Journal of Zhejiang University SCIENCE A
ISSN 1673-565X (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Adaptive foreground and shadow segmentation using hidden conditional random fields*

CHU Yi-ping†, YE Xiu-zi†‡, QIAN Jiang, ZHANG Yin, ZHANG San-yuan
(*School of Computer Science, State Key Lab. of CAD & CG, Zhejiang University, Hangzhou 310027, China*)
†E-mail: hzcyp@yahoo.com.cn; yxz@cs.zju.edu.cn
Received Aug. 15, 2006; revision accepted Oct. 26, 2006

Abstract: Video object segmentation is important for video surveillance, object tracking, video object recognition and video editing. An adaptive video segmentation algorithm based on hidden conditional random fields (HCRFs) is proposed, which models spatio-temporal constraints of video sequence. In order to improve the segmentation quality, the weights of spatio-temporal constraints are adaptively updated by on-line learning for HCRFs. Shadows are the factors affecting segmentation quality. To separate foreground objects from the shadows they cast, linear transform for Gaussian distribution of the background is adopted to model the shadow. The experimental results demonstrated that the error ratio of our algorithm is reduced by 23% and 19% respectively, compared with the Gaussian mixture model (GMM) and spatio-temporal Markov random fields (MRFs).

Key words: Video segmentation, Shadow elimination, Hidden conditional random fields (HCRFs), On-line learning
doi:10.1631/jzus.2007.A0586 **Document code:** A **CLC number:** TP391.7

INTRODUCTION

Video object segmentation is important for video surveillance and object tracking. There were many approaches for video object segmentation presented previously (Stauffer and Grimson, 2000; Stenger *et al*., 2001; Yang *et al*., 2004; Zivkovic, 2004). Since application environment may be complicated, the segmentation algorithms should tolerate camera shake, illumination changes and dynamic background outdoors. Shadows are other factors affecting video segmentation quality. In this paper, we only consider the shadows cast by intruding objects. To obtain better segmentation quality, the algorithms must correctly separate foreground objects from the shadows they cast.

Yang *et al*.(2004) proposed a simple and efficient method to segment video object, whose accumulated pixels change slowly as the background. The scheme provides two kinds of methods to update the background, with pixel level and frame level. It subtracts the background from the current frame in order to obtain the video foreground. The advantages of the method are ease of implementation and lower complexity, though it may not perform well with dynamic backgrounds containing sharp lighting changes and wavy leaves. The Gaussian mixture model (GMM) (Stauffer and Grimson, 2000; Zivkovic, 2004) is a popular approach to model the background, which can model complex dynamic background. Stenger *et al*.(2001) presented a topology free hidden Markov model and applied it to background modeling. This background modeling approach is robust to sudden changes of illumination. Both background modeling methods perform at pixel level without containing pixel neighboring constraints.

Spatial and temporal neighboring relationships of pixels are useful information for object segmentation. Markov random fields (MRFs) are desirable tools to model the spatio-temporal neighboring con-

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (Nos. 60473106, 60273060 and 60333010), the Ministry of Education of China (No. 20030335064), and the Education Department of Zhejiang Province, China (No. G20030433)

straints. Migdal and Grimson (2005) adopted MRFs to model spatio-temporal neighboring relationship, the segmentation problem is solved under the MAP framework with Gibbs sampling. Zhou *et al*.(2005) introduced an algorithm based on hierarchical MRFs, which segments the foreground objects accurately from scene with camera noise and shake. Sheikh and Shah (2005) proposed a background model as a single probability density by using a nonparametric density estimation method. The background and foreground models are used competitively in a MAP-MRF decision framework.

Conditional random fields (CRFs) introduced by Lafferty *et al*.(2001) are discriminative probabilistic models most often used for labelling or parsing of sequential data. CRFs are also applied to other areas such as image object recognition, image classification (Sha and Pereira, 2003; Kumar and Hebert, 2003; 2005). Wang and Ji (2005) and Wang Y. *et al*.(2006) introduced the CRFs to video object segmentation. Spatio-temporal neighboring constraints are modeled by CRFs model, where a filter is constructed for updating parameters of CRFs according to previous frame data. Quattoni *et al*.(2004) introduced a hidden variable into the CRFs to solve the problem of image object recognition, called hidden conditional random fields (HCRFs). HCRFs were also introduced to phone classification (Gunawardana *et al*., 2005) and gesture recognition (Wang S. *et al*., 2006).

Martel-Brisson and Zaccarin (2005) used GMM to build statistical models describing moving cast shadows on surfaces. The means of models are adopted to discriminate shadow and foreground. Porikli and Thornton (2005) applied a weak classifier as a pre-filter, and projected shadow models into a quantized color space to update a shadow flow function. The Bayesian method was adopted to update the parameters of models. Wang Y. *et al*.(2006) considered shadows as the linear transform for Gaussian component of the corresponding background. Parameters of shadow models can be computed from the corresponding background models.

In this paper, a novel adaptive video segmentation algorithm based on HCRFs is proposed, spatio-temporal neighboring constraints are modeled via HCRFs. To adjust weights of spatial and temporal neighboring relationship according to scene changes, on-line learning method is exploited to update the parameters of HCRFs. Shadows are modeled by Gaussian function and determined by the threshold in our method.

The rest of this paper is organized as follows. In the next section we will review previous work on HCRFs, and then, the HCRF models are modified for video segmentation. To adjust weights of spatial and temporal neighboring relationship adaptively, on-line learning is used for updating parameters. The method to construct shadow models will be given then. In Section 3, different experimental data are used to test the proposed method. To compare the results, other segmentation methods are also tested in our experiment. Finally in Section 4, we give our conclusion.

HCRFS FOR VIDEO SEGMENTATION

HCRFs were proposed by Quattoni *et al*.(2004) for image object recognition. $X=\{x_1, x_2, \ldots, x_m\}$ is an observation set of pixels in video sequence, with each element x_j in X being local observation. There is a hidden random variable set $H=\{h_1, h_2, \ldots, h_m\}$ corresponding to X. For $\exists h_i \in H$, its spatio-temporal neighboring variables are denoted by h_k ($k \in N_i \cup M_i$), where N_i denotes spatial neighboring relationships of site i, and M_i stands for temporal neighboring relationships with previous frame. H and its corresponding neighboring relationships form an undirected graph, which is a tree in this paper. We use E to denote the set of edges in the graph, and $(i, k) \in E$ denote that hidden variables h_i and h_k have neighboring relationship. $L \in \{0, 1, 2\}$ are labels, where 0, 1 and 2 stand for background, shadow and foreground, respectively. In this paper, each $h_i \in H$ has the same range as L.

HCRF model

HCRF models the conditional probability of a segmentation label L given the observation sequence X:

$$P(L \mid X;\theta) = \sum_H P(L,H \mid X;\theta) = \frac{1}{Z(X;\theta)} \sum_H \exp[\Psi(L,H,X;\theta)], \quad (1)$$

in which $Z(X; \theta)$ is the partition function, ensuring

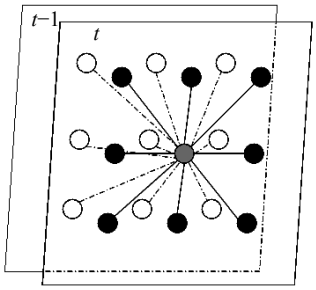

Fig.1 Spatio-temporal neighboring relationships in video sequence. The sites colored by black are spatial neighboring relationships for gray site in the same frame, while gray site in the *t*th frame and white sites in previous frame form temporal neighboring relationships

that the model is a properly normalized probability. It can be calculated by the formula $Z(X;\theta)=\sum_{L',H}\exp[\Psi(L',H,X;\theta)]$. H is not observed on training examples, and will therefore form a set of hidden variables in the model. $\Psi(L, H, X; \theta)$ is the feature function, which is defined as follows:

$$\Psi(L,H,X;\theta)=\sum_i\sum_k\theta_k^1 f_k^1(L,h_i,x)+\sum_{(i,j)\in E}\sum_k\theta_k^2 f_k^2(L,h_i,h_j,x), \quad (2)$$

where both θ_k^1 and θ_k^2 are parameters to be learned, feature function $f_k^1(L,h_i,x)$ denotes the strength depending on single hidden variable, and $f_k^2(L,h_i,h_j,x)$ depends on two hidden variables h_i and h_j.

We use GMM to classify local observation, if the single hidden variable was classified as the background, the Mahalanobias distance for the background will be used to compute feature function. We will discuss how to obtain shadow models in the subsection "Shadow model". We assume that the foreground has uniform distribution, therefore probability of the foreground in RGB color space is $\exp(-\ln 2^{24})$ (Migdal and Grimson, 2005). Feature function $f_k^1(L,h_i,x)$ is defined as

$$f_k^1(L,h_i,x)=\begin{cases}\lambda_b/2, & h_i=0,\\ \lambda_s/2, & h_i=1,\\ \ln 2^{24}, & h_i=2,\end{cases} \quad (3)$$

where λ_b is the Mahalanobias distance computed by $\lambda_b=(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the mean and variance of the matched background model in GMM, T denotes the matrix transpose operator, λ_s is the Mahalanobias distance of the shadow model.

Neighboring relationship can provide useful information for improving segmentation quality. Fig.1 shows 8-spatial neighboring and 9-temporal neighboring relationships between two consecutive frames. These neighboring relationships form cliques, in which sites are mapped to the hidden variables. The feature function for neighboring strength represents interaction of local classification between neighboring hidden variables, we define it as

$$f_k^2(L,h_i,h_j,x)=\delta(h_i,h_j), \quad (4)$$

where $\delta(\cdot)$ is the Kronecker delta function.

Labelling and parameter learning

Given the sequence X and model parameters θ, we can label pixel to be

$$\hat{L}=\arg\max_L P(L\mid X;\theta). \quad (5)$$

We can infer the model and obtain the following formula:

$$P(L\mid H,X;\theta)=\frac{1}{Z(L,X;\theta)}\exp[\Psi(L,H,X;\theta)], \quad (6)$$

where $Z(L, X; \theta)=\sum_H\exp[\Psi(L, H, X; \theta)]$, which can be computed by belief propagation. The normalization constant $Z(L, X; \theta)$ can be computed by the same inference algorithm that we use to compute $Z(X; \theta)$. In fact, calculating $Z(L, X; \theta)$ is easier than that of $Z(X; \theta)$, because $Z(L, X; \theta)$ only sums over H, while $Z(X; \theta)$ need to sum over H and L. Once we have $Z(L, X; \theta)$, the marginal likelihood can be calculated as

$$P(L\mid X;\theta)=\frac{1}{Z(X;\theta)}\sum_H\exp[\Psi(L,H,X;\theta)]=Z(L,X;\theta)/Z(X;\theta). \quad (7)$$

Eq.(5) can perform efficiently in the model by Eq.(7).

In (Wang S. *et al*., 2006), HCRFs search for the

optimal parameter values, $\hat{\theta}$=argmax$_\theta$ $l(\theta)$, by using gradient ascent with Quasi-Newton optimization technique. $l(\theta)$ is objective function for batch training; but for video segmentation, it is more suitable to learn parameters by means of on-line training method. In our experiments, we train model parameters by using segmentation results of previous frames in order to update parameters according to scene changes.

We adopt stochastic gradient descent (SGD) (Gunawardana *et al*., 2005) to update model parameters. Given training data $(L^{(i)}, X^{(i)})$, i=1, 2, …, the parameters are updated by

$$\theta_i^{(n+1)} = \theta_i^{(n)} + \eta^{(n)} \nabla_{\theta_i} \log P(L^{(i)} \mid X^{(i)}; \theta), \qquad (8)$$

where $\eta^{(n)}$ is the learning rate we set to a constant value in this paper. $\nabla_{\theta_i} \log P(L^{(i)} \mid X^{(i)}; \theta)$ is the gradient for each single training sample, θ_k^1 and θ_k^2 can be computed by

$$\nabla_{\theta_k^1} \log P(L^{(i)} \mid X^{(i)}; \theta) = \sum_{h_i \in H} P(h_i \mid L^{(i)}, X^{(i)}) f_k^1(L, h_i, x) - \sum_{h_i \in H, L'} P(h_i, L' \mid X^{(i)}) f_k^1(L', h_i, x). \qquad (9)$$

Analogously, we have

$$\nabla_{\theta_k^2} \log P(L^{(i)} \mid X^{(i)}; \theta) = \sum_{(i,j) \in E} P(h_i \mid L^{(i)}, X^{(i)}) f_k^2(L, h_i, h_j, x) - \sum_{(i,j) \in E, L'} P(h_i, L' \mid X^{(i)}) f_k^2(L', h_i, h_j, x), \qquad (10)$$

where both $P(h_i|L_i, X_i)$ and $P(h_i|L', X_i)$ can be computed by belief propagation.

Shadow model

Shadows are generated from objects moving under sunshine and light. In fact, the shadows we need to eliminate are the background illumination shaded by intruding objects. We assume that shadow decreases the luminance and changes the saturation, yet it does not affect the hue. To construct the shadow models, we transform corresponding background models by linear operator. As in (Wang Y. *et al*., 2006), we adopted linear transform method to model shadow. Given a pixel belonging to the background, which has Gaussian distribution with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$; if the pixel is shaded by intruding objects, it becomes Gaussian function $N(\boldsymbol{x}, \alpha\boldsymbol{\mu}, \beta\boldsymbol{\Sigma})$, where both α and β are linear factors. Therefore, we can compute the shadow model for each pixel from corresponding background Gaussian models.

After obtaining shadow models, we can use them to determine whether the pixel is shadow or not by the variance threshold. When pixel intensities change, the parameters of background models will be updated, as will be those of shadow models. To reduce computation, we update parameters of the shadow model only when the current pixel does not match the background model. If a pixel is unmatched to the background in GMM, the shadow model will be updated by the previous background Gaussian function. The current shadow model will be used to determine whether the pixel is classified as shadow or foreground.

Implementation

We label each local observation x_i by using GMM method, the Mahalanobias distance is calculated meanwhile. We just set θ_k^1 to 1 in the paper. Parameters θ_k^2 for neighboring constraints $(i, j) \in E$, $j \in N_i$ are initialized by $1/\|h_i - h_j\|$, where $\|h_i - h_j\|$ denotes Euclidean distance in images between two hidden variables. This distance can be pre-computed, because we adopt fixed number of neighboring relationships in our program.

The parameters θ_k^2 for temporal neighboring constraints $(i, j) \in E$, $j \in M_i$ are initialized by $\sqrt{2}/(2|M_i|)$, where $|M_i|$ denotes the number of temporal neighboring relationships for h_i in the previous frame. Because $|M_i|$ is constant, θ_k^2 can be pre-computed too. Each parameter θ_k^2 has maximum and minimum threshold, when it goes over the range in the learning procedure, the parameter will be initialized over once again.

EXPERIMENTAL RESULTS

We use Pentium IV 2.4 GHZ PC as our experimental platform, the algorithm is programmed on Visual C++ and OpenCV, which runs at about 8 frames per second with 320×240 size of test video

data. The test data contain different scenes including illumination changes, stationary and dynamic backgrounds. We have used about 12000 frames of video images to test the validity of our algorithm. To compare with our proposed method, we incorporate shadow elimination into GMM (Stauffer and Grimson, 2000) and spatio-temporal MRFs (Migdal and Grimson, 2005). In our experiments, both spatio-temporal MRFs and our proposed method use 8-spatial neighboring relationship and 9-temporal neighboring relationship. The experimental results showed that our proposed method has more accuracy in tiny details than the other two algorithms. None results have been processed by morphology filtering.

Fig.2 shows a pedestrian walking around the street. The video data have stationary backgrounds. Because GMM segments objects with pixel level, without considering spatial information, pixels in the border will be easy to be classified as shadow when the colors of foreground are similar to those of the background. The spatio-temporal MRFs have an advantage over GMM in segmentation results because of incorporation of spatial and temporal constraints. Our proposed method can update parameters according to scenes, therefore, our method segments more accurately than others in tiny details, such as the head and the bag of the pedestrian. Fig.2a is the original image of frame #535, Figs.2b and 2c are segmentation results by using GMM and MRFs respectively. The segmentation results by our proposed method are in Fig.2d. Fig.2e is the results when shadows have been eliminated from Fig.2d.

The video clip is that of customers walking around the mall, in which the backgrounds have some bits of complexity more than previous ones. Fig.3a is the original image of frame #286, and Fig.3b is the results segmented by using GMM. There are some noise points in the results segmented by GMM. Both MRFs and HCRFs can eliminate the noise points because of using spatial and temporal constraints.

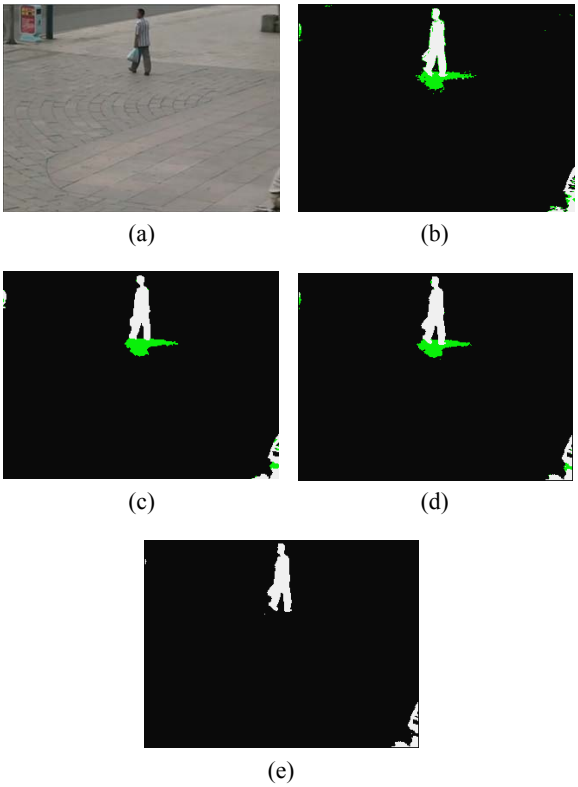

Fig.2 Segmentation results for "Pedestrian"
(a) Original image of frame #535; (b) Results segmented by GMM; (c) Segmentation results by using spatio-temporal MRFs; (d) Results by using our proposed method; (e) Results by eliminating the shadow

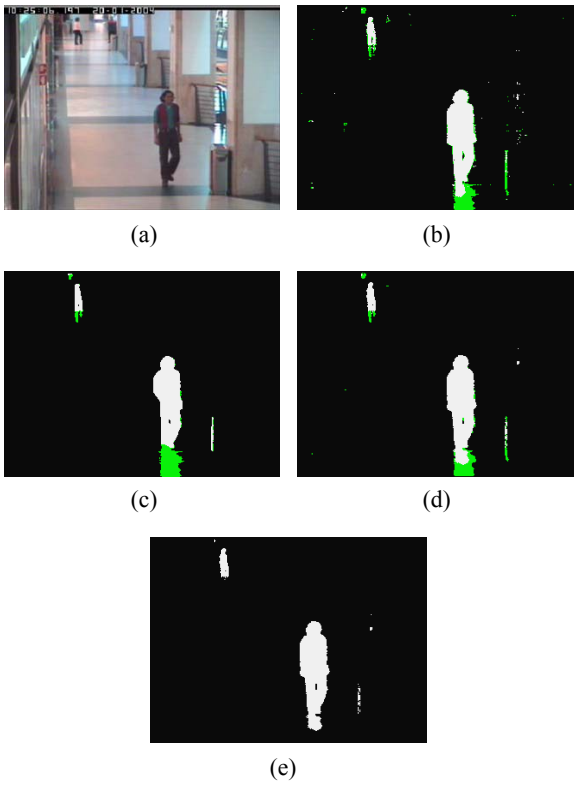

Fig.3 Segmentation results for "Mall"
(a) Original image of frame #286; (b) Results segmented by using GMM; (c) Segmentation results by using spatio-temporal MRFs; (d) Results by using our proposed method; (e) Results by eliminating the shadow

Figs.3c and 3d are the results by using MRFs and HCRFs respectively. Eliminating the shadows from Fig.3d, we will obtain Fig.3e.

Fig.4 is a video clip in library surveillance, in which strong illumination and branch of plants are contained. The video data have heavy camera noises that lead to many blobs in segmentation results. Fig.4a is the original image of frame #157, and Fig.4b is the results segmented by using GMM. The results of segmentation using spatio-temporal MRFs are in Fig.4c. Fig.4d shows the results segmented by using our proposed method, and the results of our proposed method with shadow elimination are in Fig.4e.

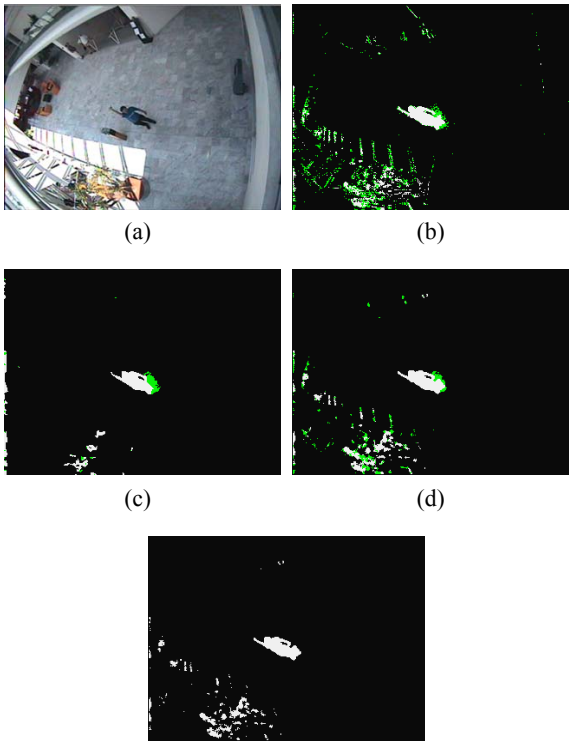

Fig.4 Segmentation results for "Library"
(a) Original image of frame #157; (b) Results segmented by GMM; (c) Segmentation results by using spatio-temporal MRFs; (d) Results by using our proposed method; (e) Resulsts by eliminating the shadow

To evaluate the error ratios of these algorithms, we segment the video sequence manually for obtaining the ground-truth images. Error ratios of GMM, MRFs and HCRFs are computed by comparing their segmentation results with ground-truth images, respectively. Table 1 shows the error ratios of the three methods. The experimental results demonstrated that the error ratio of our algorithm is less than that of GMM and spatio-temporal MRFs by 23% and 19%, respectively.

Table 1 Error ratios of GMM, MRFs and our proposed method

	GMM	MRFs	HCRFs
Error ratio	1.11	1.05	0.86

CONCLUSION

A novel adaptive video segmentation algorithm based on HCRFs is proposed, where spatial and temporal neighboring constraints are modeled via HCRFs. To adjust weights of spatio-temporal neighboring relationships according to scene changes, online learning method is adopted to update parameters of HCRFs. The results of experiments showed that the error ratio of our algorithm is less than that of GMM and spatio-temporal MRFs.

Shadow elimination is incorporated into our method. Linear transform applied to background models is used to construct the shadow models. Although sometimes, these shadow models cannot deal well with the shadows. In future work, we want to improve the approach.

References

Gunawardana, A., Mahajan, M., Acero, A., Platt, J.C., 2005. Hidden Conditional Random Fields for Phone Classification. Proc. 9th International Conference on Speech Communication and Technology. Lisbon, Portugal, p.1117-1120.

Kumar, S., Hebert, M., 2003. Discriminative Random Fields: A Discriminative Framework for Contextual Interaction in Classification. ICCV'03. Nice, France, **2**:1150-1157. [doi:10.1109/ICCV.2003.1238478]

Kumar, S., Hebert, M., 2005. A Hierarchical Field Framework for Unified Context-based Classification. ICCV'05. Beijing, China, **2**:1284-1291. [doi:10.1109/ICCV.2005.9]

Lafferty, J., McCallum, A., Pereira, F., 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Proc. Int'l Conf. Machine Learning, p.282-289.

Martel-Brisson, N., Zaccarin, A., 2005. Moving Cast Shadow Detection from a Gaussian Mixture Shadow Model. Proc. CVPR'05. IEEE Computer Society, Washington DC, **2**:643-648. [doi:10.1109/CVPR.2005.233]

Migdal, J., Grimson, E., 2005. Background Subtraction Using

Markov Thresholds. IEEE Workshop on Motion and Video Computing. Washington DC, USA, p.58-65. [doi:10.1109/ACVMOT.2005.33]

Porikli, F., Thornton, J., 2005. Shadow Flow: A Recursive Method to Learn Moving Cast Shadows. ICCV'05, **1**:891-898. [doi:10.1109/ICCV.2005.217]

Quattoni, A., Collins, M., Darrell, T., 2004. Conditional Random Fields for Object Recognition. Advances in Neural Information Processing Systems. Canada.

Sha, F., Pereira, F., 2003. Shallow Parsing with Conditional Random Fields. Proc. Human Language Technology-NAACL. Edmonton, Canada, p.213-220.

Sheikh, Y., Shah, M., 2005. Bayesian modeling of dynamic scenes for object detection. *IEEE Trans. Pattern Anal. Machine Intell.*, **27**(11):1778-1792. [doi:10.1109/TPAMI.2005.213]

Stauffer, C., Grimson, W., 2000. Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Machine Intell.*, **22**(8):747-757. [doi:10.1109/34.868677]

Stenger, B., Ramesh, V., Paragios, N., Coetzee, F., Buhmann, J.M., 2001. Topology Free Hidden Markov Models: Application to Background Modeling. Proc. Int'l Conf. Computer Vision, **1**:294-301. [doi:10.1109/ICCV.2001.10008]

Wang, Y., Ji, Q., 2005. A Dynamic Conditional Random Field Model for Object Segmentation in Image Sequences. CVPR'05. San Diego, CA, p.264-270. [doi:10.1109/CVPR.2005.26]

Wang, Y., Loe, K.F., Wu, J.K., 2006. A dynamic conditional random field model for foreground and shadow segmentation. *IEEE Trans. Pattern Anal. Machine Intell.*, **28**(2):279-289. [doi:10.1109/TPAMI.2006.25]

Wang, S., Quattoni, A., Morency, L., Demirdjian, D., Darrell, T., 2006. Hidden Conditional Random Fields for Gesture Recognition. CVPR'06. New York, **2**:1521-1527. [doi:10.1109/CVPR.2006.132]

Yang, T., Li, S.Z., Pan, Q., Li, J., 2004. Real-Time and Accurate Segmentation of Moving Objects in Dynamic Scene. ACM_VSSN'04. New York, p.10-16. [doi:10.1145/1026799.1026822]

Zhou, Y., Xu, W., Tao, H., Gong, Y.H., 2005. Background Segmentation Using Spatial-Temporal Multi-Resolution MRF. IEEE Workshop on Motion and Video Computing, **2**:8-13. [doi:10.1109/ACVMOT.2005.32]

Zivkovic, Z., 2004. Improved Adaptive Gaussian Mixture Model for Background Subtraction. ICPR'04. Cambridge, United Kingdom, **2**:28-31. [doi:10.1109/ICPR.2004.1333992]