A Multi-layer MRF Model for Video Object Segmentation

Zoltan Kato[1] and Ting-Chuen Pong[2]

[1] University of Szeged, Institute of Informatics,
P.O. Box 652, H-6701 Szeged, Hungary
Fax:+36 62 546 397
kato@inf.u-szeged.hu

[2] Hong Kong University of Science and Technology, Computer Science Department,
Clear Water Bay, Kowloon, Hong Kong, China
Fax:+852 2358 1477
tcpong@cs.ust.hk

Abstract. A novel video object segmentation method is proposed which aims at combining color and motion information. The model has a multi-layer structure: Each feature has its own layer, called *feature layer*, where a classical Markov random field (MRF) image segmentation model is defined using only the corresponding feature. A special layer is assigned to the combined MRF model, called *combined layer*, which interacts with each feature layer and provides the segmentation based on the combination of different features. Unlike previous methods, our approach doesn't assume motion boundaries being part of spatial ones. Therefore a very important property of the proposed method is the ability to detect boundaries that are visible only in the motion feature as well as those visible only in the color one. The method is validated on synthetic and real video sequences.

1 Introduction

Video object segmentation consists of labeling pixels which are associated with different moving objects or parts. Most of the existing approaches tackle the problem by assigning a label to each pixel based on its estimated motion vector. This can be achieved in different frameworks like MRF modeling [1], mixture modeling [2], etc...The evaluation of segmentation results depends on many factors and is inherently subjective. However, many applications like MPEG-4 encoding, require that detected boundaries align with actual object boundaries. Due to the aperture problem and occlusions, motion information alone may not provide such high quality contours.

There has been some attempt to combine different features (like color and motion) in order to improve segmentation quality. In [3], color, motion and spatial information is used in a joint probabilistic model. Since features are assumed to be independent, the joint probability is split into a weighted product of the corresponding three terms. The weights assigned to the color and motion part are computed as a confidence measure, which is basically derived from the probability of the motion part. The optimal segmentation is then obtained

P.J. Narayanan et al. (Eds.): ACCV 2006, LNCS 3852, pp. 953–962, 2006.

via Maximum A Posteriori (MAP) estimation. In [4], a region based approach is proposed which relies on the assumption that motion edges are a subset of spatial edges. Therefore the method first detects regions using color and then motion segmentation is based on these regions. However, the human visual system is not treating different features sequentially. Instead, as pointed out by Kersten *etal.* [5], multiple cues are perceived simultaneously and then they are integrated by our visual system in order to explain the observations. Therefore different image features has to be handled in a parallel fashion. In this paper, we attempt to develop such a model in a Markovian framework. A very important property of our approach is that it doesn't assume motion boundaries being part of spatial ones. Therefore it is able to detect boundaries that are visible only in the motion feature as well as those visible only in the color one.

2 Multi-layer Segmentation Model

Our model consists of 3 layers. At each layer, we use a first order neighborhood system and extra inter-layer cliques (Fig. 1). Let us denote the color layer by $\mathcal{S}^c$, the motion layer by $\mathcal{S}^m$ and the combined layer by $\mathcal{S}^x$. All layers are of the same size. Our MRF model is defined over the lattice $\mathcal{S} = \mathcal{S}^c \cup \mathcal{S}^x \cup \mathcal{S}^m$. For each site s, the region-type (or class) that the site belongs to is specified by a class label, ω_s, which is modeled as a discrete random variable taking values in $\Lambda = \{1, 2, \ldots, L\}$. The set of these labels $\omega = \{\omega_s, s \in \mathcal{S}\}$ is a random field, called the *label process*. Furthermore, the observed image features (color and motion) are supposed to be a realization $\mathcal{F} = \{\vec{\mathbf{f}}_s | s \in \mathcal{S}^c \cup \mathcal{S}^m\}$ from another random field, which is a function of the label process ω. Basically, the *image process* $\mathcal{F}$ represents the deviation from the underlying label process. Thus, the overall segmentation model is composed of the hidden label process ω and the observable noisy image process $\mathcal{F}$. Our goal is to find an optimal labeling $\hat{\omega}$ which maximizes the a posteriori probability $P(\omega \mid \mathcal{F})$, that is the *maximum a posteriori* (MAP) estimate [6]: $\arg\max_{\omega \in \Omega} P(\omega \mid \mathcal{F}) = \arg\max_{\omega \in \Omega} \prod_{s \in \mathcal{S}} P(\vec{\mathbf{f}}_s \mid \omega_s) P(\omega)$, where Ω denotes the set of all possible labellings. According to the *Hammersley-Clifford theorem* [6], $P(\omega \mid \mathcal{F})$ follows a Gibbs distribution:

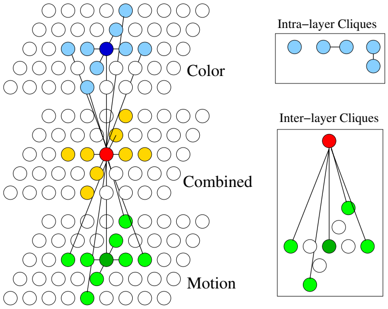


Fig. 1. Multi-layer MRF model

$$P(\omega \mid \mathcal{F}) = \frac{\exp(-U(\omega))}{Z(\beta)} = \frac{\prod_{C \in \mathcal{C}} \exp(-V_C(\omega_C))}{Z(\beta)} \tag{1}$$

where $U(\omega)$ is called the *energy function*, $Z(\beta) = \sum_{\omega \in \Omega} \exp(-U(\omega))$ is the normalizing constant and V_C denotes the *clique potential* of clique $C \in \mathcal{C}$ having the label configuration ω_C. In our model, the energy function can be further decomposed into the sum of the layer energies: $U^c + U^m + U^x$. Note that the energies of *singletons* (ie. cliques of single sites $s \in \mathcal{S}$) directly reflect the probabilistic modeling of labels without context, while higher order clique potentials express relationship between neighboring pixel labels. It is clear from Eq. (1) that the MAP estimation is equivalent to finding the global energy minimum of $U(\omega) = U^c + U^m + U^x$. Since $U(\omega)$ is a non-convex function, we have to use Simulated Annealing [6] or the ICM algorithm [7] for the minimization. In the remaining part of this section, we will define these energy functions for each layer (see Eq. (2), Eq. (5), Eq. (6)).

2.1 Color Layer

On the color layer, we use perceptually uniform CIE-L*u*v* color values where color differences can be measured by Euclidean distance. The observed image $\mathcal{F}^c = \{\vec{\mathbf{f}}^c_s | s \in \mathcal{S}^c\}$ consists of the three spectral component values (L*,u*,v*) at each pixel s denoted by the vector $\vec{\mathbf{f}}^c_s$. We assume that $P(\vec{\mathbf{f}}^c_s \mid \omega_s)$ follows a Gaussian distribution, the classes $\lambda \in \Lambda^c = \{1, 2, \ldots, L^c\}$ are represented by the mean vectors $\vec{\mu}^c_\lambda$ and the covariance matrices $\mathbf{\Sigma}^c_\lambda$. The class label assigned to a site s on the color layer is denoted by ψ_s. The energy function of the so defined MRF layer has the following form:

$$U^c = U(\psi, \mathcal{F}^c) = \sum_{s \in \mathcal{S}^c} \mathcal{G}^c(\vec{\mathbf{f}}^c_s, \psi_s) + \beta \sum_{\{s,r\} \in \mathcal{C}} \delta(\psi_s, \psi_r) + \rho^c \sum_{s \in \mathcal{S}^c} V^c(\psi_s, \eta^c_\cdot) \tag{2}$$

where $\mathcal{G}^c(\vec{\mathbf{f}}^c_s, \psi_s)$ denotes the following log Gaussian:

$$\ln\left(\sqrt{(2\pi)^3 \mid \mathbf{\Sigma}^c_{\psi_s} \mid}\right) + \frac{1}{2}(\vec{\mathbf{f}}^c_s - \vec{\mu}^c_{\psi_s}) {\mathbf{\Sigma}^c_{\psi_s}}^{-1} (\vec{\mathbf{f}}^c_s - \vec{\mu}^c_{\psi_s})^T \tag{3}$$

$\delta(\psi_s, \psi_r) = 1$ if ψ_s and ψ_r are different and -1 otherwise. $\beta > 0$ is a parameter controlling the homogeneity of the regions. As β increases, the resulting regions become more homogeneous. The last term $(V^c(\psi_s, \eta^c_\cdot))$ is the inter-layer clique potential which will be defined later in Section 2.4.

2.2 Motion Layer

Herein, we will present both an optic flow based model as well as a motion compensated color matching method.

Flow-Based Model. For this segmentation model, we use optical flow data at the motion layer. The flowfield is obtained via the algorithm proposed in [8], which provides smooth optic flow fields necessary for our MRF model. We then model each motion label by a Gaussian pdf which indicates a normally distributed noise around the mean flow. Therefore the MRF model itself is quite similar to the one outlined in the previous section. Note that this kind of modelization implicitly assumes translational motion. It is not too difficult, however, to extend our model to use parametric motion models instead of Gaussians. One such model is presented next.

Motion Compensated Model. Each region's motion is modeled by an affine model given by:

$$\begin{aligned} v_x(i,j) &= a_{x0} + a_{xx}i + a_{xy}j \\ v_y(i,j) &= a_{y0} + a_{yx}i + a_{yy}j \end{aligned} \tag{4}$$

where $v_x(i,j)$ (resp. $v_y(i,j)$) denotes the X (resp. Y) component of the flow vector at pixel (i,j). If we know the flow $\vec{v}$ at each pixel then we can warp the reference frame into the second view. When the flows are correct then the color differences between the warped and real second view must be low. Assuming n different motions in a frame, we can assign a motion label to each pixel by minimizing the warped (or motion compensated) color difference. However, we also have to deal with occlusions. Clearly, occluded pixels would have a high color difference as the warped pixel is not visible in the second frame. Therefore we allocate an additional label λ_o at the motion layer for *occlusions.* Putting these considerations together, we get the following energy function at the motion layer:

$$\begin{aligned} U^m = U(\phi, \mathcal{I}, \mathcal{I}') = &\sum_{s \in \mathcal{S}^m, \phi_s \neq \lambda_o} ||\mathcal{I}(s) - \mathcal{I}'(\vec{\mathbf{v}}(s))||^2 + \sum_{s \in \mathcal{S}^m, \phi_s = \lambda_o} V(\lambda_o) \\ &+ \beta' \sum_{\{s,r\} \in \mathcal{C}} \delta(\phi_s, \phi_r) + \rho^m \sum_{s \in \mathcal{S}^m} V^m(\phi_s, \eta^m_.) \end{aligned} \tag{5}$$

where $\mathcal{I}$ and $\mathcal{I}'$ are the reference and second frames respectively, and $V(\lambda_o)$ denotes the constant penalty for occlusion. The second and third terms are the intra- and inter-layer potentials similar to the color layer. In our experiments, we have estimated affine motion parameters using the method from [9].

2.3 Combined Layer

The combined layer only uses the motion and color features indirectly, through inter-layer cliques. A label consists of a pair of color and motion labels such that $\eta = \langle \eta^c, \eta^m \rangle$, where $\eta^c \in \Lambda^c$ and $\eta^m \in \Lambda^m$. The set of labels is denoted by $\Lambda^x = \Lambda^c \times \Lambda^m$ and the number of classes $L^x = L^c L^m$. Obviously, not all of these labels are valid for a given image. Therefore the combined layer model also estimates the number of classes and chooses those pairs of motion and color labels which are actually present in a given image. The energy function of the combined layer is of the following form:

$$U^x = U(\eta) = \sum_{s \in \mathcal{S}^x} (V_s(\eta_s) + \gamma^c V^c(\psi_., \eta^c_s) + \gamma^m V^m(\phi_., \eta^m_s)) + \alpha \sum_{\{s,r\} \in \mathcal{C}} \delta(\eta_s, \eta_r) \tag{6}$$

where $V_s(\eta_s)$ denotes singleton energies defined as

$$V_s(\eta_s) = R((10N_{\eta_s})^{-3} + \mathcal{P}(L)) \tag{7}$$

The singleton potential controls the number of classes at the combined layer: $(10N_{\eta_s})^{-3}$ penalizes small classes (N_{η_s} is the percentage of the sites assigned to class η_s), while $\mathcal{P}(L)$ includes some prior knowledge about the number of classes. Currently $\mathcal{P}(L)$ is expressed by a log Gaussian term (similar to the one in Eq. (3)) with mean value $\hat{L}$ (basically an initial guess) and variance σ (confidence in the initial guess). R is simply a weight of this term, we set it to 0.5 in our tests.

The last term of Eq. (6) corresponds to second order intra-layer cliques which ensures homogeneity of the combined layer. α has the same role as β in the color layer model and $\delta(\eta_s, \eta_r) = -1$ if $\eta_s = \eta_r$, 0 if $\eta_s \neq \eta_r$ and 1 if $\eta^c_s = \eta^c_r$ *and* $\eta^m_s \neq \eta^m_r$ or $\eta^c_s \neq \eta^c_r$ *and* $\eta^m_s = \eta^m_r$. The idea is that region boundaries present at both color and motion layers are preferred over edges that are found only at one of the feature layers.

2.4 Inter-layer Interactions

At any site s, we have an inter-layer clique $\mathcal{C}_5$ consisting of *five* interactions between two layers: Site s interacts with the corresponding site on the other layer as well as with the 4 neighboring sites two steps away (see Fig. 1). Depending on where is the site s, $V^c(\psi_\cdot, \eta^c_s)$ (s is on the combined layer) and $V^c(\psi_s, \eta^c_\cdot)$ (s is on the color layer) denote the inter-layer clique potential of the following form:

$$V^c(\psi_\cdot, \eta^c_s) = \sum_{\{s,r\}\in\mathcal{C}_5} W_r D^c(\psi_r, \eta^c_s); \qquad V^c(\psi_s, \eta^c_\cdot) = \sum_{\{s,r\}\in\mathcal{C}_5} W_r D^c(\psi_s, \eta^c_r) \tag{8}$$

where $D^c(\psi_r, \eta^c_s) =\mid \mathcal{G}^c(\vec{\mathbf{f}}^c_r, \psi_r) - \mathcal{G}^c(\vec{\mathbf{f}}^c_s, \eta^c_s) \mid$ (see Eq. (3)). $V^m(\phi_\cdot, \eta^m_s)$, $V^m(\phi_s, \eta^m_\cdot)$ and $D^m(\phi_r, \eta^m_s)$ are defined in a similar way using motion features and corresponding singleton energies. W_r is the weight of the interaction $\{s, r\} \in \mathcal{C}_5$. We assign higher weight (0.6) to the corresponding site whereas smaller weights (0.1 each) to the other 4 neighboring sites. The latter 4 sites help to ensure homogeneity on the combined layer (see Fig. 1). Note that D^c and D^m equals to the difference of the first order potentials at the corresponding feature layer. Clearly, the difference is 0 if and only if both the feature layer and the combined layer has the same label. Otherwise it is proportional to the energy difference between the two labels. γ^c (resp. γ^m) in Eq. (6) controls the influence of the inter-layer cliques. A higher value will increase the importance of the information coming from the feature layers. Furthermore, ρ^c in Eq. (2) and ρ^m in Eq. (5) controls the influence of the combined layer to the *color* and *motion* layers respectively. Therefore, depending on the ratios γ^c/ρ^c and γ^m/ρ^m, one can balance the flow of information between the *combined* and *feature* layers.

3 Experiments

The proposed algorithm has been tested on real and synthetic video sequences. The computing time was around 20 sec on a Pentium4 3GHz on 170 × 140

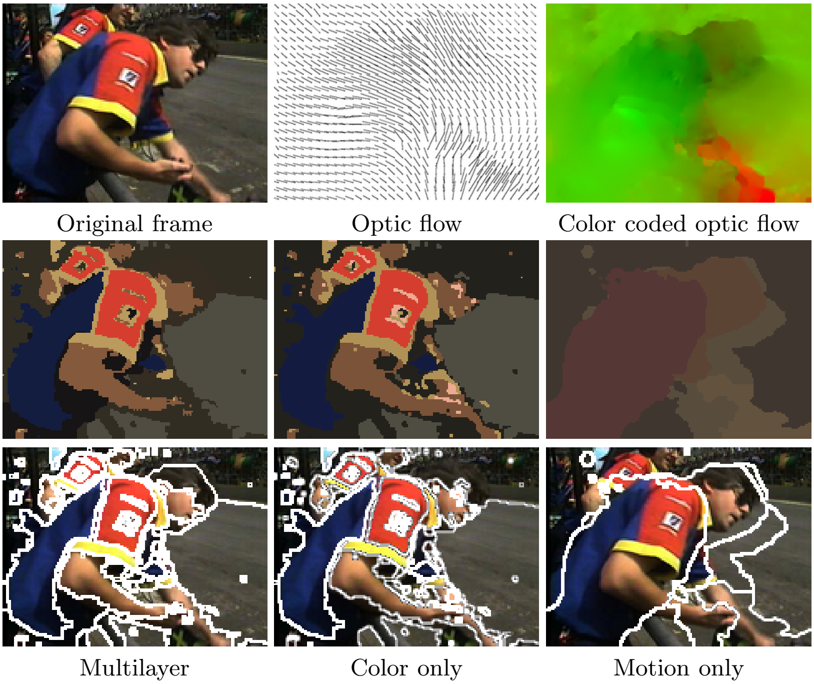
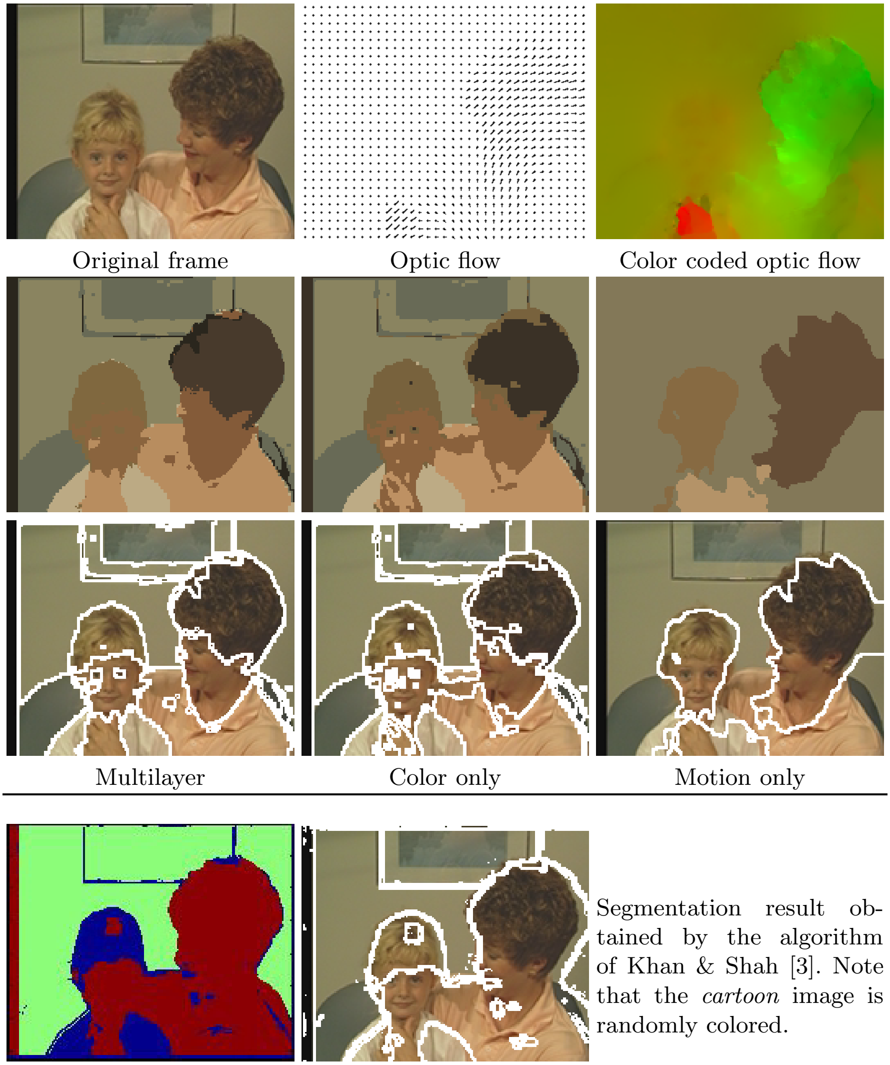


Fig. 2. Results of color only, motion only, and combined models using the *flow-based motion model*. Segmented regions are sown as a *cartoon* image (region pixels are displayed using the average color of their region) in the second row while boundaries are overlayed on the original image in the third row.

frames. Much of this CPU time is spent by the iterative optimization process (Simulated Annealing [6] or ICM [7]). However, such algorithms are known to be highly parallelizable allowing a near real time implementation on special hardware (see [10] for an example). We also compare the results to motion only and color only segmentation (basically a monogrid model similar to the one defined for the feature layers but without inter-layer cliques).

Parameter Settings. Although we do not consider parameter estimation in this paper, it is relatively easy to extend our method to handle this issue. The so called hyper parameters (the different weights of intra- and inter-layer clique-potentials) are less sensitive to the input data. We have found that one setting works for all tested sequence. Hence the only real problem is the estimation of the number of regions and the region parameters (Gaussian mean and covariance or the affine motion parameters). Since we are working on video sequences, one can naturally reuse parameters from previous frames (with some slight adjustment). As for an initial setting of the first frame, mean shift clustering has been adopted with success by many researchers [11, 12]. Once initial clusters are available, one can adopt an adaptive segmentation procedure where region parameters



Original frame

Optic flow

Color coded optic flow

Multilayer

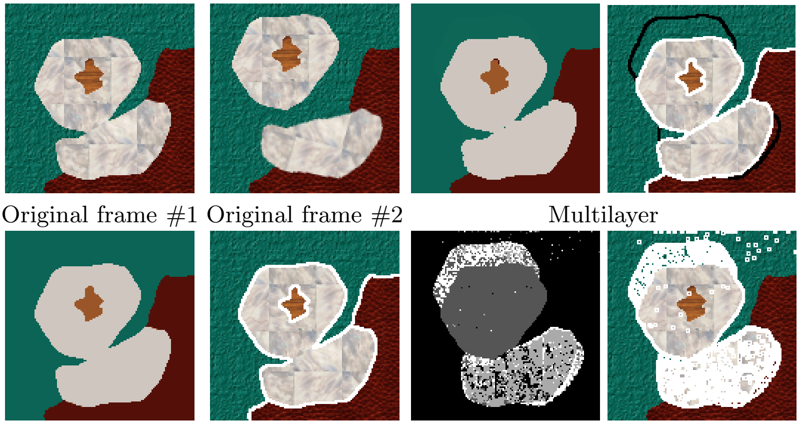
Color only

Motion only

Segmentation result obtained by the algorithm of Khan & Shah [3]. Note that the *cartoon* image is randomly colored.

Fig. 3. Results of color only, motion only, and combined models using the *flow-based motion model*. Segmented regions are shown as a *cartoon* image (region pixels are displayed using the average color of their region) in the second row while boundaries are overlayed on the original image in the third row. The last row presents the results of the method from [3].

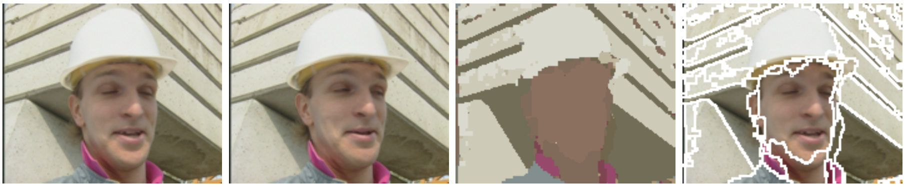
are regularly updated during the segmentation process. We have successfully applied such a technique for color textured image segmentation [12]. In the following experiments, the mean vectors and covariance matrices as well as the affine motion parameters were computed over representative regions selected by the user. The number of motion and color classes is known a priori but classes on the combined layer are estimated during the segmentation process.



Original frame #1 Original frame #2 Multilayer

Color only Motion only

Fig. 4. Results of color only, motion only, and combined models using the *motion compensated motion model*. Segmented regions are sown as a *cartoon* image (region pixels are displayed using the average color of their region) in the first column while boundaries are overlayed on the original image in the second column of the result images.



Original frame #1 Original frame #2 Multilayer

Fig. 5. Results of color only, motion only, and combined models using the *motion compensated motion model*. Segmented regions are sown as a *cartoon* image (region pixels are displayed using the average color of their region) in the first column while boundaries are overlayed on the original image in the second column of the result images.

Flow-Based Model. Fig. 2 and Fig. 3 show some segmentation results using optical flow data and Gaussian motion model. In Fig. 2, note that the head of the men can only be separated from the background using motion features. Clearly, the multi-layer model provides significantly better results compared to color only and motion only segmentations. See Fig. 3 to compare the performance of the proposed method with the one from [3] on the *Mother and Daughter* standard sequence: Some of the contours are lost by [3] (the sofa, for example) while our method successfully identifies region boundaries. In particular, our method is able to separate the hand of the mother from the face of the daughter in spite of their similar color. This demonstrates again that the proposed method is quite powerful at combining motion and color features in order to detect boundaries visible only in one of the features.

Motion Compensated Model. In Fig. 4 we present the results of a synthetic sequence using the motion compensated model. The image contains regions visible only in the color layer and boundaries visible only in the motion feature. The two white regions (one with a small painted area) are moving: the upper region is translating while the lower one is rotating around its center. Note that the moving objects are touching hence separation without motion information is not possible. Observe also that the method has detected the occluded areas (these boundaries are drawn in black). In the final segmentation, these occluded areas can be assigned to a neighboring region based on its color label. This way, a perfect segmentation can be obtained. In Fig. 5, we have used the same model on the *foreman* standard sequence. Note that the head of the men is moving hence his face is correctly separated from his neck (which is not moving). On this image, we can also see the weak point of the algorithm: when neither the color nor the motion layer can distinguish an object then it cannot be segmented. This is why the men's hat has been merged with the background: the colors are similar (white) and motion is almost impossible to detect because of the smooth homogeneous color of the hat.

4 Conclusion

We have proposed a novel multi-layer MRF segmentation model which successfully combines color and motion features. Although the current implementation doesn't estimate model parameters (except number of classes on the combined layer), it is possible to use an adaptive segmentation technique [12] to tackle this problem. Further research will concentrate on this issue as well as on using motion history in our data model.

Acknowledgment

This research was partially supported by the Janos Bolyai research fellowship of the Hungarian Academy of Sciences, the Hungarian Scientific Research Fund–OTKA T046805, and the Hong Kong research grants DAG01/02.EG16 and HIA01/02.EG04. We also thank Peter Bandi for his support in programming the algorithm.

References

1. Odobez, J.M., Bouthemy, P.: Direct model-based image motion segmentation for dynamic scene analysis. In: Proceedings of Asian Conference on Computer Vision. (1995)
2. Weiss, Y., Adelson, E.H.: A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. In: Proceedings of International Conference on Computer Vision and Pattern Recognition. (1996)

3. Khan, S., Shah, M.: Object based segmentation of video using color, motion and spatial information. In: Proceedings of International Conference on Computer Vision and Pattern Recognition. Volume II., Kauai, Hawaii, IEEE (2001) 746–751
4. Altunbasak, Y., Eren, P.E., Tekalp, A.M.: Region-based parametric motion segmentation using color information. Computer Graphics and Image Processing: Graphical Models and Image Processing **60** (1998) 13–23
5. Kersten, D., Mamassian, P., Yuille, A.: Object perception as Bayesian inference. Annual Review of Psychology **55** (2004) 271–304
6. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. on Pattern Analysis and Machine Intelligence **6** (1984) 721–741
7. Besag, J.: On the statistical analysis of dirty pictures. J. Roy. Statist. Soc., ser. B (1986)
8. Proesmans, M., Gool, L.V., Pauwels, E., Oosterlinck, A.: Determination of optical flow and its discontinuities using non-linear diffusion. In: Proceedings of Eurpoean Conference on Computer Vision. Volume 2. (1994) 295–304
9. Odobez, J.M., Bouthemy, P.: Robust multiresolution estimation of parametric motion models. Journal of Visual Communication and Image Representation **6** (1995) 348–365
10. Czuni, L., Sziranyi, T.: Motion segmentation and tracking with edge relaxation and optimization using fully parallel methods in the cellular nonlinear network architecture. Real Time Imaging **7** (2001) 77–95
11. Comaniciu, D., Meer, P.: Mean shift: A robust approach towards feature space analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence (2001)
12. Kato, Z., Pong, T.C., Song, G.Q.: Unsupervised segmentation of color textured images using a multi-layer MRF model. In: Proceedings of International Conference on Image Processing. Volume I., Barcelona, Spain, IEEE (2003) 961–964