Background Removal of Multiview Images by Learning Shape Priors

Yu-Pao Tsai, Cheng-Hung Ko, Yi-Ping Hung, and Zen-Chung Shih

***Abstract*—Image-based rendering has been successfully used to display 3-D objects for many applications. A well-known example is the *object movie*, which is an image-based 3-D object composed of a collection of 2-D images taken from many different viewpoints of a 3-D object. In order to integrate image-based 3-D objects into a chosen scene (e.g., a panorama), one has to meet a hard challenge—to efficiently and effectively remove the background from the foreground object. This problem is referred to as *multiview images (MVIs) segmentation*. Another task requires MVI segmentation is image-based 3-D reconstruction using multiview images. In this paper, we propose a new method for segmenting MVI, which integrates some useful algorithms, including the well-known graph-cut image segmentation and volumetric graph-cut. The main idea is to incorporate the shape prior into the image segmentation process. The shape prior introduced into every image of the MVI is extracted from the 3-D model reconstructed by using the volumetric graph cuts algorithm. Here, the constraint obtained from the discrete medial axis is adopted to improve the reconstruction algorithm. The proposed MVI segmentation process requires only a small amount of user intervention, which is to select a subset of acceptable segmentations of the MVI after the initial segmentation process. According to our experiments, the proposed method can provide not only good MVI segmentation, but also provide acceptable 3-D reconstructed models for certain less-demanding applications.**

***Index Terms*—Graph cut, image segmentation, Markov random field (MRF), medial axis, multiview images (MVIs), object movie, 3-D modeling, volumetric graph cuts.**

I. Introduction

CONSTRUCTING a realistic environment is an important research topic in the domain of computer graphics. Virtual reality systems involve two major classes of technique, namely geometry-based and image-based rendering. In geometry-based methods, a complete 3-D model of the environment, including all the objects within the virtual world, is constructed and rendered to simulate the virtual world. Conversely, image-based methods, collections of images taken from different viewpoints of the environment is used to generate novel views of the virtual world. Both approaches have their own advantages and weaknesses. However, image-based methods have become increasingly popular because of the ease of construction of the high quality and photorealistic environment. Additionally, in the image-based methods, the computational expense of rendering the virtual world is independent of the complexity of the objects and scenes.

Apple, Inc. [1] proposed a popular image-based method, called object movies, to capture and display 3-D objects. An object movie is composed of a collection of 2-D images taken from many different viewpoints of a 3-D object. This technique has been applied to provide many applications in the virtual reality, digital archives, digital museum, marketing, and entertainment. In this paper, we refer multiview images (MVIs) to the image set of an object movie. An MVI can be acquired in less than one hour by using automatic shooting equipment [2], [3]. This work used the AutoQTVR system developed by Texnai Inc. to capture the MVI.

MVI segmentation removing the background from the interesting object is necessary to integrate an MVI into the virtual world and to obtain satisfied rendering results. However, MVI segmentation is known to be a tedious and expensive task compared to the acquisitin of the MVI as mentioned above. In our experience, segmenting the images manually would take more than 30 man hours because an MVI generally contains hundreds of images. Additionally, the MVI segmentation task can become very time-consuming and burdensome for stereo object movies [4].

Yielding two distinct foreground and background color distributions can obviously mitigate the difficulty of MVI segmentation. Blue-screen and green-screen matting have been widely adopted in movie production to achieve this purpose. However, a black screen is preferable to capture the MVI to prevent the object from reflecting the blue or green light, particularly in the domain of digital archives and digital museums. A black screen frequently results in ambiguously shadowed regions that can significantly increase the difficulty.

Therefore, even a patient expert will become tired of the segmentation task if the usability of the designed MVI segmentation method should be examined in terms of computational expense, accuracy of the segmentation result and amount of user intervention. This work devises a new image segmentation

Manuscript received August 14, 2006; revised May 1, 2007. This work was supported in part by the National Science Council, Taiwan, R.O.C., under Grants NSC 95-2422-H-002-020 and NSC 95-2752-E-002-007-PAE, and in part by the Excellent Research Projects of National Taiwan University under Grant 95R0062-AE00-02. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Zhigang (Zeke) Fan.

Y.-P. Tsai is with the Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan, R.O.C., and also with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C.

C.-H. Ko is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, R.O.C.

Y.-P. Hung is with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan, R.O.C.; the Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan, R.O.C.; and also with the Institute of Information Science, Academia Sinica, Taipei, Taiwan, R.O.C. (e-mail: hung@csie.ntu.edu.tw).

Z.-C. Shih is with the Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan, R.O.C.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2007.904465

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

Fig. 1. Part of an equi-tilt set before applying the MVI segmentation method. Except for the leftmost image in the figure, the remainder of the images in this paper are cropped in order to show more examples. (a) $I_{0^\circ;20^\circ}$; (b) $I_{10^\circ;20^\circ}$; (c) $I_{20^\circ;20^\circ}$; (d) $I_{30^\circ;20^\circ}$; (e) $I_{40^\circ;20^\circ}$.

method that satisfies all the requirements to help the user obtain a quality MVI segmentation result in less than one man hour.

The notation of the MVI (which is a collection of images) used throughout the paper is defined as follows. Let $I_{\theta,\phi}$ denote the image taken at pan angle θ and tilt angle ϕ. An equi-tilt set $\mathcal{O}_\phi$ is defined as a subset of the images in an MVI captured at the same tilt angle ϕ, i.e.,

$$\mathcal{O}_\phi = \{I_{\theta,\phi} \mid 0 \le \theta \le 2\pi\}. \tag{1}$$

Finally, an MVI $\mathcal{O}$ is defined as

$$\mathcal{O} = \left\{\mathcal{O}_\phi \mid -\frac{\pi}{2} \le \phi \le \frac{\pi}{2}\right\} \tag{2}$$

$$= \left\{I_{\theta,\phi} \mid 0 \le \theta \le 2\pi, -\frac{\pi}{2} \le \phi \le \frac{\pi}{2}\right\}. \tag{3}$$

Fig. 1 shows a portion of an equi-tilt set that are contained in the MVI of the pottery owl.

The remainder of this paper is organized as follows. Section II discusses related work. Section III presents the overview of the proposed method. The proposed method consists of two main parts, 1) the automatic initial segmentation and 2) the rectification of the segmentation errors with the learned shape prior. The first part is introduced in Section IV, while the second part is presented in Section V. The experimental results are presented in Section VI, together with the discussion of the proposed method. Finally, conclusions are drawn in Section VII.

II. Related Works

To our knowledge, MVI segmentation is currently performed entirely by the artists. These experts mainly manipulate some industrial interactive tools (e.g., magic wand and intelligent scissors from Adobe Photoshop [5]) to remove the background of each image individually. The work flow does not utilize any information between images captured in neighboring viewing directions, and consequently is very expensive. Unfortunately, background removal in the MVI has not been widely investigated, so MVI segmentation is an obstacle to the spreading of image-based objects. This section describes the state of the art in interactive background removal tools. Video object segmentation methods related to MVI segmentation are then surveyed and discussed.

Interactive background removal tools have been developed for many years because of their practical importance. Such tools include magic wand [5], intelligent scissors [5]–[8], Bayesian matting [9], graph cut based image segmentation [10]–[13], and interactive matting based on belief propagation [14]. The color information (e.g., foreground and background color model) and contrast information (e.g., gradient and edge strength) are usually exploited to achieve the goal. The most popular research direction among the above mentioned methods is probably graph cut based image segmentation. After a user manually gives some foreground and background hard constraints on the image, the remaining of the image are automatically classified as the foreground or background immediately. These approaches are often quite successful for the single image segmentation, but hard to apply to the MVI segmentation due to the endless drudgery of manually specifying hard constraints on all images of the MVI individually.

Video object segmentation involves extracting objects from an image sequence. Both automatic [15]–[17] and semi-automatic [18]–[20] methods have been proposed. Since automatic video object segmentation might be problematic, semi-automatic approach, which allows the user to guide the segmentation algorithm, can be applied to obtain robust and accurate results. The motion field between the neighboring images can be estimated and utilized to simplify the video object segmentation. Because of the MVI is a specific type of video sources, MVI segmentation can also benefit from the video object segmentation technique. To alleviate the difficulty of individually cutting the object out of the background, perhaps the most intuitive approach is to utilize the motion field to propagating the segmentation results from some representative images to their neighboring images. If user intervention is allowed, then it can also be propagated. The information propagation scheme has been frequently used to solve many video manipulation problems to achieve different applications [10], [21], [22]. All the video object segmentation methods, based on the information propagation scheme, can provide satisfactory results in some cases, but lack elegance, because their performance mostly depends on the accuracy and robustness of the motion estimator. For constant intensity regions, these methods would fail due to the failure of the motion estimator.

Goldlücke and Magnor have addressed the issue of the problems of the 3-D reconstruction and background separation [23]. However, instead of reconstructing the volumetric representation of the object that is independent on any image plane, their approach computes the depth map with respect to an image plane. Additionally, for each image where the foreground layer needs to be determined, their approach also requires a background image taken at the same viewpoint. Since the foreground

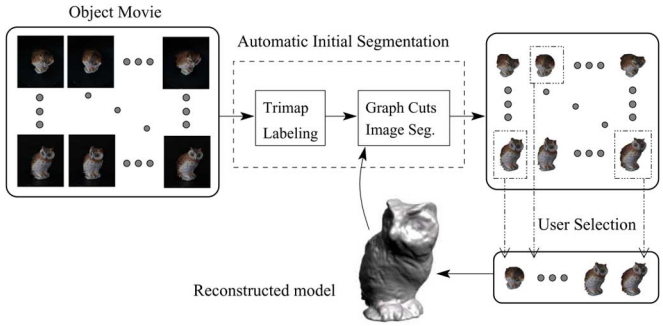


Fig. 2. Process flowchart of the proposed method.

objects should be taken away before these required background images can be captured, the practicability of their approach is significantly reduced.

III. Overview of Our Approach

The proposed approach aims to let every single image segmentation, rather than only those in neighboring viewing directions, benefit from the segmentation results of the images captured in all possible viewing directions. Besides the problem of computing a reliable motion field, one more shortcoming of the information propagation scheme is that the information can only be propagated from neighboring images, because of the error accumulation problem which is hard to avoid when computing the motion field. The proposed approach overcomes this limitation with the help of the reconstructed 3-D model, and some preliminary results have been shown in [24]. A quality motion field can also be computed between any pair of the neighboring images after the 3-D object is reconstructed.

Fig. 2 illustrates the process flowchart of the proposed approach. Given an MVI with the intrinsic and extrinsic parameters of the camera calibrated for all views [25], the proposed method starts with the automatic initial segmentation, which aims to provide some tentative segmentation results based entirely on the color and contrast information. To take the shape prior into account, the user is required to selectd a subset of acceptably segmented images. The 3-D shape is then generated from these selected images. The reconstructed 3-D model can be used to infer the shape of the object in any given 3-D configuration of a view. For each image of the MVI, a quality segmentation result can be computed by incorporating the inference of shape of the object into the segmentation algorithm, along with the original color and contrast information. The main advantage of the approach is that each time the user gives some intervention to a part of the MVI, the influence can be propagated to the whole MVI segmentation problem. Thus, if the user is still not satisfied with the MVI segmentation result, then interactive background removal tools can be utilized to refine some problematic images. This procedure can be repeated in order to refine the MVI segmentation result further.

Notably, to apply our method, camera parameters are indeed required for 3-D reconstruction. The reconstructed 3-D model may be inaccurate due to calibration error, which may then introduce errors when the shape priors are extracted from the inaccurate 3-D model. However, the final 2-D image segmentation results are not very sensitive to small errors in the shape priors, as long as the errors are within a few pixels. In our experiments, we used the method described in [25] to estimate the camera parameters, and the calibration errors are less than 3 pixels, in general.

Since the shape prior of the object is expressed by using a volumetric representation in this approach, a reliable 3-D reconstruction method is desired. The volumetric graph cuts proposed by Vogiatzis *et al.* [26] are adopted in this case. Moreover, a discrete medial axis constraint is introduced to alleviate the protrusion flattening problem in the original volumetric graph cuts algorithm.

IV. Automatic Initial Segmentation

The automatic initial segmentation presented is inspired by the graph cut image segmentation [10]. Because the fully automatic image segmentation could be very challenging, this work is not interested in successfully segmenting all the images here; instead, the aim is simply to obtain a collection of acceptably segmented images. These acceptably segmented images are then mimed for the knowledge that can be used to tackle the problem of automatically segmenting all the other difficult images in the next run. In practice, the user can specify this set of acceptably segmented images.

A. Graph Cut Image Segmentation

The background removal tool proposed by Boykov and Jolly [10] on which our MVI segmentation method is built, is described here. Graph cut image segmentation requires the user to interactively mark some pixels as being inside the foreground objects, and others as a part of the background scene. The two disjoint sets of marked pixels serve as the foreground and background hard constraints, respectively. All the other pixels are considered to be unknown, and then they can be classified into the foreground or background by Markov random field (MRF) optimization.

Each candidate segmentation is associated with an energy that considers the following properties. For each foreground pixel of the candidate segmentation, a penalty is given to reflect on whether its color fits into the foreground model. The model can be learned from the foreground pixels marked by the user. A penalty is similarly given to each background pixel based on the similarity of its color to the background model. Next, the algorithm penalizes every pair of the adjacent pixels where one is inside the foreground and the other is outside according to how likely a boundary is probable to appear between the adjacent pixels. A small penalty is generally given for the adjacent pixels that have a large difference in their colors. The algorithm determines the optimal segmentation by finding the global minimum among all segmentations that meet the specified hard constraints.

B. Trimap Labeling

Therefore, a trimap consisting of labels drawn from $\{\mathfrak{F}, \mathfrak{B}, \mathfrak{U}\}$[1] is required to activate the graph cut image segmentation. The pixels with labels $\mathfrak{F}$ and $\mathfrak{B}$ correspond to the foreground and background hard constraints, respectively. Intuitively, automatic segmentation can be achieved if the trimap can be generated automatically. To obtain the tentative

[1]Abbreviations of foreground, background, and unknown, respectively.

segmentation results here, a labeling method is presented to generate automatically the trimap for each image of the MVI.

From our observation, MVI has three basic characteristics which can help the method generate the trimap.

1) When an equi-tilt set of the MVI is captured, a large proportion of the background scene is static.
2) Only one interesting object is presented in every image of the MVI.
3) The foreground and background color distributions are distinct in most cases.

The trimap labeling method comprises $\mathfrak{B}$-labeling and $\mathfrak{F}$-labeling. Each equi-tilt set of the MVI is processed individually by the trimap labeling method. Given an equi-tilt set $\mathcal{O}_\phi$, the trimap of each image in $\mathcal{O}_\phi$ is initialized to $\mathfrak{U}$. During the $\mathfrak{B}$-labeling, pixels are examined to be labeled as $\mathfrak{B}$ based on the color difference. During the $\mathfrak{F}$-labeling, all pixels that are still labeled as $\mathfrak{U}$ are examined to be labeled $\mathfrak{F}$ based on the background model.

1) $\mathfrak{B}$-Labeling: By the first characteristic, if the color of a pixel varies barely throughout the equi-tilt set $\mathcal{O}_\phi$, then the pixel should be the background and labeled $\mathfrak{B}$. Since an equi-tilt set $\mathbf{O}_\phi$ can be treated as a short video sequence, a pixel $\mathfrak{B}$ is labeled by examining its color difference compared with the corresponding pixels in both directions of the video sequence. The pixels with a small color variation are labeled $\mathfrak{B}$.To relieve the camera noises and consider the color changes caused by the lighting, the zero-mean normalized cross correlation (ZNCC) is adopted to measure the color difference.

Most of the $\mathfrak{B}$ pixels are exactly within the background as shown in Fig. 3, but there are exceptions, such as the pixels of a uniform colored patch of the object. The concept of label consistency is then introduced. If the pixels at the same image position do not have the same label throughout the whole sequence, then they are relabeled as $\mathfrak{U}$. Finally, by the second characteristic of the MVI, mathematical morphology is applied to filter out the remained noises such that only one $\mathfrak{U}$ region exists, surrounded by the $\mathfrak{B}$ region. Notably, all the images in $\mathcal{O}_\phi$ until now had the same trimap consisting only the $\mathfrak{B}$ and $\mathfrak{U}$ labels. Fig. 4 shows an example of such a global background mask.

2) $\mathfrak{F}$-Labeling: By the third characteristic of the MVI, each pixel whose color differs widely from the background model can be labeled $\mathfrak{F}$. To learn the background model of a given image, the $\mathfrak{B}$ pixels that are reasonably close to the boundary between the $\mathfrak{B}$ and $\mathfrak{U}$ regions are collected and clustered by using K-means. Let $\boldsymbol{\mu}^i_{\theta,\phi}$ denote the mean color of the ith cluster for image $I_{\theta,\phi}$. Each pixel $\mathbf{p}$ with the label $\mathfrak{U}$ in the image $I_{\theta,\phi}$ is examined and labeled $\mathfrak{F}$ if

$$\min_{\forall i} \left| I_{\theta,\phi}(\mathbf{p}) - \boldsymbol{\mu}^i_{\theta,\phi} \right|_2 > \omega_{\mathfrak{F}} \tag{4}$$

where $\omega_{\mathfrak{F}}$ is a strict threshold to ensure that only the pixels that differ widely from the background model are labeled $\mathfrak{F}$.

Fig. 3 shows the result of the trimap labeling. The trimap of each image is used to activate the graph cut image segmentation.

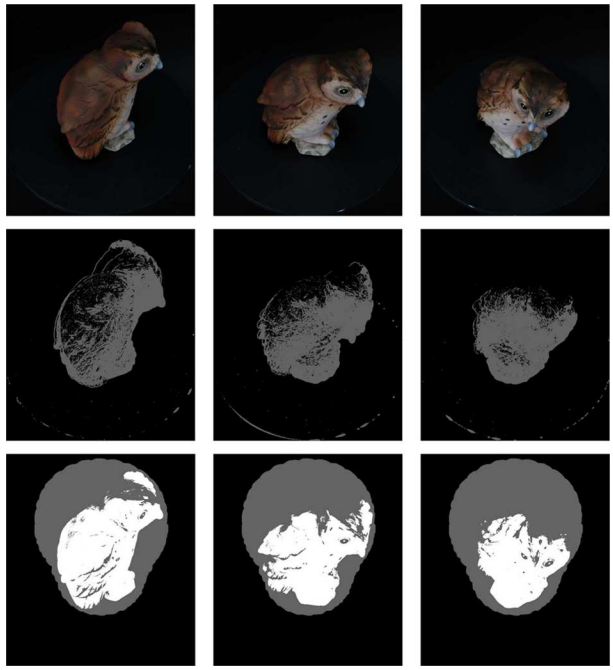
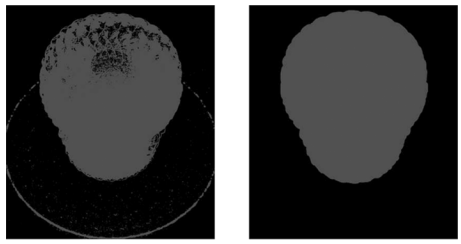


Fig. 3. Top row shows a portion of the input image sequence taken from an equi-tilt set of the pottery owl MVI. For all the images in the middle and bottom rows, the black pixels correspond to the classified background regions. The foreground regions are colored white, and the unknown regions are colored gray. The middle row shows the corresponding result during the $\mathfrak{B}$-labeling for each image. Notably, to filter out the incorrectly classified pixels and obtain the global background mask used during $\mathfrak{F}$-labeling, label consistency and mathematical morphology are used as shown in Fig. 4. Finally, the bottom rows shows the generated trimap for each image that is used to activate the graph cut image segmentation.



(a) (b)

Fig. 4. (a) Result including the label consistency concept is included; (b) global background mask obtained by applying the mathematical morphology in (a).

V. Segmentation With Shape Priors

The shape prior of the object used in our method is expressed by a volumetric 3-D model. The problem of reconstructing a volumetric 3-D model from multiple calibrated images has been widely investigated in the last decade [26]–[29]. Besides the camera calibration, these algorithms also require the silhouettes of the object in all the images. However, obtaining these silhouettes is exactly what we want to solve. The proposed method avoids this contradiction based on the observation that a subset of the MVI is sufficient for the 3-D reconstruction. Sufficient number of images that have satisfactory segmentations after the automatic initial segmentation. The user is then required to select a subset of acceptably segmented images to accomplish the 3-D reconstruction. Vogiatzis *et al.* recently proposed a graph cut-based method, called volumetric graph cuts [26], to solve the reconstruction problem. This work adopts Vogiatzis *et al.*'s algorithm to learn the shape prior.

A. Volumetric Graph Cuts

The volumetric 3-D reconstruction problem can be expressed as a labeling problem, which involves deciding whether a given voxel within the volume is inside or outside the surface of the object. The idea of the volumetric graph cuts is as follows. The true surface is assumed to be between a given base surface and a parallel inner surface. The base surface is an approximation of the true surface, encloses the true surface. In practice, the base surface can be obtained from the visual hull [27]. Each candidate surface under this assumption is then scored mainly according to whether the points on the surface are photo-consistent. The algorithm finds the optimal surface by solving the minimum cut of a corresponding weighted graph. Specifically, for each voxel $\mathbf{x} \in \mathbb{R}^3$, let $\rho(\mathbf{x})$ be the photo-consistency score of $\mathbf{x}$, where a lower value represents a better photo-consistency. For a candidate surface S, let $V(S)$ be the volume between S and the base surface. Each candidate surface is associated with the energy function consisting of the integral of the photo-consistency score $\rho(\mathbf{x})$ on the surface and the size of the volume $V(S)$. The true surface S^* is determined by finding the global minimum of the energy function $E(S)$ among all candidate surfaces S

$$S^* = \arg\min_S E(S) \tag{5}$$

where

$$E(S) = \int\int_S \rho(\mathbf{x}) \mathrm{d}A + \lambda \int\int\int_{V(S)} \mathrm{d}V. \tag{6}$$

In (6), the first integral tends toward a photo-consistent surface, while the second, called the *ballooning term*, prefers a fatter reconstructed model. The reason for preferring a fatter model is that finding the global minimum can result in a trend to remove the protrusive parts of the object. The goal of the ballooning term is to counterbalance the protrusion flattening problem. Vogiatzis *et al.* [26] describes the detailed formulation and graph construction.

B. Discrete Medial Axis Constraint

One problem with the volumetric graph cuts is that the parameter λ in (6) has to be chosen through trial and error in order to obtain a satisfactory result. Furthermore, the ballooning term could lead to a tug-of-war between the original protrusion flattening problem and the following concavity filling problem, where the concavities presented in the object are filled. For some objects, a befitting ballooning term still can not be found out to obtain a correctly reconstructed object even after an exhaustive search of the parameter λ. The phenomenon is also demonstrated in one of our experiments. The discrete medial axis constraint can be utilized to alleviate these difficulties is one of our contributions.

1) Energy Function Analysis: As is well known, solving the two terminals min-cut problem is equivalent to finding the maximum a posteriori (MAP) estimation of a MRF with two labels. The graph cut energy minimization, such as that used in the volumetric graph cuts, is widely adopted in many computer vision applications. Similar to most of the energy functions that can be minimized by the graph cut, (6) also includes the data and boundary properties. Let $\mathcal{V}$ be the set of voxels within the base surface. Let $\mathcal{N}$ be a neighborhood system defined for $\mathcal{V}$, which containing the set of all pairs of neighboring voxels. Let $\mathcal{L} = \{l_p \mid \forall\, p \in \mathcal{V}\}$ be a family of random variables defined on the set $\mathcal{V}$, in which each variable takes a label l_p from $\{\mathfrak{I}, \mathfrak{O}\}$.[2] Given a candidate surface S, a corresponding random field $\mathcal{L}$ is uniquely defined such that for any voxel p in $\mathcal{V}$

$$l_p = \begin{cases} \mathfrak{I}, & \text{if } p \text{ is within the surface } S \\ \mathfrak{O}, & \text{otherwise.} \end{cases} \tag{7}$$

In the discrete case, it can be easily proven that the energy function $E(S)$ in (6) associated with a candidate surface S can be rewritten as $E(\mathcal{L})$ which corresponds to the joint of data and boundary properties of a random field $\mathcal{L}$

$$E(\mathcal{L}) = \sum_{p \in \mathcal{V}} D(p) + \sum_{(p,q) \in \mathcal{N}} B(p,q) \tag{8}$$

where

$$D(p) = \lambda \cdot \delta(l_p, \mathfrak{I}) \tag{9}$$

$$B(p,q) = \frac{\rho(p) + \rho(q)}{2} \cdot \delta(l_p, l_q) \tag{10}$$

and

$$\delta(A,B) = \begin{cases} 1, & \text{if } A \neq B \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

Here, $D(p)$ is the penalty according to how well the voxel p fits into the given label l_p, while $B(p,q)$ indicates whether the surface is likely to pass through the edge between p and q. Additionally, $B(p,q)$ can maintain the smoothness prior such that the physical property in the neighborhood of the space offers some coherence and does not change abruptly [30].

The choice of $B(p,q)$ presents no serious difficulties unlike the choice of $D(p)$. To counterbalance the protrusion flattening problem, a simple constant penalty λ in (9) is chosen to penalize all voxels that are not inside the surface. But to achieve better performance, the definition of $D(p)$ should consider the likelihood that the voxel p is inside or outside the surface with respect to the available observations. Unfortunately, until now, it is still not clear on how to compute a good estimate of the likelihood based on the available observations. Here, we present a new definition of $D(p)$ based on the medial axis of the object, which has been proven to work well as shown in the experiments.

2) Imposing the DMA Constraint: The medial axis of the 3-D object is defined as the centers of all maximal spheres in the object that touch the shell of the object at two or more points. In practice, the medial axis is represented by a set of discrete voxels interior to the 3-D object, called discrete medial axis (DMA). The DMA of a volumetric model can be obtained by analyzing the 3-D distance field, which is computed by the distance transformation method. A good overview of these methods has been provided by Cuisenaire [31]. The local maxima in the 3-D distance field are examined to serve as the DMA. Because undesired branches might exist, which is considered to be meaningless, only the large enough connected components of the voxels in the DMA are retained.

[2]Abbreviations of being inside and outside the surface, respectively.

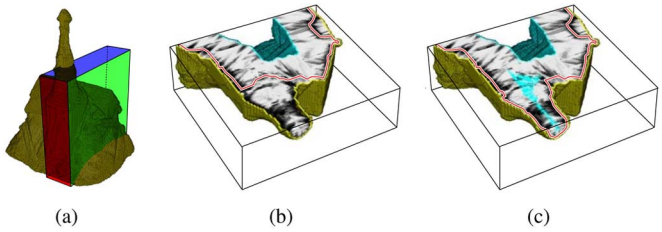


Fig. 5. Visualization and comparison of the 3-D reconstruction algorithm. Both (b) and (c) are taken from a cross-section of the visual hull for the toy house, which is shown in (a). The golden voxels correspond to the base surface in all three images. The cyan voxels denote the inner surface, which is parallel to the base surface. Additionally, the voxels in $\mathcal{V}_A$ are also colored cyan in (c). The photo-consistency scores between the base and inner surfaces are shown, where the darker region indicates a better photo-consistency. Additionally, the line within the base and inner surfaces represents the reconstructed surface of the object. In (b), without the DMA constraint, although the reconstructed surface passes through the worse photo-consistency regions, the integral of the energy on the entire surface is lower. Consequently, the protrusive part (i.e., the tower of the house) is flattened incorrectly. The image in (c) shows the correctly reconstructed surface for the same portion of the object with the DMA constraint.

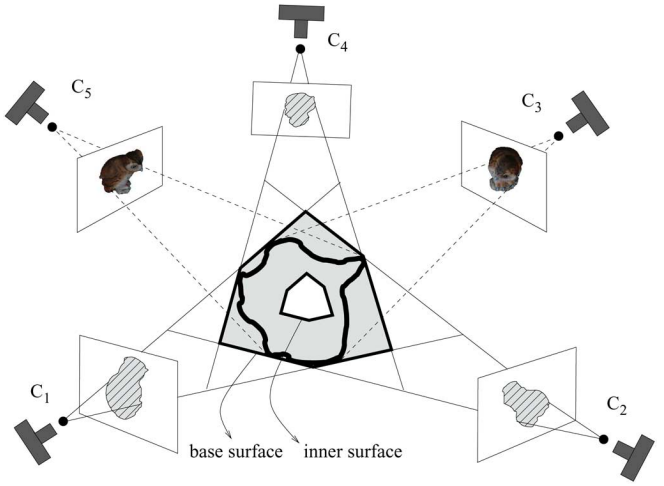


Fig. 6. C_1, C_2, and C_4 denote the views adopted to built the visual hull. Notably, the true surface of the object is assumed to be between the base and inner surfaces. Although the segmentation results of C_3 and C_5 are poor, they can be improved by incorporating the projection of the reconstructed model into the graph cut image segmentation algorithm.

Compared to the original volumetric graph cuts, we first compute the DMA of the base surface, which is assumed to be an adequate approximation of the DMA of the true surface. The DMA itself is imposed as the hard constraint of the object such that the voxels in the DMA are enforced to be inside the object, while the voxels in the neighborhood of the DMA act as the soft constraint that are very probable to be inside the object.

Specifically, let $\mathcal{V}_A$ be the set of voxels in the DMA. Let d_p be the minimum distance from the voxel p to its nearest voxel in $\mathcal{V}_A$. Computing the minimum distances for all voxels can be accelerated by using the distance transformation method to obtain an approximate solution. For each voxel within the base surface, the possibility of being inside the true surface is considered to be inversely proportional to the minimum distance. Thus, we define the new data property $D_A(p)$, into which the DMA constraint has been embedded

$$D_A(p) = \begin{cases} \infty \cdot \delta(f_p, \mathfrak{I}), & \forall p \in \mathcal{V}_A \quad (12) \\ \lambda \cdot \exp\left(\dfrac{-{d_p}^2}{2\sigma^2}\right) \cdot \delta(f_p, \mathfrak{I}), & \text{otherwise.} \end{cases} \quad (13)$$

Here, (12) guarantees that the voxels in $\mathcal{V}_A$ are always labeled as being inside the surface. Additionally, (13) encourages the voxels in the neighborhood of the DMA to be labeled as being inside the surface. Notably, the parameter λ adjusts the strength of the soft constraint, while σ^2 controls the influenced range. The energy function with the new data property $D_A(p)$ is globally minimized by using the graph cut technique similar to [26]. Fig. 5 illustrates the benefit of the DMA constraint. The visualization of the photo-consistency scores is also provided.

C. Segmentation Refinement

Besides the color and contrast information, a good inference of the shape available for any possible view of the object can provide the favorable information on solving the MVI segmentation problem. Because the camera is calibrated for all images in the MVI, a good shape prior of the object can be obtained to rectify the segmentation errors in some problematic views by projecting the reconstructed 3-D model. Fig. 6 illustrates the idea of the segmentation refinement. For each image with the discontented segmentation result, the projection of the reconstructed 3-D model under the same viewpoint is integrated to serve as the foreground hard constraints, together with the previously generated trimap. The graph cut image segmentation is then applied again to obtain the satisfied segmentation result.

Significantly, the photo-consistent reconstruction is mandatory to obtain a good shape. The visual hull can only represent an approximate geometry of the object, and tends to be fatter than the real object, regardless of whether the object is convex or concave. This characteristic of the visual hull could be more obvious when the number of images available to be used is limited. Consequently, the projection of the visual hull might introduce unreliable foreground hard constraints in the segmentation refinement. Fig. 6 also illustrates the problem when photo-consistent reconstruction is not used. Here, directly using the projection of the base surface on C_3 and C_5 imposes incorrect foreground hard constraints, and lead to failed segmentation results.

VI. Experiments

In our experiments, Each MVI consists of 360 images from ten equi-tilt sets $\mathcal{O}_{(0\pi/18)}, \mathcal{O}_{(1\pi/18)}, \ldots, \mathcal{O}_{(9\pi/18)}$. Each equi-tilt sets had 36 images captured equally from pan angle 0 to pan angle 2π with the image size 3000×2000 pixels. In all the experiments, because the lens distortion occurred in an area far from the center of the image, and the object was mostly located in the center of the image, each image is cropped to about 1000×1000 pixels before evaluating our MVI segmentation method. The experiments were performed on a 2.4-GHz Pentium 4 desktop with 1-GB memory.

The remainder of the experiments section is arranged as follows. First, the results of the automatic initial segmentation are shown. The reconstructed 3-D models, from which the shape prior can be extracted, are then shown. Following this,

Fig. 7. Results of the automatic initial segmentation corresponding to the image sequence shown in Fig. 3. The two images on the left show the segmentation results that should be selected for the 3-D reconstruction, while the other shows the segmentation result that should be excluded and refined in the next run. The red circles denote the noticeable segmentation errors in the image.

we demonstrate how to rectify the segmentation errors existing in some problematic images using the obtained shape prior.

A. Initial Segmentation Results

To reduce the response time to the user, the automatic initial segmentation can be carried out on the downsized MVI. After obtaining the initial segmentation results, the set of segmented images chosen by the user was then resized to the original image size to generate the base surface used in the 3-D reconstruction. When finding the optimal surface within the base surface, besides the selected images, all the other images in the MVI can also be considered when computing the photo-consistency scores. Additionally, the automatic initial segmentation does not need to be applied on all the equi-tilt sets of the MVI. Experimental results shows that about 3 or 4 equi-tilt sets captured in the relatively small tilt angles can yield enough satisfactory segmentation results for the 3-D reconstruction job.

First, the automatic initial segmentation was applied to the pottery owl MVI. Fig. 7 shows the results of the automatic initial segmentation for the pottery owl with respect to the image sequence as shown in Fig. 3. Because of the low contrast boundaries of the pottery owl, the black screen and the shadows caused by the lighting, automatic foreground extraction of the whole MVI could be a demanding challenge when applying methods based on color and contrast information alone. However, since different geometries, textures, and lightings are presented in different viewing directions of the pottery owl, the foreground can be automatically separated from the background in some images. For the other problematic images, the segmentation errors can be rectified in the next run by incorporating the learned shape prior into the segmentation process. To learn the shape prior, 36 segmented images were selected for the 3-D reconstruction of the pottery owl.

Fig. 8 shows the results of the automatic initial segmentation for a portion of the equi-tilt set in the toy house MVI. Since the tower of the house had mixed together with the black screen in some viewing directions in the photo studio arranged for capturing the MVI, the tower was difficult to separate from the background without the shape knowledge learnt from the other successfully segmented views. To rectify the segmentation errors, 48 segmented images were selected for the 3-D reconstruction of the toy house.

B. Learning Shape Prior

In the course of learning the shape prior for the object, 3-D reconstruction from a selected subset of the segmented images is carried out. The volume was discretized into $200 \times 200 \times 200$ voxels. In our implementation, the 3-D reconstruction task can generally be completed within 3 min, depending on the number of acceptably segmented images chosen by the user. The parameter λ in (6) is quite stable after including the DMA constraint, and, hence, remained constant in our experiments.

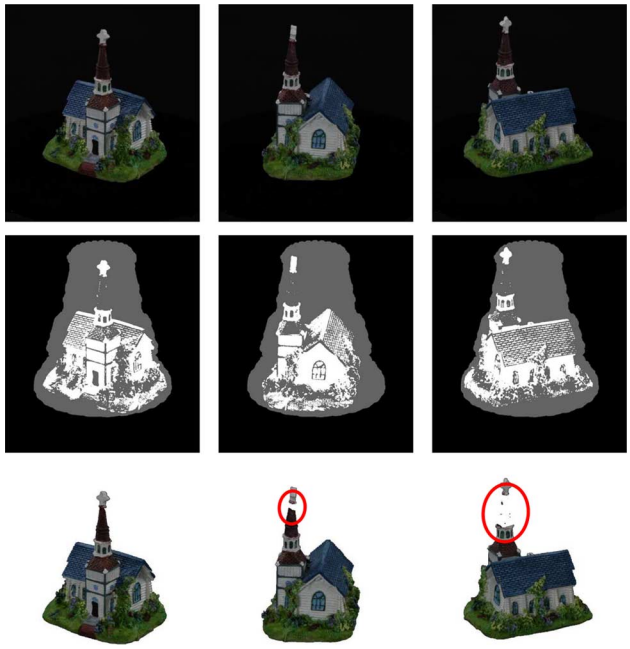


Fig. 8. Top row shows a portion of an equi-tilt set for the toy house MVI. The middle row shows the trimap labeling result for each image. Finally, the bottom row shows the results of the automatic initial segmentation. The red circles indicate the noticeable segmentation errors in each image, to be rectified in the next run.

The first experiment involved the toy house, which was also adopted to demonstrate the advantage of using the DMA constraint. The toy house was chosen deliberately because it represents a difficult 3-D reconstruction problem, due to noticeable protrusions and concavities in the object. Fig. 9 demonstrates the difficulty of reconstructing the toy house, indicated by the tug-of-war between the protrusion flattening problem and the concavity filling problem. Without the DMA constraint, even if the concavities all around the house are going to be filled, the ballooning term still cannot correctly deal with the tower even after it has been exhaustively searched. Fig. 10 shows the successfully reconstructed model of the toy house by imposing the DMA constraint to alleviate this difficulty. Notably, the algorithm can properly reconstruct both the protrusive parts, i.e., the tower and chimney of the toy house, and the concavities all around the house.

The second experiment adopted the pottery owl. Fig. 11 shows the reconstructed model. Although the ears of the pottery owl are thin and sharp, they were correctly reconstructed with the DMA constraint. Additionally, the concavities around the eyes and feet were handled properly.

C. Rectification of Segmentation Errors

This section describes the refinement of the segmentation results with the learned shape prior. Fig. 12 shows the rectification of the segmentation errors for each problematic image in Figs. 7

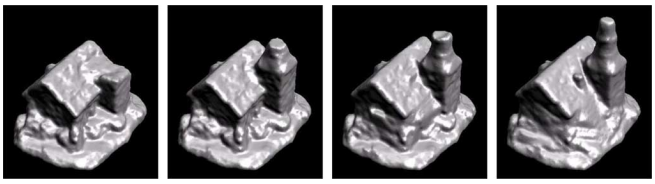


Fig. 9. Reconstructed model of the toy house by using the volumetric graph cuts algorithm without imposing the DMA constraint. The ballooning term is increased gradually from left to right. The figure indicate that reconstructing the toy house is a difficult task without the DMA constraint

.

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

Fig. 10. Image (a) shows the visual hull generated from the available silhouettes of the toy house to act as the base surface in the algorithm; (b) the DMA of the visual hull that is considered to be an approximate DMA of the toy house. Images (c) and (d) show the reconstructed model from three different viewpoints of the toy house, together with the image captured at similar viewpoint.

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

Fig. 11. (a) Visual hull of the pottery owl. (b) DMA of the visual hull. (c) Example image of the pottery owl MVI. (d) Reconstructed model of the pottery owl by using our method.

and 8, which are denoted by the red circles. Since the projection of the reconstructed model can provide a good inference of the shape for the object in each calibrated view, a robust segmentation result can be achieved even when the boundary of the object goes through the low-contrast and shadowed regions where the foreground and background color distributions can not be effectively separated. On each trimap that includes the projection of the reconstructed model, the learned shape prior provides much information about the segmentation problem that the original foreground hard constraints do not reveal.

Fig. 13 indicates that the background removal of the pottery cat MVI increases the benefit of using shape priors. Because the foreground and background color distributions are entirely mixed with each other in some difficult regions, the images are quite difficult to segment by using only the color and contrast information. Moreover, for such a troublesome MVI, segmentation errors generally appear in several consecutive images at the same time. Consequently, propagating successful segmentation results by using the motion field becomes quite unstable due to the error accumulation problem when estimating the motion field. For such a difficult object, the automatic initial segmentation might not provide enough successful segmentation results

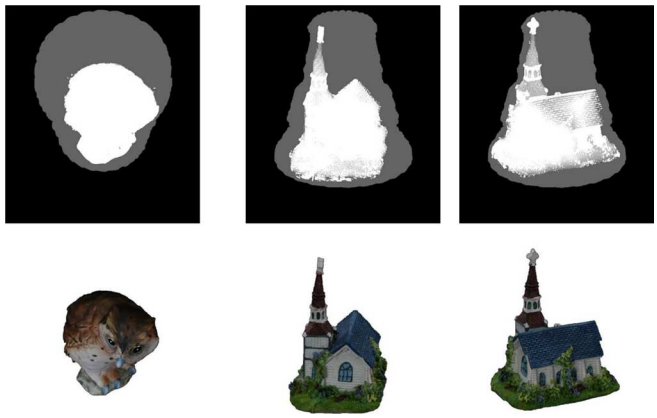


Fig. 12. Rectification of the segmentation errors for the pottery owl in Fig. 7 and the toy house in Fig. 8. Trimaps (top row): tProjection of the reconstructed model is colored white, and serves as the foreground hard constraints together with the previously generated trimap. Refinement (bottom row): Refinement of the segmentation result is shown for each image.

for the 3-D reconstruction. Here, an equi-tilt set was manually segmented by using the interactive background removal tool. Both the automatic and manual segmentation results were used to accomplish the 3-D reconstruction job. The problematic segmentation results was then refined be refined using shape priors obtained from the reconstructed 3-D model.

To measure the segmentation improvement, the proposed method was applied to the synthetic data composed of the rendering results of the 3-D model and random background noises, as depicted in Fig. 14. Since the silhouette is known in the synthetic data set, the error between the segmentation result produced by our method and the silhouette can be calculated. The Hausdroff distance was adopted to measure the segmentation errors. In our experiment, four levels of background noises were composed to the synthetic data, and 10 and 20 ground truth images were randomly selected to learn the shape prior. The results of Fig. 15 indicate that shape information is indeed critical to alleviate eliminate segmentation errors, and ensures that the segmentation method is robust to background noises. Fig. 14 shows the comparison between ground truth and the segmentation results produced by the proposed method with shape prior 2.

VII. Conclusion

The major advantage of the proposed method is it can propagate the successful segmentation results from some selected images to the whole MVI. With the new MVI segmentation method, 2-D shape is extracted from the reconstructed 3-D model and used to remove the background from the foreground object. Our work has demonstrated that significant improvement for MVI segmentation can be obtained with the proposed method.

The proposed MVI segmentation process requires only a small amount of user intervention, which is to select a subset of acceptable segmentations of the MVI after the initial segmentation process. Notice that human is much more efficient in selecting a good segmentation result than in manually delineating a precise object contour. For some very difficult objects, the automatic initial segmentation might not meet

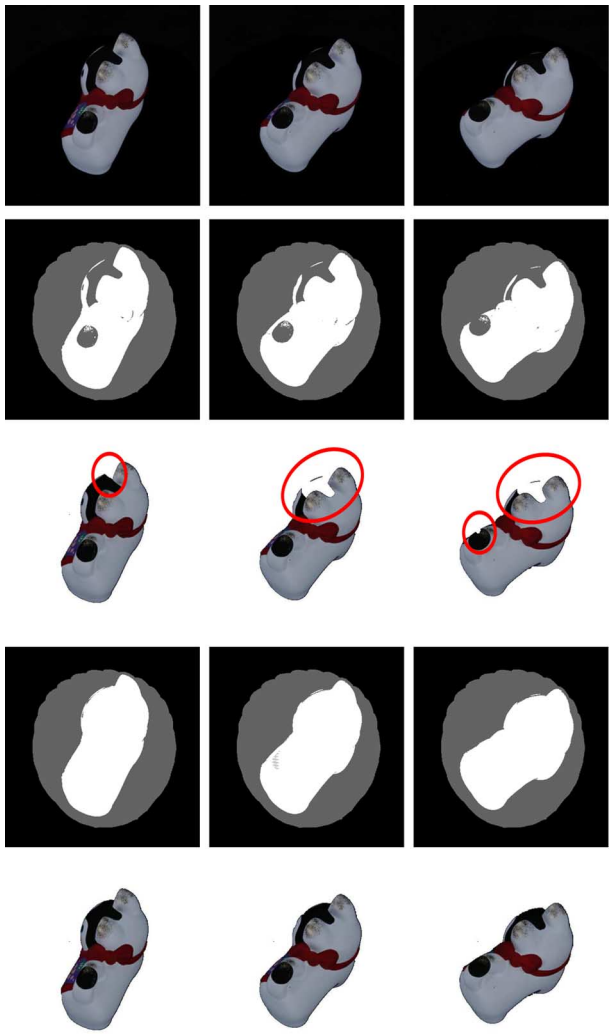


Fig. 13. First row shows three consecutive images in an equi-tilt set of the pottery cat MVI. The second row shows the result of trimap labeling. The third row shows the result of the automatic initial segmentation. In the fourth row, the projection of the reconstructed 3-D model provides the information on regions that is quite difficult to obtain by the methods based on color and contrast alone. The last row shows the refinement of the segmentation result by using shape priors.

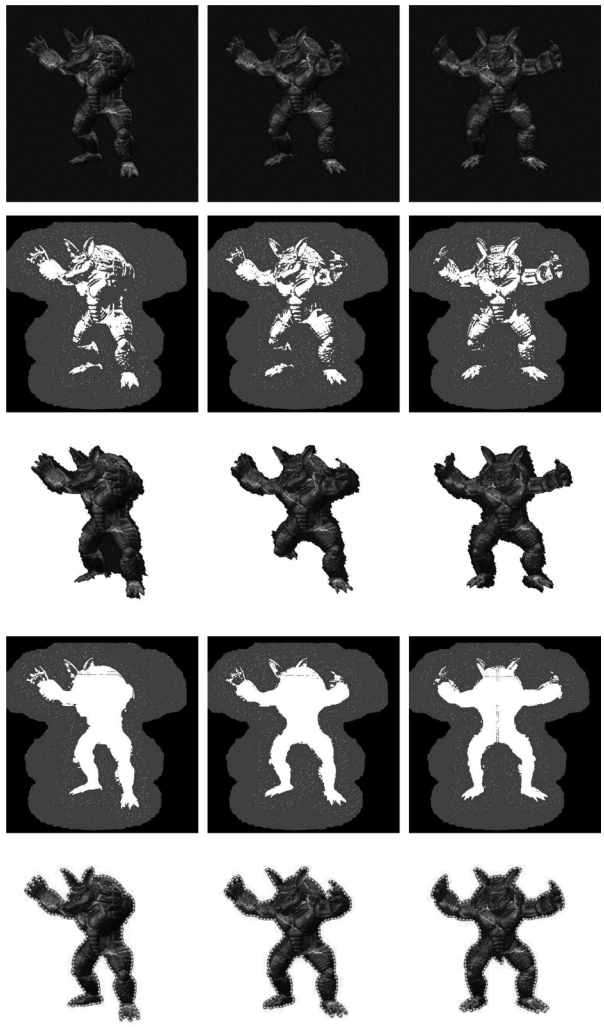


Fig. 14. First row shows three consecutive images in an equi-tilt set of the Armadillo MVI. Second row shows the result of trimap labeling. The third row shows the result of the automatic initial segmentation. In the fourth row, the projection of the reconstructed 3-D model provides the information on regions that is quite difficult to obtain by the methods based on color and contrast alone. Last row shows the refinement of the segmentation result by using shape priors, the comparison between the segmentation results produced by the proposed method and the ground truth. Red solid lines denote the contours of the ground truth, and the green dot lines denote the segmentation results produced by the proposed method.

the user's requirement. In the situation, the user can always select a subset of images and delineate the object contours for those images using some interactive image segmentation tools, e.g., the GrabCut [11], Lazy Snapping [12]. The segmentation results can then be propagated to refine the segmentation results of the remaining images using the method proposed in this paper.

While this paper only presents *binary* segmentation results, it is straightforward to apply other existing methods for alpha matting [9], [21] in order to obtain smoother boundary transition. A major limitation of the proposed method is that it cannot effectively deal with specular objects because the zero-mean normalized cross correlation (ZNCC), adopted to measure the photo consistency score, is not robust to specular reflection. Our plan is to apply to the MVI some diffuse-specular separation techniques before 3-D reconstruction. Another plan is to further reduce the user intervention by analyzing the energy of the minimum cut, after the initial segmentation, and then automatically identifying a subset of acceptable segmented images.

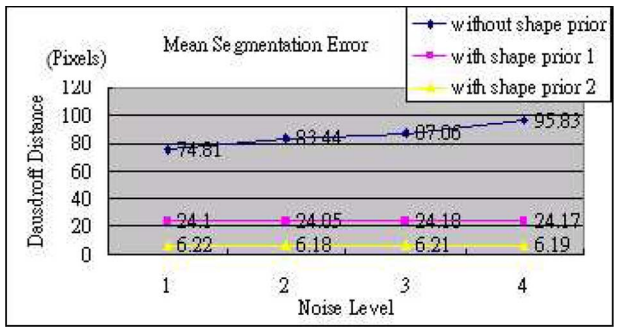


Fig. 15. Mean segmentation errors on the synthetic data. Image size is 800×600. In the experiments, the 3-D shape was reconstructed by randomly selecting ground truth images. The sShape prior 1 was learnt by using ten images, and the shape prior 2 was learnt by 20 images.

REFERENCES

[1] Apple, Inc. [Online]. Available: http://www.apple.com

[2] Kaidan, Inc. [Online]. Available: http://www.kaidan.com

[3] Texnai, Inc. [Online]. Available: http://www.texnai.co.jp

[4] C.-W. Chen, L.-W. Chan, Y.-P. Tsai, and Y.-P. Hung, "Augmented stereo panoramas," in *Proc. Asian Conf. Computer Vision*, 2006, vol. 1, pp. 41–49.

[5] Adobe, Inc. [Online]. Available: http://www.adobe.com/products/photoshop/

[6] E. N. Mortensen and W. A. Barrett, "Intelligent scissors for image composition," in *Proc. ACM SIGGRAPH*, 1995, pp. 191–198.

[7] E. N. Mortensen and W. A. Barrett, "Toboggan-based intelligent scissors with a four-parameter edge model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1999, pp. 2452–2458.

[8] Y.-P. Hung and Y.-P. Tsai, "Trail-dependent intelligent scissors based on multi-scale image segmentation," in *Proc. Asian Conf. Computer Vision*, 2002, pp. 539–544.

[9] Y.-Y. Chung, B. Curless, D. Salesin, and R. Szeliski, "A Bayesian approach to digital matting," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001, vol. 2, pp. 264–271.

[10] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images," in *Proc. IEEE Int. Conf. Computer Vision*, 2001, pp. 105–112.

[11] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph*, vol. 23, no. 3, pp. 309–314, 2004.

[12] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, "Lazy snapping," *ACM Trans. Graph*, vol. 23, no. 3, pp. 303–308, 2004.

[13] D. Freedman and T. Zhang, "Interactive graph cut based segmentation with shape priors," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 755–762.

[14] J. Wang and M. F. Cohen, "An iterative optimization approach for unified image segmentation and matting," in *Proc. IEEE Int. Conf. Computer Vision*, 2005, pp. 936–943.

[15] I. Patras, E. A. Hendriks, and R. L. Lagendijk, "Video segmentation by map labeling of watershed segments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 326–332, Mar. 2001.

[16] D. Wang, "Unsupervised video segmentation based on watersheds and temporal tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 539–546, 1998.

[17] Y. Wang, K.-F. Loe, T. Tan, and J.-K. Wu, "Spatiotemporal video segmentation based on graphical models," *IEEE Trans. Image Process.*, vol. 14, pp. 937–947, 2005.

[18] C. Gu and M.-C. Lee, "Semiautomatic segmentation and tracking of semantic video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 572–584, 1998.

[19] H. Luo and A. Eleftheriadis, "An interactive authoring system for video object segmentation and annotation," *Signal Process.: Image Commun.*, vol. 17, no. 7, pp. 559–572, Aug. 2002.

[20] B. Marcotegui, P. Correia, F. Marqués, R. Mech, R. Rosa, M. Wollborn, and F. Zanoguera, "A video object generation tool allowing friendly user interaction," in *Proc. IEEE Int. Conf. Image Processing*, 1999, vol. 2, pp. 391–395.

[21] Y.-Y. Chuang, A. Agarwala, B. Curless, D. Salesin, and R. Szeliski, "Video matting of complex scenes," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 243–248, 2002.

[22] Y.-P. Tsai, Y.-P. Hung, Z.-C. Shih, J.-J. Su, and S.-R. Tsai, "Background removal system for object movies," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2004, vol. 1, pp. 608–611.

[23] B. Goldlücke and M. A. Magnor, "Joint 3-D-reconstruction and background separation in multiple views using graph cuts," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2003, vol. 1, pp. 683–688.

[24] C.-H. Ko, Y.-P. Tsai, Z.-C. Shih, and Y.-P. Hung, "A new image segmentation method for removing background of object movies by learning shape priors," in *Proc. IEEE Int. Conf. Pattern Recognition*, 2006, vol. 1, pp. 323–326.

[25] P.-H. Huang, Y.-P. Tsai, W.-Y. Lo, S.-W. Shih, C.-S. Chen, and Y.-P. Hung, "A method for calibrating a motorized object rig," in *Proc. Asian Conf. Computer Vision*, 2006, vol. 1, pp. 379–388.

[26] G. Vogiatzis, P. H. S. Torr, and R. Cipolla, "Multi-view stereo via volumetric graph-cuts," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, vol. 2, pp. 391–398.

[27] A. Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 2, pp. 150–162, Feb. 1994.

[28] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *Int. J. Comput. Vis.*, vol. 38, no. 3, pp. 199–218, 2000.

[29] G. Zeng, S. Paris, L. Quan, and F. Sillion, "Progressive surface reconstruction from images using a local prior," in *Proc. IEEE Int. Conf. Computer Vision*, 2005, pp. 1230–1237.

[30] S. Z. Li, *Markov Random Field Modeling in Computer Vision*. New York: Springer-Verlag, 1995.

[31] O. Cuisenaire, "Distance transformations: Fast algorithms and applications to medical image processing," Ph.D. dissertation, Univ. Catholique de Louvain, Louvain, Belgium, 1999.

Yu-Pao Tsai received the B.Sc. degree in computer science from the National Chengchi University, Taiwan, R.O.C., in 1993, and the M.Sc. degree in computer and information science from the National Chiao Tung University, Hsinchu, Taiwan, in 1999, where he is currently pursuing the Ph.D. degree in computer and information science.

His current research interests include image-based rendering, video object segmentation, and virtual reality.

Cheng-Hung Ko was born in Taiwan, R.O.C., in 1982. He received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, R.O.C., in 2004 and 2006, respectively.

He is currently with MediaTek, Inc., Taiwan. His research interests include image processing, pattern recognition, and computer vision.

Yi-Ping Hung received the B.S. degree in electrical engineering from the National Taiwan University, Taipei, Taiwan, R.O.C., in 1982, and the M.S. degree from the Division of Engineering, the M.S. degree from the Division of Applied Mathematics, and the Ph.D. degree from the Division of Engineering, Brown University, Providence, RI, in 1987, 1988, and 1990, respectively.

He is currently a Professor in the Graduate Institute of Networking and Multimedia and the Department of Computer Science and Information Engineering, National Taiwan University. From 1990 to 2002, he was with the Institute of Information Science, Academia Sinica, Taiwan, where he became a tenured Research Fellow in 1997 and where he is now an Adjunct Research Fellow. He served as a Deputy Director of the Institute of Information Science from 1996 to 1997.

Dr. Hung received the Young Researcher Publication Award from Academia Sinica in 1997. He has served as the program cochairs of ACCV'00 and ICAT'00, as the workshop cochair of ICCV'03, and he has been a member in the editorial board of the *International Journal of Computer Vision* since 2004. His current research interests include computer vision, pattern recognition, image processing, virtual reality, multimedia, and human-computer interaction.

Zen-Chung Shih was born on February 10, 1959, in Taipei, Taiwan, R.O.C. He received the B.S. degree in computer science from Chung-Yuan Christian University, Taiwan, in 1980, and the M.S. and Ph.D. degrees in computer science from the National Tsing Hua University, Taiwan, in 1982 and 1985, respectively.

Currently, he is a Professor in the Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan. His current research interests include procedural texture synthesis, non-photorealistic rendering, global illumination, and virtual reality.