Identifying Foreground from Multiple Images*

Wonwoo Lee[1], Woontack Woo[1], and Edmond Boyer[2]

[1] GIST U-VR Lab., 500-712, S. Korea
{wlee, wwoo}@gist.ac.kr
[2] LJK - INRIA Rhône-Alpes, Montbonnot, France
Edmond.Boyer@inrialpes.fr

Abstract. In this paper, we present a novel foreground extraction method that automatically identifies image regions corresponding to a common space region seen from multiple cameras. We assume that background regions present some color coherence in each image and we exploit the spatial consistency constraint that several image projections of the same space region must satisfy. Integrating both color and spatial consistency constraints allows to fully automatically segment foreground and background regions in multiple images. In contrast to standard background subtraction approaches, the proposed approach does not require any *a priori* knowledge on the background nor user interactions. We demonstrate the effectiveness of the method for multiple camera setups with experimental results on standard real data sets.

1 Introduction

Identifying foreground regions in single or multiple images is a necessary preliminary step of several computer vision applications in object tracking, motion capture or 3D modeling for instance. In particular, several 3D modeling applications optimize an initial model obtained using silhouettes extracted as foreground image regions. Traditionally, foreground regions are segmented under the assumption that the background is static and known beforehand in each image. This operation is usually performed on an individual basis, even when multiple images of the same scene are considered. In this paper, we take a different strategy and propose a method that simultaneously extract foreground regions in multiple images without any *a priori* knowledge on the background. The interest arises in many applications where multiple images are considered and where background information are not available, for instance when a single image only is available per viewpoint.

The approach we propose relies on a few assumptions that are often satisfied. First, the region of interest should appear entirely in several images. Second, in each image, the background colors should be consistent, i.e. the background is homogeneous to some extent, and differ from the foreground colors. Under these

* This project is funded in part by ETRI OCR and in part by the Korea Research Foundation Grant funded by the Korean Government(MOEHRD)(KRF-2006-612-D00081).

Y. Yagi et al. (Eds.): ACCV 2007, Part II, LNCS 4844, pp. 580–589, 2007.

assumptions, we iteratively segment in each image the 2 regions such that one, the background, satisfies color consistency constraints, and the second, the foreground, satisfies geometric consistency constraints with respect to other images. To initiate the iterative process, we use the first assumption above to identify regions in the images that are necessarily background. Such regions are simply image regions which are outside the projections of the observation volume common to all considered viewpoints. These initial regions are then grown iteratively by adding pixels that inconsistently belong to background and foreground regions in other images. We adopt an EM scheme for that, where background and foreground models are updated in one step, and images are segmented in another step using the new model parameters. Some important features of the approach are as follows. The method is fully automatic and does not require *a priori* knowledge of any type nor user interactions. In addition, a single camera at different locations or several cameras can be considered. In the latter, cameras do not need to be color calibrated since geometric and not color consistency is enforced between viewpoints.

The remainder of the paper is as follows. In section 2, we review existing segmentation methods. Sections 3 and 4 detail the implementation of the proposed method. Experimental results and conclusions are given in sections 5 and 6, respectively.

2 Related Works

Background subtraction methods usually assume that background pixel values are constant over time while foreground pixel values vary at some time. Based on this fact, several approaches have been proposed which take into account photometric information: greyscale, color texture or image gradient among others, in a monocular context. For non-uniform backgrounds, statistical models are computed for pixels. Several statistical models have been proposed to this purpose, for instance: normal distributions used in conjunction with the Mahalanobis distance [1], or mixture of Gaussian to account for multi-value pixels located on image edges or belonging to shadow regions [2,3]. In addition to these models, and to enforce smoothness constraints over image regions, graph cut methods have been widely used. After the seminal work of Boykov and Jolly [4], many derivatives have been proposed. GrabCut reduces the user interaction required for a good result by iterative optimization [5]. Li *et al.* proposed a coarse to fine approach in Lazy Snapping. It provides a user interface for boundary editing [6]. Freedman *et al.* [7] exploit the shape prior information to reduce segmentation error in the area where both the foreground and background have the similar intensities . The current graph cut based methods shows good results with both static images and videos, but user interaction are often required to achieve good results.

All the aforementioned approaches assume a monocular context and do not consider multi-camera cues when available. An early attempt in that direction was to add stereo information, i.e. depth information obtained using 2

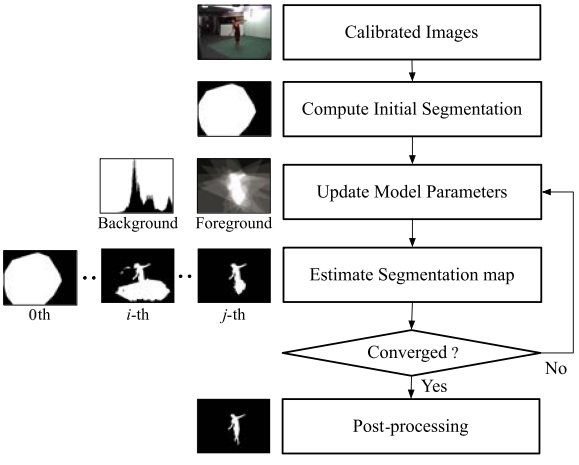

Fig. 1. Overall procedure of the proposed foreground extraction method

cameras, to the photometric information used for classification into background and foreground [8]. Incorporating depth information makes the process more robust, however it does not account for more than 2 camera consistencies. Zeng and Quan [9] proposed a method which estimates the silhouette of an object from the unknown background. They exploit the relationship between a region of an image and the visual hull. The approach requires however a good color segmentation, since foreground regions are identified based on the regions. Sormann *et al.* [10] applied the graph cut method to multiple view segmentation problem. They combine the color and the shape prior for robust segmentation from a complex background but user interactions are still required.

Our contribution with respect to the aforementioned approaches is to provide a fully automatic method that does not require static background prior knowledge, or user interaction. The different steps of the method are depicted in the Fig 1 and explained in the following sections.

3 Probabilistic Modeling

3.1 Definitions

We represent the input color images as $\mathcal{I}$ and the segmentation map as $\mathcal{S}$. τ is the prior knowledge of the scene. The knowledge of the background and foreground are noted as $\mathcal{B}$ and $\mathcal{F}$, respectively. $\mathcal{F}$, $\mathcal{B}$, and $\mathcal{S}$, are unknown variables, and $\mathcal{I}$ is the only known variable among all the variables. For a pixel, $\mathcal{S}$ has a value of either 0 for the background or 1 for the foreground. We use superscript i to represent a specific view. Subscript $\mathbf{x}$ indicates a pixel located at $\mathbf{x}(u, v)$. $\mathcal{I}^i_{\mathbf{x}}$ means the color values of the pixel $\mathbf{x}$ in the ith image.

3.2 Joint Probability Decomposition

With the defined variables, we represent the problem as a Bayesian network. Before we infer the probability of the segmentation map, we need to compute the joint probability of the variables. From the dependencies among the variables we decompose the joint probability as the equation 1.

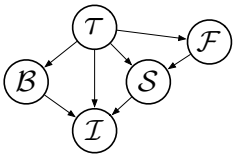

Fig. 2. Bayesian network representing the dependencies among the variables

$$Pr\left(\mathcal{S},\mathcal{F},\mathcal{B},\mathcal{I},\tau\right) = Pr\left(\tau\right) Pr\left(\mathcal{B}|\tau\right) Pr\left(\mathcal{F}|\tau\right) \\ Pr\left(\mathcal{S}|\mathcal{F},\tau\right) Pr\left(\mathcal{I}|\mathcal{B},\mathcal{S},\tau\right) \quad (1)$$

$Pr\left(\tau\right)$, $Pr\left(\mathcal{B}|\tau\right)$, and $Pr\left(\mathcal{F}|\tau\right)$ are the prior probabilities. Since we give no constraints on them, we assume that they have uniform distributions. We don't need to consider these priors to infer the probability of the segmentation map. Thus, they are ignored from now on. $Pr\left(\mathcal{S}|\mathcal{F},\tau\right)$ is the spatial consistency term. It represents the probability of the segmentation map, when the foreground information is available. The term $Pr\left(\mathcal{I}|\mathcal{B},\mathcal{S},\tau\right)$ is the image likelihood term. It tells us how much the image is related to the background information we know.

3.3 Spatial Consistency Term

Although each camera sees its own background different from the other's, the foreground should be consistent among all the views under our assumption. For a pixel in the ith image $\mathcal{I}^i$, the spatial consistency represents how much the other views agree that the pixel belongs to the foreground. It is referred from the segmentation maps $\mathcal{S}^k$ where $i \neq k$.

To compute the spatial consistency, we exploit the silhouette calibration ratio proposed in [11]. The silhouette ratio computes the probability of a pixel to be foreground from the silhouettes of the other views. We use modified silhouette calibration ratio $R_{\mathbf{x}}$ with a Gaussian distribution to give more penalty to the low silhouette calibration ratio value.

$$R_{\mathbf{x}} = e^{-(1-C_{\mathbf{x}})^2/\sigma^2} \quad (2)$$

where $C_{\mathbf{x}}$ is the silhouette calibration ratio corresponding to $\mathbf{x}$. σ is the standard deviation through which we can control the slope of the probability curve.

σ determines how much penalty is given to the a silhouette calibration ratio value. We give a tolerance to the silhouette calibration ratio through σ, since the knowledge about the foreground inferred from the other views can be wrong. σ

is computed from the number of cameras m we allow to miss, the corresponding silhouette calibration ratio C_m and the expected probability q.

$$\sigma = \sqrt{-\frac{C_m^2}{ln\,(q)}} \tag{3}$$

Since $\mathcal{S}_{\mathbf{x}}^i$ has the value of either 0 or 1, the spatial consistency term can be decomposed into two terms, when $\mathcal{S}_{\mathbf{x}}^i = 0$ and $\mathcal{S}_{\mathbf{x}}^i = 1$. When $\mathcal{S}_{\mathbf{x}}^i = 0$, the foreground information inferred from the other views does not give any clue about the background. In this case, we assume the spatial consistency has a uniform distribution $\mathcal{P}_b$. When $\mathcal{S}_{\mathbf{x}}^i = 1$, the spatial consistency term follows the inferred foreground information. Thus, we refer the modified silhouette calibration ratio as the spatial consistency term. Consequently, $Pr\left(\mathcal{S}_{\mathbf{x}}^i|\mathcal{F}^i, \tau\right)$ is computed as the equation 4.

$$Pr\left(\mathcal{S}_{\mathbf{x}}^i|\mathcal{F}^i, \tau\right) = \begin{cases} \mathcal{P}_b & \text{if } \mathcal{S}_{\mathbf{x}}^i = 0 \\ e^{-(1-C_{\mathbf{x}})^2/\sigma^2} & \text{if } \mathcal{S}_{\mathbf{x}}^i = 1 \end{cases} \tag{4}$$

3.4 Image Likelihood Term

The image likelihood term $Pr\left(\mathcal{I}_{\mathbf{x}}^i|\mathcal{B}^i, \mathcal{S}_{\mathbf{x}}^i, \tau\right)$ measures the similarity between a pixel's color $\mathcal{I}_{\mathbf{x}}^i$ and the background information we know. If a pixel is assumed to belong to the background, $\left(\mathcal{S}_{\mathbf{x}}^i = 0\right)$, the likelihood of the pixel color is computed from the the statistical color model of the background colors. To represent the color model of the background, several different methods, such as Gaussian mixture model or histogram, can be used.

When the pixel is considered as the foreground $\left(\mathcal{S}_{\mathbf{x}}^i = 1\right)$, the knowledge of the background color distribution does not provide any information to us. As we make no assumptions on the colors of the foreground, we set the image likelihood term to a uniform distribution $\mathcal{P}_f$ in this case. Consequently, the image likelihood term is defined as the following equation.

$$Pr\left(\mathcal{I}_{\mathbf{x}}^i|\mathcal{B}^i, \mathcal{S}_{\mathbf{x}}^i, \tau\right) = \begin{cases} \mathcal{H}_B\left(\mathcal{I}_{\mathbf{x}}^i\right) & \text{if } \mathcal{S}_{\mathbf{x}}^i = 0 \\ \mathcal{P}_f & \text{if } \mathcal{S}_{\mathbf{x}}^i = 1 \end{cases} \tag{5}$$

where $\mathcal{H}_B$ represents the statistical model of the background colors.

3.5 Segmentation Map Inference

After the joint probability distribution is defined, it is possible to infer the segmentation map from the given conditions by exploiting Bayes' rule. What we want to know is the probability distribution of the segmentation map $\mathcal{S}$, given the variables, $\mathcal{F}$, $\mathcal{B}$, $\mathcal{I}$, and τ.

For a pixel $\mathcal{I}_{\mathbf{x}}^i$, the probability of the segmentation map is inferred as the equation 6.

$$Pr\left(\mathcal{S}^i_{\mathbf{x}}|\mathcal{F}^i,\mathcal{B}^i,\mathcal{I}^i_{\mathbf{x}},\tau\right) = \frac{Pr\left(\mathcal{S}^i_{\mathbf{x}},\mathcal{F}^i,\mathcal{B}^i,\mathcal{I}^i_{\mathbf{x}},\tau\right)}{\sum_{\mathcal{S}^i_{\mathbf{x}}=0,1} Pr\left(\mathcal{S}^i_{\mathbf{x}},\mathcal{F}^i,\mathcal{B}^i,\mathcal{I}^i_{\mathbf{x}},\tau\right)} = \frac{Pr\left(\mathcal{S}^i_{\mathbf{x}}|\mathcal{F}^i,\tau\right)Pr\left(\mathcal{I}^i_{\mathbf{x}}|\mathcal{B}^i,\mathcal{S}^i_{\mathbf{x}},\tau\right)}{\sum_{\mathcal{S}^i_{\mathbf{x}}=0,1} Pr\left(\mathcal{S}^i_{\mathbf{x}}|\mathcal{F}^i,\tau\right)Pr\left(\mathcal{I}^i_{\mathbf{x}}|\mathcal{B}^i,\mathcal{S}^i_{\mathbf{x}},\tau\right)} \tag{6}$$

4 Iterative Optimization with Graph-Cut

To compute the optimal segmentation maps of all the views, we exploit the graph-cut method. In the same manner of the method proposed in [5], we iteratively update the unknown variables, $\mathcal{F}^i$ and $\mathcal{B}^i$, and estimate $\mathcal{S}^i$ for each view.

We build a graph $\mathcal{G}^i$ for every image $\mathcal{I}^i$. The pixels of $\mathcal{I}^i$ become the nodes of $\mathcal{G}^i$ and each pixel has edges connected to its eight neighbors. There are two special nodes in $\mathcal{G}^i$, the source S and the sink T which are connected to every nodes in the graph. Computing min-cut minimizes the the segmentation energy defined in the equation 7.

$$E_{total} = \sum_{\mathbf{x}\in\mathcal{I}^i} \lambda_1 E_p\left(\mathbf{x}\right) + \sum_{(\mathbf{x},\mathbf{y})\in N, \mathcal{S}_{\mathbf{x}}\neq\mathcal{S}_{\mathbf{y}}} \lambda_2 E_n\left(\mathbf{x},\mathbf{y}\right) \tag{7}$$

where E_p is the prior energy term and E_n is the neighborhood energy term. N is the sets of neighboring pixels.

The prior energy represents that how much a pixel is close to the foreground or the background. The neighborhood energy is for the smoothness of the segmentation. If two pixels have large color difference, there is high probability of the existence of the segmentation boundary. Thus, $Pr\left(\mathcal{S}^i_{\mathbf{x}}|\mathcal{F}^i,\mathcal{B}^i,\mathcal{I}^i_{\mathbf{x}},\tau\right)$ is used as the prior energy. We define the neighborhood energy as the equation 8 in this work.

$$E_n\left(\mathbf{x},\mathbf{y}\right) = \frac{1}{1+D\left(\mathbf{x},\mathbf{y}\right)} \tag{8}$$

where $D(\mathbf{x},\mathbf{y})$ is a color difference measurement between two neighboring pixels, $\mathbf{x}$ and $\mathbf{y}$. According to our experiments, the neighborhood energy is not limited to the equation 8. It may have different form, if it is designed to have small value with large color difference and large value with the similar colors of the two pixels.

We assign the capacity $w\left(\mathbf{x},\mathbf{y}\right)$ to every edge between the node $\mathbf{x}$ and $\mathbf{y}$ in $\mathcal{G}^i$. The prior energy assigns the capacities to the edges connected to S and T. For the edges connected to S, we set the capacities as shown in the equation 9. If a pixel is already known as background, we assign infinity to the edge. If not, the capacity follows the inferred probability with with $\mathcal{S}^i_{\mathbf{x}} = 0$,

$$w\left(\mathbf{x},S\right) = \begin{cases} \infty & \text{if } \mathbf{x} \text{ is known as the background} \\ Pr\left(\mathcal{S}^i_{\mathbf{x}}=0|\mathcal{F}^i,\mathcal{B}^i,\mathcal{I}^i_{\mathbf{x}},\tau\right) & \text{otherwise} \end{cases} \tag{9}$$

If T is one of vertices of an edge, the capacity is set as the inferred probability with $\mathcal{S}^i_{\mathbf{x}} = 1$.

$$w(\mathbf{x}, T) = Pr\left(\mathcal{S}^i_{\mathbf{x}} = 1 | \mathcal{F}^i, \mathcal{B}^i, \mathcal{I}^i_{\mathbf{x}}, \tau\right) \tag{10}$$

For the edges between two neighboring pixels, we assign the scaled neighborhood energy to it.

$$w(\mathbf{x}, \mathbf{y}) = \lambda_n E_n(\mathbf{x}, \mathbf{y}) \tag{11}$$

where λ_n is a scale factor.

After the convergence of the iterative optimization process, there can be misclassified pixels. To remove remaining errors, we perform a graph-cut based segmentation again as a post-processing. In the post-processing, the image likelihood term is modified as the equation 12. We use the foreground color model instead of the uniform distribution $\mathcal{P}_f$, in the same manner of the conventional graph cut methods.

$$Pr\left(\mathcal{I}^i_{\mathbf{x}} | \mathcal{B}^i, \mathcal{S}^i_{\mathbf{x}}, \tau\right) = \begin{cases} \mathcal{H}_B\left(\mathcal{I}^i_{\mathbf{x}}\right) & \text{if } \mathcal{S}^i_{\mathbf{x}} = 0 \\ \mathcal{H}_F\left(\mathcal{I}^i_{\mathbf{x}}\right) & \text{if } \mathcal{S}^i_{\mathbf{x}} = 1 \end{cases} \tag{12}$$

where $\mathcal{H}_B$ and $\mathcal{H}_F$ represents the color model of the background and the foreground, respectively.

5 Experimental Results

To demonstrate the effectiveness of our method, we performed experiments with the 'Dancer' and 'Temple' data sets which contain one foreground object. The 'Dancer' data set consists of 8 images with the size of 780x582. The 'Temple' data set is the selected 10 images with the size of 640x480.

Fig. 3. Interim segmentation results with and without the spatial consistency. One of the input images(left) Segmentation result with the color difference only (middle) Segmentation result with the color difference and the spatial consistency(right).

Fig 3 shows an intermediate segmentation results with and without the spatial consistency. In the input image, the foreground contains self shadows under the arms and the color of its hair is similar to colors of the background. Thus, some part of the foreground is lost, when only the color difference criterion is used. In contrast, segmentation with the spatial consistency preserves the part of the foreground well.

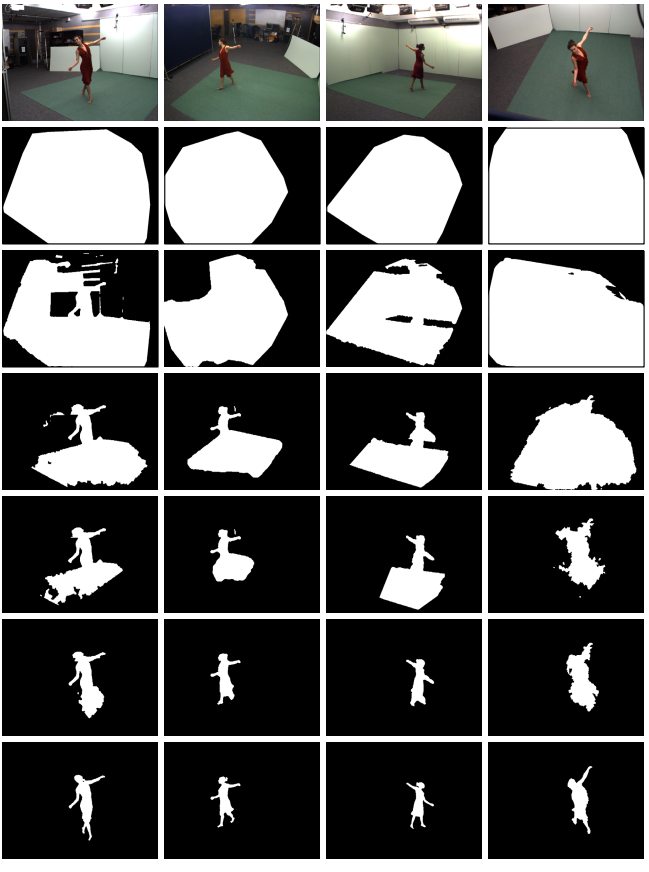

Fig. 4. Foreground extraction results with the 'Dancer' data set

Fig 4 shows the results with the 'Dancer' data set. The first row shows the selected images among the 8 input images. The initial segmentation computed from the intersection of the viewing volumes are depicted in the second row. The extracted foreground after each iteration is shown from the row 3 to 6. The initial segmentations in the row 2 converge to the results in the row 6 by the iterative optimization we presented in the previous section. Thanks to the spatial consistency, our method removes the background even though there are hard edges between the green mat and the gray floor. After the post-processing, we obtain the final segmentation maps depicted in the last row.

Fig 5 shows the experimental results with the 'Temple' data set. Since the Temple data set has almost black background, the segmentation of the foreground looks easier. However, the color similarity between the foreground and the background causes errors. As shown in the Fig 5, our method extracts the foreground successfully. Note that only the selected results among 10 images are presented because of the lack of the space.

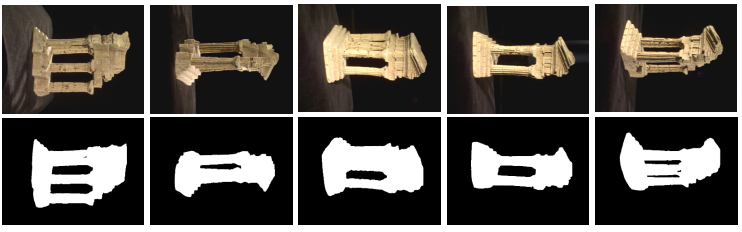

Fig. 5. Foreground extraction results with the 'Temple' data set

Table 1 shows the performance of the proposed method. To measure the performance, we computed the hit rate of the segmentation results from the ground truth. There exist errors in the segmentation, but it still shows good performance with the hit rates over 90 %.

Table 1. Segmentation performance (unit: %)

View ID	1	2	3	4	5	6	7	8	9	10
Temple	98.38	98.31	98.20	98.51	98.73	98.25	98.51	98.20	98.05	98.09
Dancer	93.93	97.74	96.44	93.14	92.41	94.94	97.20	94.93	-	-

Fig 6 shows the visual hulls reconstructed from the silhouettes of the ground truth that is obtained by manual hand operation and the silhouette obtained by the proposed method, respectively. They are not identical because of the segmentation errors, but the visual hull computed from our result still preserves the 3D shape of the foreground well.

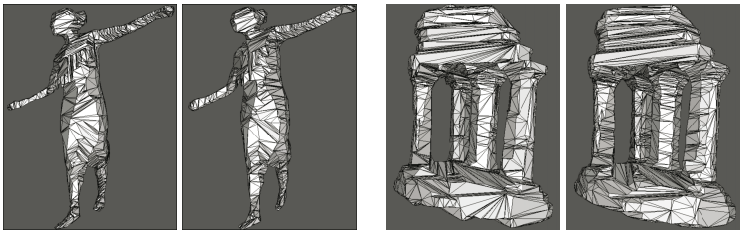

Fig. 6. Visual hulls reconstructed from the silhouettes of the ground truth(left) and the silhouette obtained by our method(right)

6 Conclusions

In this paper, we proposed a novel method of foreground extraction from multiple images. Our method integrates the spatial consistency with the color information

for robust estimation of the foreground from the unknown background. As shown in the experimental results, the spatial consistency provides an important clue to the separation of the foreground from the background. Since the proposed method requires neither the pre-knowledge of the scene nor the user interaction, it is more close to an automatic method than others. As a future work, an interesting issue is to extend the current work to the multi-view video sequence by exploiting both spatial and temporal constraints.

References

1. Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-Time Tracking of the Human Body. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 780–785 (1997)
2. Rowe, S., Blake, A.: Statistical mosaics for tracking. IVC 14, 549–564 (1996)
3. Friedman, N., Russell, S.: Image Segmentation in Video Sequences: A Probabilistic Approach. In: Proc. Thirteenth Conf. on Uncertainty in Artificial Intelligence (1997)
4. Boykov, Y.Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: International Conference on Computer Vision, vol. 1, pp. 105–112 (2001)
5. Rother, C., Kolmogorov, V., Blake, A.: Grabcut-interactive goreground extraction using iterated graph cuts. In: ACM SIGGRAPH, vol. 24, pp. 309–314 (2004)
6. Li, Y., Sun, J., Tang, C.K., Shum, H.Y.: Lazy snapping. In: ACM SIGGRAPH, vol. 23, pp. 303–308 (2004)
7. Freedman, D., Zhang, T.: Interactive graph cut based segmentation with shape priors. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 755–762. IEEE Computer Society Press, Los Alamitos (2005)
8. Gordon, G., Darrell, T., Harville, M., Woodfill, J.: Background Estimation and Removal Based on Range and Color, 459–464 (1999)
9. Zeng, G., Quan, L.: Silhouette extraction from multiple images of an unknown background. In: Hong, K.S., Zhang, Z. (eds.) Asian Conference on Computer Vision. Asian Federation of Computer Vision Societies, vol. 2, pp. 628–633 (2004)
10. Sormann, M., Zach, C., Karner, K.: Graph cut based multiple view segmentation for 3d reconstruction. In: The 3rd International Symposium on 3D Data Processing, Visualization and Transmission (2006)
11. Boyer, E.: On using silhouettes for camera calibration. In: The 7th Asian Conference on Computer Vision, pp. 1–10. Springer, Heidelberg (2006)