Model-Based Vision System for Reconstructing 3D Objects from Multiple Images

Zong-Sheng Chen, Chia-Hsiang Wu, and Yung-Nien Sun*

Department of Computer Science and Information Engineering,

National Cheng Kung University, Tainan, Taiwan, R.O.C.

Abstract

In this paper, we propose a model-based 3D reconstruction method that is suitable for feature-less objects. Typical objects with very few features are usually seen in industrial applications, such as heap of stones/minerals whose 3D reconstruction is very essential in automated storage and transportation. The shape of the heap is usually similar to a cone; hence, we use a parametric cone as the basic geometric model. At the beginning, we use background subtraction to detect the silhouette of the object for each image. The silhouettes are then used to estimate the parameters of the model. We adjust the size, position, and orientation of the estimated model by minimizing the overlapping ratio of its back-projections and detected silhouettes in images. At last, the model is voxelized to facilitate redundant area removal, and surface coloring is accomplished with visibility testing. For scenes with multiple objects, occlusion problem is formulated as a registration problem and solved by iterative closet point algorithm. Experimental results show that the proposed method is usually fast and the quality of the recovered objects is satisfactory.

I. Introduction

Three-dimensional reconstruction from multiple images is a challenging problem in computer vision, especially for objects without significant features. For example, reconstruction of heaps of minerals/stones that have uniform color and texture is very essential in automated storage and transportation in industry. Although a three-dimensional laser scanner can build dense and accurate shape, it is too expensive and not popular. In this paper, we propose a method that uses low-cost image acquisition equipment, e.g. commercial digital cameras, and a generic parametric geometric model to reconstruct feature-less 3D objects.

To reconstruct heaps of minerals/stones, we select cone as the geometric model. We extract necessary information such as object silhouettes from images to estimate the model parameters. Because the object is usually not a perfect cone, the initial model is further adjusted toward an actual shape. The estimated model is back-projected onto each image and the overlapping ratio of the back-projected regions and initial object silhouettes is minimized to adjust the position, orientation, and scale of the model. The resultant model is then voxelized and refined to eliminate the redundant area. For multiple objects, occlusion problem usually occurs. We use the non-occluded images to build the initial model, project it on the occluded images to estimate the object silhouette, and optimize the model parameters to best fit the actual object.

The organization of this paper is as follows. The related studies are described in Section II. Parameter estimation of the model is explained in Section III. We refine the recovered model in Section IV. Multi-objects reconstruction is introduced in Section V. Experimental results are shown in section VI, followed by conclusions.

II. Related studies

A typical research in automatic operation systems for yard machine is proposed in [1]. The system mounts laser sensors on the arm of a reclaimer, and the distance between the pile and the machine can be detected to determine landing points and to avoid collision. In recent years, three-dimensional laser scanners were developed with laser radar concepts to detect the shape and height of the pile [2]. Because a contemporary 3D scanner has high accuracy, it is more reliable to determine the landing point automatically and to avoid collision. However, it is very expensive. In computer vision, methods have been proposed to reconstruct the 3D objects from images. Feature-based approach uses corresponding points across images to estimate the shape of the object. If the surface features are not obvious enough to be detected, structure light illumination can be applied as assistance [3-5]. Some researchers use voxel-based approach in which the object silhouette is back-projected to space to carve an initial shape such as a bounding box [6]. In [7], a group of feature points is used to estimate a parametric surface. If position and orientation of the light source is known, the brightness of an image is useful to reconstruct the 3D object [8].

III. Model fitting

Geometric Cone Model

Our goal is to build 3D model of the viewed scene using multiple images. In order to speed up the computation time and to increase the accuracy of reconstruction, we developed a generic model to fit the object, as shown in Fig. 1. There are four parameters:

*: corresponding author

slant height s, open angle ϑ, height h, and the base radius r. For a perspective projection camera model, object length appeared in images are varied in different positions, but the projected open angle of a cone is approximately invariant. We first estimate the open angle from the images, and calculate the other parameters. The novelty of the proposed method is that we combine a generic model and at least one 3D point to reconstruct the object. With the geometric model, we can estimate the shape, number, position, and size of object(s) in the imaged scene. Even for some specific situations that only a few images, e.g. 2-3 images, can be acquired from limited viewpoints so that some surface of the object has no any image information available (such as the back of object), with the proposed model-based reconstruction method, the reconstruction result is still very close to the real object. Before extracting the image information to construct the cone model, we have to do camera calibration to estimate the intrinsic and extrinsic parameters of cameras, and correct the image distortion.

Camera Calibration

The camera is mathematically modeled by its intrinsic, extrinsic, and distortion parameters. The intrinsic parameters include focal length f, principal point $c(c_x,c_y)$, and skew as well as image distortion. The extrinsic parameters describe the translation T and rotation R between camera coordinate and world coordinate systems. Distortion is used to correct image distortion. Popular camera calibration techniques include [9-12] in which [12] is used in our system.

Silhouette extraction

Object silhouette is typically done by using background subtraction, i.e. image differencing between the foreground image F_{im} and the background image B_{im}. Let P_f and P_b be points representing the same location in the foreground and the background images, respectively. If their color difference is less than a pre-defined threshold, the point is considered as a point belonging to the object. After background subtraction, there might have some noise or small non-object regions due to surface scattering or shadow of the object. If the number of pixels in a region is less than 500, it is considered as a non-object region and removed.

Open angle and top point estimation

The open angle is estimated from the object silhouette in each image. With the object silhouette point cloud, two fitted edges representing object's slant height are automatically obtained by Hough transform. Their intersection is regarded as the projection of the top of a cone-shape object, and the included angle of the edges is considered as an estimate of the open angle ϑ. The included angle can be estimated in each image, and we select the median of these values as the open angle. The 3D position of the top point is obtained by stereo reconstruction technique using the intersection points of the fitting edges in images.

Height and base radius estimation

To estimate the height and base radius of the model, we need at least one 3D point that is obtained by feature detection followed by feature matching. The original color images are transformed to be gray level images, followed by feature detection [13]. Let P_L and P_R be the sets of the feature points in the left and right images, respectively. A feature in the left image is defined as $p_L(x_L, y_L) \in P_L$, and we find correspondence of p_L using block matching with image rectification. Therefore, with the matched point pair and camera parameters, we can compute the 3D position of the feature points.

Now, we collect three kinds of data: object silhouette, 3D and 2D positions of the top point, and the open angle. We use Fig. 2 to illustrate how to estimate the height and base radius. In Fig. 2(a), $A=[X_A,Y_A,Z_A]^T$ is the top point and $B=[X_B,Y_B,Z_B]^T$ is a surface feature in the 3D space, and $a=[x_a,y_a]^T$ and $b=[x_b,y_b]^T$ are their projection points in the image, respectively, as shown in Fig. 2(b). Point C is the intersection of $\overrightarrow{AB}$ and the bottom of the cone. Because A, B and C are collinear, a, b and c will also be collinear. Point $c=[x_c,y_c]$, the projection of $C=[X_C,Y_C,Z_C]^T$, is the intersection of $\overrightarrow{ab}$ and the object silhouette. Let $C=A+t\overrightarrow{AB}$, we can solve for C by letting the projection of $A+t\overrightarrow{AB}$ be c. Since the projection point c has already been estimated, C can be computed. Therefore the height h is obtained by $\overline{AC}\times\cos\frac{\vartheta}{2}$, and the base radius r is $\overline{AC}\times\sin\frac{\vartheta}{2}$.

Simplification of initial model

In real world, many cone-shape objects have rounded top. In this situation, we have to simplify the initial cone model. This simplification begins by fitting an edge over of the top of the object silhouette, such as line p_lp_r shown in Fig. 3(a). The object center C in the image can be solved using the average of the point cloud of the object silhouette, and the slant heights, L_l and L_r, have been given by Hough transform. Top point T is the intersection of L_l and L_r. Line $\overline{TC}$ will in general have two intersections with the object silhouette. We select the upper point p, and use its neighbors to solve a fitted line L by least-squares method. Next, we back project L into 3D space. In Fig. 3(b), intersections of L and the back-projected initial model are p_l (x_l, y_l) and $p_r(x_r,y_r)$. In the camera

coordinate system, p_l and p_r are located at (x_l, y_l, f) and (x_r, y_r, f), respectively, where f is the focal length. With (x_l, y_l, f), (x_r, y_r, f), and the origin O_c of the camera coordinate system, we can define a plane. By using the normal N ($=\overrightarrow{O_c p_r} \times \overrightarrow{O_c p_l}$) of this plane, the sign of inner product of N and a surface point provides a simple and effective way to judge the surface point is above- or below-plane which are respectively positive and negative. Those unnecessary above-plane points are then removed from the model.

IV. Model Refinement

Ideally, the back-projection of the estimated model onto the images should match the object silhouette. However, mismatch usually happened due to the imaged object is not an ideal cone, which indicates the estimated model needs further adjustment.

Minimization of mismatch

We search for a scale rigid transformation that consists of seven parameters: translation $[T_x, T_y, T_z]$, scale $[S_x, S_y, S_z]$, and rotation $Ry(\theta)$ against Y axis, to minimize the mismatch of the back-projected model. We define an objective function by the overlapping ratio of the back-projected regions and the object silhouettes. The objective function could be defined by the Hausdorff distance of these regions. Hausdorff distance is defined as

$$h(A,B) = \max_{a \in A} \left\{ \min_{b \in B} \{d(a,b)\} \right\}$$

where $h(A,B)$ means the Hausdorff distance of point set A and set B; $d(a,b)$ represents the distance of a and b. Using the Hausdorff distance as the objective function needs to calculate the back-projected silhouettes. During minimization, the objective function will be executed iteratively and it will spend a lot of time to calculate the back-projected regions. Therefore, we adopt the idea of Hausdorff distance and design a simpler objective function for fast adjustment of the model. We only project the top point and base points of the model to images. In Fig. 4(a), projection of the top point are *top,* and those of base points are represented as the solid curve between *b1* and *b2*, which are determined by the leftmost and rightmost points from the projected pixels of the base points. In addition, *b3* is determined by the lowest point of the projected pixels. Then, the back-projected model is well approximated by *top*, *b1*, *b2*, and *b3*, as the dashed lines, between *top* and *b1* and *top* and *b2*, and the curve connecting *b1*, *b2*, and *b3*, as shown in Fig. 4(a). On the other hand, consider the object silhouette in Fig. 4(b) in which *C* is the center of object silhouette, *p1* and *p2* are the intersection points of the horizontal line passing through *C* and the object silhouette. The vertical line passing through *C* intersects the object silhouette at *p3*. Our goal is to minimize the distance between the object silhouette and the back-projected model. Intuitively, the distance can be defined as the summation of Euclidean distances between the object silhouette point and their nearby back-projection points. This approach needs intensive computation on back-projecting all vertices of the model and searching for the nearby points of object silhouette. The time-consuming approach can be simplified by using the distance of *p1*, *p2*, *p3* and the back-projected model. The distance $L1_k$ between *p1* and the object silhouette is calculated as the distance from this point to the line connecting *top* and *b1*. Since our target object is usually close to a cone, it is reasonable to use the point-to-line distance as the measurement of object silhouette and the back-projection. Similarly, the distance $L2_k$ between *p2* and the line connecting *top* and *b2* is calculated. $L3_k$ is the vertical distance between *p3* and *b3*. So, the objective function is defined as

$$\sum_{k=1}^{n} \left(L1_k + L2_k + L3_k \right) \tag{1}$$

where n is the number of images.

The error of the objective function is minimized by Powell's method, which is chosen because it does not need the gradient values of the objective function and the results is quite satisfactory.

Carving and coloring

Anisotropic scaling of the scaled rigid transformation estimated in the previous subsection allows the model to be deformed and no longer constrained to be a standard cone in which the standard circular cross section is deformed to be similar to an ellipse. The deformed model is partitioned into a set of small cubes, called voxels, to facilitate refinement. If the object is shifted by $[T_x, T_y, T_z]$ to be located at an object-centered coordinate system, a point $p(x_0, z_0)$ in a circular cross section of a cone model is transformed to $p'(x_0', z_0')$ by $[S_x, S_z]$ and $Ry(\theta)$

$$\begin{bmatrix} x_0' \\ z_0' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} Sx & 0 \\ 0 & Sz \end{bmatrix} \begin{bmatrix} x_0 \\ z_0 \end{bmatrix}. \tag{2}$$

For a standard circular cross section of a cone model, each point inside the model will satisfy $x^2 + y^2 < r^2$, where r is the radius of the circle. Substitute this inequality to (2), the points inside the cross section of the deformed model with be defined by.

$$\left(\frac{1}{Sx} (\cos\theta x_0' + \sin\theta z_0') \right)^2 + \left(\frac{-1}{Sz} (\sin\theta x_0' + \cos\theta z_0') \right)^2 < r^2 \tag{3}$$

Using the scaled rigidly transformed model, we build and project its voxels into the images and check whether the back-projected voxel is inside the object silhouette or not. To determine color of voxels, we

use ray-tracing to check the visibility of surface voxels in each image and determine its color accordingly. Ideally, the color of a voxel v in all visible images should be identical. In practice, factors such as the view angle, surface material, and environment lighting could cause color variations in images. Given all pixel colors in visible images, we choose the median value of them as the color of voxel v.

V. Multiple objects

The above method is suitable for scenes with a single object. As the number of objects increases, immediate problems are how to decide the number of objects and the corresponding relationship of objects among different images, as well as object occlusion. The following discussion assumes satisfactory background subtraction and the objects are not occluded in at least two images. The number of objects, *ObjectNum*, is obtained by $\max_{i=1}^{k}(num_i)$, where k is the number of images and num_i is the number of individual regions in the ith image. If $num_i < ObjectNum$ for some i, there exists occluded object(s) in the ith image. The object corresponding relationship among different images could be determined by object surface color, silhouette size, and other object characteristics. In our experiments, silhouette size is used. With the number of objects and their corresponding relationship across images, the next step is dependent on whether the occlusion image exists or not.

If there is no occlusion appeared in all images, the reconstruction is accomplished by repeating *ObjectNum* times of the previous single object algorithm. In practice, occlusion usually occurs, and it is more complicated to analyze the scene. It is quite difficult to detect object silhouette directly from the occlusion image. In this situation, our multi-object reconstruction consists of three steps as follows. First, we build initial models from images without occlusion. Next, they are projected onto the occlusion images. In general, the object silhouettes can be approximated by the projected regions. Hence, we align these regions as initial guesses of actual silhouettes into the boundary of foreground objects. In this paper, we choose the Iterative Closest Point (ICP) algorithm [14] to accomplish the registration. Then, the boundaries of the registered projected regions represent the estimated object silhouettes. Since the registered boundary pixels usually do not constitute a close boundary, so edge linking, which consists of dilation followed by thinning over the registered pixels, is performed. Finally, with the estimated object silhouettes, the recovered models are further adjusted by the Powell's method. When there are multiple objects appeared in the scene, most of the images would contain occluded objects. With the technique to estimate object silhouettes, these images are able to contribute the reconstruction. Hence, the recovered objects are more accurate and more reliable than just using relatively fewer non-occlusion images.

VI. Experimental results

All the testing data used in our experiments are 24 bits color images of size 320×240. The first example is a pile of sand, as shown in Fig. 5(a), captured in outdoor environment using a Canon G1 digital camera from three viewpoints. We compare the geometric parameters including height, base radius and volume of the real object and the reconstruction result in Table III. In this table, base radius is not a fixed value because the reconstructed model is no longer a standard cone model after scaled rigid transformation and carving. The other cone parameters, including slant heights and open angle, are dependent on height and radius, so they are not listed. The actual volume is estimated by $\frac{1}{3}\pi r^2 h$, where r is calculated by the average over the range of the radius. The estimated volume is the summation of product of the number of voxels of the model and the voxel size ($2\times2\times2$ mm^3). The 3D reconstruction results are shown in Fig. 5(b). The second example uses three objects to evaluate the usefulness in a more complicated scene with multiple objects. The objects labeled as I, II, and III, respectively, are arranged from left to right in Fig. 6(a), and the reconstruction results are shown in Fig. 6(b).

VII. Conclusion

In this paper, we reconstruct objects without significant color or texture variation from multiple images. The characteristics of the geometric model are used to formulate the reconstruction as a parameter estimation problem. Experimental results demonstrate the reconstructed model is good enough even the color or texture of the objects are very similar, and the cost of our method is much lower than using 3D laser scanner. During 3D reconstruction, only visually accessible surface will be reconstructed, no matter the reconstruction is from images or a laser scanner. If the object information is known in advance, the inaccessible surface can be compensated like the proposed method. We use a geometric cone model as the basis of our method. Using this idea, an extension that uses more kinds of geometric models or even mixture of different basic models will increase the applicability and flexibility in various situations.

Acknowledgement

This work was supported by the National Science Council of Taiwan, R.O.C. under grant NSC-93-2213-E-006-041.

References

[1] Y. Sato, M. Sugawara, Y. Tagawa, et al. "Automatic operation system for yard ma-

chines," *Kawasaki Steel Technical Report*, no 13, Set, 1985.

[2] C. Chio, K. Lee, K. Shin, K. S. Hong, and H. Ahn. "Automatic Landing Method of a Reclaimer on the Stockpile," *IEEE Trans. Syst., Man, Cybern. C*, vol. 29, no 1, pp. 308-314, Feb. 1999.

[3] Y. Xu, C. Xu, Y. Tian, S. Ma; M. Luo. "3D face image acquisition and reconstruction system," in *Proc. IEEE Conf. Instrum. Meas. Technol.*, vol. 2, pp. 18-21, May. 1998.

[4] H. Kawasaki and R. Furukawa, "Entire model acquisition system using handheld 3D digitizer," in *Proc. 2nd Int. Conf. 3D Data Processing, Visualization, and Transmission.*, pp. 6–9, Sep. 2004.

[5] V. Sequeira, J.G. M. Goncalves, M. I. Ribeiro, "3D reconstruction of indoor environments," in *Proc. Int. Conf. Image Processing.*, vol. 1, pp. 16–19, Sep. 1996.

[6] A. Y. Mülayim, U. Yilmaz, and V. Atalay, "Silhouette-based 3-D model reconstruction from multiple images," *IEEE Trans. Syst., Man, Cybern. B*, vol. 33, no. 4, pp. 582–591, Aug. 2003.

[7] G.-Q. Wei and G. Hirzinger, "Parametric shape-from-shading by radial basis functions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 4, pp. 353–365, Apr. 1997.

[8] H. Maitre and W. Luo, "Using models to improve stereo reconstruction," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, no. 2, pp. 269–277, Feb. 1992.

[9] Y. I. Abdel-Aziz, H. M. Karara, "Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry," in *Proc. Proceedings of the Symposium on Close-Range Photogrammetry*, 1971, pp. 1–18.

[10] R.Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. Robotics and Automation*, vol. 3, no. 4, pp. 323–344, Aug. 1987.

[11] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Proc. Int. Conf. Computer Vision,* Corfu, Greece, pp. 666–673, Sep. 1999.

[12] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, no. 11, pp.1330–1334, Nov. 2000.

[13] C. Tomasi and T. Kanade, "Detection and tracking of point features," Carnegie Mellon University, Tech. Rep. CMU-CS-91-132, April 1991.

[14] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 14, no. 2, pp. 239–256, 1992.

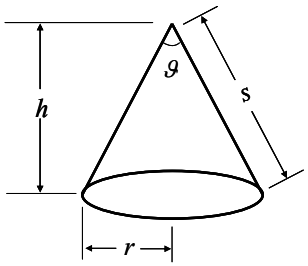
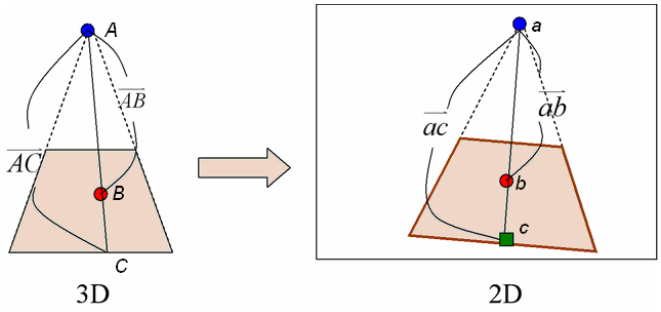
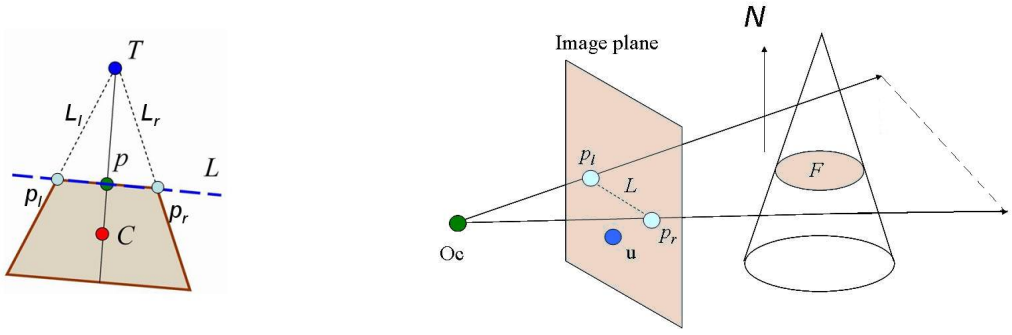


Fig. 1. Geometric cone model.



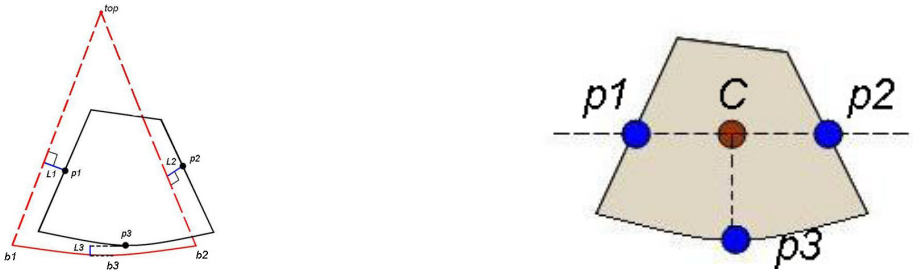
(a) (b)

Fig. 2. 3D cone model and its projection on an image.



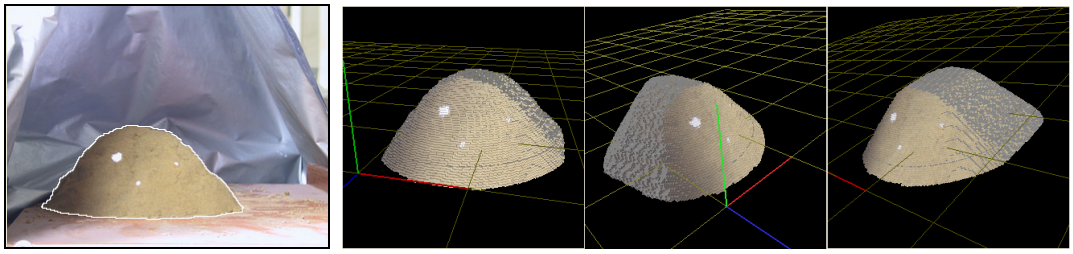
(a) (b)

Fig.3. Model simplification.



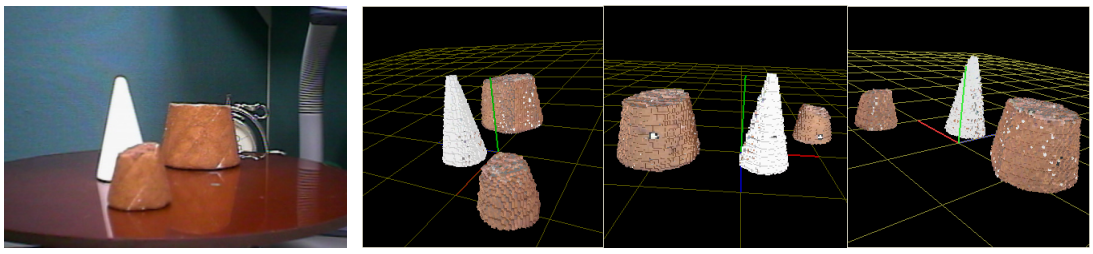
(a) (b)

Fig. 4. Mismatch minimization.



(a) (b)

Fig. 5. First example. (a) The object and (b) reconstruction result.



(a) (b)

Fig. 6. Second example. (a) The object and (b) reconstruction result.

Table I. Experimental evaluation of the first example

	Height(mm)	Base radius range(mm)	Volume(mm^3)
Reconstructed object	95.024	115.567~140.748	2,226,152
Real object	90	110~140	2,299,527

Table II. Experimental evaluation of the second example

	Object index	Height (mm)	Base radius range (mm)	Volume (mm^3)
Reconstructed model	I	111.707	30.726~30.726	114,096
	II	50.457	29.733~29.733	80,408
	III	78.444	50.692~50.692	419,992
Real Object	I	112	28~29	99,767
	II	53	28~29	93,025
	III	77	49~50	479,288