VISUAL LOCALIZATION AND SEGMENTATION BASED ON FOREGROUND/BACKGROUND MODELING

Hanzi Wang, Tat-Jun Chin and David Suter
School of Computer Science, the University of Adelaide, Adelaide, SA, 5005, Australia.

ABSTRACT

In this paper, we propose a novel method to localize (or track) a foreground object and segment the foreground object from the surrounding background with occlusions for a moving camera. We measure the likelihood of a target position by using a combination of a generative model and a discriminative model, considering not only the foreground similarity to the target model but also the dissimilarity between the foreground and the background appearances. Object segmentation is treated as a binary labeling problem. A Markov Random Field (MRF) is employed to add a spatial smooth prior on the foreground/background patterns. We demonstrate the advantages of the proposed method on several challenging videos and compare our results with the results of several other popular methods. The proposed method has achieved good results.

Index Terms— Visual tracking, video segmentation, particle filters, appearance modeling, occlusions

1. INTRODUCTION

Visual localization (or tracking) has many potential applications. Most of the visual tracking methods directly model foreground object appearance and track foreground objects without modeling the background scene. For example, [6] used the mean shift algorithm to iteratively search for a region which maximizes the similarity measure between this region and the target object model. In [8, 14, 15], particle filters were employed to simultaneously track multiple hypotheses and recursively approximate the posterior Probability Density Function (PDF) in the state space, with a set of randomly sampled particles. These methods work well for typical cases. However, they often fail when (1) the foreground appearance dramatically changes, (2) serious occlusion occurs, or (3) the background scene includes significant color distractors.

The surrounding background context is a useful cue to track foreground objects and segment them from the surrounding background scene. In recent years, a few methods have been proposed to utilize both foreground and background information to track/localize objects with a moving camera [1, 5, 10-13]. Nguyen et al. [13] localized an object by maximizing a texture-based discriminant function between a foreground object and the surrounding background. Collins et al. [5] proposed a feature selection approach which selects a feature that best discriminates a foreground object from the background in the vicinity. Lin et al. [10] proposed a Fisher Linear Discriminant based framework to find a discriminative generative model that best separates the target object from the background. Avidan [1] trained an ensemble of weak classifiers which distinguish an object from the background context. Lu et al. [11, 12] employed image segmentation approaches as a pre-processing step and classified the segments (or 'superpixels') into either foreground or background using the maintained foreground/background models consisting of a set of randomly selected patches. All of [1, 5, 12] localized an object in a confidence map using the mean shift algorithm. However, mean shift converges to a local maximum. Thus, these methods are sensitive to background color distractors, clutter, occlusions, scaling, and quick moving objects.

In this paper, we propose a new method which can accurately track and segment an object in a video sequence, even under serious occlusions. Our work exploits information from both the foreground and the background. We model both foreground and background regions with the spatial-color mixture of Gaussians (SMOG), yielding two SMOG models. The two SMOG models are used to generate a foreground Probability Response Map (PRM) and a background PRM for each testing region: consisting of one foreground region and one background region. The two PRMs are combined to form a Confidence Map (CM). We treat object segmentation as a binary labeling problem. The PRMs are fed to a Markov Random Field (MRF) to generate the foreground/background segmentation. The segmentation results are in turn utilized to update/learn the foreground and background SMOG models. We do not require a user to guide the system for segmentation.

2. LOCALIZATION WITH FOREGROUND/BACKGROUND MODELING

In this section, we develop our method for visual localization. We use abbreviations FG/BG as foreground/background.

2.1. FG/BG Appearance Modeling with SMOG

We employ SMOG [15] to model both FG and BG

appearances. Let $\mathbf{C}=(r,g,\mathrm{I})$ be the color feature of a pixel: $r=R/(R+G+B)$; $g=G/(R+G+B)$; $\mathrm{I}=(R+G+B)/3$ and $\mathbf{S}=(x,y)$ be the spatial feature of a pixel (i.e., the 2D image coordinate of that pixel). Each pixel is represented by a 5D feature vector $\mathbf{x}=(\mathbf{S}_{\mathbf{x}},\mathbf{C}_{\mathbf{x}})=(x,y,r,g,\mathrm{I})$. The target object Λ^F is represented by a Gaussian mixture in the joint spatial-color space. Given the parameters of Gaussian mixtures $\Theta^F=\{\theta_i^F\}_{i=1,...,n_F}=\{\omega_i^F,\mu_i^{\mathbf{S},F},\mu_i^{\mathbf{C},F},\Sigma_i^{\mathbf{S},F},\Sigma_i^{\mathbf{C},F}\}_{i=1,...,n_F}$, the SMOG model of a FG object can be written as follows:

$$\Lambda^F=\sum_{i=1}^{n_F}\omega_i^F\Lambda_i^F=\sum_{i=1}^{n_F}\omega_i^F N(\mu_i^{\mathrm{S},F},\mu_i^{\mathrm{C},F},\Sigma_i^{\mathrm{S},F},\Sigma_i^{\mathrm{C},F}) \quad (1)$$

where ω_i^F is the mixture proportions ($0\le\omega_i^F\le 1$, $\sum_{i=1}^{n_F}\omega_i^F=1$); $\mu_i^{\mathbf{S},F}/\mu_i^{\mathbf{C},F}$ and $\Sigma_i^{\mathbf{S},F}/\Sigma_i^{\mathbf{C},F}$ are respectively the mean and the covariance of the *i*th Gaussian component of the object model in the spatial/color feature space.

Let Λ^B be the appearance of the background context. The surrounding background Λ^B is modeled by SMOG with the parameters of Gaussian mixtures $\Theta^B=\{\theta_j^B\}_{j=1,...,n_B}=\{\omega_j^B,\mu_j^{\mathbf{S},B},\mu_j^{\mathbf{C},B},\Sigma_j^{\mathbf{S},B},\Sigma_j^{\mathbf{C},B}\}_{j=1,...,n_B}$:

$$\Lambda^B=\sum_{j=1}^{n_B}\omega_j^B\Lambda_j^B=\sum_{j=1}^{n_B}\omega_j^B N(\mu_j^{\mathrm{S},B},\mu_j^{\mathrm{C},B},\Sigma_j^{\mathrm{S},B},\Sigma_j^{\mathrm{C},B}) \quad (2)$$

The PDF of a data feature $\mathbf{x}$ belonging to Λ^F or Λ^B is:

$$p(\mathbf{x}\mid\Theta^F)=\sum_i\omega_i^F N(\mathbf{S}\mid\mu_i^{\mathbf{S},F},\Sigma_i^{\mathbf{S},F})N(\mathbf{C}\mid\mu_i^{\mathbf{C},F},\Sigma_i^{\mathbf{C},F}) \quad (3)$$

$$p(\mathbf{x}\mid\Theta^B)=\sum_i\omega_i^B N(\mathbf{S}\mid\mu_i^{\mathbf{S},B},\Sigma_i^{\mathbf{S},B})N(\mathbf{C}\mid\mu_i^{\mathbf{C},B},\Sigma_i^{\mathbf{C},B}) \quad (4)$$

where $N(\mathrm{z}\mid\mu_i,\Sigma_i)=(2\pi)^{-\frac{d}{2}}|\Sigma_i|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\mathcal{D}^2(z,\mu_i,\Sigma_i)\right]$; $\mathcal{D}^2(z,\mu_i,\Sigma_i)\equiv(\mathrm{z}-\mu_i)^T\Sigma_i^{-1}(\mathrm{z}-\mu_i)$ is the Mahalanobis distance for a *d*-dimension feature space.

Once we have $p(\mathbf{x}\mid\Theta_t^F)$ and $p(\mathbf{x}\mid\Theta_t^B)$ for all pixels in a testing region, we can generate a confidence map (CM) which can be used to coarsely label pixels in the testing region. Let $p_t^F(\mathbf{x})\equiv p(\mathbf{x}\mid\Theta_t^F)$ and $p_t^B(\mathbf{x})\equiv p(\mathbf{x}\mid\Theta_t^B)$. The confidence map $\mathbb{C}_t$ is defined as:

$$\mathbb{C}_t(\mathbf{x})=\frac{p_t^F(\mathbf{x})-p_t^B(\mathbf{x})}{p_t^F(\mathbf{x})+p_t^B(\mathbf{x})} \quad (5)$$

2.2. Likelihood Measurement with the FG/BG Models

Given a testing window Ω $(=\Omega^F\bigcup\Omega^B)$ which consists of a target candidate region Ω^F (i.e., the inner rectangle) and the surrounding BG region Ω^B (i.e., the area between the inner and outer rectangle), we score Ω by considering three criteria: **(1)** the similarity between the target candidate Ω^F and the FG model Λ^F ; **(2)** the discriminability between the target candidate region Ω^F and the BG model Λ^B ; **(3)** the discriminability between the surrounding BG region Ω^B and the FG model Λ^F .

Criterion 1: Denote $\Lambda^{\dagger F}\left(=\{\omega_i^{\dagger F},\mu_i^{\dagger\mathrm{S},F},\mu_i^{\dagger\mathrm{C},F},\Sigma_i^{\dagger\mathrm{S},F},\Sigma_i^{\dagger\mathrm{C},F}\}_{i=1,...,n_F}\right)$ as the target candidate model. The parameters of $\Lambda^{\dagger F}$ can be derived from the region Ω^F and the FG model Λ^F by an approach similar to [15]. The similarity measure between $\Lambda^{\dagger F}$ and Λ^F is written as:

$$\Phi(\Lambda^{\dagger F},\Lambda^F)= \quad (6)$$
$$\sum_{i=1}^{n_F}\min(\omega_i^{\dagger F},\omega_i^F)\exp\left\{-\frac{1}{2}(\mu_i^{\dagger\mathrm{S},F}-\mu_i^{\mathrm{S},F})^T\left(\hat{\Sigma}_i^{*\mathrm{S},F}\right)^{-1}(\mu_i^{\dagger\mathrm{S},F}-\mu_i^{\mathrm{S},F})\right\}$$

where $\left(\hat{\Sigma}_i^{*\mathrm{S},F}\right)^{-1}=\left(\Sigma_i^{\dagger\mathrm{S},F}\right)^{-1}+\left(\Sigma_i^{\mathrm{S},F}\right)^{-1}$.

Criterion 2: Let $\bar{\Lambda}^{\dagger B}\left(=\sum_{j=1}^{n_B}\bar{\omega}_j^{\dagger B}N(\bar{\mu}_j^{\dagger\mathrm{S},B},\bar{\mu}_j^{\dagger\mathrm{C},B},\bar{\Sigma}_j^{\dagger\mathrm{S},B},\bar{\Sigma}_j^{\dagger\mathrm{C},B})\right)$ be the model derived from the FG candidate region Ω^F and the BG model Λ^B . We define the discriminant function between $\bar{\Lambda}^{\dagger B}$ and Λ^B as:

$$DS(\bar{\Lambda}^{\dagger B}\,\|\,\Lambda^B)= \quad (7)$$
$$1-\sum_{j=1}^{n_B}\min(\bar{\omega}_j^{\dagger B},\omega_j^B)\exp\left\{-\frac{1}{2}(\bar{\mu}_j^{\dagger\mathrm{C},B}-\mu_j^{\mathrm{C},B})^T\left(\hat{\Sigma}_j^{*\mathrm{C},B}\right)^{-1}(\bar{\mu}_j^{\dagger\mathrm{C},B}-\mu_j^{\mathrm{C},B})\right\}$$

where $\left(\hat{\Sigma}_j^{*\mathrm{C},B}\right)^{-1}=\left(\Sigma_j^{\dagger\mathrm{C},B}\right)^{-1}+\left(\Sigma_j^{\mathrm{C},B}\right)^{-1}$.

Criterion 3: Let $\bar{\Lambda}^{\dagger F}\left(=\sum_{i=1}^{n_F}\bar{\omega}_i^{\dagger F}N(\bar{\mu}_i^{\dagger\mathrm{S},F},\bar{\mu}_i^{\dagger\mathrm{C},F},\bar{\Sigma}_i^{\dagger\mathrm{S},F},\bar{\Sigma}_i^{\dagger\mathrm{C},F})\right)$ be the model derived from the surrounding BK region Ω^B and the FG model Λ^F . The discriminant function between $\bar{\Lambda}^{\dagger F}$ and Λ^F is written as:

$$DS(\bar{\Lambda}^{\dagger F}\,\|\,\Lambda^F)= \quad (8)$$
$$1-\sum_{i=1}^{n_F}\min(\bar{\omega}_i^{\dagger F},\omega_i^F)\exp\left\{-\frac{1}{2}(\bar{\mu}_i^{\dagger\mathrm{C},F}-\mu_i^{\mathrm{C},F})^T\left(\hat{\Sigma}_i^{*\mathrm{C},F}\right)^{-1}(\bar{\mu}_i^{\dagger\mathrm{C},F}-\mu_i^{\mathrm{C},F})\right\}$$

where $\left(\hat{\Sigma}_i^{*\mathrm{C},F}\right)^{-1}=\left(\Sigma_i^{\dagger\mathrm{C},F}\right)^{-1}+\left(\Sigma_i^{\mathrm{C},F}\right)^{-1}$.

The discriminant measure between $\bar{\Lambda}^{\dagger B}$ and $\bar{\Lambda}^{\dagger F}$ is as:

$$DS(\bar{\Lambda}^{\dagger B},\ \bar{\Lambda}^{\dagger F})=\frac{1}{2}\left(DS(\bar{\Lambda}^{\dagger B}\,\|\,\Lambda^B)+DS(\bar{\Lambda}^{\dagger F}\,\|\,\Lambda^F)\right) \quad (9)$$

The likelihood function in our method is defined as:

$$\mathrm{L}(\mathrm{Y}_t\mid\mathrm{X}_t^j)\propto\exp\left\{\frac{1}{2\sigma_O^2}\left(\alpha\Phi_j^*(\Lambda_t^{\dagger F},\Lambda_t^F)+(1-\alpha)DS_j^*(\bar{\Lambda}_t^{\dagger B},\bar{\Lambda}_t^{\dagger F})\right)\right\} \quad (10)$$

where σ_O is the observation variance. α is a coefficient. $\Phi_j^*(\Lambda_t^{\dagger F},\Lambda_t^F)$ and $DS_j^*(\bar{\Lambda}_t^{\dagger B},\bar{\Lambda}_t^{\dagger F})$ are normalized and calculated as follows:

$$\Phi_j^*(\Lambda_t^{\dagger F},\Lambda_t^F)=\frac{\Phi_j(\Lambda_t^{\dagger F},\Lambda_t^F)-\min_i\left\{\Phi_i(\Lambda_t^{\dagger F},\Lambda_t^F)\right\}}{\max_{j'}\left\{\Phi_{j'}(\Lambda_t^{\dagger F},\Lambda_t^F)-\min_i\left\{\Phi_i(\Lambda_t^{\dagger F},\Lambda_t^F)\right\}\right\}} \quad (11)$$

$$DS_j^*(\bar{\Lambda}_t^{\dagger B},\bar{\Lambda}_t^{\dagger F})=\frac{DS_j(\bar{\Lambda}_t^{\dagger B},\bar{\Lambda}_t^{\dagger F})-\min_i\left\{DS_i(\bar{\Lambda}_t^{\dagger B},\bar{\Lambda}_t^{\dagger F})\right\}}{\max_{j'}\left\{DS_{j'}(\bar{\Lambda}_t^{\dagger B},\bar{\Lambda}_t^{\dagger F})-\min_i\left\{DS_i(\bar{\Lambda}_t^{\dagger B},\bar{\Lambda}_t^{\dagger F})\right\}\right\}} \quad (12)$$

2.3. Video segmentation

Video object segmentation is an important and challenging task, and it has been widely investigated in recent years (e.g., [4, 9, 16]). Visual object localization and segmentation are closely related to each other. On one hand, when one accurately segments an object from the background, it is easier to localize the object and update/learn the object appearance model without mistakenly using background pixels; on the other hand, when one correctly localizes an object, there is less chance for color distractors in the background causing problems in object segmentation. In this section, we propose a method to combine both object localization and object segmentation. We treat the segmentation problem as a binary labeling issue. We employ a Markov Random Field [7] with two-valued clique potentials to add spatial smooth priors on the foreground and background regions. We model an image as a two-terminal grid graph $G = \langle \mathbf{X}, \mathbf{E} \rangle$, consisting of a set of nodes $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$, a set of undirected edges $\mathbf{E}$, and two terminals which partition the nodes $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ in the graph G into two disjoint subsets. In our case, the nodes correspond to image pixels and the edges are the pairs of neighboring pixels $\mathcal{N} = \{\mathcal{N}_1, ..., \mathcal{N}_n\}$, where $\mathcal{N}_i$ is the neighboring pixels of the node $\mathbf{x}_i$. Each node contains a random variable $\mathbf{L}_i$ which belongs to one of a set of possible labels $\{l_1, ..., l_k\}$. We label each pixel in the testing region as either foreground (**F**) or background (**B**), i.e., $\mathbf{L}_i \in \{\mathbf{F}, \mathbf{B}\}$. MRF requires that the random variable at $\mathbf{x}_i$ only depends on the random variables of the neighboring nodes $\mathcal{N}_i$. According to the Bayes' law, the most likely configuration of the field should minimize the posterior energy function:

$$E(\mathbf{L}) = \sum_{\mathbf{x}_i \in \mathbf{X}} \mathrm{D}_i(\mathbf{L}_i) + \sum_{\mathbf{x}_i} \sum_{\mathbf{x}_j \in \mathcal{N}_i} \mathrm{V}_{i,j}(\mathbf{L}_i, \mathbf{L}_j) \qquad (13)$$

where $\mathrm{D}_i(\mathbf{L}_i)$ is a data penalty function and $\mathrm{V}_{i,j}$ is an interaction potential.

With the SMOG foreground/background models (Λ_t^F and Λ_t^B) and the estimated target location $\hat{\mathrm{X}}_t$, we define the data penalty function as follows:

$$\mathrm{D}_i(\mathbf{L}_i \in \mathbf{F}) = -\log p(\mathrm{x} \mid \Theta^F) \text{ and } \mathrm{D}_i(\mathbf{L}_i \in \mathbf{B}) = -\log p(\mathrm{x} \mid \Theta^B) \quad (14)$$

For the interaction potential, we employ a simple but efficient generalized Potts model [3]. To get the global minimum of the energy function in Eq. (13), we employ the min-cut algorithm [2]. Once we generate the foreground/background segmentation, we can utilize the results to update the foreground and background models.

3. EXPERIMENTS

We test our method on several challenging video sequences with a moving camera. We employ the Particle Filtering (PF) framework to estimate the target location status. We use a first-order AR dynamic model and 30 particles in the PF module. We also compare with the methods in [5] and [12], denoted as M1 and M2 respectively, for visual localization.

We first evaluate our method on a challenging video sequence. This video sequence includes 899 frames and the lady's face was significantly occluded by a book several times.

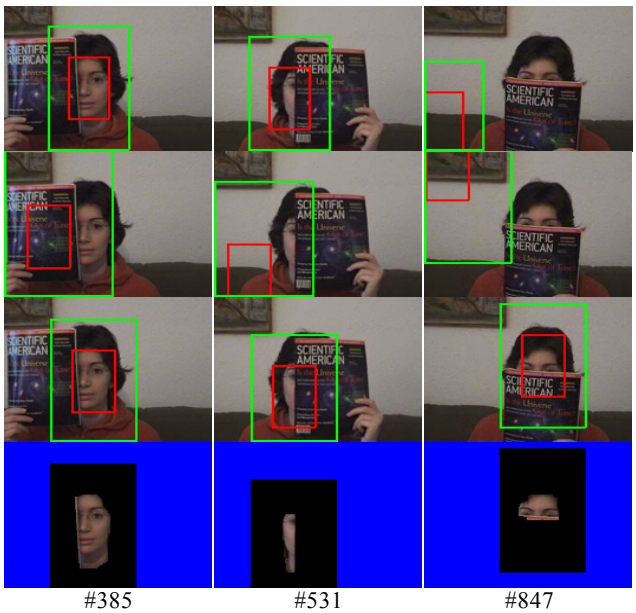

Figure 1: An example for tracking and segmenting a face under significant occlusions. First and second rows are the tracking results by M1 and M2; Third and fourth rows are respectively the tracking and segmentation results by the proposed method.

As shown in Figure 1, both M1 and M2 fail to track the face for the whole sequence. While our method successfully localizes the face and segment the face from the surrounding and occluding background. Even when serious occlusion happens, for example, at the frame 531, the proposed method can still accurately segment the face from the background and the occluding book. The success of our method in segmentation shows that the method has effectively learned/updated the FG/BG models and adapted the BG model to the occluding background, which leads to the correct localization of the face in the following frames.

In the second experiment, we evaluate the proposed method by using another challenging 501-frame long video sequence. The girl's head experienced rotation, scaling, illumination changes, and occlusions by a man's head around the end of the sequence. The background contains clutters and significant color distractors in both the surrounding background and the occluding background (i.e., the man's head). Figure 2 shows the results, from which we can see that only the proposed method successfully localizes (and segments) the head throughout the video sequence. Note the proposed method works well even when the significant color distractor (i.e., the man's head) occludes the girl's head (see the frame 443 in Figure 2).

In the third experiment, we show the generality of the proposed method for video object segmentation. We enlarge

the testing region to contain the whole image. The video was captured with a rapidly panning camera. Thus the background changes dramatically. Figure 3 shows some tracking/segmentation results obtained by the proposed method, from which we can see that the proposed method can effectively learn the BG model, adapting to the rapidly changing background, and accurately localize and segment the human from the background.

Figure 2: An example showing the robustness and effectiveness of the proposed method for tracking and segmenting a human head under rotation, scaling, illumination changes, color distractors, clutters, and occlusions. First and second rows are the tracking results by M1 and M2; Third and fourth rows are the tracking and the segmentation results by the proposed method.

Figure 3: Human localization and segmentation from a 111-frame long video [4]. The background is rapidly panning from the left to the right. We show three pairs of example at #16, #65 and #108. For each pair, the human localization result is shown on the top and the human segmentation on the bottom.

4. CONCLUSION

We have proposed an effective method for tracking and segmenting video objects. We utilize the information of foreground appearance and of the surrounding background scene, and model both the foreground and the background appearances with a spatial-color mixture of Gaussians. We further propose a new objective function which considers the similarity between the foreground object and the foreground model, and the discriminability between the foreground object and the surrounding background. We consider the problem of segmentation as a binary labeling issue. A MRF is employed to add the spatial smoothness prior on the foreground and background patterns. Because the proposed method can segment foreground objects from the background scene even with serious occlusions, it can effectively learn/update the foreground/background models. We have tested our method on challenging video sequences and compared the results with those of two other popular methods ([5] and [12]): showing that the proposed method is more robust and achieve better results in object localization; and the proposed method has also achieved promising results on video segmentation.

5. ACKNOWLEDGEMENT

The authors would like to thank Dr. Le Lu and Dr. Konrad Schindler for their valuable discussions and helps. This work was partially supported by the Australian Research Council (ARC) under the project DP0878801.

6. REFERENCES

1. Avidan, S., *Ensemble Tracking.* TPAMI, 2007. **29**(2): p. 261-271.
2. Boykov;, Y. and V. Kolmogorov, *An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision.* TPAMI, 2004. **26**(9): p. 1124-1137.
3. Boykov;, Y., O. Veksler; and R. Zabih. *Markov Random Fields with Efficient Approximations*. in *CVPR*. 1998. 648-655.
4. Chuang, Y.-Y., et al., *Video Matting of Complex Scenes*. *SIGGRAPH*. 2002. 243-248.
5. Collins, R.T., Y.X. Liu and M. Leordeanu, *On-Line Selection of Discriminative Tracking Features.* TPAMI, 2005. **27**(10): p. 1631-1643.
6. Comaniciu, D., V. Ramesh and P. Meer, *Kernel-based Object Tracking.* TPAMI, 2003. **25**(5): p. 564 - 577.
7. Geman, S. and D. Geman, *Stochastic Relaxation, Gibbs Distribution and the Bayesian Restoration of Images.* TPAMI, 1984. **6**: p. 721-741.
8. Isard, M. and A. Blake, *Condensation-Conditional Density Propagation for Visual Tracking.* IJCV, 1998. **29**(1): p. 5-28.
9. Li, Y., J. Sun and H.-Y. Shum. *Video Object Cut and Paste*. in *SIGGRAPH*. 2005. 595-600.
10. Lin, R.-S., D. Ross, J. Lim and M.-H. Yang. *Adaptive Discriminative Generative Model and Its Applications*. in *NIPS*. 2004.
11. Lu, L. and G.D. Hager. *Dynamic Background/Foreground Segmentation From Images and Videos using Random Patches*. *NIPS*. 2006.
12. Lu, L. and G.D. Hager. *A Nonparametric Treatment on Location/Segmentation Based Visual Tracking*. in *CVPR*. 2007.
13. Nguyen, H.T. and A.W.M. Smeulders, *Robust Tracking Using Foreground-Background Texture Discrimination.* IJCV, 2006. **69**(3): p. 277-293.
14. Perez, P., C. Hue, J. Vermaak and M. Gangnet. *Color-Based Probabilistic Tracking*. in *ECCV*. 2002. 661-675.
15. Wang, H., et al., *Adaptive Object Tracking Based on an Effective Appearance Filter.* TPAMI, 2007. **29**(9): p. 1661-1667.
16. Wang, J., et al., *Interactive Video Cutout*. *SIGGRAPH*. 2005. p. 574-583.