Interactive 3-D Modeling System Using a Hand-held Video Camera

Kenji Fudono[*1], Tomokazu Sato[*2] and Naokazu Yokoya[*2]

[*1]Victor Company of Japan
[*2]Nara Institute of Science and Technology, Japan

Abstract. Recently, a number of methods for 3-D modeling from images have been developed. However, the accuracy of a reconstructed model depends on camera positions and postures with which the images are obtained. In most of conventional methods, some skills for adequately controlling the camera movement are needed for users to obtain a good 3-D model. In this study, we propose an interactive 3-D modeling interface in which special skills are not required. This interface consists of "indication of camera movement" and "preview of reconstruction result." In experiments for subjective evaluation, we verify the usefulness of the proposed 3D modeling interfaces.

1 Introduction

In recent years, 3-D models of real objects have been often used for several purposes such as entertainment, education, and design. Generally, these 3-D models are constructed by experts who have special skills and devices for 3-D modeling. On the other hand, high-quality 3-D graphics have become very familiar to general people because 3-D graphics are available even on a cellular telephone today. Such a situation gives an increased demand to import 3-D models of real objects to personal web pages, games, and so on. For this purpose, simple 3-D modeling methods for real objects are necessary for users who have no special skills and devices to model the 3-D objects. To reconstruct 3-D models of real objects, several methods have been developed in the literature; methods using a video camera[1, 2], methods using a laser rangefinder[3], and methods using structured lights[4]. However, the accuracy of reconstructed 3-D models depends on the way of measurement. Thus, measurement skill is necessary to obtain good 3-D models.

To remove the difficulty in 3-D measurement, several support systems to obtain good 3-D models have been investigated[5–7]. These systems indicate how to move a measuring device based on a result of a reconstructed model. The indications allow users who have no special skills of modeling to get good 3-D models in a short time. However, it is difficult for personal users to use such systems, because these systems are designed for special and expensive devices such as a laser rangefinder. Although 3-D modeling systems that use only cheap devices have been developed[1, 2, 8–10], such systems do not indicate how to move cameras. There is also a problem that these systems take a long time to

reconstruct the models due to expensive computational cost. It is difficult for users to efficiently learn how to move the camera.

In this study, we propose an interactive 3-D modeling system by which users can obtain the model efficiently. The proposed system realizes two new functions: "indication of camera movement" and "real-time preview of reconstruction result". Users without special skills can easily obtain 3-D models by following the indication from the system. Users can also get a good 3-D model in a short time owing to a real-time preview of reconstructing a model.

2 Interactive Modeling System

In this section, we first describe a design policy and outline of the proposed interactive modeling system. Each process of the interactive modeling system for personal users is then detailed.

2.1 Design Policy and System Outline

The purpose of the proposed modeling system is to allow personal users to get good 3-D models efficiently. To realize this purpose, the following three requirements should be satisfied:
(a) realization of real-time modeling using cheap devices,
(b) realization of real-time indication of camera movement,
(c) realization of real-time preview of reconstruction results.
Our modeling system assumes that an object is located on a marker sheet and users move a hand-held video camera by following the indication from the system. To satisfy the requirements above, the system provides the following functions:
(1) Real-time Modeling Using a Hand-held Video Camera: While users capture objects by using a hand-held video camera, the system reconstructs 3-D models of the objects in real-time. This function satisfies the requirement (a).
(2) Capturing Support Interface: The system estimates the best view position from which images can be captured to acquire good 3-D models. The motion path from the current camera position to the best view position is shown to user on a computer display. User can easily obtain a 3-D model by following the indication of camera motion provided by the system. This function satisfies the requirement (b).
(3) Preview of Reconstruction Model: Preview of generating a 3-D model of the object is displayed and updated in every frame. User can preview reconstruction results without any waiting time. This function satisfies the requirement (c).

Figure 1 shows a flow diagram of the proposed system. The system consists of three phases: the phase A is a real-time process for 3-D modeling and preview, the phase B is an intermittent process for computationally expensive texture generation and the best view decision, and the phase C is a refinement process to acquire a detailed modeling result.

The phase A reconstructs a 3-D shape of the object in real-time. First, the position and posture of the hand-held camera are estimated by recognizing markers (A-1). A silhouette image which discriminates object regions and background

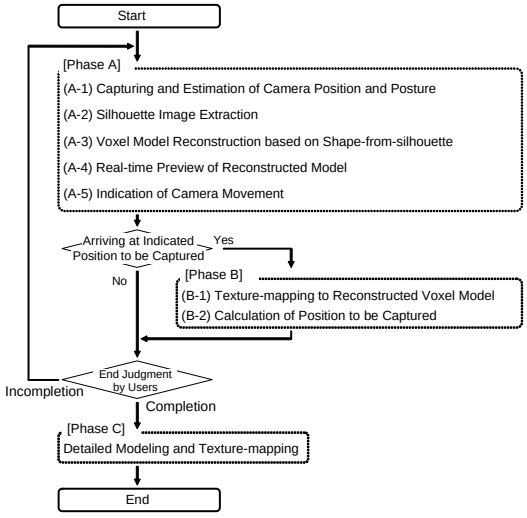

Fig. 1. Flow of interactive modeling system.

regions is generated (A-2). A voxel model is then reconstructed based on shape-from-silhouette (A-3). Preview of the reconstructed model is generated (A-4). Finally, the best view position is indicated (A-5). The phase B intermittently performs processes that are difficult to perform in real-time. First, texture-mapping to the reconstructed model is performed (B-1). The new best view position is then calculated (B-2). Above-mentioned phases A and B are repeated until users decide that further capturing is unnecessary. The phase C reconstructs a more detailed model than that reconstructed in the process (A-3) by using the whole captured image sequence. Note that the intrinsic parameters of the hand-held video camera are assumed to be known in this paper. To acquire a good silhouette image, it is also assumed that a marker sheet is located under a target object and wall and table have the same color.

2.2 Capturing and Estimation of Camera Position and Posture

In this process (A-1), the current position and posture of the hand-held video camera are estimated by using recognized markers in a captured image. In this section, first, markers are extracted from the input image that is captured by the hand-held video camera. Next, extracted markers are identified based on the patterns of the markers. Coordinates of markers on both world coordinate and image coordinate are recognized by identifiers of markers, and finally the position and posture of the camera are estimated from the coordinate values of the markers.

Extraction and Recognition of Markers

Figure 2 shows the marker sheet. Circular markers printed in this sheet are those proposed by Naimark et al.[11]. Each marker has 6 bits identifier that makes it possible to discriminate one another. Moreover, one marker has 3 identifiable

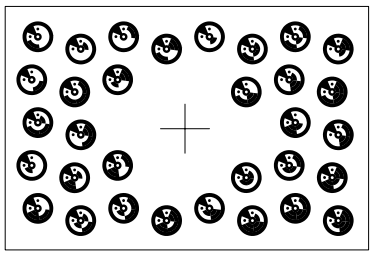

Fig. 2. Marker sheet.

points. The system gets multiple identifiable points by extracting and recognizing the markers from the captured frame, and acquires the coordinates in image and world coordinate systems.

Estimation of Camera Position and Posture
The camera position and posture are estimated by solving the Perspective n-Point (PnP) problem from the relation between image coordinates and world coordinates using a standard computer vision technique[12]. Three parameters (X, Y, Z) as a camera position and three parameters (pitch, roll, and yaw angles) as a camera posture are actually calculated by solving the PnP problem.

2.3 Silhouette Image Extraction

In this process (A-2), a silhouette image is extracted from a captured frame by using the estimated position and posture information of the camera. The silhouette image is used as an input for the shape-from-silhouette process (A-3). In this section, first, colors of background wall and desk are detected from the input image by using the camera position and posture information. Background regions are then extracted based on the detected background colors, and a silhouette image is generated.

Detection of Background Colors
Backgrounds consist of several regions; marker sheet, table, and wall behind the object. In this study, it is assumed that the base color of marker sheet, the region of table, and the region of wall surface have basically the same color. However, there may be a little difference in each color. To determine the background colors, firstly, the colors of the unprinted subregions around the extracted markers on the marker sheet are determined. Then the table regions are extracted based on the camera position and posture information, and the colors of the table regions are also detected.

Extraction of Object Regions
An input image is divided into object and background regions by using the differences of the brightness and the chromaticness of background colors detected in the previous step. After the detection of background and object regions by paper and wall colors, a silhouette image is generated by merging extracted object regions.

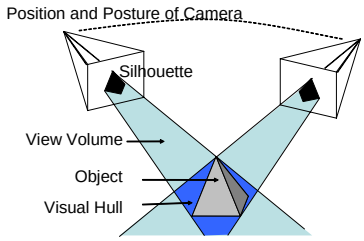

Fig. 3. Silhouette constraint.

2.4 Voxel Model Reconstruction based on Shape-from-silhouette

In this process (A-3), a 3-D voxel model is reconstructed based on shape-from-silhouette. In this section, first, the framework of shape-from-silhouette is briefly summarized. A method of voxel model reconstruction is then described.

Shape-from-silhouette
Shape-from-silhouette is a 3-D reconstruction method which is based on the silhouette constraint[13]. As shown in Figure 3, shape-from-silhouette approach reconstructs a 3-D model by assuming that "a target object is included in a view volume determined by the object's silhouette from camera center of the projection to the space." Intersections of view volumes generated from multiple camera positions are called visual hulls. A shape of visual hull is an approximated shape of the underlying object captured by multiple cameras.

Voxel Model Reconstruction
As one of the shape-from-silhouette methods, we employ a method that sets a cuboid in a voxel space which comprises the object preliminarily. The shape of the voxel model approximates a shape of object model by gradually carving the voxels of the cuboid which are outside of the view volume[14]. To reconstruct the voxel model efficiently, we use the parallel volume intersection method based on plane-to-plane projection proposed by Wu et al.[15].

2.5 Real-time Preview of Reconstructed Model

In this process (A-4), the reconstructed voxel model is rendered and updated in every frame. The user can look at the 3-D model by using the mouse operation. The user can also confirm the progress of the reconstruction by this preview of the generated model.

2.6 Indication of Camera Movement

In this process (A-5), how to move a camera is indicated for user by superposition to present the best view position from which the target object should be captured. The best view position is calculated in the intermittent process (Phase B). In this study, we prepare two types of indications: " (1) Indication of Rotation Movement" and " (2) Indication of Up-and-down Movement." In our system, the best view camera position is expressed by longitude and latitude on a virtual sphere which is located at the center on the marker sheet. As shown

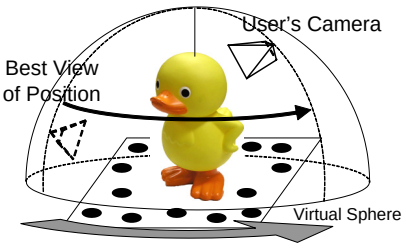

Fig. 4. Rotation of Marker Sheet under Object.

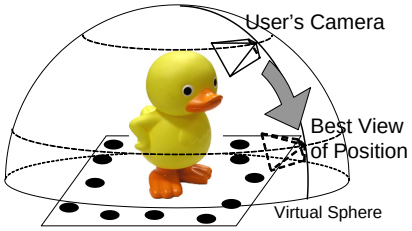

Fig. 5. Up-and-down Movement of Camera.

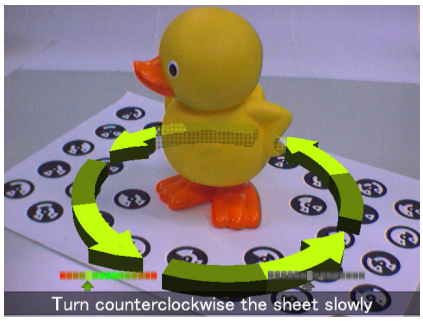

Fig. 6. Rotation Arrows.

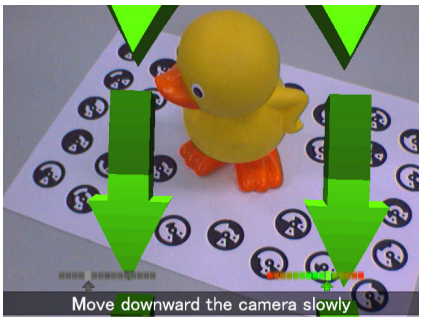

Fig. 7. Arrows for Camera Movement.

in Figure 4, the longitude of the camera position and the best view position are matched by rotating the marker sheet under the target object. Subsequently, as shown in Figure 5, the latitude of the camera position and the best view position are matched by up-and-down camera movement. Indications (1) and (2) are not shown at the same time. The indication (1) is shown first. When the indication (1) is finished, the indication (2) is then shown. Each indication is explained below in some details.

Indication of Rotation of Marker Sheet under Object

First, the system calculates the shortest rotation direction from the current camera position to the best view position. Then, as shown in Figure 6, the system shows rotation arrows by superposition on the sheet. The amount of rotation is shown on the arrows using color and the indicator at the bottom of a screen. Indication (1) is finished when the longitude difference between the camera and the best view position becomes sufficiently small.

Indication of Up-and-down Movement of Camera

As shown in Figure 7, the system shows arrows for a camera movement by superposition in a real scene. The amount of movement is shown using color and the indicator at the bottom of a screen. This indication is finished when the latitude difference between the camera and the best view position becomes sufficiently small. When a user completes to move the camera to the best view position by following indications, the system goes to the phase B.

2.7 Texture-mapping to Reconstructed Voxel Model

In this process (B-1), voxels are painted by projecting colors of an input image to a reconstructed model. The procedure of texture-mapping is detailed below.

Detection of Surficial Voxels
Only surficial voxels of a reconstructed model should be painted. In this process, voxels that are sorrounded by the other voxels are removed to detect surficial voxels $V_i(i = 1, ...,$the number of surficial voxels).

Visibility Test
In this section, surficial voxel V_i that is visible from each captured position $C_j(j = 1, ...,$ the number of captured frames) is detected. If there is no voxels between a surficial voxel V_i and a captured position C_j, a surficial voxel V_i is visible from C_j. Visibility tests for all surficial voxels are performed by all the captured frames.

Coloring Voxel
A surficial voxel V_i visible from a captured position C_j is projected to an image plane of a captured position C_j, and the color of the surficial voxel is set by the color of projected pixel on the image plane. If the surficial voxel is visible from multiple captured positions, the color of the surficial voxel is set to the average color of projected pixels on the image planes.

2.8 Calculation of Position to be Captured

As shown in Figure 8, some uncolored surficial voxels exist because they are invisible from all the input images. In this process (B-2), the system computes the best view position from which the most uncolored voxels can be observed. Users can get a good 3-D model efficiently by following the indication from the system. First, to reduce the computational cost, candidates of positions to capture are enumerated. The best indication position to capture is then chosen from the candidates.

Candidates Enumeration
To calculate the position from which the most number of uncolored voxels can be observed, it is necessary to count visible voxels from all the positions and postures of the camera. This is computationally expensive. In our system, the candidates of the best view positions are enumerated. The candidates are vertices of a geodesic dome as shown in Figure 9. The geodesic dome is located on the center of the marker sheet. The radius of the dome is set so that the whole of initial voxel model can be captured by a camera. Each posture of candidates faces the center of the dome.

Determination of Best View Position
Uncolored surficial voxels are counted for all the candidate positions. By the result of uncolored voxel count, the system selects the best view position from which the most number of uncolored surficial voxels are visible.

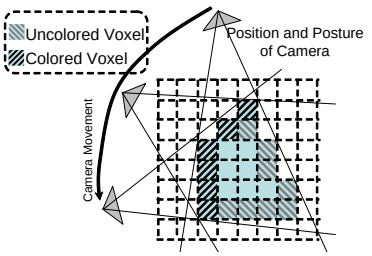

Fig. 8. Colored and Uncolored Voxels.

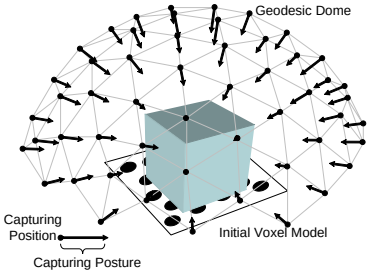

Fig. 9. Part of Candidates of Best View.

2.9 Detailed Modeling and Texture-mapping

When users decide further capturing is unnecessary by viewing a preview of a reconstructed model, the system goes to the phase C. In the phase C, more detailed 3-D model is generated by the off-line processing. The detailed modeling process generates the 3-D model by using the shape-from-silhouette method in higher resolution of a voxel space than that in the phase A. The system also performs more accurate texture-mapping than the process (B-1) by using area information at the visibility test.

3 Experiment

To verify the validity of the proposed system for personal users who have no special skills for modeling, we have carried out experiments with the proposed modeling system. In experiments, a prototype system is developed using a PC (CPU: pentium4 3.2GHz, Memory: 2GB) and a hand-held video camera (capture resolution: 640 $\times$ 480 pixels, frame rate: 30 fps). Intrinsic parameters of the camera were estimated by using the Tsai's method [16] in advance. A marker sheet was printed on A3 paper by a laser printer. Figure 10 shows the modeling environment. Figure 11 shows a modeling object. The voxel space for the real-time modeling is constructed of 64 $\times$ 80 $\times$ 64 voxels. Fifteen examinees used our system. Seven of 15 examinees are inexperienced in modeling real objects.

After the trials of 3-D modeling by examinees, we sent out questionnaires about the accuracy of the reconstructed model and capturing labor. Table 1 shows results of the experiments and the questionnaires. Figure 12 shows an example of reconstructed detailed model. The voxel space of detailed model is constructed of 150 $\times$ 150 $\times$ 150 voxels (voxel size: 0.86 $\times$ 1.39 $\times$ 0.98mm). The average frame rate of the phase A was 7.6 fps. The phase B process took 389 milliseconds on an average. The phase C process took 360 seconds on an average. In experiments, examinees could get good 3-D models by capturing in about 150 seconds, and we verified the usefulness of the system. However, some problems were found in the indication interfaces.

Fig. 10. Modeling Environment.

Fig. 11. Modeling Object.

4 Conclusion

In this paper, we have proposed an interactive modeling system using a hand-held video camera. The proposed system has new two functions: "indication of camera movement" and "real-time preview of reconstruction result." In experiments, we have verified that users who have no special skills for modeling can get a good 3-D model easily in a short time. In future work, the system should be evaluated by more examinees who have no modeling skills, and the indication interface should be reformed.

References

1. NTT DATA SANYO SYSTEM. Cyber modeler handy light. http://www.nttd-sanyo.co.jp/, 2002.
2. UZR GmbH & Co KG. imodeller 3D. http://www.imodeller.com/en/, 2001.
3. Leica Geosystems HDS LLC. Hds2500. http://hds.leica-geosystems.com/, 2000.
4. KONIKA MINOLTA. Vivid 910. http://konicaminolta.jp/, 2002.
5. J. E. Banta, L. M. Wong, C. Dumont, and M. A. Abidi. A next-best-view system for autonomous 3D object reconstruction. *IEEE Trans. Systems, Man and Cybernetics*, Vol. 3, No. 5, pp. 589–598, 2000.
6. K. Haga and K. Sato. A shape measurement with support light by a handy projector. *Proc. the 9th Pattern Measurement Symp. on Society of Instrument and Control Engineers (SICE)*, pp. 35–38, 2004 (In Japanese).

Table 1. Results of Experiments and Questionnaires.

Items		Score
Capturing Time	[second]	150.0
Accuracy of Reconstructed Model	[1:Bad - 4:Good]	3.3
Capturing Labor	[1:Tired - 4:Untired]	3.2

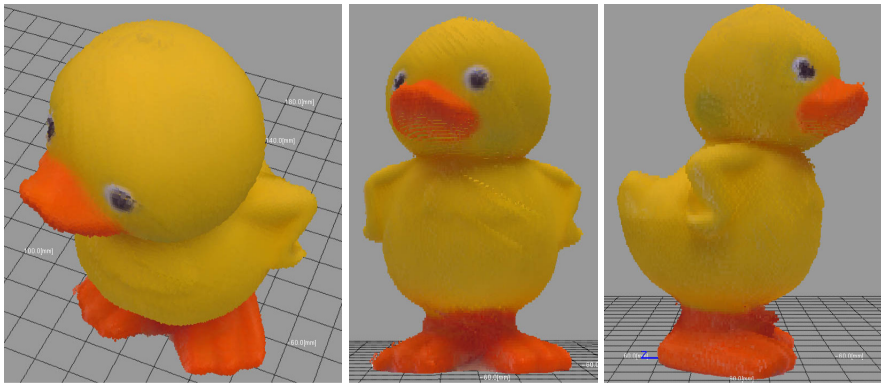

Fig. 12. Reconstructed Detailed Model.

7. M. Matsumoto, M. Imura, Y. Yasumuro, Y. Manabe, and K. Chihara. Support system for measurement of relics based on analysis of point clouds. *Proc. the 10th Int. Conf. on Virtual Systems and Multimedia (VSMM)*, p. 195, 2004.
8. L. Zhang, B. Curless, A. Hertzmann, and S. M. Seitz. Photometric method for determining surface orientation from multiple images. *Proc. the 9th IEEE Int. Conf. on Computer Vision (ICCV)*, Vol. 1, pp. 618–625, 2003.
9. G. G. Slabaugh, W. B. Culbertson, T. Malzbender, M. R. Stevens, and R. W. Schafer. Methods for volumetric reconstruction of visual scenes. *Int. Journal on Computer Vision (IJCV)*, Vol. 57, No. 3, pp. 179–199, 2004.
10. H. Kim and I. Kweon. Optimal photo hull recovery for the image-based modeling. *Proc. the 6th Asian Conf. on Computer Vision (ACCV)*, Vol. 1, pp. 384–389, 2004.
11. L. Naimark and E. Foxlin. Circular data matrix fiducial system and robust image processing for a wearable vision-inertial self-tracker. *Proc. the 1st IEEE/ACM Int. Symp. on Mixed and Augmented Reality (ISMAR)*, pp. 27–36, 2002.
12. R. Klette, K. Schluns, and A. koschan, editors. *Computer Vision: Three-dimensional Data from Image*. Springer, 1998.
13. H. Baker. Three-dimensional modeling. *Proc. the 5th Int. Joint Conf. on Artificial Intelligence (IJCAI)*, Vol. 2, pp. 649–655, 1977.
14. Y. Kuzu and V. Rodehorst. Volumetric modeling using shape from silhouette. *Proc. the 4th Turkish-German Joint Geodetic Days*, pp. 469–476, 2001.
15. X. Wu, T. Wada, S. Tokai, and T. Matsuyama. Parallel volume intersection based on plane-to-plane projection. *IPSJ Trans. on Computer Vision and Image Media*, Vol. 42, No. SIG6(CVIM2), pp. 33–43, 2001.
16. R. Y. Tsai. An efficient and accurate camera calibration technique for 3D machine vision. *Proc. Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 364–374, 1986.