ROBUST 3D MODELING FROM SILHOUETTE CUES

Enliang Zheng, Qiang Chen, Xiaochao Yang, Yuncai Liu,

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University
Shanghai 200240, P. R. China

ABSTRACT

We consider the problem of 3D modeling under the environments where colors of the foreground objects are similar to the background, which poses a difficult problem of foreground and background classification. A purely image-based algorithm is adopted in this paper, with no prior information about the foreground objects. We classify foreground and background by fusing the information at the pixel and region levels to obtain the similarity probability map, followed by a Bayesian sensor fusion framework to infer the space occupancy grid. The estimation of the occupancy allows incremental updating once a new observation is available, and the contribution of each observation can be adjusted according to its reliability. Finally, three parameters in the algorithm are analyzed in detail and experiments show the effectiveness of this method.

Index Terms—Classification, robust, 3D modeling, Bayesian framework

1. INTRODUCTION

Traditional silhouette-based 3D reconstruction [1] is advantageous in simplicity and computational efficiency. However, the application of these methods is limited under experimental environments due to their sensitivities to noise. If one pixel is misclassified as none-silhouette, all voxels along the viewing line that connects the pixel and the camera center are labeled unoccupied, regardless of all the other observations. Under the environments where the objects and background colors are similar, the simple way of background subtraction cannot extract ideal silhouettes.

Only a few literatures incorporate 3D reconstruction under the environments where perfect silhouettes are unavailable. Snow et al. [2] used graph cuts to minimize formulation of the voxel occupancy problem, which incorporates spatial smoothness. Franco et al. [3] introduced the notion of occupancy grid from robotics community and applied the Bayesian framework to fuse the multi-view silhouette cues. Though these algorithms are more robust than traditional SFS for 3D modeling, they fail in the case of many classification errors. In [4], a learned class-specific prior is applied to reduce the effect of classification errors, but with limited generality.

Recently many works resort to the photo-consistency information for 3D modeling [5]. These approaches are confined to the disadvantage of computational complexity, because they have to deal with the visibility relationships of points on the objects' surfaces. Also these algorithms are more sensitive to camera calibration errors than the silhouette-based algorithms. Though Kutulakos et al. [5] declared no necessity of strict foreground objects segmentation, no experiments were conducted under complex environments.

In this paper, we improve the foreground and background classification by fusing the information at pixel and region levels, and succeed combining the classification results by modifying the framework presented in [3]. The contributions in this paper are the proposing a novel and robust way of foreground and background classification, and the 3D reconstruction under environments where the colors of the foreground matches the background from most views. This paper is organized as follows: Section 2 describes the classification of background and foreground on the image level. Section 3 presents the Bayesian framework for voxel occupancy inference. In Section 4, we discuss about the parameters in detail and show the results. The conclusion and future work are presented in Section 5.

2. CLASSIFICATION OF BACKGROUND AND FOREGROUND

This section introduces the way of classifying foreground and background by fusing the information at pixel and region levels to calculate the similarity probability map (SPM). The similarity probability for each pixel shows its likelihood of representing background. At first, some notations used in this paper are introduced. Let I_r denote the image captured by camera r, and I_r^p denote the color feature vector at position p in image r, r=1,2…n.

2.1 Pixel level Bayesian classification

The posterior probability of I_r^p representing background can be calculated by Bayesian theory [6]:

$$\mathrm{P}(b_r^p = 1 \mid I_r^p) = \frac{\mathrm{P}(I_r^p \mid b_r^p = 1) P(b_r^p = 1)}{\sum_{b_r^p} \mathrm{P}(I_r^p \mid b_r^p) P(b_r^p)} \qquad (1)$$

where b_r^p=1 means that the pixel p in image r represents background, while b_r^p=0 represents foreground.

In this paper, no assumptions about the priority is taken, so we get $P(b_r^p=1)= P(b_r^p=0)=1/2$. $P(I_r^p \mid b_r^p =1)$ is the probability of I_r^p if detection of background happens. The classical Gaussian model is used to formulate the background: $P(I_r^p \mid b_r^p =1)=N(I_r^p \mid \mu_r^p , \sigma_r^p)$, where μ_r^p, σ_r^p are the parameters of the Gaussian function. $P(I_r^p \mid b_r^p =0)$ is the probability of I_r^p if the pixel p in image r reports a foreground object detection. We take no assumptions about the color features of foreground objects: $P(I_r^p \mid b_r^p =0)=U(I_r^p)$, where $U(\cdot)$ represents the uniform distribution.

2.2. Region level classification

No ideal results can be achieved from the method presented in the previous subsection if the colors between the foreground and background objects are ambiguous. Some recent works [7] applied the notion of super-pixel as a significant step to fulfill their algorithms. That is the image is over-segmented into small regions according to some local features. Its main advantage is that a large number of pixels can be reduced to a relatively small number of super-pixels, hence making algorithms tractable. Since it is less likely for a super-pixel in the foreground object to be identical with that of background, it will obtain better classification if the super-pixels in two images are compared.

We choose the mean shift [8] out of some excellent image segmentation algorithms for two reasons. First, mean shift provides discontinuity preserving smoothing, which eliminates the image sensor noise, hence ensuring the correct segmentation. Second, it is much faster compared to normalized cut [9] and Pb detector [10]. This is important because the number of multi-view images at one time instant is often large.

Let (R_1, R_2) denotes the corresponding super-pixels from I_r and B_r (the image without foreground objects). The corresponding super-pixels are the regions with the same shape at the same position in I_r and B_r. We define $f(x, i)$ as the number of pixels in color histogram bin i of region x. The similarity between region R_1 and R_2 is defined as:

$$sim(R_1, R_2) = \left(\frac{\sum_i f(R_1, i) \cdot f(R_2, i)}{\left(\sum_i f(R_1, i)^2 \cdot \sum_i f(R_2, i)^2 \right)^{\frac{1}{2}}} \right)^{\alpha} \quad (2)$$

The parameter α will be discussed in Section 4. The value of *sim*, which shows the similarity between R_1 and R_2, ranges from 0 to 1. If R_1 and R_2 are similar, *sim* is close to 1, and vice versa. Here we resort to (2) instead of some classical dissimilarity measures such as Chi-Square distance, because in order to fuse the information acquired from the pixel level classification, the similarity between 0 and 1 is needed.

2.3. Similarity probability map

SPM shows each pixel's likelihood of representing background. That is if the similarity probability of a pixel is close to 0, it is more likely to represent the foreground. The SPM at position p in image r can be defined as:

$$SPM_r^p = \min(P(b_r^p = 0 \mid I_r^p), sim(R_1, R_2)) \quad (p \in R_1) \quad (3)$$

There are two reasons for this definition: First, note the Bayesian classification will usually misclassify the foreground as background because of color ambiguities. However, it does not tend to misclassify the background as foreground. It may happen because of noise, but a few misclassifications do not affect the later process of voxel reconstruction. Second, this formula reduces the impact of occasionally false segmentation around the boundary of foreground objects.

3. FUSION OF THE SPMs

The Bayesian framework, which is based on occupancy grid, is applied to fuse the SPMs of different images. The occupancy grid is a multi-dimensional tessellation of space into cells, where each cell stores a probabilistic estimate of its state. It is extensively used in the area of robotics [11] and firstly introduced for 3D modeling by [3]. In order for fusing SPMs of different images and clear, easy analysis, we modify the framework presented in [3].

The volume of interest is subdivided into m voxels with equal sizes. We denote the state of voxel i, i=1,2…m as S_i. We define S_i=1 if the voxel is occupied and S_i=0 otherwise. Since the two states of each voxel are exclusive and exhaustive, $P(S_i=1)+P(S_i=0)=1$. The purpose in this section is to find the posterior probability of each voxel being occupied after observations of all the n images $P(S_i=1 \mid \{I\}_n)$, $\{I\}_n=\{I_1,I_2…I_n\}$. The determination of an optimal estimation of occupancy grid is an incremental fusion of sensory information. That is the probability of occupancy is updated once a new image is available. Given the current estimation of a voxel i after observing r-1 images $P(S_i=1 \mid \{I\}_{r-1})$, $\{I\}_{r-1}=\{I_1,I_2…I_{r-1}\}$ and a new observation of I_r, estimation of the voxel can be updated by Bayesian theory:

$$P(S_i = 1 \mid \{I\}_r) = \frac{P(I_r \mid S_i = 1) P(S_i = 1 \mid \{I\}_{r-1})}{\sum_{S_i} P(I_r \mid S_i) P(S_i \mid \{I\}_{r-1})} \quad (4)$$

In formula (4), the posterior probability $P(S_i=1 \mid \{I\}_{r-1})$ serves as a priori to calculate the posterior probability of the voxel state when I_r is available.

Without any information about the initial prior state probability, we assume $P(S_i=1 \mid I_0)=P(S_i=1)=1/2$. One voxel's projection covers a region on the image. In order to reduce computation complexity, we use the projection of each voxel's center to represent the whole region. This hypothesis holds in the case the voxel is distant from cameras and the voxel size is small. We denote the feature vector at the voxel's projection p in image I_r as I_r^p. Therefore we have $P(I_r \mid S_i)= P(I_r^p \mid S_i)$.

Two important hidden variables O_r^p and $Dect_r$ are introduced. Let L_r^i denotes the viewing line connecting

voxel i and the camera r's center. O_r^p models some other object on L_r^i (In this paper, the object refers to part of the foreground object, either in front of or behind the voxel i along L_r^i). As to $Dect_r$, mainly due to silhouette extraction errors, I_r^p does not always correctly reflect the voxel occupancy on L_r^i. $Dect_r$=1 means image r reports a foreground object anywhere along L_r^i. $Dect_r$ is used to improve the framework's robustness to noise. Actually, as will be shown in Section 4, $Dect_r$ models the reliability of image r when updating the occupancy grid, which is very important in the whole framework. We propose the following formula:

$$\begin{aligned} P(I_r^p \mid S_i) &= \sum_{O_r^p} P(I_r^p \mid O_r^p, S_i) P(O_r^p) \\ &= \sum_{O_r^p} \left(\sum_{Dect_r} P(I_r^p \mid Dect_r) P(Dect_r \mid O_r^p, S_i) \right) P(O_r^p) \end{aligned} \tag{5}$$

Without prior information about O_r^p, we assume $P(O_r^p=1)=P(O_r^p=0)=1/2$. The four parametric distributions of $P(Dect_r \mid O_r^p, S_i)$ are set as follows:

$$\begin{aligned} P(Dect_r=1 \mid 0,0) &= PFA_r & P(Dect_r=1 \mid 1,0) &= PD_r \\ P(Dect_r=1 \mid 0,1) &= PD_r & P(Dect_r=1 \mid 1,1) &= PD_r \end{aligned} \tag{6}$$

Where PFA_r is the false alarm rate and PD_r is the detection rate. The term $P(Dect_r=1 \mid O_r^p=0, S_i=0)$ is the probability that I_r^p falsely reports a foreground object on the viewing line, when in fact there is none. Other three terms in (6) is the probability that I_r^p correctly detects an object in the viewing line L_r^i. Meaningful values of PD_r are close to 1, while PFA_r is generally close to 0. Different values of PD_r and PFA_r can be used for each image according to its reliability. As for the term $P(I_r^p \mid Dect_r)$, as SPM_r^p is normalized to 1, we can obtain the following:

$$\begin{aligned} P(I_r^p \mid Dect_r = 0) &= SPM_r^p \\ P(I_r^p \mid Dect_r = 1) &= 1 - SPM_r^p \end{aligned} \tag{7}$$

4. RESULTS AND DISCUSSION

4.1. Discussion About α

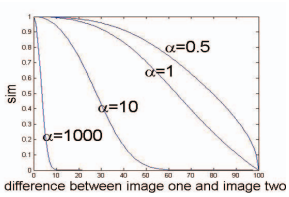
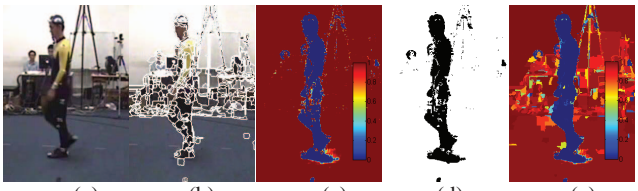


Figure 1. Results of *sim* with different α.

There is a close connection between the parameter α and the quality of images. To illustrate this, we assume one simple condition of two ideal images with the same sizes. Image one II_1, which is uniformly painted by color1, is the foreground, Image two II_2 simulates background in a dynamic process: Originally the whole image is painted by color1, and then gradually an increasing part of the image is changed to color2. The *sim* is calculated by formula (2) with different α. Figure 1 clearly shows when α increases, a smaller percentage of color difference between II_1 and II_2 can result in the same *sim*. If α is close to positive infinite, even one pixel difference between these two regions can make *sim* close to 0, which means the two regions compared are defined to be totally different. In reality too large value of α will result in misclassification of background as foreground. In our experiments, we set α between 1 and 3.

4.2. The SPM of Images

We take Figure 2(a), with some typical characteristics, as an example to show the effectiveness of classification. First, the image is blurred due to human motion and low quality of cameras. Second, the color of the person's clothes is similar to the background. And last, from this view the black back of the computer monitor in the scene is merged with the foreground, which challenges the algorithm of segmentation. Figure 2(b) shows the super-pixel obtained by mean shift. By simply thresholding Figure 2(c) with 0.5, we can see clearly various artifacts and holes in Figure 2(d). Note the SPM in Figure 2(e). There is a considerable improvement of the classification compared to Figure 2(c). Though there are still small holes in the silhouette, observations reveal the similarity probability of the holes is around 0.5, which is more meaningful in the step of occupancy grid estimation compared to closing to 1. (In Figure 2(c), the pixel probability is either close to 0 or 1).

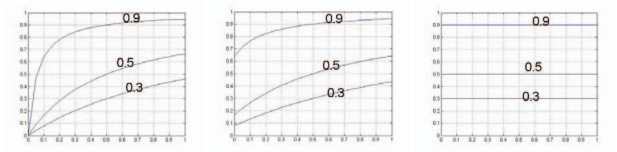


(a) (b) (c) (d) (e)

Figure 2. (a) The image of frame 42 from one view. (b) Super-pixels obtained by mean shift. (c)Pixel level Bayesian classification. (d) The silhouette achieved by thresholding (c) with 0.5. (e) The similarity probability map.

4.3. Analysis of the Bayesian fusion framework

Detailed Analysis of the Bayesian fusion framework is presented in this section. In Figure 3(a), PD_r and PFA_r are set to 1 and 0 respectively. Careful analysis of formula (5) and (6) reveals it equals to abandoning the hidden variable $Dect_r$, and directly relating I_r^p to S_i or O_r^p. Figure 3(a) shows in the condition that the true state of a voxel is occupied, there is no opportunity to recover the correct voxel information if 1-SPM_r^p is falsely set to 0 from any one view. This is similar to the traditional SFS. If we set PDr=0.5 and $PFAr$=0.5 (See Figure 3(c).), which means no observations are reliable, the value of $P(I_r^p \mid Dect_r=1)$ has no impact on $P(S_i=1 \mid \{I\}_{r-1})$. In another word, this image is abandoned and makes no contributions to the update of grid occupancy. At last, from all these three figures as a series, it is shown when PD_r decreases and PFA_r increases, the slope of the curve gradually becomes smaller,which means image r contributes



(a) (b) (c)

Figure 3. In all these three figures, y axis represents P(S_i=1| $\{I\}_r$), and x axis 1-SPM_r^p (or P(I_r^p | $Dect_r$=1)). Different curves represent different values of P(S_i=1| $\{I\}_{r-1}$). (PD_r ,PFA_r) is set (a):(1, 0), (b): (0.9, 0.1) and (c): (0.5, 0.5) respectively for all three figures.

less to the update of the occupancy grid. It provides one hint that is not discussed in this paper: If the reliability of background and foreground classification can be automatically evaluated for image r, we can relate the reliability to the parameters PD_r and PFA_r.

4.4. 3D model

In our experiments, 15 calibrated cameras (704*576, at 25Hz) have a common scene of a region of 2m*3m. The volume of interest is divided into 200*300*200 voxels. We illustrate our algorithm using frame 42 of the walking sequence. Figure 4 shows the images used are blurred and difficult to extract ideal silhouettes. The surface model can be obtained by extracting the isosurface from the occupancy grid. As becomes apparent in Figure 5, the surface model produced by our algorithm is much smoother and more integral than the one calculated by [3]. The model surface (in Figure 5(b)) is swollen instead of being improved when the isosurface threshold value is lowered. The actual reason is that 3D scene contents cannot be well presented because of numerous classification errors.

5. CONCLUSIONS

This work aims to solve the problem of 3D modeling under environments where ideal silhouettes are unavailable. We propose a novel method of foreground and background classification by fusing the information at pixel and region levels, which successfully solves the problem of color ambiguities. The SPMs are directly used to incrementally update the occupancy grid, hence avoiding the hard decisions about silhouette extraction. The experiments of our algorithm confirm the effectiveness of this method and show a considerable improvement.

ACKNOWLEDGMENTS

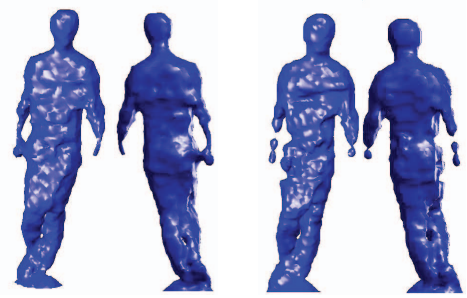
Supported by National Natural Science Foundation of China (grant no. 60675017), and 973 Program of China (grant no. 2006CB303103). We thank Li Guan for helpful comments.

REFERENCE

[1] G. Cheung, T. Kanade, JY. Bouguet, M. Holler. A real time system for robust 3D voxel reconstruction of human motions. CVPR, vol.2, pp.714 – 720, June 2000.
[2] D. Snow, P. Viola, R. Zabih. Exact voxel occupancy with graph cuts. CVPR, vol.1, pp:345 – 352, June 2000.
[3] J. Franco, E.Boyer. Fusion of multi-View silhouette cues using a space occupancy grid. ICCV, vol. 2, pp: 1747- 1753, Oct. 2005.
[4] K.Grauman, G.Shakhnarovich, T. Darrell. A bayesian approach to image-based visual hull reconstruction. CVPR, 2003.
[5] KN. Kutulakos, SM. Seitz. A theory of shape by space carving. IJCV, vol. 38, pp.199-218, July 2000.
[6] L. Li, W. Huang, IYH. Gu, Q. Tian. Statistical modeling of complex backgrounds for foreground object detection. TIP, 2004
[7] X. Ren, J. Malik. Learning a classification model for segmentation. ICCV, vol.1, pp.10 – 17, Oct. 2003.
[8] D Comaniciu, P Meer. Mean shift: a robust approach toward feature space analysis. PAMI, vol.24, pp.603-619, May 2002.
[9] J. Shi, J. Malik. Normalized cuts and image segmentation. PAMI, vol.22, pp.888-905, August 2000,.
[10] DR. Martin, CC. Fowlkes, J. Malik. Learning to detect natural image boundaries using local brightness, color and texture cues. PAMI, vol. 26, pp.530–549, May 2004.
[11] A. Elfes. Occupancy grids: a probabilistic framework for robot perception and navigation. PhD thesis, Carnegie Mellon Univ.1989.

(a) (b) (c)

Figure 4. (a): Four of the 15 images of frame 42, seen from different views. (b) : The corresponding background of the left four images. (c): The corresponding silhouettes computed from these images. Morphological operations are used to reduce noise



(a) (b)

Figure 5. Comparison of our reconstruction scheme with [1]. (a) Two views of the isosurface of probability 0.90 generated by our algorithm. We set PD_r =0.9, PFA_r =0.1 for all images. (b) Two views of the isosurface of probability 0.60 generated by the algorithm presented in [4]. Parameters are carefully chosen to get the best isosurface: P_d =0.9, P_{fa} =0.1, k=5.