parameters θ^* are estimated according to the MLE estimation, i.e.,

$$\theta^* = \arg\max_{\theta} L(\theta) \tag{6}$$

We use the stochastic gradient descent method [28] to find the optimal parameters θ^*. The gradient of the log-likelihood $L(\theta)$ is calculated as follows:

$$\begin{aligned} \frac{\partial L(\theta)}{\partial \mu} &= \sum_{i=1}^{m}[\sum_{u}\sum_{v\in\mathcal{N}_{\mathrm{u}}} y_u^{(i)}y_v^{(i)} g_{uv}(x^{(i)}) \\ &\quad -E_{P(y|x^{(i)};\theta)}(\sum_{u}\sum_{v\in\mathcal{N}_{\mathrm{u}}} y_u y_v g_{uv}(x^{(i)}))] \\ \frac{\partial L(\theta)}{\partial \gamma} &= \sum_{i=1}^{m}[\sum_{u}\sum_{v\in\mathcal{M}_{\mathrm{u}}} y_u^{(i)}y_v^{(i)} g_{uv}(x^{(i)}) \\ &\quad -E_{P(y|x^{(i)};\theta)}(\sum_{u}\sum_{v\in\mathcal{M}_{\mathrm{u}}} y_u y_v g_{uv}(x^{(i)}))] \end{aligned} \tag{7}$$

where $E_P[\cdot]$ denotes the expectation with respect to the distribution P. For example,

$$\begin{aligned} &E_{P(y|x^{(i)};\theta)}(\sum_{u}\sum_{v\in\mathcal{N}_{\mathrm{u}}} y_u y_v g_{uv}(x^{(i)})) \\ &= \sum_{y} P(y|x^{(i)};\theta)\sum_{u}\sum_{v\in\mathcal{N}_{\mathrm{u}}} y_u y_v g_{uv}(x^{(i)}) \\ &= \sum_{u} P(y_u|x^{(i)};\theta)\sum_{v\in\mathcal{N}_{\mathrm{u}}} y_u y_v g_{uv}(x^{(i)}) \end{aligned} \tag{8}$$

The summation is performed over all possible configurations of the labeling variables y. The term $P(y_u|x^{(i)};\theta)$ is the marginal probability of the label y_u given the observation $x^{(i)}$ and the model parameters θ.

The parameter estimation includes the following steps:

1. Given $\{x^{(i)}, y^{(i)}\}$, randomly initialize θ_0;
2. k=1, do the following until the maximum iteration is reached or the change of weights is small enough:
   - calculate the unary potential and the pairwise potentials according to Eq. (1), Eq. (2) and Eq. (3);
   - calculate the marginal probability $P(y_u|x^{(i)};\theta_{k-1})$ by Sum-product loopy belief propagation [22];
   - calculate the gradient $\frac{\partial L(\theta)}{\partial\theta}$ according to Eq. (7) and Eq. (8);
   - update θ_k by stochastic gradient descent [28];
   - $k = k + 1$;
3. Return θ_k.

3.2. Labeling Inference

After all the parameters are estimated, the model in Eq. (1) is used to segment video sequences. Given three consecutive image frames $\{x^{t-1}, x^t, x^{t+1}\}$, we segment the frame x^t according to the Maximum Posterior Marginal (MPM) criterion. Each site u is assigned a label that maximizes its posteriori marginal probability, i.e.,

$$y_u^* = \arg\max_{y_u\in\mathcal{Y}} P(y_u|x^{t-1}, x^t, x^{t+1};\theta) \tag{9}$$

The marginal probability $P(y_u|x^{t-1}, x^t, x^{t+1};\theta)$ is calculated by Sum-product loopy belief propagation (LBP) [22]. The labeling inference is summarized as follows:

1. Given $\{x^{t-1}, x^t, x^{t+1};\theta\}$,
   - calculate the unary potential and the pairwise potentials according to Eq. (1), Eq. (2) and Eq. (3);
   - calculate the posteriori marginal probability $P(y_u|x^{t-1}, x^t, x^{t+1};\theta)$ using LBP;
   - assign the optimal label to the pixel u by Eq. (9);
2. Return the labeling result y.

4. Experiments

We have tested the proposed approach on three standard video sequences. The first sequence is the "Mother and Daughter" sequence that includes 150 frames. Since the difference between consecutive frames is normally very small due to the frame rate, we segment one frame every five frames to make the problem a little more difficult.

We use the color features in CIELAB color space as the pixelwise features used in Eq.(1) because the CIELAB color space is perceptually close to the human vision system. To train the model, we manually label a few image frames. However, we use very few training data. Only three manually labeled images are actually used to train

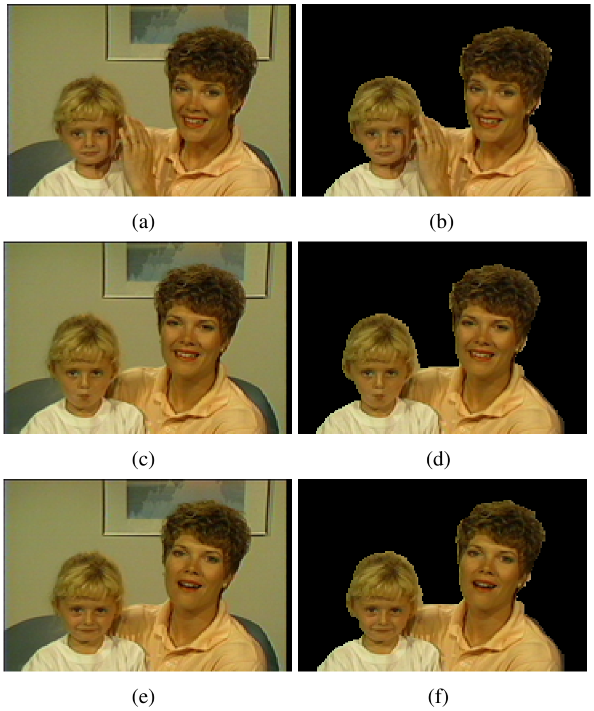

Figure 2. Examples of the original images and the corresponding segmentation masks for the "Mother and Daughter" sequence. a)the original #16 image; b)the segmentation of the #16 image; c)the original #71 image; b)the segmentation of the #71 image; e)the original #121 image; f)the segmentation of the #121 image.

both the three-layer perceptron classifier and the parameters of the pairwise potentials. This is the minimum number of data needed for training the spatial-temporal CRF model because we need at least three consecutive frames to estimate the temporal pairwise potentials.

Figure 2 shows several segmentation results on the "Mother and Daughter" sequence. Although there are significant movements of the mother's hand and her head, the proposed approach accurately finds out the foreground (i.e. the mother and the daughter) in these images.

The second sequence is the "Foreman" sequence that includes 250 frames. The human in this sequence has much larger and diversified movements, which makes it more challenging to segment these images. We still use the minimum number of three consecutive frames for training our spatio-temporal CRF model. We segment one frame every five frames to make the difference of two consecutive segmentations apparent. This setup makes the segmentation problem harder because the human will have more significant movements in consecutive segmentations.

Figure 3 shows the typical segmentation results on the "Foreman" sequence. Despite the large movement of the person, he is still successfully segmented in these results. Small errors exist on the boundary of this person, especially on the boundary of the safety helmet. This is because the color features at these places are very close to the color features in the nearby background, which makes the segmentation difficult.

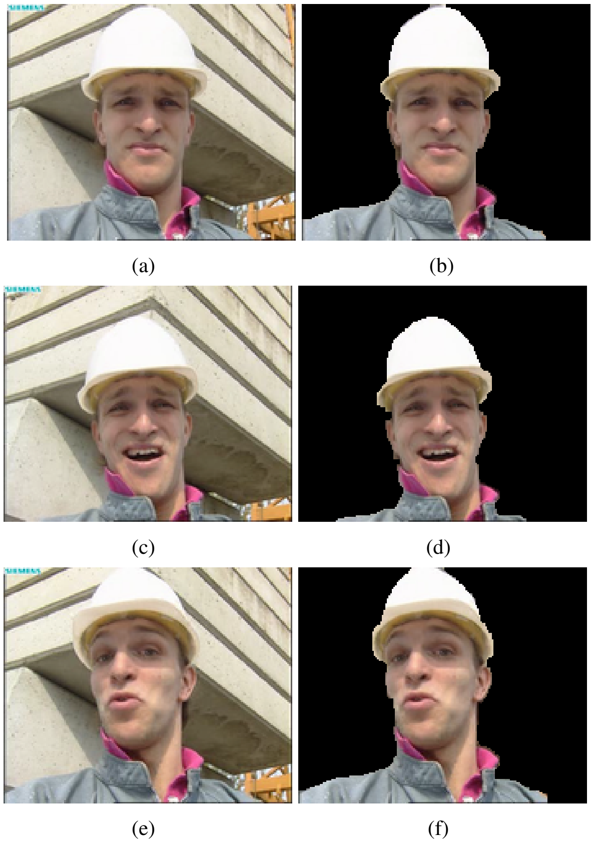

Figure 3. Examples of the original images and the corresponding segmentation masks for the "Foreman" sequence. a)the original #15 image; b)the segmentation of the #15 image; c)the original #130 image; b)the segmentation of the #130 image; e)the original #185 image; f)the segmentation of the #185 image.

The third sequence is the "Silent" sequence that includes 200 frames. The person in this sequence has quick hand movements and the background looks more complex. These reasons make it challenging to segment these images. We did similar experiments as above. Figure 4 shows the typical segmentation results for the "Silent" sequence. The person is also successfully segmented in these frames.

To quantitatively evaluate the segmentation results, we calculate the average rates of wrongly labeled pixels (i.e., the error rates) in the test images and the standard deviation of errors. The quantitative results for the three test sequences are summarized in Table 1. We achieve accurate segmentation results for all these video sequences.

Tsaig *et al*. [27] also perform experiments on the three video sequences used in this paper. Our segmentation results are qualitatively comparable (if not better) to their results. They did not give the quantitative results, preventing us from more detailed comparison with their approach.

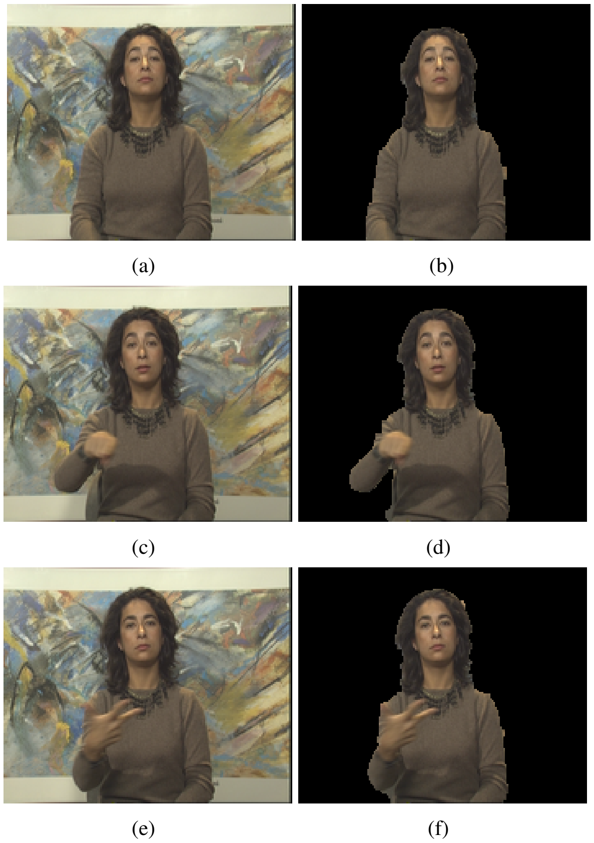
(a) (b)
(c) (d)
(e) (f)

Figure 4. Examples of the original images and the corresponding segmentation masks for the “Silent“ sequence. a)the original #8 image; b)the segmentation of the #8 image; c)the original #168 image; b)the segmentation of the #168 image; e)the original #178 image; f)the segmentation of the #178 image.

Table 1. The average segmentation error rates and their standard deviations for all video sequences that have been tested using our spatial-temporal Conditional Random Fields model.

image sequence	average error	standard deviation
Mother and Daughter	1.56%	0.12%
Foreman	2.0%	0.56%
Silent	1.62%	0.23%

Wang *et al*. [29] report an average error rate of 2.2% on the “Mother and Daughter“ sequence. Compared with their approach, our model is much simpler. Our approach requires very few training data. Most of important, our model does not use any generative models and keeps the discriminative nature of the Conditional Random Fields model. Our segmentation results can also rival their results according to the error rates.

We also redid the experiments using the Conditional Random Fields model with only the pairwise potentials in the spatial domain. We compared the segmentation results with those produced by the whole model. Figure 5 shows some apparent differences between these results. The spatial-temporal CRF model produces visually better segmentation than the CRF model that only includes the spatial pairwise potentials. These results demonstrate the importance of the temporal links in Figure 1.

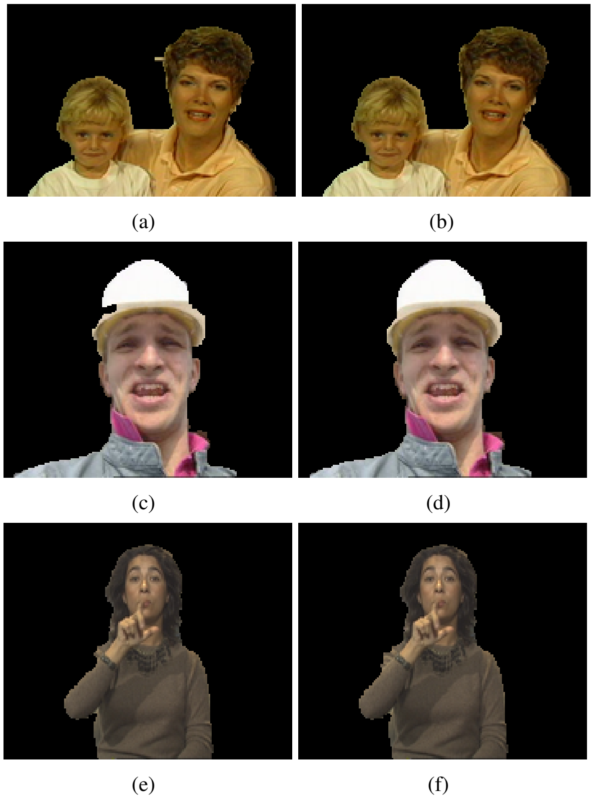
(a) (b)
(c) (d)
(e) (f)

Figure 5. Comparison of segmentation results. The first column was produced by the model with only the spatial pairwise potentials. The second column was produced by the whole model, i.e., the model with both the spatial and the temporal pairwise potentials. a) and b): the segmentation masks of #141 image in the “Mother and Daughter“ sequence. The apparent difference lies on the head of the mother; c) and d): the segmentation masks of #155 image in the “Foreman“ sequence. The apparent difference lies on the safety helmet of the person. e) and f): the segmentation masks of #98 image in the “Silent“ sequence. The apparent difference lies on the right shoulder of the person.

5. Summary

In this paper, we present a spatial-temporal Conditional Random Fields model for segmenting video sequences. This model exploits both the spatial relationship and the temporal relationship among the labeling random variables. It also keeps the discriminative nature of the CRF model. We tested the proposed model on three standard video sequences and achieved accurate segmentation results in all these sequences.

The future work includes studying the use of more temporal links than the current model, which makes the temporal relationship stronger. We want to see how different choices of the temporal links can influence the performance of the model. We will also exploit other features such as the optical flow to complement the color features and make the model more powerful. On the other hand, although the computational complexity of loopy belief propagation is linear in the number of nodes, the computation in the pixel based CRF model is still high. A multi-scale CRF model may help ameliorate this problem.

References

[1] P. Andrey and P. Tarroux. Unsupervised segmentation of markov random field modeled textured images using selectionist relaxation. *PAMI 1998*, 20(3):252–262.

[2] P. Awasthi, A. Gagrani, and B. Ravindran. Image modeling using tree structured conditional random fields. *International Joint Conferences on Artificial Intelligence, 2007*.

[3] P. Brault and A. Mohammad-Djafari. Bayesian segmentation and motion estimation in video sequences using a markov-potts model. *WSEAS Transactions on MATHEMATICS, 2004*.

[4] M. M. Chang, A. M. Tekalp, and M. I. Sezan. Simultaneous motion estimation and segmentation. *IEEE Transactions on Image Processing*, 6(9):1326–1333, 1997.

[5] A. Criminisi, G. Cross, A. Blake, and V. Kolmogorov. Bilayer segmentation of live video. *CVPR 06*.

[6] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *PAMI*, 6(6):721–741, 1984.

[7] P. E. Gill, W. Murray, and M. H. Wright. Practical optimization. *New York: Academic Press, 1981*.

[8] J. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. *unpublished, 1971*.

[9] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. *CVPR, 2004*, 2.

[10] X. He, R. S. Zemel, and D. Ray. Learning and incorporating top-down cues in image segmentation. *ECCV 2006*.

[11] K. Held, E. R. Kops, B. J. Krause, W. M. W. III, R. Kikinis, and H.-W. Mller-Gartner. Markov random field segmentation of brain mr images. *IEEE Trans. on Medical Imaging*, 16(6), 1997.

[12] F. Jiao, S. Wang, C.-H. Lee, R. Greiner, and D. Schuurmans. Semi-supervised conditional random fields for improved sequence segmentation and labeling. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, 2006*.

[13] S. Khan and M. Shah. Object based segmentation of video using color, motion and spatial information. *CVPR 2001*, 2.

[14] S. Kumar and M. Hebert. Discriminative random fields. *International Journal of Computer Vision*, 68(2):179–201, 2006.

[15] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. 18th International Conf. on Machine Learning, 2001*.

[16] S. Z. Li. Markov random field modeling in image analysis. *Springer,2001*.

[17] Y. Liu, J. Carbonell, V. Gopalakrishnan, and P. Weigele. Protein quaternary fold recognition using conditional graphical models. *International Joint Conference in Artificial Intelligence, 2007*.

[18] A. McCallum, K. Rohanimanesh, and C. Sutton. Dynamic conditional random fields for jointly labeling multiple sequences. *NIPS 2003 Workshop on Syntax, Semantics, Statistics*.

[19] V. Murino and A. Trucco. Edge/region-based segmentation and reconstruction of underwater acoustic images by markov random fields. *CVPR 1998*.

[20] A. S. Ogale, C. Fermller, and Y. Aloimonos. Motion segmentation using occlusions. *PAMI*, 27(6):988– 992, 2005.

[21] D. K. Panjwani and G. Healey. Markov random field models for unsupervised segmentation of textured color images. *PAMI*, 17(10):939 – 954, 1995.

[22] J. Pearl. Probabilistic reasoning in intelligent systems: networks of plausible inference. *Morgan-Kaufmann Publishers Inc. 1988*.

[23] P. Pérez. Markov random fields and images. *CWI Quarterly*, 11(4):413–437, 1998.

[24] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. *Advances in Neural Information Processing Systems, 2004*.

[25] X. Ren and J. Malik. Tracking as repeated figure/ground segmentation. *Computer Vision and Pattern Recognition 2007*, 2007.

[26] C. Sutton, K. Rohanimanesh, and A. McCallum. Dynamic conditional random fields: factorized probabilistic models for labeling and segmenting sequence data. *ICML 2004*.

[27] Y. Tsaig and A. Averbuch. Automatic segmentation of moving objects in video sequences: a region labeling approach. *IEEE Trans. on Circuits and Systems for Video Technology*, 12(7):597–612, 2002.

[28] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy. Accelerated training of conditional. random fields with stochastic gradient methods. *Proceedings of the 23rd international conference on Machine learning, 2006*, 148.

[29] Y. Wang and Q. Ji. A dynamic conditional random field model for object segmentation in image sequences. *CVPR, 2005*, 1.

[30] Y. Wang, K.-F. Loe, T. Tan, and J.-K. Wu. Spatiotemporal video segmentation based on graphical models. *IEEE Trans on Image Processing*, 14(7):937–947, 2005.

[31] Y. Wang, K.-F. Loe, and J.-K. Wu. A dynamic conditional random field model for foreground and shadow segmentation. *PAMI*, 28(2):279–289, 2006.

[32] J. Weinman, A. Hanson, and A. McCallum. Sign detection in natural images with conditional random fields. *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop*.

[33] S. X. Yu, T. S. Lee, and T. Kanade. A hierarchical markov random field model for figure-ground segregation. *Lecture Notes in Computer Science, 2001*, 2134:118–133, 2001.