Shape from silhouette under varying lighting and multi-viewpoints

Maxim Mikhnevich  
*Computer Vision and Systems Laboratory*  
*Laval University*  
Quebec, Canada  
Maxim.Mikhnevich.1@ulaval.ca

Patrick Hébert  
*Computer Vision and Systems Laboratory*  
*Laval University*  
Quebec, Canada  
Patrick.Hebert@gel.ulaval.ca

Abstract—We present a new segmentation method for visual hull (VH) reconstruction from an image sequence. The camera moves on a hemisphere around an object set on a turntable with complex and unknown background. Even though the background is assumed to be constant, it cannot be modeled from all possible viewpoints. A series of images of a moving object under different lighting conditions are captured and analyzed. In this analysis, information from space, time, and light domains are fused into an MRF framework. The method is validated by VH reconstruction of challenging objects with different optical properties.

Keywords—shape from silhouette; segmentation; visual hull; multi-view; graph cut;

I. INTRODUCTION

Shape from silhouette (SFS) is a classic technique in computer vision for 3D object reconstruction. The advantage of this method is that it requires neither constant object appearance nor the presence of textured regions. Instead, it relies on an object silhouette extracted in each view. In this paper we take an SFS approach in order to reconstruct a wide range of objects.

We are interested in automatic silhouette extraction since we work with dense datasets; a semi-automatic method would be inadequate in such a scenario. For this purpose a roboticized system was designed, which allows us to rotate an object on a turntable, and position the camera viewpoint on a hemisphere, as illustrated in figure 2.

Several methods have been proposed to obtain high-quality silhouettes such as chroma keying [1] or flash matting [2]. Chroma keying relies on a uniform and known background. As a result, this method can only be used with predefined camera viewpoints. In our project an object can be captured from any viewpoint on a hemisphere, therefore the usage of a controlled background or modeling the background priorly is either not possible or impractical. Flash matting is also not appropriate since one of the requirements is the presence of a distant background.

The roboticized system further integrates a light system. It gives us freedom to control a set of light sources on the hemisphere, besides the camera viewpoint. Although the method is presented in this context, it is applicable to any scene with constant background where a dense set of images is captured.

The method works as follows: for a fixed viewpoint we observe the evolution of each pixel intensity in time while the object turns; this signal is further analyzed to extract background likelihood as a function of time. Since the background is assumed to be constant, background likelihood depends on constancy of the intensity evolution with time. Due to the specific topology and geometry of an object, the constancy must be evaluated both locally and globally. Combining these two constancy measurements, the background likelihood can be estimated. It can be observed, in figure 1(c), that the intensity profile (blue) is constant and uniform in the interval from 0 to 140 and from 240 to 360, where the signal corresponds to the background. On the contrary, the intensity profile in the interval from 140 to 240 changes, the value of each point is different to its neighbors and thus such a signal should be classified as an object. This is reflected in the estimated background likelihood (red curve).

Furthermore, due to possible shadow the same analysis is repeated for different lighting conditions. Having an ability to control an individual light source, we show that it is better to exploit light sources alternately instead of turning them all on at the same time. Finally color bleeding is a view dependent effect, and it is taken into account by processing several viewpoints.

These ideas are integrated into a Markov Random Field (MRF) optimization framework, and the optimization is performed through graph cut. The final result is a set of silhouettes and a visual hull (VH).

The contributions of our work are:

- the analysis of the pixel in time signal which leads to the automatic estimation of the background likelihood.
- the fusion of information from space, time, light domains and its integration into an MRF framework.

The paper is organized as follows: In section II we will give an overview of the related work. Section III will briefly describe the acquisition process. In sections IV and V the details of the estimation of background and object likelihoods as well as the segmentation framework will be given. In
Figure 1: Teapot images captured at different times and under different lighting. Images (a) and (b) are captured under lighting #1 and images (d) and (e) under lighting #2. For each of these two cases, an intensity profile (blue curve) and estimated background likelihood (red curve) are depicted in (c) and (f) respectively. While at time 50 the imaged pixels \( x_{50} \) and \( y_{50} \) belong to the background and are located in a constant interval in an intensity profile, pixels \( x_{145} \) and \( y_{145} \) belong to the object and are located within varying intensity intervals. Segmentation then consists in identifying these intervals.

Section VI some experimental results will be demonstrated and discussed.

II. RELATED WORK

The concept of VH was first introduced by Baumgart [3]. This concept is strongly linked to image segmentation into background and foreground, since the object’s silhouettes are required for VH construction. One of the classical and simplest techniques for silhouette extraction is chroma keying [1]. The idea is to take a picture of an object against uniform or known background, then by using color thresholding or background subtraction a silhouette of an object is obtained. This approach was used in many 3D reconstruction systems for silhouette extraction [4], [5]. Even though this method is easy to implement and it provides fairly good results, there are some drawbacks. First the environment must be modified in order to set up the background. This limits possible camera placement on a hemisphere since the background has to be visible from all viewpoints. Second, if an object part has the same color as the background this may lead to holes in the object’s silhouette.

Chroma keying was extended in several works where an active background system was used [4], [6] instead of a static background. As an active background, authors set up several monitors around an object. Different image patterns are displayed on these monitors and a reference picture is taken without an object. In the next step the procedure is repeated with an object between the camera and the background screens. Then all images are used in a multi-background matting equation to compute an alpha matte [4] or environmental matte [6]; environmental matte describes not only object opacity but also its ability to refract and reflect the light for each pixel. Even though such an approach allows the extraction of the alpha matte of an object made from material with complex optical properties such as glass, it seriously complicates an acquisition system. Actually, it is not possible to move the camera with respect to the background screens, because images with and without an object have to be aligned at the pixel level.

Active lighting based approaches make it possible to directly estimate the geometry of an object without silhouette estimation. These methods can be classified into several subcategories: range scanners, structured light, photometric stereo. Laser range scanning [7] and structured light [8] are limited in the type of materials that can be scanned, such that they may perform poorly on shiny and transparent surfaces or the materials that have low albedo.

An interesting active lighting approach for recovering object geometry is photometric stereo [9], [10]. One of the particularities of this type of method is an explicit assumption about the reflectance model of an object, as a result applying this method to an object whose reflectance does not satisfy a simple diffuse model proved to be less reliable. We consider that photometric stereo may be used as a complementary technique to SFS, in order to increase the quality of the reconstructed geometry.

Flash matting [2] resembles active lighting techniques. This method is dedicated to foreground segmentation. The idea is to use two images: flash and no-flash. It is assumed that the appearance of a foreground object in a flash image is different from that of a no-flash image. However, the background remains similar in both images. The main requirement in this method is that the background has to be sufficiently far from the object, such that it is not affected by the camera flash. Unfortunately, this condition cannot be satisfied adequately in our experimental environment.

Another method for image segmentation is to ask a user
to initialize the segmentation process [11]. Here, the idea is to consider object segmentation as an energy minimization problem on a graph, and then apply a min-cut algorithm to find the global minimum. In order to construct an energy function, a priori knowledge about an image is supplied by a user. A user labels some parts of an image as background or foreground, this information is used to learn object and background models and construct an energy function. In our work we also use a min-cut algorithm to obtain optimal segmentation, however we concentrate our attention on automatic estimation of object and background likelihoods.

Semi-automatic image segmentation in a single image by graph cut was extended to automatic silhouette extraction in multiple views in [12], [13]. In [13] images were collected from different views, then object and background models were estimated, and finally a graph cut framework was applied for each image individually. A similar technique was designed in [12], the main difference to the one mentioned previously is that the energy minimization is performed on a 3D volume. Although these methods may work well, there is still one disadvantage: an explicit color model of an object and background is required; thus if the same color belongs to the object and background model, the result may lead to over or under silhouette estimation. In our work we try to avoid explicit color modeling of an object and background in order to overcome this limitation.

III. IMAGE ACQUISITION AND NOTATION

An image sequence is acquired using a roboticized turntable setup, as shown in figure 2. We fix the camera view and capture $N_L$ images under light sources located at different positions. Then the turntable is rotated and image acquisition is repeated. This process is repeated $N_T$ times, and as a result for each position of an object there are $N$ images, and there are $N_T$ different object locations. The total number of images acquired for one camera viewpoint is $N_L \cdot N_T$. These images are organized into a 4D volume. The dimensions of this volume are: $U$, $V$, $T$ and $L$. $U$ and $V$ are spatial dimensions, they represent the image domain. The third dimension is time $T$, parameterizing object displacement on the turntable. The last one is the lighting condition $L$: all of the light sources are different in position but identical in intensity. Thus $I(u, v, t, l)$ is an intensity of a pixel $(u, v)$ at time $t$ under lighting condition $l$ for one viewpoint. We also define two subsets of $I$, where $I_L$ is a subset with all the images for one turntable position, collected under all available lighting, $I_T$ is a subset where all the images with fixed lighting but for multiple turntable positions are gathered, and $I_{l, t}$ is a single image under light source $l$ in position $t$. The acquisition process is repeated for other camera viewpoints.

Figure 2: Acquisition setup.

IV. BACKGROUND LIKELIHOOD ESTIMATION

Estimation of background likelihood is based on a constant background assumption, a condition which is satisfied in the $I_T$ subset. Coordinates $U$ and $V$ are fixed in the $I_T$ subset and all the pixel values are collected along the $T$ axis. These values form an intensity profile defined as $I_T(u, v) = X = \{I_T(u, v, t_1), I_T(u, v, t_2), ..., I_T(u, v, t_{N_T})\} = \{x_1, x_2, ..., x_i, ..., x_{N_T}\}$, depicted by a blue curves in figures 3(b) and 3(c). This profile can be interpreted in the following way: a larger likelihood is attributed to a pixel value located inside a uniform interval. A pixel value inside a region which changes its intensity quite rapidly, is assigned low background likelihood.

The background likelihood of a pixel value from an intensity profile is estimated in both a local and global step. In a local step, first, pixel values in the intensity profile $X$ are sorted in ascending order based on the intensity. A new sorted profile is defined as $X'$ (figure 3(a) blue curve). Then for each point in $X'$, the minimum standard deviation ($\sigma$) is calculated within a large window size $w_g$ (in the experiments $w_g = 90$), see red curve in figure 3(a). Finally, estimated minimum $\sigma$ values are reordered based on the correspondence between $X$ and $X'$ (figure 3(b) red curve).

The necessity of the global step is dictated by the possibility that an object may contain holes or gaps between its parts. In such a case the points inside the intensity profile are mixed between object foreground and background, (see figure 5). Thus by sorting points according to their intensity we group background and object points separately.

A minimum $\sigma$ within window size $w$ for a point $x_i$ is defined as:

$$S(x_i) = \min_{j \in [i-w+1, i+w]} \sigma_w(x_j),$$

where $\sigma_w(x_j)$ is the standard deviation calculated on the subset $\{x_j, x_{j+1}, ..., x_{j+w-1}\}$. $S(x_i)$ describes the constancy of a point $x_i$ in a region with the size $w$ and it has two properties. The smaller it is, the more a point is
constant. The larger the window size, the more significant the measurement is.

In the second step, a similar procedure is carried out. It is applied now to $X$ and the minimum $\sigma$ with a small window size $w_l$ (in the experiments $w_l = 10$) is estimated (see figure 3(c)). The idea of this part is to evaluate the local property. It is possible that at some positions an object point may have the same color intensity as a background and thus during the estimation of the minimum $\sigma$ at the global step it will be low. However, the intensity values in $X$ in a small interval around this point will be different, since the object changes its position at each time step. Thus the measurement of the minimum $\sigma$ with $w_l$ will result in a higher value.

The local and global measurements of $\sigma$ are transformed into background likelihood based on the following criteria. A point that has a small minimum $\sigma$ both during a local and global step is more likely to belong to the background. On the other hand a point that has a high minimum $\sigma$ at least in one step is more likely to belong to an object.

The whole algorithm for background likelihood estimation can be summarized as follows:

1) Sort all the points from intensity profile $X' = \text{sort}(X)$
2) Estimate global constancy of each point $x'_i \in X'$, thus compute $S'_g$ as follows:
   $$S'_g(x'_i) = \min_{j \in [i - w_g + 1, i + w_g]} \sigma_{w_g}(x'_j).$$
3) Based on the correspondence between $X'$ and $X$, reorder $S'_g$ in order to obtain $S_g$
4) Estimate local constancy of each point $x_i \in X$, thus compute $S_l$ as follows:
   $$S_l(x_i) = \min_{j \in [i - w_l + 1, i + w_l]} \sigma_{w_l}(x_j).$$
5) Compute background likelihood for each point in $x_i \in X$ as follows:
   $$P_B(x_i) = \frac{w_g}{\exp(S_l(x_i) + S_g(x_i))}.$$  

Equation 4 has the property such that it tends to 0 when $S_l + S_g \to \infty$, meaning that the point is inside a varying region and most likely belongs to an object; and it tends to the window size when $S_l + S_g \to 0$, meaning that the point is inside a stable region and most likely belongs to a background.

A. Shadow and color bleeding

When direct light cannot light a turntable due to object occlusion, a shadow (relatively uniform area with low intensity) appears. While the object moves, the shadow follows it, thus violating the assumption concerning constant background. Since shadows produce a moving region, the resulting background likelihood will be low. In order to overcome this problem we use several light sources located at different positions. A reasonable question arises here, why not use all the light sources at the same time in order to illuminate an object from all sides and thereby eliminate shadows. It would be beneficial to switch on all the light sources when it is possible to provide a uniform light distribution in the scene. Unfortunately it is mostly unfeasible to provide perfectly uniform lighting in a scene. On the other hand, a sequential usage of the light sources located at different points provides a control on an object’s shadows, and thus helps to cope with this effect. In figure 6(a) the picture of the can is taken with all the light sources turned on, and we can observe some shadow penumbras caused by non-uniform distributions of the lights.

When the object consists of shiny reflective material such as aluminum, it reflects light on the turntable and changes its appearance, see figure 6(b). This phenomenon depends on the object position and changes the color observed on a small area of the turntable near the object. The estimated background likelihood at such a point is close to 0 as bleeding moves with the object and prompts a small deviation in intensity. A similar problem occurs with circular targets on the turntable, since their color is different from the turntable and thus helps to cope with this effect. In figure 6(b) the turntable, since their color is different from the turntable and they move with an object. Although it is difficult to cope with this problem in a single view, it can be solved using multiple views while recovering the VH, since color bleeding is not consistent with different views.

B. Fusion space-time light volume for each frame

The estimation of background likelihood for each space-time volume $I_T$ was described above. This is enough if the scene is lit uniformly by global ambient light during an acquisition process. However if several directional light sources are used (as in our case), then the fusion process is applied in order to incorporate information from different light sources. The difficulty of the fusion is caused
by contradictory estimation of background likelihood from different light sources. For example with one light source, some parts of an object can be in the shadow or self-shadow which results in higher likelihood for background, and under another light source the same part of an object can be illuminated and thus have a low probability for the background. This is demonstrated in figure 1, under light source \#1 a point \textit{x}_{1,45} is attributed a low background likelihood due to shadow. However under light source \#2 the same point obtained high background likelihood. The same happens if we consider specular reflection. Under one light source it could lead to a low background likelihood and under another one the same area could be little affected by the reflection which may result in a high value. Thus, in order to choose an appropriate light source we use a simple but effective rule. For a given view and pixel we consider all the images under different lighting, and for each pixel we find the one that corresponds to a maximum intensity. Figures 4(a) and 4(b) show a maximum image along with the corresponding indices of the light sources respectively. Based on the light source index we choose the background likelihood for each point in the \textit{UV} space, and it gives us a good approximation of the background likelihood (figure 4(c)). Thus, the final background likelihood is estimated as follows:

\[
\max_{\text{ind}} = \arg \max (I_L(u, v, l)),
\]

\[
P_B \text{ final} = P_B(u, v, t, \max_{\text{ind}}).
\]  

\[\text{(5)}\]

**C. Background to object likelihood**

Background pixel likelihood is estimated through equation 5. The object likelihood cannot rigorously be estimated as \(1 - P_B\), since all the information about each pixel is not available. Thus we select all the values close to 0 (less than a threshold \(R\)) in background likelihood and use these values as an approximation of object likelihood. We define a small value \(f\) for these pixels in order to indicate that there is a possibility for an object and to other pixels, we assign the value \(\frac{1}{10}\) indicating that the likelihood that they represent an object is much smaller. The object likelihood is estimated as follows:

\[
P_O = \begin{cases} f & : P_B \text{ final} < R \\ \frac{1}{10} & : \text{otherwise}. \end{cases}
\]

\[\text{(6)}\]

**V. Segmentation as an Optimization Process**

In the previous section the estimation of prior background and object likelihoods was described, and now the whole segmentation process can be defined. The goal of segmentation is to assign for each pixel \(p\) in image \(I_{t,l}\) a label \(m_p\) which can be the object or the background. Segmentation is addressed in an MRF framework; more precisely, it is performed by energy minimization of \(E\) through graph cut [11].

\[
E(M) = \lambda \sum_p P(m_p) + \sum_{m_p,m_q \in N} B(m_p,m_q),
\]

\[\text{(7)}\]

where \(P(m_p)\) is the prior knowledge that each pixel belongs to object and background; \(B(m_p,m_q)\) is a boundary property of an image and it reflects the strength of the connection between neighboring pixels; \(M\) is the set of all labels in an image \(I_{t,l}\); each element can be either background or object with values \(\{0, 1\}\); \(\lambda\) controls the importance of prior knowledge versus boundary term (\(\lambda \in [0, \inf]\)); \(N\) is the neighborhood which defines connectivity between pixels.

The boundary \(B(m_p,m_q)\) term characterizes the relationships between neighboring pixels. It can be considered as a penalty between two neighboring pixels. If pixels belong to the same object, then the connection between them has to be strong. If pixels are from different objects, then \(B(m_p,m_q)\) should give a value close to 0 for the graph cut to make a cut between them. We have multiple images of the same view but under different lighting conditions. Thus in order to take into account this information, we modify this term from [11]. For a given pair of pixels \(p\) and \(q\), we use the sum of the squared differences over all lighting:

\[
B(m_p,m_q) = \exp \left( -\frac{\sum_i^N (I_{p,i} - I_{q,i})^2}{2\gamma^2} \right) \cdot \frac{1}{D(p,q)}.
\]

\[\text{(8)}\]
where $I_{p,i} = I(u_q, v_q, t, l_i)$ and $I_{q,i} = I(u_q, v_q, t, l_i)$ are intensities for pixel $p$ and $q$ at time $t$ under lighting $l_i$, $D(p, q)$ is the Euclidean distance between two pixel sites, and $\gamma$ is a scale factor.

The prior knowledge term $P(m_p)$ in equation 7 defines a preference for pixel $p$ to be object and background:

$$P(m_p) = \begin{cases} P_O & m_p = 1 \ (\text{object}) \\ P_B & m_p = 0 \ (\text{background}). \end{cases} \quad (9)$$

VI. EXPERIMENTS

A calibrated set of images acquired with the roboticized setup was used in the experiments. The background behind the object is complex and not uniform: it includes walls, some parts of the setup and turntable with circular targets and varies with viewpoint. These aspects make it impractical to use a background subtraction technique.

Several experiments were performed with different types of objects. A few are presented here. In the datasets one camera view and 30 different light sources were used. A turntable was rotated with an object 360 times by 1 degree and a gray level image was captured each time. Some results are shown in figures 7, 8, and 9. The first row of the columns number 1 and 2, $\text{max}_{\text{ind}}$ and $P_{B\ \text{final}}$ from equation 5 are depicted. In the third row the results of the segmentation are shown. In the third and fourth columns of the first row the original images are shown. Finally in the last row, the reconstructed VH is depicted.

Figures 7(i) and 7(j) show the results of the segmentation of a cola can. It is made from aluminum and presents a specular reflection and non constant appearance while moving on a turntable. In figure 7(i) it can be seen that part of the background was also included into the object’s silhouette. This is due to color bleeding as we pointed out in section IV-A. Notwithstanding, any false segmentation part that is not consistent with other views will be removed while recovering VH from the intersection of visual cones, see figures 7(g) and 7(h).

In figure 8 the results of the vase segmentation are presented. This vase is made of clay, which was not painted on the outside and thus it is almost textureless. Inside it is covered by lacquer. Such covering results in specular reflections. In figure 8(j) we can see that some parts of the background were also included into the silhouette. This is due to the location of the targets close to the vase. Such parts are not consistent with other views and will be automatically removed in the VH construction stage, see figures 8(g) and 8(h).

In figure 9 the result of the Rubik’s Cube segmentation are presented. Each side of the cube reflects a lot of light such that a strong color bleeding effect appears in certain locations. It can be noticed that in figures 9(e) and 9(f) the bottom boundary of the cube’s background likelihoods are a bit fuzzy due to this phenomenon. Nevertheless, the reconstructed VH has sharp boundaries, see figures 9(g) and 9(h).

Note that the turntable background likelihood is lower than the wall background likelihood (see figures 7(e), 7(f), 8(e), 8(f), 9(e), 9(f)). This is due to the intensity variation during turntable rotation. Despite this, the turntable background likelihood is still significantly higher than the object background likelihood.

VII. CONCLUSION

In this paper a new method for object silhouette extraction on a turntable setup was proposed. The method avoids background modeling since the camera viewpoint is moved freely on a hemisphere. We assume that the background, if it is observed, is constant. Also, several light sources
Figure 7: Results of cola can segmentation. (a), (b) depict $\text{max}_{\text{ind}}$ while (e), (f) depict $P_{B_{\text{final}}}$ calculated from equation 5; (i), (j) are the segmentation results for a given camera viewpoint; (c), (d) are raw images; (g), (h) show the reconstructed VH calculated after integrating silhouettes from several camera viewpoints. The viewpoints of the VH are specified in terms of azimuth (az) and elevation (el).

located at different positions were used to cope with different light phenomena such as shadows and color bleeding. It has been shown here that this assumption and varying lighting allows silhouettes to be extracted from an object with complex reflectance properties and without any prior knowledge. A dense set of images that is required to obtain an object segmentation is however imposed by the proposed approach. As part of our future work, we intend to refine the reconstructed VH using captured photometric information in order to improve 3D interpretation of a wide variety of objects.

REFERENCES


Figure 8: Results of vase segmentation. (a), (b) depict $\text{max}_{\text{ind}}$ while (e), (f) depict $P_{B_{f inal}}$ calculated from equation 5; (i), (j) are the segmentation results for a given camera viewpoint; (c), (d) are raw images; (g), (h) show the reconstructed VH calculated after integrating silhouettes from several camera viewpoints. The viewpoints of the VH are specified in terms of azimuth (az) and elevation (el).

Figure 9: Results of the Rubik’s Cube segmentation. (a), (b) depict $\text{max}_{\text{ind}}$ while (e), (f) depict $P_{B_{f inal}}$ calculated from equation 5; (i), (j) are the segmentation results for a given camera viewpoint; (c), (d) are raw images; (g), (h) show the reconstructed VH calculated after integrating silhouettes from several camera viewpoints. The viewpoints of the VH are specified in terms of azimuth (az) and elevation (el).