Foreground object segmentation for moving camera sequences based on foreground-background probabilistic models and prior probability maps

Jaime Gallego, Pascal Bertolino

FOREGROUND OBJECT SEGMENTATION FOR MOVING CAMERA SEQUENCES BASED ON FOREGROUND-BACKGROUND PROBABILISTIC MODELS AND PRIOR PROBABILITY MAPS

Jaime Gallego, Pascal Bertolino

GIPSA-lab

{Jaime.Gallego-Vila,pascal.bertolino}@gipsa-lab.grenoble-inp.fr

ABSTRACT

This paper deals with foreground object segmentation in the context of moving camera sequences. The method that we propose computes a foreground object segmentation in a MAP-MRF framework between foreground and background classes. We use region-based models to model the foreground object and the background region that surrounds the object. Moreover, the global background of the sequence is also included in the classification process by using pixel-wise color GMM. We compute the foreground segregation for each one of the frames by using a Bayesian classification and a graph-cut regularization between the classes, where the prior probability maps for both, foreground and background, are included in the formulation, thus using the cumulative knowledge of the object from the segmentation obtained in the previous frames. The results presented in the paper show how the false positive and false negative detections are reduced, meanwhile the robustness of the system is improved thanks to the use of the prior probability maps in the classification process.

Index Terms— Object segmentation, SCGMM, moving camera segmentation, spatial prior probability maps.

1. INTRODUCTION

Foreground segmentation in video sequences is a major challenge in the image processing area that attracts great interest among the scientist community, since it makes possible the detection of the objects that appear in the sequences under analysis, and allows us to achieve a correct performance of high level applications which use foreground segmentation as an initial step.

The main challenges to overcome when performing a foreground object segmentation are, among others, camouflage situation between the foreground object to segment and the background, changes in the color regions due to the presence of shadow or highlight effects or the existence of dynamic background in the scene to segment. When we have to perform a foreground object segmentation in a moving camera sequence, all these situations increase in a special manner the complexity scenario, since there is no possibility to perform an exact background learning at a pixel-wise level (typical of the static camera setups [15, 13]), to apply well known pixel-wise techniques for computing the background subtraction. Moreover, moving camera scenarios present an additional complexity due to the camera translation, rotation and zoom effects, which added to the objects' movements, can produce strong color and shape modifications of the image regions.

The authors would like to thank the French *Région Rhônes-Alpes* for its funding of this work in the context of the ReadPlay project.

1.1. Previous work

In the recent years, the researchers have followed different strategies to achieve a correct segmentation of the foreground object regions. Methods based on camera motion estimation compute camera motion and, after its compensation, they apply an algorithm defined for fixed camera. [7] proposes a multi-layer homography to rectify the frames and compute pixel-wise background subtraction.

Techniques based on the evolution of different features of the image along the frames are being of great interest in the community to achieve supervised and unsupervised video segmentations. [2] estimates a dense optical flow, [16, 8] uses the agglomerative clustering approach on supervoxels while [14] computes spatio-temporal graph-cuts. A few new approaches rely on multiple per-frame figure-ground segmentations: [9] utilizes motion saliency to detect the right segments to track, then run successive graph cuts on clips propagating from the most confident key segment. [11] proposes video segmentation by simultaneously tracking multiple holistic figure-ground segments, initialized from a pool of segment proposals generated from a figure-ground segmentation algorithm.

Finally, foreground segmentation proposals based on probabilistic models achieve correct results when specific objects of the sequence have to be tracked, isolated from other regions of the video. In these approaches, the objects to segment are characterized by using probabilistic models to classify the pixels belonging to the object. [10] proposes a non parametric method to approximate, in each frame, a *pdf* of the objects bitmap, while in [6] we used a Bayesian classification framework between the foreground object and the background that is surrounding it.

1.2. Proposed method

In this paper, we propose a foreground object segmentation system for moving camera sequences that deals with the segmentation methods based on probabilistic models. The system explained in the paper is based on the method that we proposed in [6]. We have added strong improvements that make the segmentation more robust in front of object modifications, while avoiding the drift of the probabilistic models along the sequence. As in [6], we use in this method a Bayesian Maximum a Posteriori - Markov Random Field (MAP-MRF) framework between the foreground (fg) and the background (bg) classes, where each one of the models rivals one another to model the pixels of each one of the frames but, in this approach, we differentiate between two types of background:
-Near Background: it is the background that is surrounding the object inside a Region of Interest (ROI). It is necessary to model the details of the close background to maintain the limits of the object.
-Global Background: it is based on the most relevant background

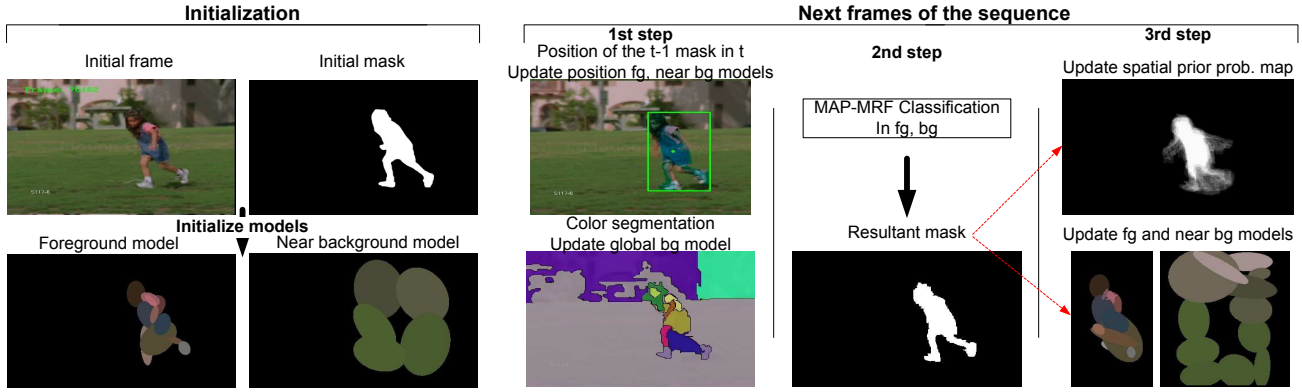

Fig. 1. Work flow of the proposed system. Probabilistic models are represented in the images, where each one of the ellipses represents one Gaussian in the spatial domain, and each one is colorized with the mean color that each one is modeling. Green rectangle in updating position shows the estimation of the object's position in frame t.

regions, according to color homogeneity, which are present in each frame. Characterize this background is important in order to avoid drifts of the foreground model when global background regions, not modeled by the near background model, appear close to the object.

Moreover, in order to increase the robustness of the overall system, we define the spatial prior probability maps for the foreground and background classes, computed with the last segmentation results obtained in the previous frames. In this way, we include the information of the anterior detections in each classification step, thus avoiding possible drifts of the model in the spatial domain.

Figure 1 shows the work flow of the system. As we can observe, an input mask of the object that we want to segment is needed to initialize the foreground and the near background models. Next, for each frame of the sequence, three main steps are applied:
The first step is devoted to update the global background model with the biggest color regions of the background, and to displace the foreground and near background models to the new spatial position of the object in order to improve the spatial modeling before the classification. To this end, we localize the position of the $t-1$ foreground segmentation in t.
In the second step, the classification among foreground and background classes is performed. The Bayesian probability between foreground, near background and global background is computed, for each one of the pixels, by taking into account the spatial prior probability maps. The resultant classification is regularized with the neighborhood information using MRF framework.
Finally, the resultant foreground object mask is used in the third step to update the foreground and near background models, as well as to update the prior probability maps.

The details of these processes will be explained in the following sections. The remainder of the paper is organized as follows. Section 2 describes the probabilistic models proposed for the foreground, near background and global background classes, and the classification process. Section 3 explains the spatial prior probability maps computed with the previous resultant masks. Finally, some results and conclusions are presented in Section 4 and Section 5 respectively.

2. PROBABILISTIC MODELS AND CLASSIFICATION

Since foreground and near background classes have to model specific color-spatial regions of the frames, which present spatial continuity along the sequence, analogously to [6] we use region-based spatial-color Gaussian mixture models (SCGMM) to probabilistically model the foreground and the near background regions. On contrast, since we want the global background to be present in each one of the pixels of the image, we use pixel-wise GMM to characterize the color regions that form this background class. The color domain used in the formulation is denoted as $c = (r, g, b)$, while the spatial domain is denoted as $s = (x, y)$. The combination of both color and space domains are defined as the joint domain-range representation $z = (r, g, b, x, y)$.

2.1. Foreground and Near Background models

Since in this kind of sequences the foreground and background are constantly moving and changing, an accurate model at a pixel level is difficult to build and update. For this reason, we use a region based Spatial Color Gaussian Mixture Model (SCGMM), as in [17, 6], because foreground objects and background regions are better characterized by color and position. Thus, the foreground and background pixels are represented in a five dimensional space, and the likelihood of pixel i is then,

$$\begin{aligned} P(z_i|l) &= \sum_{k=1}^{K_l} \omega_k G_l(z_i, \mu_k, \Sigma_k) \\ &= \sum_{k=1}^{K_l} \omega_k \frac{1}{(2\pi)^{5/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(z_i-\mu_k)^T \Sigma_k^{-1}(z_i-\mu_k)} \end{aligned} \quad (1)$$

where $z_i \in \mathbb{R}^5$ is the i-th pixel value ($i = 1, ..., N$), l stands for each class: $l \in \{\text{fg, near bg}\}$, ω_k is the mixture coefficient, $\mu_k \in \mathbb{R}^5$ and $\Sigma_k \in \mathbb{R}^{5\times5}$ are, respectively, the mean and covariance matrix of the k-th Gaussian distribution.

2.1.1. *Initialization*

The initialization and the updating processes of these models are done according to [6]. For the initialization, a first input mask of the object, manually defined, or obtained from a previous detection algorithm, is used to determine the number of Gaussians that will compound each one of the models. Hence, the histogram of the foreground and near background regions is analyzed to determine the color regions that appear inside the Region of Interest. Once the number of Gaussians are defined, the initialization of parameter estimation can be reached via Bayes' development, with the EM algorithm [4].

2.1.2. Updating

As we can observe in the work-flow of the system (Figure 1), there exist two updates for these region-based models:

Spatial updating before the classification. With the objective to improve the characterization of the regions before the classification step, the $t-1$ foreground and background region-based models are spatially displaced to the estimated position of the object in the frame t. Hence, we estimate the new position of the object by computing the mean square color distance between the pixels of the $t-1$ mask in the new frame t. We perform the analysis inside the region of interest, which allows us to make an exhaustive search without increasing significantly the computation time. Since the process is similar to a block matching algorithm, any optimization strategy can be utilized to speed up the process.

Spatial and color updating after the classification step. Analogously to [6], a complete updating process is applied by using the segmentation mask obtained in the classification. This foreground mask is used in order to adapt the foreground model to the new regions detected. The complementary mask is used to update the near background regions.

2.2. Global Background model

Moving camera sequences make not possible a precise background learning of the scene since it is changing along the frames, and occlusion situations are constantly present. These drawbacks produce that new background regions appear suddenly, close to the object to segment, before the near background could model them. These situations can lead to false positive detections that can be rapidly assumed to the foreground model, thus originating the drift of the models.

With the objective to reduce the false positive detections and the consequent drift of the models that can appear in these situations, we improve the probabilistic modeling of the background by creating the global background model. We propose to include the most relevant color regions of each frame in the background class, by creating one color Gaussian for each one of the Q detected regions, thus creating a GMM in the color $c = (r, g, b)$ domain. This GMM is applied in each one of the pixels of the image. Hence, the likelihood of the model for the pixel i is:

$$P(c_i|\text{global bg}) = \sum_{q=1}^{Q} \omega_q G_{\text{global bg}}(c_i, \mu_{c,q,i}, \Sigma_{c,q,i}), \qquad (2)$$

where $c_i \in \mathbb{R}^3$ is the pixel's color value, ω_q is the weight assigned to the region q. For simplicity, we impose the same weight to each Gaussian. $\mu_{c,q,i} \in \mathbb{R}^3$ is the color mean, of the region q and $\Sigma_{c,q,i} \in \mathbb{R}^{3\times3}$ is its covariance matrix. $G_{\text{global bg}}(c_i, \mu_{c,q,i}, \Sigma_{c,q,i})$ is the Gaussian likelihood in the $c = (r, g, b)$ domain.

Since we want this color model to be comparable with the five dimensional domain, we extend this model to a five dimensional domain by including the spatial component [5].

$$P(z_i|\text{global bg}) = \sum_{k=1}^{N} \frac{1}{N} \sum_{q=1}^{Q} \omega_q G_{\text{global bg}}(z_i, \mu_{z,k,q}, \Sigma_{z,k,q}), \quad (3)$$

where:

$$G_{\text{global bg}}(z_i, \mu_{z,k,q}, \Sigma_{z,k,q}) = \delta(s_i - \mu_{s,k,q}) P(c_i|\text{global bg}), \quad (4)$$

$\mu_{z,k,q} \in \mathbb{R}^5$ is the Gaussian mean for the k-th spatial Gaussian and the q-th color Gaussian, $\Sigma_{z,k} \in \mathbb{R}^{5\times5}$ is the covariance matrix, $s_i \in \mathbb{R}^2$ is the spatial pixel's coordinate and $\mu_{s,k,q} \in \mathbb{R}^2$ is the spatial mean of the Gaussian in the position k and modeling the color q.

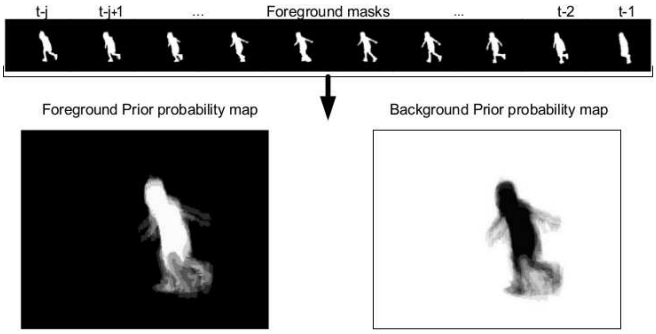

Fig. 2. Spatial prior probability maps for Foreground and Background classes. The brighter the higher the prior probability value. Background prior map is computed with the complementary resultant masks.

Hence, we use N color GMM, (one for each pixel of the image) each one centered (in space) at each pixel position ($\mu_{s,k}$) with a zero spatial variance.

The model is updated at each frame with the segmentation of the color regions. This color segmentation can be computed with classical segmentation methods like mean-shift [3] or pyramid segmentation [12] or by analyzing the histogram of the background regions. The number of Gaussians is defined according to the number of regions that present a bigger size than a certain threshold. In our experiments $1/3$ of the overall image area is enough to model the most relevant regions.

2.3. Bayesian Foreground/Background classification

The pixel classification is done at frame t, using a Maximum A Posteriori (MAP) decision. The priors are obtained according to the spatial prior probability maps that will be explained in next Section 3. Hence, a pixel i is assigned to the class $l' \in \{\text{fg, near bg, global bg}\}$ that maximizes $P(l'_i|z_i) \propto P(z_i|l'_i)P(l'_i)$.

Since we can assume that near background and global background pixels will be treated in the same way for the final segmentation mask, we combine them into the background ones according to the following criterion:

$$P(\text{bg}|z_i) = max(P(\text{near bg}|z_i), P(\text{global bg}|z_i)) \qquad (5)$$

Analogously to [5, 17], we consider a MRF framework in order to take into account neighborhood information that can be solved using standard graph-cut algorithm [1].

3. SPATIAL PRIOR PROBABILITY MAPS

In order to preserve the spatial shape of the object along the sequence, we propose to use the cumulative knowledge of the object obtained from the J previous masks. Since in a normal video sequence there is a high degree of overlapping between consecutive frames, and the objects to segment present a moderate degree of change, we can take into account the history of the object segmentation into the classification process. To this end, we use a LIFO queue with the last J segmentation masks, and normalize the spatial domain of each mask by using the centroid position in each frame obtained in Section 2.1.2, thus allowing a correct overlapping of the J masks. As we can observe in Figure 2, the spatial prior probability maps present a value between $(0, 1]$, for each one of the pixels, according to the occupancy that each one of the J masks presents.

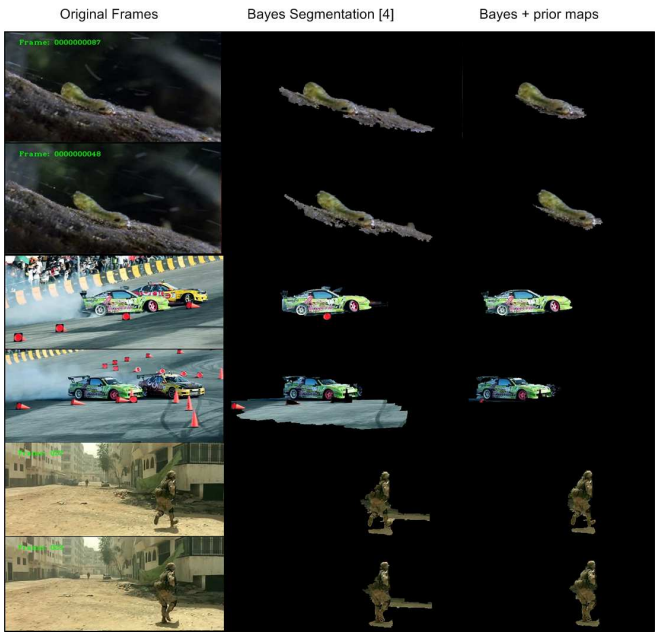

Fig. 3. Qualitative results. Frames belong to SegTrackv2 data base.

The spatial prior probability $P_i(l)$ is formulated as:

$$P_i(l) = \frac{1}{J}\sum_{j=1}^{J} M_{l,i}(t-j), \tag{6}$$

where $P_i(l)$ is the prior probability for pixel, $l \in \{\text{fg}, \text{bg}\}$, $M_{l,i}(t-j) \in \{1, 0\}$ is the i-th binary value for the mask obtained for class l, in frame $t-j$ and J is the total number of masks that we use for compounding the Prior map. In our experiments, $J = 10$ frames yields correct results for preserving the shape information of the object. Also, we avoid the zero prior value by adding a low bound of 0.1.

4. RESULTS

We have evaluated our proposal by analyzing the *SegTrackv2* database published in [11], and some well known sequences commonly used in the state of the art, which present strong difficulties to achieve a correct foreground object segmentation due to the presence of foreground-background similarity, slow motion effects, occlusions or changes in the object.

The qualitative evaluation is done comparing the segmentation results with our previous proposal [6] in some representative frames of the sequences: *worm, drift-1 and soldier*. These segmentation results are displayed in Figure 3. As we can observe in the third column, the foreground segmentation that we propose achieves a correct segmentation results by improving the stability of the probabilistic models and reducing the drift of the models. Hence, we avoid the false positive detections that appear by using [6] (Second column).

Quantitative results obtained by analyzing the database *SegTrackv2* ([11]) are displayed in table 1. We compare our results with the segmentation proposals presented in [6, 9, 11], which have proved to achieve good results when segmenting complex sequences. The metric used for the comparison is the *intersection over union*, computed as: $\frac{\text{TP}}{\text{TP+FN+FP}}$, where TP are the True positive detections, FN are the false negatives and FP are the false positive detections. As we can see, the method proposed in this paper (*Bayes p.maps* column) improves in general the previous approach presented in [6]. It achieves the best scores in sequences where rigid objects or objects with small modifications have to be segmented, which is the case of *drift* or *penguin* sequences. Also, our segmentation can present high scores, when there are not strong camouflage situations between the object and the background, as occurs in *monkeydog-1* or *bird of paradise* sequences. In contrast, when there are strong camouflage between foreground and background, the segmentation can present a big number of false negative detections, which occurs in *hummingbird-1* and *hummingbird-2*. The results of the segmentation methods proposed in [9, 11], show that the unsupervised segmentation can achieve correct segmentation results when there is not too much similarity between the object and background regions, obtaining a better refinement of the regions. The binary file of our software, as well as more qualitative and quantitative results will be accessible in our web pages [1] [2]. The computational cost of our system is 0.5 frames/second analyzing a standard sequence and using an Intel i5 2.3GHz processor and 6 GB RAM. Since the proposed system performs several time consuming computations at pixel-wise level, the parallelization of the code using GPU parallel processing will speed up the overall process.

Table 1. **Overall SegTrackv2 Data Base Comparison Results metric**. In bold type the results corresponding to the best *intersection over union* scores.

Sequence	Segmentation Technique			
	SPT+CSI [11]	Key seg. [9]	Bayes [6]	Bayes p.maps
Girl	**89.2**	87.7	87.82	87.86
Birdfall	**62.5**	49.0	29.13	59.60
Cheetah-1	37.3	**44.5**	23.31	30.12
Cheetah-2	**40.9**	11.7	16.47	20.51
Parachute	93.4	**96.3**	94.03	93.62
Monkeydog-1	71.3	74.3	75.60	**80.17**
Monkeydog-2	18.9	4.9	48.02	**48.29**
Penguin-1	51.5	12.6	83.18	**95.41**
Penguin-2	76.5	11.3	80.35	**89.35**
Penguin-3	75.2	11.3	79.43	**81.07**
Penguin-4	57.8	7.7	73.80	**80.62**
Penguin-5	66.7	4.2	72.75	**76.34**
Penguin-6	50.2	8.5	**82.20**	77.97
Driftcar-1	74.8	63.7	47.14	**85.41**
Driftcar-2	60.2	30.1	33.60	**72.39**
Hummingbird-1	**54.4**	46.3	38.60	26.22
Hummingbird-2	72.3	**74.0**	62.42	59.20
Frog	72.3	0	73.79	**74.54**
Worm	82.8	**84.4**	0.28	53.34
Soldier	**83.8**	66.6	62.36	76.71
Monkey	**84.8**	79.0	70.67	75.06
BirdParadise	94.0	92.2	92.52	**95.35**
BMX-1	85.4	87.4	**88.64**	86.60

5. CONCLUSIONS

We have proposed in this paper a novel foreground segmentation system for moving camera sequences, based on the use of the region-based spatial-color GMM to model the foreground object to segment and the near background regions that surrounds the object. Moreover, a pixel-wise GMM in the color domain is utilized to model the global background of the scene. The system is proposed in a MAP-MRF framework between the three classes, where the prior probability for each one of the classes is computed by taking into account the J previous segmentation masks, thus computing a spatial prior probability maps for the foreground and the background. The results presented in this paper show that the proposed system achieves a correct object segmentation reducing the false positives, and false negatives detections also in those complicated scenes where camera motion, object changes and occlusions are present.

[1] http://www.jaimegallego.com.es/icip2014_bayesian_prior

[2] http://www.gipsa-lab.grenoble-inp.fr/~pascal.bertolino/projects/readplay1/

6. REFERENCES

[1] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

[2] T. Brox and J. Malik. Large displacement optical flow: descriptor matching in variational motion estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 33(3):500–513, 2011.

[3] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002.

[4] A. Dempster, N. Laird, D. Rubin, et al. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[5] J. Gallego, M. Pardàs, and G. Haro. Enhanced foreground segmentation and tracking combining bayesian background, shadow and foreground modeling. *Pattern Recognition Letters*, 33(12):1558–1568, 2012.

[6] J. Gallego, M. Pardàs, and M. Solano. Foreground objects segmentation for moving camera scenarios based on scgmm. In *Lecture Notes on Computer Science: Computational Intelligence for Multimedia Understanding*, pages 195–206. Springer, 2012.

[7] Y. Jin, L. Tao, H. Di, N. Rao, and G. Xu. Background modeling from a free-moving camera by multi-layer homography algorithm. *Proc. IEEE Int. Conf. on Image Processing (ICIP)*, pages 1572–1575, 2008.

[8] J. Lee, S. Kwak, B. Han, and S. Choi. Online video segmentation by bayesian split-merge clustering. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 856–869. Springer, 2012.

[9] Y. J. Lee, J. Kim, and K. Grauman. Key-segments for video object segmentation. In *IEEE Int. Conf. on Computer Vision (ICCV)*, pages 1995–2002. IEEE, 2011.

[10] I. Leichter, M. Lindenbaum, and E. Rivlin. Bittrackera bitmap tracker for visual tracking under very general conditions. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 30(9):1572–1588, 2008.

[11] F. Li, K. T., A. Humayun, D. Tsai, and J. M. Rehg. Video segmentation by tracking many figure-ground segments. In *IEEE Int. Conf. on Computer Vision (ICCV)*, 2013.

[12] R. Marfil, L. Molina-Tanco, A. Bandera, J. A. Rodríguez, and F. Sandoval. Pyramid segmentation algorithms revisited. *Pattern Recognition*, 39(8):1430–1451, 2006.

[13] C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2:246–252, 1999.

[14] T. Wang and J. Collomosse. Probabilistic motion diffusion of labeling priors for coherent video segmentation. *IEEE Trans. on Multimedia*, 14(2):389–400, 2012.

[15] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 19(7):780–785, 1997.

[16] C. Xu, C. Xiong, and J. J. Corso. Streaming hierarchical video segmentation. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 626–639. Springer, 2012.

[17] T. Yu, C. Zhang, M. Cohen, Y. Rui, and Y. Wu. Monocular video foreground/background segmentation by tracking spatial-color Gaussian mixture models. In *IEEE Workshop on Motion and Video Computing*, pages 5–5, 2007.