Markovian Framework for Foreground-Background-Shadow Separation of Real World Video Scenes

Csaba Benedek[1] and Tamás Szirányi[2]

[1] Pázmány Péter Catholic University, Department of Information Technology, H-1083 Budapest, Práter utca 50/A, Hungary
benedek@digitus.itk.ppke.hu
[2] Analogical Computing Laboratory, Computer and Automation Institute, Hungarian Academy of Sciences, H-1111 Budapest, Kende u. 13-17, Hungary
sziranyi@sztaki.hu

Abstract. In this paper we give a new model for foreground-background-shadow separation. Our method extracts the faithful silhouettes of foreground objects even if they have partly background like colors and shadows are observable on the image. It does not need any a priori information about the shapes of the objects, it assumes only they are not point-wise. The method exploits temporal statistics to characterize the background and shadow, and spatial statistics for the foreground. A Markov Random Field model is used to enhance the accuracy of the separation. We validated our method on outdoor and indoor video sequences captured by the surveillance system of the university campus, and we also tested it on well-known benchmark videos.

1 Introduction

Detection of foreground objects is a crucial task in visual surveillance systems. If we can retrieve the accurate shapes of the objects, their high-level description becomes much easier, so it is favorable e.g. in detection of people or activity analysis.

In the present paper, we exploit information from pixel-level estimation and neighborhood connection, while motion and structure are not considered. Based on the present results, more sophisticated segmentation methods can be developed by using tracking [12], object model matching [13], or edge information [4] [14]. However, all these developments can be preceded by an exact model on generating still background and reasonable shadow/foreground classes.

For foreground separation based on pixel intensity, Stauffer and Grimson [10] proposed an adaptive, real time algorithm, but it cannot handle some important problems. Shadows become part of moving objects, and since some parts of the objects may have similar color to the background, holes appear often in the silhouettes. The above mentioned problems can be observed on the silhouette images of Figure 1.

P.J. Narayanan et al. (Eds.): ACCV 2006, LNCS 3851, pp. 898–907, 2006.

Fig. 1. Results of foreground detection with Stauffer-Grimson algorithm. Left: School Entrance in the afternoon ('SE pm') video, right: 'Highway' test sequence.

Usually shadows have to be handled separately, because they do not belong to moving objects but their color properties are different from the background. [8] gives an overview on the state-of-the-art methods.

Classification of background, shadow and foreground areas is basically a Bayesian approach [1]. For this reason we must have statistical information about the a priori and conditional probabilities of the different clusters and the observable pixel values. The spatial interaction constraint of the neighbouring pixels can be modelled by Markov Random Fields (MRF) [5].

Previously published Bayesian models are lack of some information. They skipped shadow modelling [7][15], or the conditional probabilities of the shadow and foreground processes were oversimplified functions [9][14]. Therefore these methods are less effective on complex lighting conditions. Our goal was to develop a model with correct estimation of shadow in different lightning and coloring effects, and to detect foreground pixels of different colored and textured objects. Namely, the present paper is based on the former results, introducing more adequate models for conditional probabilities.

For validation we used real surveillance videos and also the benchmark sequences from [8]. Our model was successful in experiments with non-ideal conditions, like motley background and low contrast.

2 Markov Model

Since the work of Geman and Geman [5] there are several examples where MRFs are used for solving image-labeling problems. We used a similar model to that in [2] to classify the pixels of the video images into the following three classes: foreground (fg), background (bg) and shadow (sh). The *definitions* are the following:

S - set of pixels (or sites)

$X = \{x_s \mid s \in S\}$, - set of image data (x_s is the value of pixel s)

L={bg,sh,fg} - labels or classes.

$\Omega = \{\omega_s \mid s \in S\}$ - global labeling ($\omega_s \in L$ is the label of pixel s).

$p_k(s) = P(x_s|\omega_s = k), k \in L$ - conditional probability density function. E.g. $p_{\mathrm{bg}}(s)$ is the probability of that the background process generates the color value x_s at pixel s.

According to the model the optimal labeling is the following:

$$\widehat{\Omega} = \mathrm{argmin}_{\Omega} \sum_{s \in S} -\log p_k(s) + \sum_{r,s \in S} V(\omega_r, \omega_s) \tag{1}$$

where $V(\omega_r, \omega_s) = 0$ if s and r are not neighboring pixels, otherwise:

$$V(\omega_r, \omega_s) = \begin{cases} -\beta \text{ if } \omega_r = \omega_s \\ +\beta \text{ if } \omega_r \neq \omega_s \end{cases}$$

Our task is to define the $p_k(s)$ density functions, set the constant $\beta > 0$, and choose the energy optimization technique which finds the best or at least a good suboptimal labeling according to 1. We describe exactly how to get the $p_k(s)$ probability terms in Sections 3.1, 3.2 and 3.3. In Section 6, we show the applied MRF-optimization methods. In the following color images are considered, so the pixel value is a three dimensional vector: $x_s = [x_r(s), x_g(s), x_b(s)]$.

3 Probability Model Elements

3.1 Background Probabilities

The distribution of the color values for a given background pixel is modeled by Gaussian density function with mean value $\mu_{\text{bg}}(s)$ and covariance matrix $\Sigma_{\text{bg}}(s)$. [10] proposed an effective algorithm to determine the model parameters from the color video-flow. In [14] a similar method has already been successfully used in the MRF model. The covariance matrix is in the form of $\Sigma_{\text{bg}} = \sigma_{\text{bg}}^2 \cdot I$, where I is the 3×3 identity matrix. With this simplification we avoid matrix inversion and determinant recovering during the calculation of the probabilities:

$$p_{\text{bg}}(s) = \frac{1}{\sqrt{(2\pi)^3 \cdot \sigma_{\text{bg}}^3(s)}} \exp\left(-\frac{\|x_s - \mu_{\text{bg}}(s)\|^2}{2\sigma_{\text{bg}}^2(s)}\right) \tag{2}$$

3.2 Shadow Probabilities

[6] appointed since a shadowed pixel represents the background surface under different illumination, the effect of illumination on pixel appearance is typical for a situation. The effect was approximated by a diagonal A matrix as a multiplicative term in the RGB color space, and the shadow probabilities were directly derived from the background model:

$$p_{\text{sh}}(s) = \eta\left(x_s, A \cdot \mu_{\text{bg}}(s), A^2 \cdot \Sigma_{\text{bg}}(s)\right)$$

where $\eta(.,.,.)$ marks Gaussian density function.

In case of motley background each surface may have different reflection properties, therefore the approximation of the darkening factor with a global constant causes considerable model error. In [14] a heuristic additional shadow noise parameter was used to correct the deviation term, but in practical surveillance videos, a more sophisticated method is needed.

Instead of modelling the probability density functions of the shadowed values independently at each pixel location s, we modelled the density of the darkening ratios globally in the image. We considered one global transformation, however

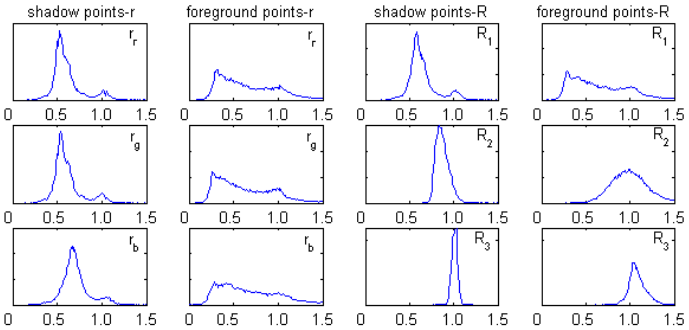

Fig. 2. Histograms for r_r, r_g, r_b, R_1, R_2 and R_3 values of shadowed and foreground points from 'SE pm' sequence

in case of images with multiple lighting and separated scene areas, the transformation parameters should be estimated in each subregion separately. With notation $\mu_{\text{bg}}(s) = [b_r(s), b_g(s), b_b(s)]$ we introduce vector containing ratios of the color values in the background and in the shadow for each pixel and for each color channel: $r(s) = [r_r(s), r_g(s), r_b(s)]$, where

$$r_r = \frac{x_r}{b_r}, \quad r_g = \frac{x_g}{b_g}, \quad r_b = \frac{x_b}{b_b}.$$

In Figure 2 the first and second columns show the histogram of the occurring r_r,r_g, and r_b values for manually marked shadowed and foreground points of the School entrance in the afternoon (SE pm) sequence. We also executed this experiment on other videos with similar results. We can observe, if we neglect the small second peaks, the 1 dimensional ratio values in shadow have approximately Gaussian distribution. However, Table 1 shows that the correlation between the elements of vector r is high, so if we model the shadowed r ratios with Gaussian distribution, the covariance matrix cannot be considered diagonal. Therefore we have searched for further quantities, and found the following ones: $R = [R_1, R_2, R_3]$

$$R_1 = \frac{r_r + r_g + r_b}{3}, \quad R_2 = \frac{r_r}{r_b}, \quad R_3 = \frac{r_g}{r_b},$$

In Figure 2 and Table 1 we can observe R_1, R_2, and R_3 values are generated also approximately by Gaussian distribution, but their correlation is definitely smaller. Therefore we characterize shadow via R values. The resulting shadow

Table 1. Average of the absolute values of nondiagonal elements in the autocorrelation matrix for r and R values of shadowed points

	Corr(r)	Corr(R)
SE pm:	0.967	0.374
Highw:	0.987	0.360

probability term for pixel s, and parameters of our shadow model are the following:

$$p_{\mathrm{sh}}(s) = \eta\left(R(s), \mu_{\mathrm{sh}}, \Sigma_{\mathrm{sh}}\right) \tag{3}$$

$$\mu_{\mathrm{sh}} = [\mu_{\mathrm{sh},1}, \mu_{\mathrm{sh},2}, \mu_{\mathrm{sh},3}], \quad \Sigma_{\mathrm{sh}} = diag\{\sigma^2_{\mathrm{sh},1}, \sigma^2_{\mathrm{sh},2}, \sigma^2_{\mathrm{sh},3}\}. \tag{4}$$

3.3 Foreground Probabilities

The description of background and shadow characterizes the scene and lighting properties so it is possible to collect statistical information about them in time. Unfortunately, the color distribution of foreground areas is unpredictable in the same way. However it is often inappropriate to model the foreground by uniform distribution, like in [9][14]. Figure 3 shows some resulting segmented images after applying MRF optimization for our background and shadow model but using uniform foreground distribution. Since the objects may have large background or shadow-like connected parts, big holes appear in the silhouettes, and the suggested Markovian model cannot remove these errors.

Instead of temporal statistics we used spatial color information to overcome this problem. First we assume that a pre-processing step is able to locate most of the foreground pixels. That process, which we introduce in Section 4, gives a preliminary foreground mask to the algorithm. Denote F the set of pixels marked as foreground elements in that mask. We have two assumptions for a given foreground pixel:

- In the neighborhood there are some foreground pixels
- The color of the pixel matches to the color distribution of set of the neighbouring foreground pixels.

In the following V_s denotes the set of the neighbouring pixels around s, considering rectangular neighborhood with window size v. F_s is the set of neighbouring pixels determined as 'foreground' by the preprocessing step: $F_s = F \cap V_s$. To deal with textured or multi level foreground components, the estimated probability density function of the color channels for F_s is in the following form:

$$f_{F_s,x_s}(x) = w_s \cdot \eta(x, \mu_{\mathrm{fg}}(s), \Sigma_{\mathrm{fg}}(s))) + (1 - w_s) \cdot f(x)$$

Namely, we divide the neighborhood pixels in two clusters: the ones, whose color-distance from x_s is smaller than a threshold, are characterized by one Gaussian term, while $f(x)$ is the residual density function with constraint: $f(x) = 0$, if

Fig. 3. Results of using MRF model with uniform foreground distribution

$\|x_s - x\| < \tau$, $0 < w_s < 1$. Accordingly, the color values of the site s are statistically characterized by the distribution of its neighborhood in the color domain:

$$p_{\text{fg}}(s) = f_{F_s,x_s}(x_s) = w_s \cdot \eta(x_s, \mu_{\text{fg}}(s), \Sigma_{\text{fg}}(s)). \tag{5}$$

To approximate the foreground model parameters we compose a subset of F_s by

$$F_s^D = \{r \mid r \in F_s,\ \|x_s - x_r\| < \tau\}.$$

Empirical mean value and deviation of the pixel values in F_s^D estimate the parameters $[\mu_{\text{fg}}(s), \Sigma_{\text{fg}}(s)]$. Weight w_s is calculated as a ratio of the cardinality of sets F_s^D and F_s. We also used an extra term to keep the probability low, if there are any or only a few pre-classified foreground pixels in the neighborhood.

4 Preliminary Foreground-Shadow-Background Classifier

The foreground model introduced in Section 3.3 needs a pre-processing step, which is able to find most of the foreground pixels. To achieve this task we used a deterministic classifier which uses the existing background and shadow model parameters from Section 3. The background matching step is the same as it was used in [10]. Pixel s is classified as background, if:

$$\|x_s - \mu_{\text{bg}}(s)\|^2 < 2c \cdot \sigma_{\text{bg}}^2(s)$$

Non-background the pixels are matched to the shadow constraints and labeled as shadow, if

$$(R_i(s) - \mu_{\text{sh},i})^2 < 2c/3 \cdot \sigma_{\text{sh},i}^2, \quad i \in \{1, 2, 3\}$$

Other way the pixel gets foreground label.

5 Parameter Settings

Our method has scene dependent and condition dependent parameters. *Scene dependent* parameters can be considered constant in a specific field, and are influenced by e.g. camera settings, expected size and shape of the objects or reflection properties. We give strategies how to set these parameters given a territory of a surveillance camera. *Condition dependent* parameters vary in time in a scene, we used adaptive algorithms to follow them.

The background parameter estimation and update procedure is automated, based on the work of [10]. It has a parameter (α in [10]), which controls the speed of model update. In our experiences it was set uniformly to 0.02.

5.1 Foreground Model Parameters

The foreground parameters are scene dependent constants. Window size s depends on the expected size of the objects in the scene. If T_B is the approximate average territory of the objects bounding boxes, we used $v = 1/3\sqrt{T_B}$.

The threshold parameter τ defines the maximum distance in the RGB color space between pixels generated by one Gaussian process. We used outdoors $\tau = 50$, indoors $\tau = 20$.

5.2 Shadow Parameters

The parameters are defined by Eq. 4. Except of window-less rooms with constant lightning, $\mu_{\mathrm{sh},1}$, the average background luminance darkening factor in shadow is strongly condition dependent. Outdoors, it can vary from 0.4 in sunburst to 0.9 in overcast weather. We observed the other shadow parameters (5 scalar values more) being approximately constant in time, letting us to estimate them once in a scene.

We built an adaptive algorithm to follow the changes of $\mu_{\mathrm{sh},1}$. For a given image we collected histogram from the R_1 values of those pixels, which are marked as non background point by the Stauffer-Grimson algorithm. If the image contains considerable shadowed parts, a peak appears in the histogram near the desired $\mu_{\mathrm{sh},1}$ value. Figure 4 shows 3 typical situations from the video 'SE pm', where the optimal $\mu_{\mathrm{sh},1}$ was definitely 0.68. On the first image, a large shadow is observable, and the peak in the histogram is very significant. On the second one, the peak is still in the right place, however it is smaller. On the third image there is small shadow and the histogram is flat. Denote $h[k]$ the location of the peak in the histogram of the k-th image, $v[k]$ is the maximum value, $\overline{v}[k]$ is the average value. $h[k]$ can be a good estimation for $\mu_{\mathrm{sh},1}$, if peak-value $v[k]$ is high and significant: $\frac{v[k]}{\overline{v}[k]}$ is high. We define the update process by the following:

$$\mu_{\mathrm{sh},1}[k+1] = \rho \cdot h[k] + (1-\rho) \cdot \mu_{\mathrm{sh},1}[k], \qquad \rho = \alpha \cdot v[k] \cdot \frac{v[k]}{\overline{v}[k]}$$

where $\alpha = 0.001$ is a constant factor, and we perform the parameter update only, if there are enough non-background points in the image.

We tested this method on videos recorded by the 'School entrance' camera in case of ten different lightning conditions, and appointed it can follow the lightning changes caused by clouds well, or in case of randomly chosen $\mu_{\mathrm{sh},1}$ it finds the correct value quite fast. However the performance of the adaption was lower round noon, when the shadows are smaller, and the corresponding darkening ratio is not so dominant in the statistics.

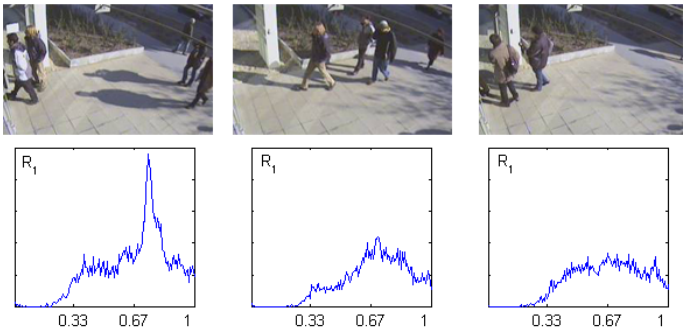

Fig. 4. Three images from sequence 'SE pm' and the corresponding histograms for the R_1 values of the non-background pixels

6 MRF Optimization and Speed of the Algorithm

The presented algorithm segments the video images via MRF optimization. First, the probability terms $p_{\mathrm{bg}}(s)$, $p_{\mathrm{sh}}(s)$, $p_{\mathrm{fg}}(s)$ are calculated for each pixel s, according to (2)(3)(5). The second level is to find a good labeling considering the energy term of (1). The results showed on Figure 5 were made using the Modified Metropolis method [2], which is not real time on a sequential architecture, however [11] have already suggested a fast parallel implementation for a special array processor.

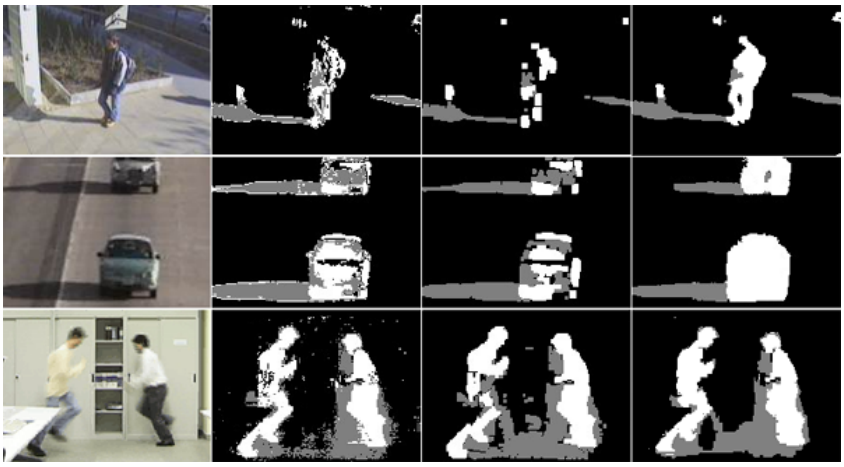

Fig. 5. *Segmentation results.* 1st column: video image, 2nd: result of the preliminary classifier, 3rd: pre. classifier result enhanced by morphology, 4th: MRF result. *Images* are from the following videos: a) Sequence 'SE pm', b) 'Highway', c) 'Laboratory'.

A well-known quick deterministic optimization method for MRF is the ICM algorithm, which gives a good sub-optimal solution in a few (2-5) iteration of steps with linear complexity. Although the quality of the segmentation produced by ICM is significantly worse than the we got by MMD, it is still enough for connected component based object detection.

We have tested out method on color videos with the resolution 320×240. The running speed was 2 fps using Intel Pentium 4 2400 MHz Processor.

7 Results

Model verification was made through manually generated ground truth sequences. Since the goal is foreground detection, the crossover between shadow and background does not count for errors.

Denote with TP (*true positive*) the number of correctly identified foreground pixels of the evaluation sequence. Similarly we introduce TN for well classified

Table 2. *Evaluation result.* SG: Stauffer-Grimson algorithm (without shadow filtering), Pre: preliminary classifier, Mor: the output of pre. enhanced by morphology, MMD: the result got by our MRF model, with MMD optimization. 'SE am' sequence was recorded in the morning by the campus' camera and contains large shadows.

	Fg. detection rate (D) %				Fg. accuracy rate (A) %			
Sequence	SG	Pre.	Mor.	MMD	SG	Pre.	Mor.	MMD
SE am	83.7	78.6	72.7	93.1	38.3	76.8	88.0	86.9
SE pm	82.9	67.6	66.7	80.7	62.5	79.3	88.4	90.1
Highw	87.4	56.5	43.9	83.1	55.9	78.2	88.8	88.5
Lab.	95.3	88.7	94.7	93.2	54.3	89.8	92.4	93.8

non-foreground points, FP for misclassified non-foreground points, and FN for misclassified foreground points.

Evaluation metrics: D is the foreground detection rate, A is the accuracy of the detection.

$$D = \frac{TP}{TP + FN} \qquad A = \frac{TP}{TP + FP}$$

The results in Table 2 are valid without postprocessing. The applied MRF model increased significantly the foreground detection and accuracy rate, compared to the deterministic step. We tried to reach homogenous regions by applying morphology on the output of the deterministic classifier but at the same time the D and A ratios became much worse. The improvement is remarkable in the difficult scenes, while on the 'Laboratory' benchmark sequence the simpler methods gave also very good results. Some examples for segmented images are in Figure 5.

8 Conclusion and Future Work

We introduced a realistic model of shadow effects and a new foreground probability calculus for segmenting videos by MRF model optimization. We measured significant improvements versus previous methods in real world videos, where the background and foreground is textured, and the color ranges of the different clusters are strongly overlapping. Our future work is to improve the automated parameter estimation process, and to speed up energy calculation of the foreground model. We want to complete our method with texture analysis, and exploit the advantages using more adequate color spaces (CIE-L*a*b* or CIE-L*u*v*). We will try to deal with difficult situations like shadow in the shadow and reflection from glass doors.

References

1. Cs. Benedek, T. Szirányi: A Markov Random Field Model for Foreground-Background Separation, Joint Hungarian-Austrian Conference on Image Processing and Pattern Recognition (HACIPPR), Veszprém, Hungary, May 11-13, (2005)
2. M. Berthod, Z. Kato, S. Yu, J. Zerubia: Bayesian image classification using Markov Random Fields. Image and Vision Computing 14 (1996) 285-295

3. R. Cucchiara, C. Grana, G. Neri, M. Piccardi, and A. Prati: The Sakbot System for Moving Object Detection and Tracking. Video-Based Surveillance Systems-Computer Vision and Distributed Processing (2001) 145-157
4. L. Czúni, T. Szirányi: Motion Segmentation and Tracking with Edge Relaxation and Optimization using Fully Parallel Methods in the Cellular Nonlinear Network Architecture. Real-Time Imaging Vol.7, No.1, (2001) 77–95
5. S. Geman and D. Geman: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence (1984) 721-741
6. I. Mikic, P. Cosman, G. Kogut and M. M. Trivedi: Moving Shadow and Object Detection in Traffic Scenes, Proc. ICPR, (2000) 321-324
7. N. Paragios, V. Ramesh. A MRF-based Real-Time Approach for Subway Monitoring. In IEEE Conference in Computer Vision and Pattern Recognition (CVPR), (2001) 1034-1040
8. A. Prati, I. Mikic, M. M. Trivedi, R. Cucchiara: Detecting moving shadows: algorithms and evaluation. PAMI(25), (2003) 7, pp. 918–923
9. J. Rittscher, J. Kato, S. Joga and A. Blake: A Probabilistic Background Model for Tracking Proc. European Conf. Computer (2000)
10. C. Stauffer and W. E. L. Grimson: Learning Patterns of Activity Using Real-Time Tracking, IEEE Trans. Pattern Anal. Mach. Intell. (2000) 22(8): 747-757
11. T. Szirányi, J. Zerubia: Markov Random Field Image Segmentation using Cellular Neural Network , IEEE Tr. Circuits and Systems (1997) I., V.44, pp.86-89,
12. A. Yilmaz, X. Li, M. Shah Object Contour Tracking Using Level Sets. Asian Conference on Computer Vision, ACCV 2004, Jaju Islands, Korea, (2004)
13. P. Viola, M. Jones: Rapid Object Detection Using a Boosted Cascade of Simple Features, Proc. IEEE Conf. Computer Vision and Pattern Recognition, (2001)
14. Y. Wang, T. Tan, and K.-F. Loe:A Dynamic Hidden Markov Random Field Model for Foreground and Shadow Segmentation Seventh IEEE Workshops on Application of Computer Vision, Breckenridge, Colorado, (2005)
15. Yue Zhou, Yihong Gong, and Hai Tao: Background segmentation using spatial-temporal multi-resolution MRF, IEEE Motion05, (January 2005)