A New Spatio-Temporal MRF Framework for Video-based Object Segmentation

Rui Huang, Vladimir Pavlovic, Dimitris Metaxas

A New Spatio-Temporal MRF Framework for Video-based Object Segmentation

Rui Huang, Vladimir Pavlovic, and Dimitris N. Metaxas

Department of Computer Science
Rutgers University, Piscataway, NJ 08854, USA
{ruihuang, vladimir, dnm}@cs.rutgers.edu

Abstract. In this paper we propose a general framework for video-based object segmentation using a new spatio-temporal Markov Random Field (MRF) model. Video-based object segmentation has the potential to improve the performance of static image segmentation by fusing information over the temporal scale. Built upon a spatio-temporal MRF model, our method offers three advantages in a unified framework. First, our model is defined on a flexible graph structure induced by the local motion information instead of the regular 3D grid, allowing the model nodes to be connected to more reliable temporal neighbors. Second, the segmentation task is considered in unison with foreground/background modeling, leading to more accurate appearance models suitable for object segmentation. Third, the inclusion of shape priors as a top-down high-level object constraints as guides for the bottom-up low-level image cues leads to improved segmentation. The object segmentation is solved as an MRF-MAP inference problem by the Loopy Belief Propagation (LBP) algorithm effectively and efficiently. Estimation of the model parameters can be accomplished simultaneously with the inference problem, using an Expectation Maximization (EM) algorithm. In our experiments, we show promising object segmentation results by combining multiple modules in the same framework.

1 Introduction

Video-based object segmentation is an important computer vision problem with many applications such as human-computer interaction, video surveillance, video indexing and retrieval, etc. Even though the modern static image segmentation algorithms have shown fairly good results with mild user interactions, it is still very difficult in practice to use these methods to segment video data frame by frame. This is because the relatively low quality and high quantity of most video data often degrade the performance of many static image segmentation algorithms, both in accuracy and running time, and sometimes forbid user interactions. On the other hand, the temporal dependencies carried in the video sequences usually provide more useful cues for potentially better and faster segmentation.

Markov Random Fields (MRFs) are a widely-used model for image analysis tasks because of their ability to deal with the noise and capture the context of images (i.e., dependencies among neighboring image pixels). They have long been successfully used in image segmentation [1] [2] [3], especially in recent years

with the newly developed efficient optimization algorithms [4] [5] such as Graph Cuts [6] [7] and Loopy Belief Propagation (LBP) [8] [9].

One particular advantage of the MRF model is that it can be easily extended to represent high dimensional data. A spatio-temporal MRF is usually obtained by adding to the regular MRFs one additional dimension, which represents time. For example, in the video-based segmentation problem, a regular 2D MRF is used to model a single frame, and all the 2D MRFs can be stacked into one 3D spatio-temporal MRF to model the whole sequence of 2D images. A spatio-temporal MRF model naturally combines the spatial and temporal aspects of a video sequence and allows one to easily explore and integrate multiple cues for video-based object segmentation.

There are four main modules in our spatio-temporal MRF framework, which involve the four most important aspects of the segmentation problem. First, the bottom-up appearance model captures the low-level cues computed from the input data (e.g., intensity, color, texture, motion field, etc.). Second, the top-down prior model brings in the high-level object priors (e.g., topology, shape, color, texture, etc.) usually learned from the training data to guide the bottom-up approaches. Note that some features such as color and texture can be used as both low-level and high-level cues, but in different ways. In the bottom-up approaches, color and texture are only used for discovering homogeneous regions, while in the top-down approaches, specific distributions of color and texture learned from the training data are sought to separate different regions. Third, the spatial constraint term imposes spatial region smoothness on the segmentation in the image plane. This constraint greatly helps eliminate the inconsistencies in the segmentation caused by the noise or other processes. Fourth, the temporal constraint term determined by the motion information imposes temporal smoothness on the otherwise unrelated static image frames. All these modules can be systematically incorporated into our framework. Furthermore, when some of these aspects are not known in advance, one can typically employ the Expectation Maximization (EM) algorithm to estimate the model parameters and perform segmentation simultaneously.

2 Related work

2.1 Combining top-down and bottom-up approaches

It has been established that a combined top-down and bottom-up segmentation approach outperforms either of the one-direction methods [10], [11]. The top-down approaches can usually obtain a coarse segmentation efficiently using the high-level object-specific prior information, which can then be significantly refined by the bottom-up approaches using the low-level image cues. MRFs are a powerful tool to deal with the low-level vision problems [9], and many efforts have been made to incorporate high-level cues into MRFs in addition to the simple Ising/Potts prior. These high-level priors can be generic shape and topology constraints, e.g., deformable models used in [12], or more object-specific shape models such as layered pictorial structures used in [13] and the stick figure model for human body used in [14]. Many of these shape prior models are formulated in the form of distance maps, commonly used in the level-set literature [15], for a

proper probabilistic interpretation. In most cases, the shape prior is not given in advance, but has to be estimated at the same time of performing segmentation. This is usually solved iteratively using the EM algorithm.

2.2 Combining spatial and temporal constraints

Background subtraction is an effective approach to detect moving regions in image sequences. The, usually low-level, features of each pixel in the background scene is modeled by a mixture of Gaussian distributions [16] [17], or a non-parametric model [18]. During the background subtraction process, false detection may occur due to the random noise or small movements of the background scene or the camera which are not captured by the background model. This can be suppressed by an additional stage of processing using spatial contextual information [18]. The other line of research addresses this problem using the spatio-temporal MRF model [19], [20], [21], [22]. The spatial smoothness is intrinsically improved due to the Ising/Potts prior in MRFs, and the temporal information is usually taken into account in the form of the MRF observation model. However, most of these methods only consider the image differences between consecutive frames as the observations, that is, they only model the motion information instead of the object or background appearance information. Hence these methods are more suitable for motion detection instead of object segmentation. Another problem is many of these models are defined on a regular 3D grid neighborhood system, i.e., each node in the current frame is connected to the nodes with the same image coordinates in the previous and next frames, which may belong to different regions due to large motions. Therefore the regular Ising/Potts prior along the time dimension may oversmooth different regions. In [20] the image frames are divided into small patches and each patch is connected to the corresponding patches, determined by patch matching, in the neighboring frames. A similar idea is suggested in [23], where the optical flow is used to detect pixel correspondences in the time dimension, and the model nodes are connected to their real temporal neighbors defined by these correspondences instead of the 3D grid.

2.3 Relation to our work

Our goal is to build a spatio-temporal MRF model for video-based object segmentation. Instead of using the motion information between the image frames for the observations as in the traditional spatio-temporal MRF models, we use such information (e.g., optical flow) to generate our model structure. Therefore our model is defined on a more flexible structure instead of the regular 3D grid, allowing the nodes in the model to be connected to more reliable temporal neighbors. The observation model of our framework is based on the foreground/background appearances, similar to the static image segmentation methods and the background modeling methods mentioned above, hence our model is more suitable for object segmentation. Furthermore, we incorporate a shape model into the otherwise bottom-up-only approach to improve the performance. To the best of the authors' knowledge, this is the first work to integrate all these different aspects for video-based object segmentation.

3 Spatio-temporal MRFs for video-based segmentation

A spatio-temporal MRF model is constructed by stacking the regular MRFs that are used to model the data at different times to form a one dimensional higher MRF model. In the video-based object segmentation setting, the spatio-temporal MRF model is three dimensional, though it is possible to have even higher dimensional models. More specifically, given a video sequence of resolution H by W and length T, the 3D spatio-temporal MRF model can be depicted by a graph that consists of $N = H \times W \times T$ nodes representing a set of random variables $\mathbf{x} = \{x_1, ..., x_N\}$. Each random variable x_i represents the label of the corresponding pixel i in the image sequence, i.e., $x_i \in L$, where L is a set of region labels, e.g., $L = \{fg, bg\}$. Each node is connected to a number of other nodes according to a neighborhood (clique) system. The connections among the nodes depict the probabilistic dependencies among the corresponding random variables, defined by compatibility functions (clique potentials). The observation at each node represents the features (e.g., intensity, color, texture, optical flow, etc.) observed at the corresponding pixel, denoted by $\mathbf{y} = \{y_1, ..., y_N\}$. Note that feature y_i can be computed not only from the single pixel i, but often from a neighborhood centered at that pixel.

The segmentation problem can be viewed as a problem of inferring the MAP solution of the MRF model:

$$\mathbf{x}_{\text{MAP}} = \arg\max_{\mathbf{x}} P(\mathbf{x}|\mathbf{y}) \tag{1}$$

which is equivalent to an optimization problem of minimizing the following energy function:

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} \phi_i(x_i) + \sum_{(i,j) \in \mathcal{N}} \psi_{ij}(x_i, x_j) \tag{2}$$

where $\mathcal{V} = \{1, 2, ..., N\}$ is the pixel/node index set, $\mathcal{N}$ is the edge set among the nodes, and $\phi_i(x_i)$ and $\psi_{ij}(x_i, x_j)$ are the unary (association) and pairwise (interaction) potential functions, respectively. Note that we discarded $\mathbf{y}$ since it is known and can be encoded into the potential functions.

The association potential function $\phi_i(x_i)$ is usually used to model the local information that can be exploited to infer the label x_i. We define this function as the sum of two different terms, corresponding to the bottom-up low-level and top-down high-level information, respectively.

$$\phi_i(x_i) = \phi_i^{low}(x_i) + \phi_i^{high}(x_i) \tag{3}$$

where the first term

$$\phi_i^{low}(x_i, y_i, \theta_{low}) = -logP(y_i|x_i, \theta_{low}) \tag{4}$$

is the traditional observation model which constrains the label to be consistent with the local low-level features, and the second term

$$\phi_i^{high}(x_i, \theta_{high}) = -logP(x_i|\theta_{high}) \tag{5}$$

is the high-level prior term which constrains the label to be consistent with our, usually object-specific, knowledge about this image location. θ_{low} and θ_{high} are the bottom-up and top-down model parameters, respectively.

The interaction potential function $\psi_{ij}(x_i, x_j)$ is a prior term used to impose region smoothness in the traditional MRFs. In the spatio-temporal MRFs, though, we argue that the temporal dimension has a different physical meaning from the spatial dimensions, hence the temporal smoothness constraints should also be different. In the image plane, when one has no knowledge about the locations of the region boundaries, the Ising/Potts model imposes spatially generic region smoothness on the whole image. Along the time dimension, however, the additional dynamic information in the video sequences implies where the discontinuities occur. For example, if one knows a point is moving from location (x, y) to $(x + \delta x, y + \delta y)$ from time t to time $t + 1$ (see the top row of Fig. 1), then there is a discontinuity between nodes (x, y, t) and $(x, y, t+1)$, even though they are a pair of neighbors in the traditional 3D MRF structure (the bottom left of Fig. 1). Instead, one should impose region smoothness between nodes (x, y, t) and $(x + \delta x, y + \delta y, t + 1)$ (the bottom right of Fig. 1). In our framework, we use such a more flexible neighborhood defined by the optical flow, and impose different smoothness constraints for spatial and temporal neighbors.

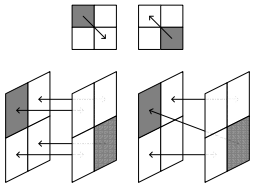

Fig. 1. Definition of temporal neighbors. Top: The optical flow in two consecutive frames; Bottom left: Define temporal neighbors by regular grid; Bottom right: Define temporal neighbors by optical flow.

More precisely, to achieve both spatial and temporal smoothness, we divide the above edge set $\mathcal{N}$ into two subsets of spatial connections $\mathcal{N}_s$ and temporal connections $\mathcal{N}_t$. The sum of the interaction potential functions hence becomes the sum of two different terms, corresponding to the spatial smoothness and temporal smoothness, respectively:

$$\sum_{(i,j)\in\mathcal{N}} \psi_{ij}(x_i, x_j) = \sum_{(i,j)\in\mathcal{N}_s} \psi_{ij}^{spa}(x_i, x_j) + \sum_{(i,j)\in\mathcal{N}_t} \psi_{ij}^{temp}(x_i, x_j) \tag{6}$$

$\mathcal{N}_s$ is defined in each frame similarly to the traditional 2D MRFs, i.e., each node is connected to its closest neighbors on the regular 2D grid. The size of the neighborhood is usually determined by different applications.

$$\psi_{ij}^{spa}(x_i, x_j, \theta_{spa}) = \begin{cases} 0 & x_i = x_j \\ \theta_{spa} & x_i \neq x_j \end{cases} \tag{7}$$

where θ_{spa} controls the strength of the smoothing effect. $\mathcal{N}_t$, on the other hand, should be defined by proper pixel correspondences in the neighboring frames.

Once $\mathcal{N}_t$ is defined, one can use an Ising/Potts model with different parameters for the temporal smoothness.

$$\psi_{ij}^{temp}(x_i, x_j, \theta_{temp}) = \begin{cases} 0 & x_i = x_j \\ \theta_{temp} & x_i \neq x_j \end{cases} \tag{8}$$

The exact MAP inference in MRFs is computationally infeasible, and various techniques have been used for approximating the MAP estimation. In our method, we use the LBP algorithm for a few preferable advantages (more discussion later). The MRF model parameters (i.e., the parameters in the potential functions) can be learned using the EM algorithm.

4 A Practical Implementation

In this section, we show an example of how to fully exploit various cues and combine them together using our framework to perform segmentation on human walking video sequences.

4.1 Incorporating bottom-up information

The bottom-up information is incorporated by the local observation model Eq. (4). In this case, we choose to use the normalized RGB color features (i.e., the chromaticity coordinates) suggested by [18] as the observation $\mathbf{y}$.

The foreground $P(y_i|x_i = fg, \theta_{fgd})$ is modeled with a mixture or Gaussian with 10 components in the normalized RGB space. The background model $P(y_i|x_i = bg, \theta_{bgd})$ is a single Gaussian at each background pixel. While there are more sophisticated background models such as the adaptive mixture model [17], the Non-parametric model [18], etc., part of our goal is to show how the simple modules can be combined and improved by our framework, so we opted for the simpler models. The foreground mixture model can be estimated from the initialization and re-estimated along with the segmentation procedure as in [12]. In other words, the pixels that are segmented as the foreground will be used to update the foreground model. A particular advantage of the LBP algorithm for the MRF-MAP inference is that a belief between 0 and 1, instead of a binary label, is assigned to each variable x_i, which allows one to update the model parameters with soft weights. The background model can be learned from the training data or, if the clean background data is not available, estimated in the same way as the foreground model.

4.2 Incorporating top-down information

The top-down information is incorporated by Eq. (5). We apply a 2D distance map based shape prior model in this case. More specifically, the shape prior we used are normalized distance maps, which are also the probability maps of the foreground given the shape prior:

$$P(x_i = fg|\theta_{shape}) = \frac{1}{1 + \exp(-\theta_{mag} \times \theta_{dist}(i))} \tag{9}$$

where θ_{mag} controls the magnitude of the shape prior, and θ_{dist} is the distance map computed from the silhouette image of the chosen shape prior. The probability maps of the background given the shape prior is simply:

$$P(x_i = bg|\theta_{shape}) = 1 - P(x_i = fg|\theta_{shape}) \tag{10}$$

On the one hand, we want to accurately model all the possible poses of the parallel frontal walking, but on the other hand, we do not want to assume the training data from a specific subject is available. Therefore we generated two sets of silhouette images using the standard male and female figure models in *Poser*® as the training data. Like other model parameters, in most cases, one needs to choose the appropriate shape prior for a particular frame, in other words, estimate the pose at the same time of performing segmentation. A simultaneous segmentation and 3D pose estimation method could be quite complicated [14]. Since we have relatively sparse samples (60 frames for one full walking cycle), we simply use the closest training instance to the current segmentation result as the shape prior for the next iteration of segmentation. However, because the usual distance metrics (e.g., Euclidean distance) in the original image space are not necessarily accurate descriptions of the distance between shapes and are very sensitive to the segmentations, one might need to employ some manifold embedding approaches to the image space as in [24] or use an explicit shape matching method, which is what we used in our implementation. Contour-based shape models are usually both effective and efficient in this scenario. We chose the profile hidden Markov model based shape model [25] due to its efficiency and robustness to missing parts and local distortions. More specifically, from a segmented silhouette image, one can easily obtain the shape contour, which is then compared to the profile models of the prior shape contours. The prior shape contour that is closest to the segmented shape contour is used to generate the distance map θ_{dist}, from which the normalized distance maps are computed according to Eq. (9). Since the shape prior images are generated from the standard figure models, and there are always differences in walking styles between the priors and the testing data, an annealing schedule is applied to the parameter θ_{mag} to weaken the top-down model and allow the bottom-up model to better refine the segmentation results. Note that the bottom-up model is re-estimated and improved over time, so its impact should be strengthen over the top-down model in the later stage of the procedure.

4.3 Incorporating Spatial constraints

We used the traditional region smoothness term as defined in Eq. (7). More complicated data-dependent terms have been used in [26] [13], which assigned different θ_{spa} values depend on the image locations and the local features, yet in our case we found the simple term works well enough.

4.4 Incorporating Temporal constraints

The temporal constraints can be simply defined in the same way as the spatial constrains, with different value of θ_{temp}. One more important and difficult task is the construction of $\mathcal{N}_t$. It has been shown in [23] that the temporal neighbors can be defined by using optical flow algorithms to detect the pixel correspondence in neighboring frames. The problem, however, is that multiple optical flows from different nodes in one frame could point to the same node in the next frame. This causes the unstable structure of the MRF model, i.e., each node in the network could have different numbers of neighbors, which forbids a simple and efficient

LBP algorithm. In our implementation, given any two consecutive frames, we compute the optical flows in both directions, i.e., one from frame t to frame $t+1$, and the other from frame $t+1$ to frame t [27]. Only those pairs of nodes that have matched flows are connected as temporal neighbors. Two nodes from neighboring frames at the same image coordinates are also considered temporal neighbors if they both do not have any optical flow, which is useful for imposing smoothness in the static background. Fig. 1 shows a simple example of this strategy. Note that some of the nodes could have no temporal neighbors at all.

4.5 Sequential loopy belief propagation

We have argued that a video sequence should be treated as a 3D spatio-temporal data instead of a batch of independent images. However, to process a whole video sequence as a single 3D volume can also be problematic, especially in computational requirements, considering the time dimension can be virtually infinite. In practice, most of the spatio-temporal frameworks apply a small window on the time dimension, i.e., work on $k(k \geq 2)$ consecutive frames at one time. Most of the time a small window can be used in the image plane as well. So only a small (but still 3D) part of the entire model is being processed at one time. The LBP algorithm is used for the inference in the small chunk of 3D data. In each step, the first frame is considered correctly segmented in the last step and the following frames are segmented, which will then be used as initialization in the next step. With this strategy, one only needs to initialize the very first frame of the video sequence. We usually start with the forward sweep with $k = 2$, i.e., two frames at one time, and after the forward propagation, backward sweep is sometimes performed to get smoother results. The whole procedure is then repeated for a few times till convergence. This procedure is essentially a limited sequential loopy belief propagation on the whole 3D data. One can increase the number k for a more aggressive message passing scheme, which may converge faster but at the cost of more memory and possible suboptimal results.

5 Experiments

5.1 Synthetic data

The first experiment is carried out on the synthetic data similar to the one used in [22]. The background is a 64×64 image whose pixels have uniformly distributed intensities between 0 and 1, and the foreground (moving object) is a 16×16 patch generated in the same way as the background. First, this is a perfect example to show the importance of the dynamic information to segmentation in video sequences. It is impossible to detect this type of camouflaged object without going through the time dimension, as shown in the first row of Fig. 2, on which we overlaid the groundtruth (we refer the readers to the attached video sequences). Second, while it might be easy for regular spatio-temporal MRFs based on image differences to recover the moving object in this noisy sequence, it is very hard for most appearance-based methods, as pointed out in [22]. Since the observation model in our framework is based on the foreground/background appearances instead of image differences, it would also be hard for our method if we ignored the temporal constraints. For example, the second row of Fig. 2 shows the segmentation results of our model defined on the regular 3D grid

structure. Since the background is static, the pixel-based Gaussian background model handles the background fairly well, but the Gaussian mixture foreground model cannot separate the object from the background clearly enough. However, this problem is rectified by the special optical flow induced temporal neighborhood structure of our model, as shown in the last row of Fig. 2. We argue that the optical flow induced temporal neighborhood system essentially encoded the image difference information used by traditional spatio-temporal MRF methods such as [19] [22]. On the other and, we even get better results than those in [22] because of the decent appearance model. Note the small holes and rough boundaries in their results, which is acceptable in tracking and motion detection, but not in segmentation.

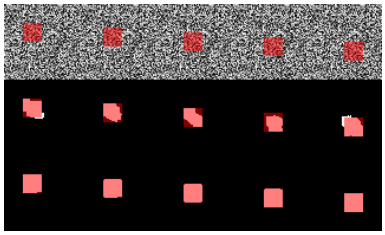

Fig. 2. Segmentation with temporal constraints. From top to bottom: (1) Input sequence (with groundtruth overlaid); (2) Segmentation results from regular 3D MRFs with appearance-based observation model (544 pixels misclassified); (3) Results from spatio-temporal MRFs with reliable temporal constraints induced by optical flow (32 pixels misclassified). From left to right: #1, #3, #5, #7, #9 out of the total 9 frames.

5.2 Real data

In this experiment, we show the segmentation results on a human walking sequence. Note that the subject is walking in place, so the major part of the upper body is only moving slightly, which makes this sequence very hard for the traditional spatio-temporal MRFs based on the image differences. On the other hand, the moving belt of the treadmill and the highly cluttered background are very hard for simple background modeling methods. To achieve good segmentation performance, one has to combine multiple cues. The first row in Fig. 3 shows image frames from the input sequence. The second row depicts the beliefs from the bottom-up observation model described in Sec. 4.1, i.e., the pixel-based Gaussian background model and the 10 components Gaussian mixture foreground model. This result can be considered as a result from a variant of ordinary background subtraction. After incorporating the spatial and temporal smoothness constraints, we can obtain the results in the third row. This procedure has the similar effect of the false detection suppressing process used in [18], and eliminates most of the random noise. Finally we incorporate the shape prior model and obtain the satisfying results in the last row of Fig. 3 using the complete framework.

Finally we show the segmentation results on a real world video sequence from the Caviar Database [28]. This sequence is more difficult than the previous one in that the image quality is lower, the motion of the subject is larger, and the reflection on the floor is severer. Another difficulty is that we do not have

clean background data for the background modeling, hence the background is initialized by averaging over the whole sequence. In the top row of Fig. 4, we show the results with all the cues combined, overlaid on the input images. We also show the estimated prior shapes which are embedded in the corresponding images. The the prior shapes are noticeably different from the test images, even though the estimated poses are close enough. Therefore, one really needs to rely on the bottom-up model to capture the fine details, especially in these low quality images. In other words, it is important to update the bottom-up model along with the segmentation. In the second row we show the improvement of the estimation of the background model. The left image is the averaged background, the middle image is the re-estimated mean of the background model based on the segmentation, and the right image is the difference between the former two. Note that the re-estimated background is much cleaner than the initial one, which has visible artifacts such as the phantoms of the human foot in multiple locations. The bottom row shows the comparison of the variance maps of the initial averaged background and the final re-estimated background. One can clearly see that the variance becomes significantly smaller after model parameter update, which means a stabler background model.

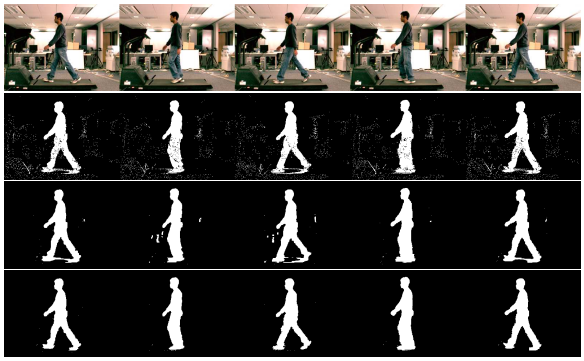

Fig. 3. From top to bottom: (1) Input sequence; (2) Segmentation results using only low-level information; (3) Results after imposing spatio-temporal constraints; (4) Final results after incorporating shape prior.

6 Discussions

In this paper we have presented a general framework for video-based object segmentation using spatio-temporal Markov random fields. This framework allows us to incorporate both top-down and bottom-up information, and impose reliable spatial and temporal constraints. The loopy belief propagation algorithm provides an effective and efficient solution for the inference problem. Moreover, one can perform segmentation and estimate the model parameters simultaneously using the EM algorithm.

While we showed results of one specific implementation that combined several relatively simple modules in our framework, one can easily replace one or more of these modules with other methods for various applications. Improvements can be made in some different aspects. For example, currently, the optical flow induced

temporal constraints has not been taken special care of. That is, we compute the optical flow as a preprocessing and do not update it like other model parameters. There could be problems caused by the failed optical flow, as shown in [23]. We are further exploring the possibility to refining the optical flow estimation and changing the temporal neighborhood structure dynamically. Another particularly interesting problem is how one can utilize the dynamic information carried in the shape prior model. Currently we use some simple assumptions to reduce the search space when we estimate the shape prior. Theoretically, the training shape sequences should follow the similar dynamics of the objects in the image sequences. Better predictions can be achieved by learning the dynamics. We are also interested in the shape manifold learning [24]. A properly learned manifold and distance metric can help us work with the shape images without resorting to an explicit shape model for the shape matching problem.

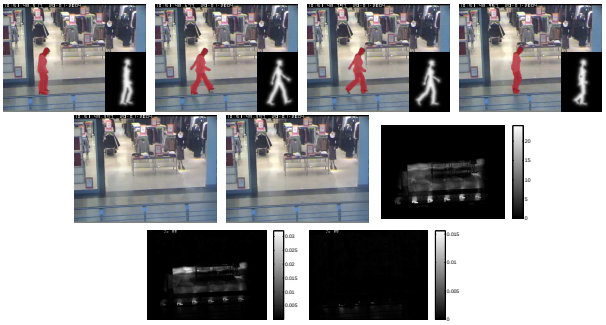

Fig. 4. From top to bottom: (1) Input sequence with output overlaid and the estimated shape prior shown in the bottom right; (2) Initial and re-estimated background models and their difference; (3) Variances of the initial and re-estimated background models.

References

1. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transaction on Pattern Analysis and Machine Intelligence **6**(6) (November 1984) 721–741
2. Besag, J.E.: On the statistical analysis of dirty pictures. Journal of the Royal Statistical Society Series B **48**(3) (1986) 259–302
3. Marroquin, J., Mitter, S., Poggio, T.: Probabilistic solution of ill-posed problems in computational vision. Journal of the American Statistical Association **82**(397) (March 1987) 76–89
4. Tappen, M., Freeman, W.: Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters. In: Proceedings of ICCV. Volume 2. (October 2003) 900–907
5. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A comparative study of energy minimization methods for markov random fields. Draft submitted to IEEE TPAMI (2007)
6. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(11) (November 2001) 1222–1239

7. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts? IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(2) (2004) 147–159
8. J. Yedidia, W.F., Weiss, Y.: Generalized belief propagation. In: Proceedings of NIPS. (2000)
9. Freeman, W., Pasztor, E., Carmichael, O.: Learning low-level vision. International Journal of Computer Vision **40**(1) (October 2000) 25–47
10. Borenstein, E., Sharon, E., Ullman, S.: Combining top-down and bottom-up segmentation. In: Proceedings of IEEE Workshop on Perceptual Organization in Computer Vision. (2004)
11. Levin, A., Weiss, Y.: Learning to combine bottom-up and top-down segmentation. In: Proceedings of ECCV. (2006)
12. Huang, R., Pavlovic, V., Metaxas, D.N.: A graphical model framework for coupling MRFs and deformable models. In: Proceedings of CVPR. (2004)
13. Kumar, M.P., Torr, P.H.S., Zisserman, A.: Obj cut. In: Proceedings of CVPR. (2005)
14. Bray, M., Kohli, P., Torr, P.H.S.: Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In: Proceedings of ECCV. (2006)
15. Rousson, M., Paragios, N.: Shape priors for level set representations. In: Proceedings of ECCV. (2002)
16. Friedman, N., Russell, S.: Image segmentation in video sequences: A probabilistic approach. In: Proceedings of Uncertainty in Artificial Intelligence. (1997) 175–181
17. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: Proceedings of CVPR. (1999)
18. Elgammal, A., Duraiswami, R., Harwood, D., Davis, L.S.: Background and foreground modeling using non-parametric kernel density estimation for visual surveillance. Proceedings of the IEEE **90**(7) (July 2002) 1151–1163
19. Luthon, F., Caplier, A., Liévin, M.: Spatiotemporal mrf approach to video segmentation: application to motion detection and lip segmentation. Signal Processing **76**(1) (1999) 61–80
20. Kamijo, S., Ikeuchi, K., Sakauchi, M.: Segmentations of spatio-temporal images by spatio-temporal markov random field model. In: Proceedings of EMMCVPR. (2001)
21. Wang, Y., Loe, K.F., Tan, T., Wu, J.K.: Spatiotemporal video segmentation based on graphical models. IEEE Transactions on Image Processing **14**(7) (2005) 937–947
22. Yin, Z., Collins, R.: Belief propagation in a 3d spatio-temporal mrf for moving object detection. In: Proceedings of CVPR. (2007)
23. Larsen, E.S., Mordohai, P., Pollefeys, M., Fuchs, H.: Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In: Proceedings of ICCV. (2007)
24. Etyngier, P., Segonne, F., Keriven, R.: Shape priors using manifold learning techniques. In: Proceedings of ICCV. (2007)
25. Huang, R., Pavlovic, V., Metaxas, D.N.: Embedded profile hidden Markov models for shape analysis. In: Proceedings of ICCV. (2007)
26. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In: Proceedings of ICCV. (2001)
27. Ogale, A.S., Aloimonos, Y.: A roadmap to the integration of early visual modules. International Journal of Computer Vision: Special Issue on Early Cognitive Vision **72**(1) (January 2007) 9–25
28. CAVIAR: The caviar database http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.