Time-Varying Mesh Generation Based on Iterative Feedback between Silhouette Extraction and Geometry Modeling

Toshihiko Yamasaki[*, †], Kentaro Yamada[*], and Kiyoharu Aizawa[‡, *]
[*]Dept. of Information and Communication Engineering, [‡]Interfaculty Initiative in Information Studies, The University of Tokyo
[†]MSR IJARC Fellow
E-mail: {yamasaki, kyamada, aizawa}@hal.t.u-tokyo.ac.jp

Abstract— **This paper proposes a shape-from-silhouette-based 3D geometry modeling scheme that is robust to defective silhouette extraction. The silhouette extraction is a significantly important task because it directly affects the quality of the generated 3D models. Our proposed approach improves the silhouette extraction and as a result the generated 3D models by the iterative feedback between the 3D modeling using the improved silhouette images and the silhouette updating using the rendered 3D images. Namely, the generated 3D models are utilized to obtain better foreground/background seeds for graph-cuts-based silhouette extraction and vice versa. In this paper, we propose two modeling approaches aiming at more accurate seeds generation. As a result, 3D modeling featuring much less loss/surplus of voxels has been made possible. Experimental results demonstrated that the loss of voxels can be reduced from 2.1% to 0.90% and the surplus of voxels can be reduced from 9.4% to 1.2% as compared to the initial 3D model. In addition, the error rate is the minimum among the conventional approaches.**

I. INTRODUCTION

Generating dynamic Three-Dimensional (3D) mesh sequences of human performances using multiple cameras has been investigated actively in the last 15 years [1]-[11]. In most cases, frames are generated independently of each other because of the nonrigid nature of human bodies and clothes. Therefore, the vertices and the connections are not always time-consistent. In this paper, we shall refer to such data as Time-Varying Meshes (TVMs).

Shape-from-silhouette (or volume intersection) is a fundamental process in generating TVMs to obtain the convex hull of the 3D objects. Because the shape-from-silhouette algorithm is directly affected by the foreground/background segmentation, a well-controlled mono-tone background is often employed [4]-[9]. However, these studios tend to be large and fixed. On the other hand, we have been developing an easy-to-setup TVM studio with a natural background.

A lot of foreground/background segmentation algorithms have been proposed so far [12]-[14] but it is still a very difficult problem. Although some 3D model refinement algorithms for high-quality TVM generation have been developed [3][4][10][11], these algorithms are designed only to eliminate unnecessary voxels, not to recover erroneously removed voxels. Therefore, the misclassification of the foreground object region as the background is a critical problem. It is not to mention that the excess number of voxels due to the dilation process to solve such a problem is difficult to remove even with [3][4][10][11]. Therefore, the purpose of this paper is to develop a TVM generation algorithm with smaller number of loss and surplus of voxels even with a natural background.

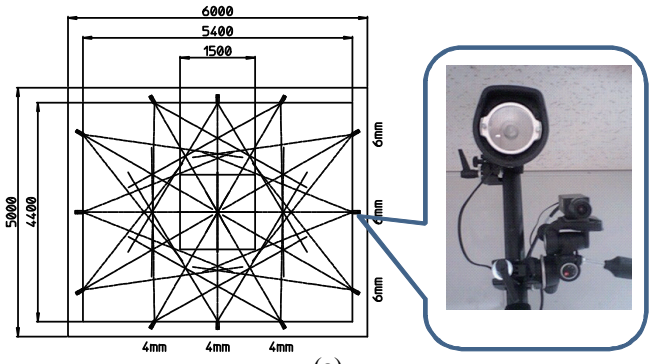
This paper presents a robust TVM generation algorithm based on the iterative feedback between the silhouette extraction and the 3D modeling. Namely, the generated 3D models are rendered and used as a seed for the graph-cuts algorithm [15][16] for better silhouette extraction. The improved silhouette images are used to reconstruct the 3D models. This iterative process is repeated until the geometrical shape of the 3D models converges. As a result, both the loss and the surplus of voxels can be suppressed drastically as compared to conventional algorithms. Experimental results showed that the loss and the surplus of voxels were reduced to 0.9% and 1.2% of the total number of voxels, respectively.

The rest of this paper is organized as follows. Section 2 reviews related works for the robust 3D model reconstruction. Section 3 describes our TVM studio and our proposed algorithm is presented in Section 4. Experimental results are demonstrated in Section 5. Finally, concluding remarks are given in Section 6.

II. RELATED WORKS

Toyoura et al. [6] proposed a silhouette extraction using a random pattern background. By using small patches of a random color pattern, the probability of the foreground color coinciding with that of the background in all viewpoints is made very small. Even when the color of the background is close to that of the foreground object in a certain view, the background color from a different view is far from that of the foreground object. Therefore, misclassification of the foreground as the background can be suppressed. This approach can reduce the loss of voxels but on the other hand tends to yield surplus voxels. In addition, a proper design of a random pattern background depending on the size of the studio is required.

Kim et al. [7] introduced a reliability map of foreground/background segmentation. When the summation of the reliability score from all the views exceeds a certain threshold value, the voxel is regarded as the foreground object. However, this approach also tends to yield superfluous voxels.



(a)

(b)

Fig. 1. Our TVM studio: (a) floor plan, (b) a view from a certain camera.

An object silhouette extraction with error detection and correction using multi-viewpoint images were proposed by Nobuhara et al. [17]. In their approach, the two constraints were introduced: "intersection" that assume that projection of the visual hull on every viewpoint should be equal to the silhouette on each viewpoint and "projection" that implies that projection of the visual hull should have outline which matches with apparent edges of captured image on each viewpoint. This algorithm required several hundreds of iteration and took 0.5~3 days to process only a single frame. Therefore, it is not feasible for our purpose.

An alternative approach is using graph-cuts in the 3D space instead of improving the silhouette extraction [2]. In [2], the data term was the sum of the values attached to the voxels where the value is based on the observed intensities of the pixels that intersect it and the smoothness terms is defined as the number of empty voxels adjacent to filled ones. However, the accuracy of the modeling was not discussed in [2]. The graph-cuts algorithm was also employed to refine the generated 3D model in addition to the shape-from-silhouette especially for refining the concave part of the objects [10][11]. This process is used only for removing unnecessary voxels: the loss of voxels deriving from erroneous silhouette extraction cannot be recovered.

On the other hand, we repeat the silhouette extraction and the 3D modeling iteratively to improve each other results. The rendering results of the generated 3D voxels are used for seed generation for better foreground/background segmentation in the next step modeling. Therefore, the loss of voxels can be reduced while suppressing the surplus of voxels even with natural backgrounds. In addition, the computational cost is not very large because the number of required iterations is quite small as discussed in Section 5.

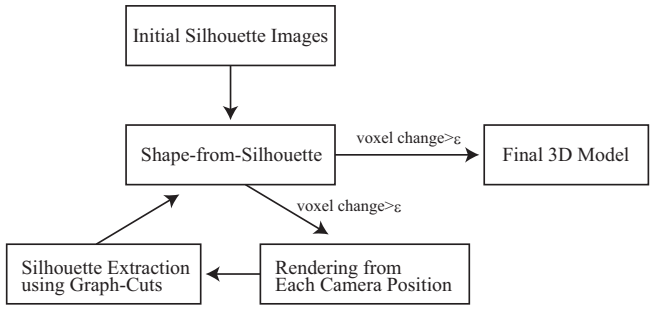


Fig. 2. Flowchart of the proposed algorithm.

III. Our TVM Studio

Our TVM studio is illustrated in Fig. 1. The studio consists of 12 sets of a capturing unit: camera with 1360×1024 resolution and camera-link interface, light, and personal computer (Intel Core2 Duo 2.4GHz, 4GB memory, RAID-0 HDD operating at 3Gb/s) attached to a pole. All the cameras are synchronized by an external signal generator. The frame rate is up to 34 fps. The system was setup in our laboratory room (Fig. 1 (b)). No special background such as blue sheet is utilized. Only the computers are covered with clothes because they are shiny and affect the silhouette extraction. The camera calibration is done with Tsai's method [18].

The system is easy-to-setup and portable. Disassembling and setting up the studio again can be done in a few hours. The size of the studio is about 6m × 5m but it is flexible depending on the size of the object and the area for the object to move around.

IV. Algorithms for Robust TVM Generation

A. *Flow of the Algorithm*

The flowchart of our TVM generation algorithm is shown in Fig. 2. In the initial step, conventional silhouette extraction and 3D modeling is conducted. Then, we proceed to the iterative processing between the silhouette refinement using the rendered images and the 3D model reconstruction with the error compensation. When the generated 3D model converges and is not very different from that of the previous step, the iteration is terminated and the final 3D mesh is obtained.

For higher-quality modeling, especially for reconstructing the concave parts, sophisticated model refinement algorithms after the shape-from-silhouette are required such as deformable mesh [3], stereo matching, [4] and graph-cuts in the 3D space [10][11]. However, such model refinement process is

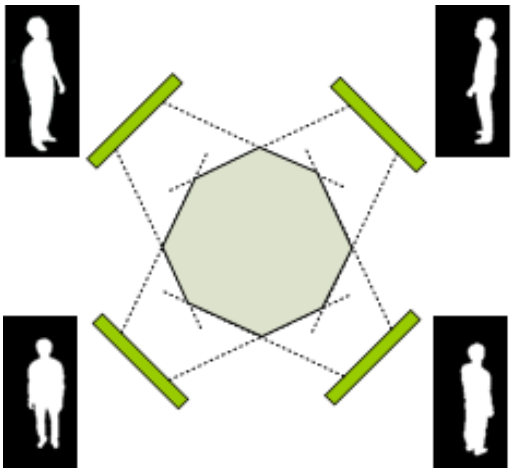


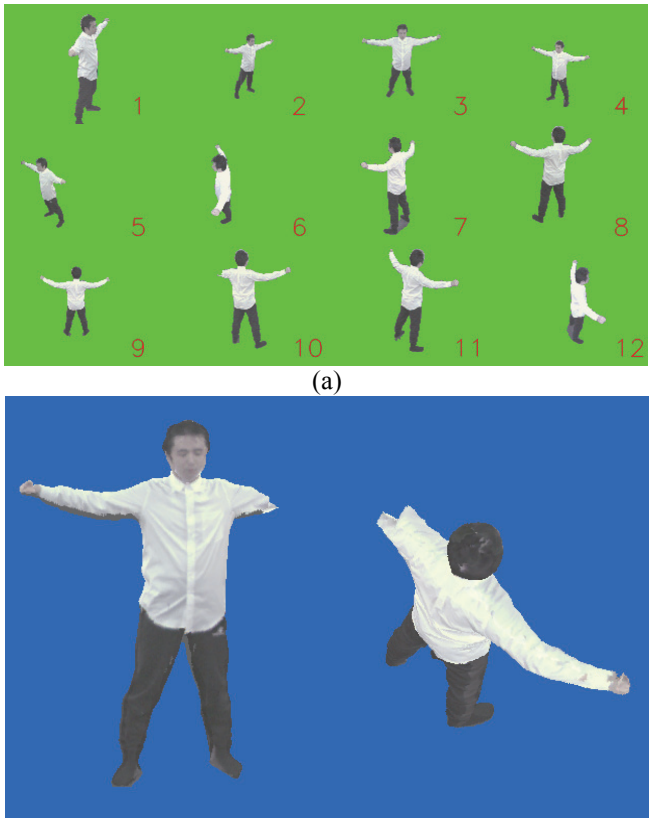
Fig. 3. Shape-from-silhouette algorithm.

out of scope of this paper. Our target is generating shape-from-silhouette-based 3D mesh models with less loss of voxels while suppressing surplus of voxels so that such refinement algorithms work better.

B. Shape-from-Silhouette with Error Compensation

The shape-from-silhouette is a 3D modeling algorithm by taking the intersections of visual cones of all the cameras surrounding the object as shown in Fig. 3. In other words, if a voxel, which is a small 3D region like a pixel in a 2D image, is seen from all the cameras, the voxel remains. Otherwise, the voxel is removed. In this manner, a visual hull of the 3D object is estimated. Then, various refinement algorithms [3][4][10][11] are applied for modeling convex parts or smoothing the model. One of the most significant disadvantages of this approach is that when the voxel is invisible from even a single camera due to erroneous silhouette extraction, the voxel is eliminated. On the other hand, the probability of the non-object voxel to be visible to all the cameras is quite low because the voxel can be labeled as non-object by other cameras. Such loss of voxels degrades the visual quality of the model. An example is shown in Fig. 4. In this case, the left arm in the camera #10 is missing due to erroneous silhouette extraction and the error affects the generated 3D model very much. Note that the error in Fig. 4 is an actual result, not a simulation. The refinement algorithms [3][4][10][11] cannot recover such loss of voxels because they are designed to eliminate unnecessary voxels, not to add necessary voxels. Therefore, two kinds of error (loss) compensation algorithms are proposed in this paper.

One is a voting-based modeling method. Here, we assume the number of cameras in the studio as n and m is an integer ranging from 1 to n-1. If the voxel is visible from n-m cameras, the voxel is left. Typically, m is set as 1~2 because the probability of the voxel that belongs to the object to be invisible from more than two or three cameras is quite low. If we increase m, the generated 3D model would expand more than necessary. If the error in silhouette extraction occurs in many camera views, we should reconsider the silhouette extraction



(a)

(b)

Fig. 4. Example of loss of voxels: (a) error in silhouette extraction only in camera #10, (b) generated 3D model in which left arm is not reconstructed properly.

algorithm itself. In this approach, one 3D model is generated for a single frame independent of the value m.

The other approach is modeling with the other (n-1) camera views. When generating the background/foreground seeds for the i-th camera view, the (n-1) camera views excluding the i-th camera view are used for the modeling. And the generated 3D model is rendered from the i-th camera position only for improving the i-th silhouette. Therefore, we need to conduct the 3D modeling for all the n camera views. This approach implicitly assumes that the segmentation error does not occur in multiple views at the same time, which is reasonable in most cases. Important to note here is that such errror can occur in multiple parts. The restriction here is that a voxel is misclassified as a non-object region by not more than a single camera. Modeling with the other (n-2) camera or less views is not a reasonable approach because the number of models to generate becomes quite large: $n\times(n\text{-}1)$.

In the iteration process, 3D model reconstruction is conducted multiple times. In particular, the cost for the modeling with (n-1) camera views approach becomes quite expensive as the number of cameras increases. To save the computational cost, the 3D modeling in the iteration can be done with rough spatial resolution and only the final modeling should be carried out with finer spatial resolution. Another option is iterate the refinement process only once because the modeling accu-

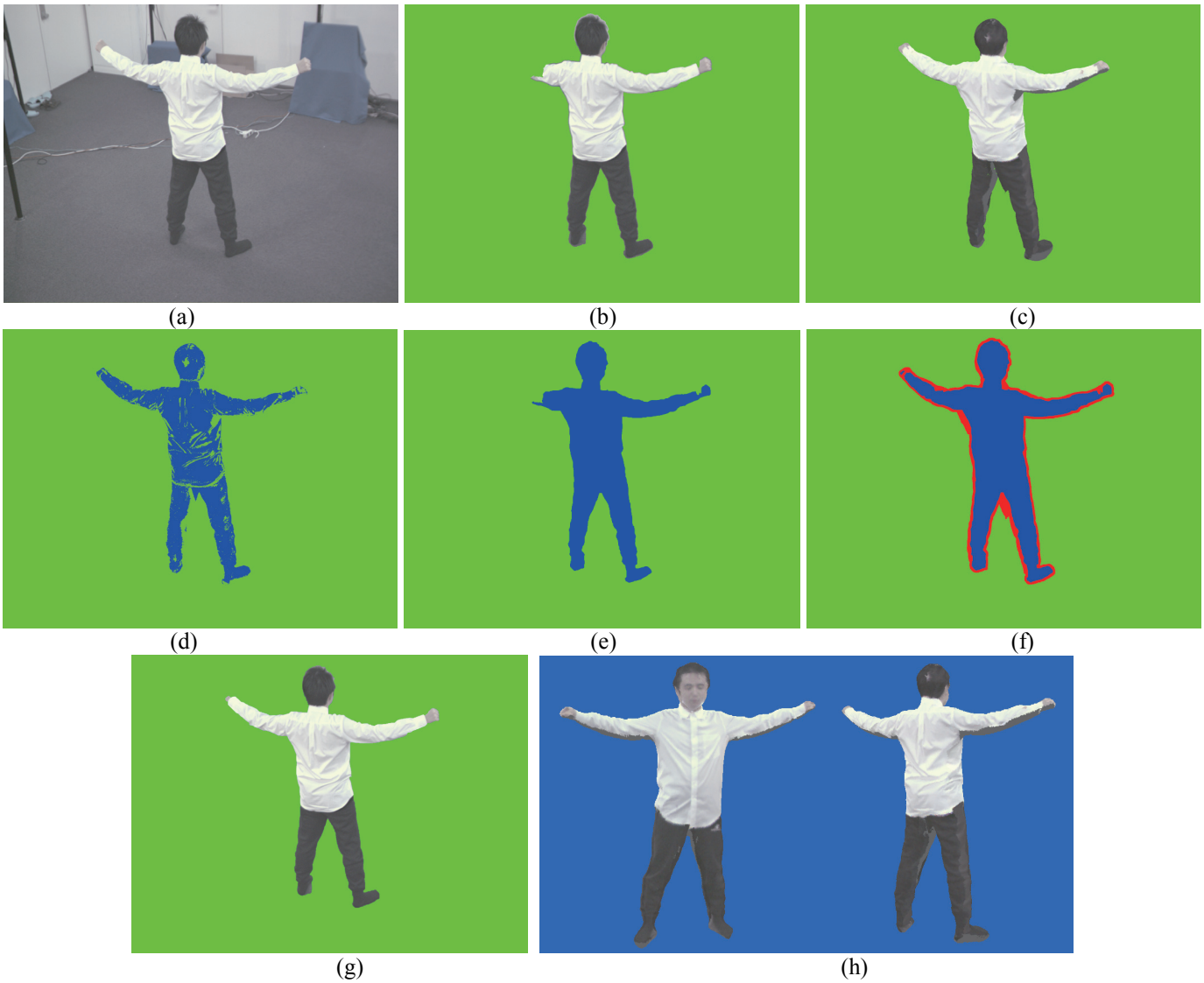


Fig. 5. Silhouette updating using the rendered 3D model: (a) original captured image, (b) initial silhouette by background subtraction and graph-cuts, (c) generated 3D model with error compensation algorithm described in 4B, (d) close-color map between (a) and (c), (e) eroded silhouette using (b), (f) updated seeds for graph-cuts, (g) updated silhouette, (h) updated 3D model.

racy becomes high enough by a single iteration as demonstrated in Section 5.

C. *Silhouette Extraction and Updating*

In the initial silhouette extraction, conventional background subtraction with the graph-cuts is employed. The background and foreground regions with high confidence are generated as follows:

$$\begin{cases} \text{if } \left|Y(x,y) - Y_{BG}(x,y)\right| > Th1, \text{then } (x,y) \text{ is foreground} \\ \text{else if } \left|Y(x,y) - Y_{BG}(x,y)\right| < Th2, \text{then } (x,y) \text{ is background} \\ \text{otherwise}, \textit{unknown} \end{cases}$$

Here, $Y(x,y)$ is the chroma value of the pixel at (x,y) and $Y_{BG}(x,y)$ is that of the background model. $Th1$ and $Th2$ are predefined threshold values where $Th1 > Th2$ in order to extract background and foreground regions with high confidence. When $|Y(x,y)-Y_{BG}(x,y)|$ is between $Th1$ and $Th2$, the pixel is left as unknown. Then, the background/foreground maps are fed to the graph-cuts algorithms as seeds. The silhouette extraction results are shown in Fig. 4(a).

In the iteration process, we assume that the erroneous loss of voxels is compensated by either way described in 4B. The silhouette refinement for each camera view is conducted using three images: the original captured image (Fig. 5(a)), the silhouette image in the previous step (Fig. 5(b)), and the rendered 3D image from the camera position (Fig. 5(c)). The background seed is generated by the logical AND operation between the background regions in the previous silhouette

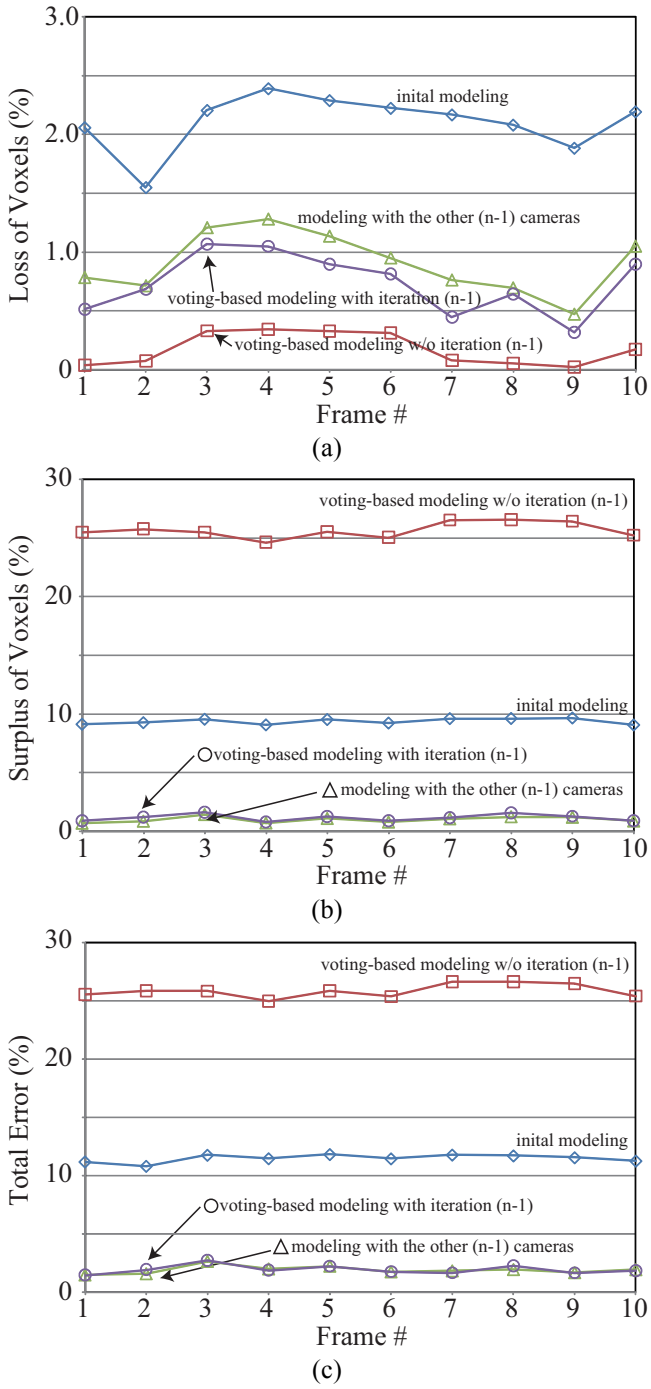


Fig. 6. Modeling accuracy: (a) surplus of voxels, (b) loss of voxels, (c) total error.

image (Fig. 5(b)) and the rendered image (Fig. 5(c)). A similar color region (Fig. 5(d)) between the original captured image (Fig. 5(a)) and the rendered image (Fig. 5(c)) and the eroded silhouette image in the previous step (Fig. 5(e)) are logically summed to form a foreground seed. As a result, the seeds for the background and the foreground for the graph-cuts in the next step are generated as demonstrated in Fig. 5(f). In the figure, the green, blue, and red regions represent the background, foreground, and unknown regions, respectively. The updated silhouette is shown in Fig. 5(g). This procedure is applied to each camera view independently. The updated silhouette images are utilized for the 3D modeling again. An example of the updated 3D model after a single feedback loop is shown in Fig. 5(h).

Table 1. Averaged modeling accuracy over the 10 frames.

	Loss	Surplus	Total Error
Initial model	2.1%	9.4%	11.5%
Voting-based w/o iteration (*n*-1 cameras)	0.18%	25.7%	25.9%
Voting-based with iteration (*n*-1 cameras)	0.73%	1.2%	1.9%
Voting-based with iteration (*n*-2 cameras)	0.64%	2.0%	2.7%
Modeling with the other (*n*-1) camera views	0.90%	0.99%	1.9%
Toyoura et al. [6]	11.3%	2.7%	14.0%

V. Experimental Results

The experiments were conducted using the TVM studio with 12 cameras as described in Section 3. Consecutive 10 frames of video (12 cameras × 10 frames = 120 images) were selected and ground truth data of the silhouettes were generated by hand. Then, 10 frames of ground truth data of TVMs were generated by the shape-from-silhouette algorithm. Our shape-from-silhouette program is based on [4] by the courtesy of Tomiyama et al. The stereo matching in [4] was disabled in the experiments

The loss and surplus of voxels and the total error for each frame is shown in Fig. 6. The accuracy of the initial 3D model and the voting-based modeling without the iteration process are also shown. Besides, the mean accuracy is summarized in Table 1. The modeling performance by Toyoura et al. [6] is also shown for comparison. Note to mention is that the experimental setup and the target models are very different from [6]. The voting-based modeling without the iteration is better in terms of loss of voxels but tends to yield much more surplus voxels than the others. In fact, the total error gets worse than the initial 3D model. On the other hand, the proposed algorithms yield a good performance both in terms of loss and surplus of voxels. The total error is less than 2% for both the voting-based modeling by (*n*-1) cameras and the modeling with other (*n*-1) camera views. When we increase the number *m* in the voting-based method, the loss of voxels is reduced more but on the other hand the surplus voxels increases rapid-

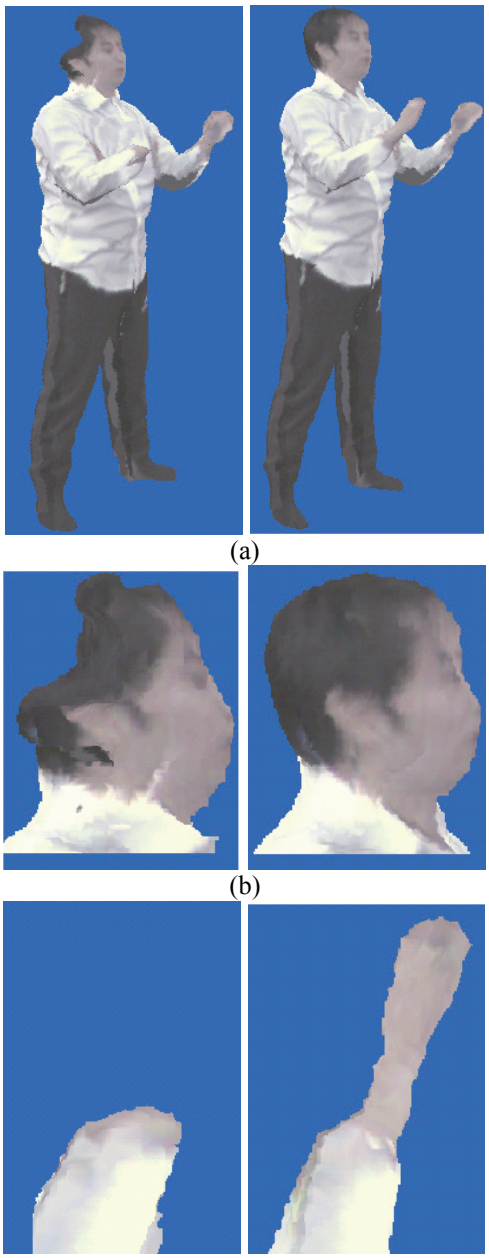


Fig. 7. Example of the improved 3D model: (a) whole image, (b) close-up of head, (c) close-up of right arm. The images on the left are those of the initial 3D model and the images on the right are those of the refined model.

ly. Therefore, *m* should be kept as small as possible depending on the silhouette extraction performance.

An example of the improved modeling by the voting-based modeling by (*n*-1) cameras is shown in Fig. 7 (frame #9). The back of the head and the right hand are missing in the initial 3D model and they are recovered by our proposed algorithms. The errors in the silhouette extraction in the two regions occurred in different views.

An example of the unsuccessful case by the voting-based modeling is shown in Fig. 8. In the figure, the right arm is properly reconstructed but the left arm is still missing because the voxels in the silhouette extraction for the left arm failed in two cameras at the same time. In such a case, voting-based modeling with (*n*-*m*) cameras (*m*>1) is feasible. In addition, such an error can be detected by monitoring the abrupt change in the number of voxels in the successive frames.

Fig. 8. The case where our proposed algorithm fails to recover the missing part.

The modeling performance improvement as a function of the number of iterations is shown in Fig. 9. In this experiment, the convergence verification was disabled. Zero means the initial 3D model. It is shown that the model converges with a small number of iterations (1~2 times). This means the seed regeneration for the graph-cuts is accurate enough in the first iteration process. Therefore, only a single feedback is sufficient in most cases.

The processing time for the initial 3D model generation and the voting-based modeling without iterations were both about 2.5 seconds. On the other hand, the voting-based modeling with iterations and the modeling with (*n*-1) camera views took 35 seconds and 45 seconds, respectively. In this paper, no code optimization was conducted. Parallel processing using GPUs or dedicated hardware is our future work for higher-speed modeling.

VI. Conclusions

In this paper, we have presented an iterative refinement algorithm for the silhouette-from-based 3D modeling. By the cross-feedback between the 3D model reconstruction with the updated silhouette and the silhouette extraction using the ren-

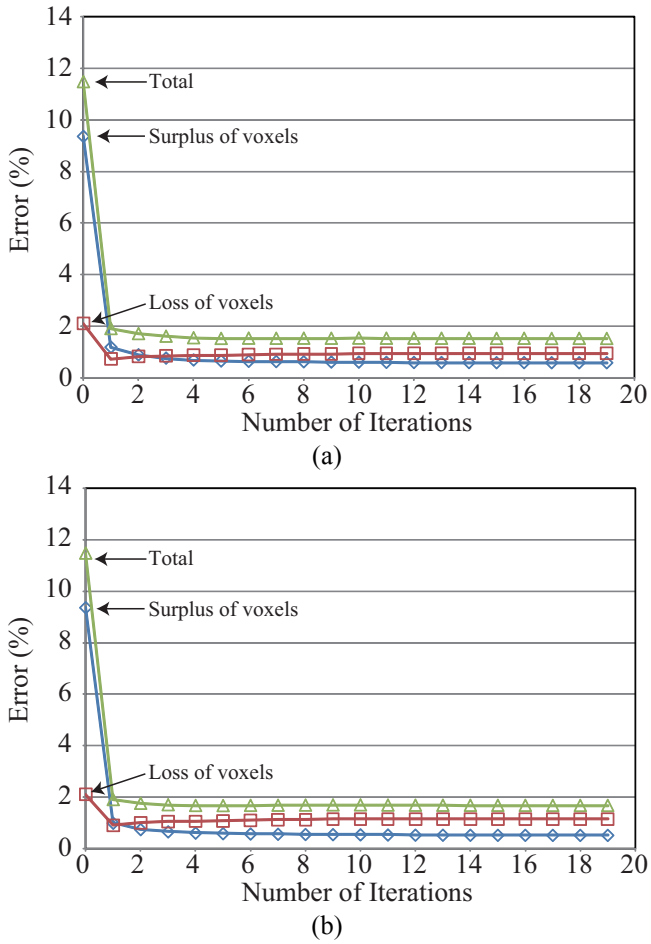


Fig. 9: Model refinement as a function of the number of iterations: (a) voting-based modeling by (n-1) cameras (b) modeling with the other (n-1) cameras. .

dered image, the loss and surplus of voxels can be kept very small. We have also proposed two shape-from-silhouette algorithms with error compensation to recover miss segmentation of the background/foreground. Experimental results demonstrated that the loss of voxels was reduced from 2.1% to 0.90% and the surplus of voxels was reduced from 9.4% to 1.2%, respectively.

ACKNOWLEDGMENT

This work is supported by the Microsoft Institute for Japanese Academic Research Collaboration (IJARC). We would like to thank Dr. Tomiyama et al. for providing us their 3D modeling source code.

REFERENCES

[1] T. Kanade, P. Rander, and P. Narayanan, "Virtualized reality: constructing virtual worlds from real scenes," IEEE Multimedia, vol. 4, no. 1, pp. 34-47, Jan./March 1997.

[2] D. Snow, P. Viola, and R. Zabih, "Exact voxel occupancy with graph cuts," IEEE Conference on Computer Vision and Pattern Recognition, vol.1, pp.345-352, 2000.

[3] T. Matsuyama, X. Wu, T. Takai, and T. Wada, "Real-time dynamic 3-d object shape reconstruction and high-fidelity texture mapping for 3-d video," In IEEE Trans. Circuit And System For Video Technology, Vol. 14, No. 3, pp. 357-369, 2004.

[4] K. Tomiyama, Y. Orihara, M. Katayama, and Y. Iwadate, "Algorithm for dynamic 3D object generation from multiviewpoint images," Proc. SPIE, vol. 5599, pp. 153-161, 2004.

[5] J. Starck and A. Hilton, "Surface Capture for Performance-Based Animation," IEEE Computer Graphics and Applications, Vol. 27, No.3, pp. 21-31, May-June 2007.

[6] M. Toyoura, M. Iiyama, K. Kakusho, and M. Minoh, "Silhouette extraction with random pattern backgrounds for the volume intersection method," The 6th International Conference on 3-D Digital Imaging and Modeling (3DIM 2007), pp.225-232, 2007.

[7] H. Kim, R. Sakamoto, I. Kitahara, N. Orman, T. Toriyama, K. Kogure, "Compensated Visual Hull for Defective Segmentation and Occlusion," 17th International Conference on Artificial Reality and Telexistence, pp.210-217, 2007.

[8] D. Vlasic, I. Baran, and W. Matusik, "Articulated mesh animatio from multi-view silhouettes," ACM Transactions on Graphics (ACM SIGGRAPH2008), #97, 2008.

[9] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.P. Seidel, and S. Thrum "Performance capture from sparse multi-view stereo," ACM Transactions on Graphics (ACM SIGGRAPH2008), #98, 2008.

[10] K. Hisatomi, K. Tomiyama, M. Katayama and Y. Iwadate, "3D reconstruction using graph cut with view-dependent polygon texture blending," 5th European Conference on Visual Media Production (CVMP 2008), p. 18, London, Nov. 2008.

[11] T. Tung, S. Nobuhara, and T. Matsuyama, "Simultaneous super-resolution and 3D video using graph-cuts," IEEE Computer Society Conference on Computer Vision and Pattern Recognition , pp. 1-8, 2008.

[12] A. McIvor, "Background subtraction techniques," in Proc. Image Video Comput., pp. 147–153., 2000

[13] M. Piccardi, "Background subtraction techniques: a review," Proc. 2004 IEEE International Conference on Systems, Man and Cybernetics, Vol. 4, pp. 3099- 3104, 2004.

[14] Y. Benezeth, P.M. Jodoin, B. Emile, H. Laurent, and C. Rosenberger, "Review and evaluation of commonly-implemented background subtraction algorithms," Proc. IEEE 19th International Conference on Pattern Recognition (ICPR 2008), pp. 1-4, 2008.

[15] Y. Boykov and M-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images," ICCV, vol. I, pp. 105-112, 2001.

[16] C. Rother, V. Kolmogorv, A. Blake,""GrabCut":interactive foreground extraction using iterated graph cuts," ACM Trans. Graphics (SIGGRAPH '04), vol.23, no.3, pp.309-314, 2004.

[17] S. Nobuhara, Y. Tsuda, T. Matsuyama, and I. Ohama, "Multi-viewpoint silhouette extraction with 3d context-aware error detection, correction, and shadow suppression," 4th European Conference on Visual Media Production (CVMP2007), pp. 1-9, 2007.

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