Robust Background Subtraction Method Based on 3D Model Projections with Likelihood

Hiroshi Sankoh, Akio Ishikawa, Sei Naito, Shigeyuki Sakazawa

KDDI R&D Laboratories Inc.
2–1–15, Ohara, Fujimino-shi, Saitama 356–8502, Japan
{sankoh, ao-ishikawa, sei, sakazawa}@kddilabs.jp

***Abstract*—We propose a robust background subtraction method for multi-view images, which is essential for realizing free viewpoint video where an accurate 3D model is required. Most of the conventional methods determine background using only visual information from a single camera image, and the precise silhouette cannot be obtained. Our method employs an approach of integrating multi-view images taken by multiple cameras, in which the background region is determined using a 3D model generated by multi-view images. We apply the likelihood of background to each pixel of camera images, and derive an integrated likelihood for each voxel in a 3D model. Then, the background region is determined based on the minimization of energy functions of the voxel likelihood. Furthermore, the proposed method also applies a robust refining process, where a foreground region obtained by a projection of a 3D model is improved according to geometric information as well as visual information. A 3D model is finally reconstructed using the improved foreground silhouettes. Experimental results show the effectiveness of the proposed method compared with conventional works.**

I. INTRODUCTION

A free viewpoint video provides a new visual experience, in which audiences can see scenes from anywhere in 3D space.[1] In the free viewpoint video system, the virtual viewpoint can be moved such as back-and-forth and around as well as up-and-down among objects in a field where cameras cannot be mounted. It gives audiences an immersive feeling, and we call these view-changing experiences "walk-through" and "fly-through".

There are two main categories for generating a free viewpoint video. One is a model-based method[2] and another is an image-based method.[3] In order to realize the visual experiences mentioned above, the former method is more suitable than the latter since the former does not have restrictions on virtual viewpoint positions in 3D space, if a 3D model can be appropriately reconstructed.[2] It should be noted that 3D model accuracy has a large impact on the video quality, and the purpose of our study is to reconstruct the 3D model with high accuracy.

A typical method for acquiring a 3D model of interesting objects (e.g., humans) is a shape from silhouette[4] which reconstructs a visual hull in a 3D voxel space using silhouettes of objects extracted from camera images. Therefore, an accurate silhouette is necessary to generate a high-quality 3D model. There have been the numerous related works on the silhouette extraction. Most of them classify each pixel of a camera image into background region or foreground region using only visual information from a single viewpoint.[5][6] There is a substantial difficulty of the works in the case the foreground area and the background area have the same color, and it is highly probable that the extracted silhouette image includes both of unwanted regions and missed parts which correspond to false positives and false negatives caused by misclassification in extraction process, respectively. In this paper, we propose a robust background subtraction method using multi-view images, instead of using only a single camera image. The method applies the likelihood of background to each pixel of a camera image, and derives the integrated likelihood of each voxel in a voxel space, which is considered to be integrated information of multi-view images. The background region is determined based on the likelihood of voxel space with local adaptation to minimize its energy functions. Furthermore, the proposed method also applies a robust refining process, in which each silhouette is improved based on projections of the 3D model to each viewpoint.

The rest of the paper is organized as follows. Section II overviews related works, and Section III details our robust background subtraction method based on 3D model projections with likelihood. Section IV presents experimental results and comparison with conventional methods. Finally, the paper is concluded in Section V.

II. RELATED WORKS

There has been some research to improve a shape from silhouette method regarding suppression of false positives and false negatives in silhouette extraction. In order to reduce the number of false negatives, papers [7] and [8] proposed the method to relax a condition for the foreground determination. These methods count the number of viewpoints in which a voxel is projected outside the silhouette, and identifies the voxel included in the foreground object when the number is below the threshold. Especially, the paper [8] calculates the likelihood of misclassification for each voxel based on the ratio of false positives and false negatives in each silhouette image, and decides the threshold used in the paper [7] mentioned above.

However, these methods do not have a removal process for false positives. Therefore, the silhouette extraction process shall exclude unwanted regions precisely. Additionally, the threshold for reducing false negatives is not continuous but a discrete value based on the accepted number of viewpoints outside the silhouette. It might cause false positives in resultant visual hull, since the threshold cannot be set sensitively.

MMSP'10, October 4-6, 2010, Saint-Malo, France.
978-1-4244-8112-5/10/$26.00 ©2010 IEEE.

Another approach to refine silhouettes and a visual hull is introducing intersection and projection consistency into 3D space and multi-view images, respectively.[9] This method refines each silhouette and visual hull by cross-referencing both of them. Furthermore, the input data of this method is a rough visual hull that includes all objects. Thus it does not need background subtraction to be applied, and is independent of the accuracy of silhouette extraction. However, in order to remove unwanted regions from each silhouette, this method only uses the edge information of camera images acquired by region segmentation. Therefore, the accuracy of visual hull and silhouette images strongly depends on the performance of region segmentation. In particular, it is very difficult to extract foreground objects with sufficient precision when similar color features are found in both the foreground and the neighboring background. Consequently, the removal process of unwanted regions does not work properly, and the final visual hull and silhouettes might include false positives and false negatives. Additionally, the refinement process for dealing with false negatives is not considered, and the false negatives remain in the final result.

There is another issue with a shape from silhouette method itself, which may include inaccurate concave surfaces. That is, the reconstructed visual hull is just a convex polyhedron in which the real object is inscribed. Some solutions are proposed in the paper [10], and we do not pursue this issue in this paper.

III. Proposed Method

To overcome the problems mentioned in II, we propose a robust background subtraction method considering the likelihood of background in each viewpoint instead of extracting binary silhouettes. We derive integrated likelihood, and the background region is determined in the voxel space. Furthermore, we refine visual hull and corresponding silhouette images based on geometric information in 3D space so that robustness for both false negatives and edge features can be improved.

The proposed method has a series of procedures as shown in Fig. 1, and is summarized into two stages. These are the determination process for the visual hull based on the likelihood and the refinement process reflecting geometric information in voxel space. The proposed scheme takes multi-view synchronized images as input and provides a reconstructed visual hull as output.

A. Introduction of likelihood as background object

Conventional shape from silhouette methods need a binary silhouette image for every viewpoint; but it is highly probable that the binary silhouette images include false positives and false negatives. Such problems often arise if images include objects whose pixel color values are close to those of the background region. In order to overcome this problem, we analyse the pixel-wise likelihood as a background object for input camera images instead of employing the binary segmentation process. Furthermore, we derive integrated likelihood in a voxel space.

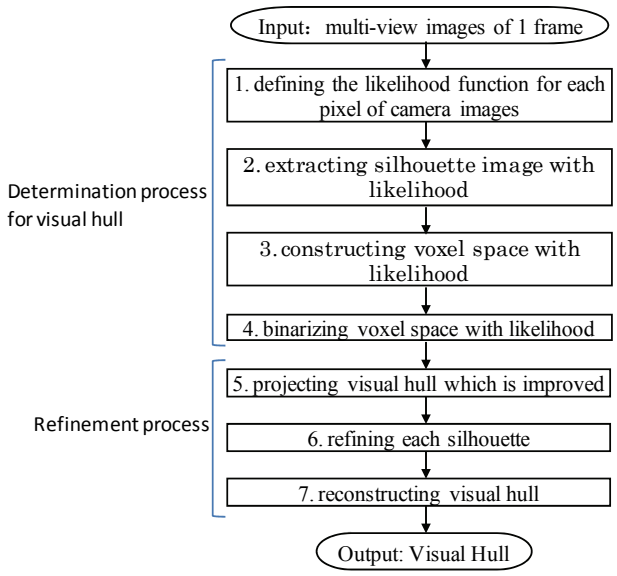


Fig. 1. Flowchart of the proposed method

First of all, it is assumed that camera images for a certain length of time without any foreground objects are provided in every camera position. We represent each pixel value as multi-dimensional vector $\mathbf{x}$ in a particular color space. On the assumption that pixel values are approximated by normal distribution, the likelihood function as a background object is defined as $f(\mathbf{x};\mathbf{u},\boldsymbol{\Sigma})$ by the following equation. In the equation, $\mathbf{u}$ and $\boldsymbol{\Sigma}$ are an average vector and a covariance matrix of pixel value $\mathbf{x}$ for some frames, respectively.

$$f(\mathbf{x};\mathbf{u},\boldsymbol{\Sigma}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{u})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{u})\right). \quad (1)$$

We acquire a silhouette image with likelihood in each viewpoint based on the function. Then, each voxel v in the voxel space is projected to viewpoint n $(n = 1,...,N)$, and voxel likelihood ρ_v is calculated based on likelihood $f(\mathbf{x}(v^{(n)}))$ of projected pixel value $\mathbf{x}(v^{(n)})$. Here, we construct the voxel space with likelihood by calculating the average value of $f(\mathbf{x}(v^{(n)}))$ to achieve a representative feature of all viewpoints using equation (2).

$$\rho_v = \frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}(v^{(n)})). \quad (2)$$

Conventional works proposed the extended method, which reduces false negatives by ignoring some viewpoints for which voxel v is projected outside the silhouette.[7][8] Compared with these works, our proposed method is able to control false negatives not discretely such as the number of viewpoints but continuously based on the likelihood.

B. Binarization of voxel space with likelihood

The simplest way to binarize voxel space with likelihood is to employ the thresholding of a single voxel. However, voxels whose likelihood is close to the threshold might be misclassified. To deal with such a problem, we define the

energy function considering the adjacency relationship in a 3D space, and binarize the voxel space with likelihood by minimizing the energy function using a graph-cut algorithm.

Energy function $E(\mathbf{v};\boldsymbol{\alpha}_\mathbf{v})$ is defined by equation (3) where $\mathbf{v}=(v_1,...,v_i,...)$ indicates indices of all voxels in a 3D space, and $\boldsymbol{\alpha}_\mathbf{v}=(\alpha_{v_1},...,\alpha_{v_i},...)$ identifies whether each voxel belongs to the foreground or the background region. $U(\mathbf{v};\boldsymbol{\alpha}_\mathbf{v})=\sum_i U(v_i;\alpha_{v_i})$ is a data term that depends on only likelihood ρ_{v_i} of voxel v_i, and gives the energy value as shown in equation (4). Here, b_v and th_ρ are positive constants.

$$E(\mathbf{v};\boldsymbol{\alpha}_\mathbf{v})=U(\mathbf{v};\boldsymbol{\alpha}_\mathbf{v})+V(\mathbf{v};\boldsymbol{\alpha}_\mathbf{v}). \tag{3}$$

$$U(v_i;\alpha_{v_i})=\begin{cases}-\log_{b_v}(\rho_{v_i}) & (\alpha_{v_i}=0)\\ \max(th_\rho+\log_{b_v}(\rho_{v_i}),0) & (\alpha_{v_i}=1).\end{cases} \tag{4}$$

$V(\mathbf{v};\boldsymbol{\alpha}_\mathbf{v})=\sum_{(i,j)\in N_v} V(v_i,v_j;\alpha_{v_i},\alpha_{v_j})$ is a smoothing term featuring the difference between ρ_{v_i} and ρ_{v_j} of a pair of adjacent voxels v_i and $v_j((i,j)\in N_v)$. It gives the energy value related to α_{v_i} and α_{v_j} by equation (5) where $dis_v(\cdot)$ is the Eucidean distance of adjacent voxels, and N_v indicates all the combinations of a pair of adjacent voxels v_i and v_j. In the equation, λ_v and κ_v are positive constants, and κ_v is calculated with the expectation operation $<\cdot>$ related to N_v as follows. [5] [6]

$$V(v_i,v_j;\alpha_{v_i},\alpha_{v_j})=\begin{cases}\dfrac{\lambda_v\exp(-\kappa_v(\rho_{v_i}-\rho_{v_j})^2)}{dis_v(i,j)} & (\alpha_{v_i}\neq\alpha_{v_j})\\ 0 & (\alpha_{v_i}=\alpha_{v_j}).\end{cases} \tag{5}$$

$$\kappa_v=(2<(\rho_{v_i}-\rho_{v_j})^2>)^{-1}. \tag{6}$$

The value of data term U decreases in proportion to the likelihood of a voxel that is classified as foreground. In addition, the value of smoothing term V decreases in proportion to the difference between likelihood ρ_{v_i} and ρ_{v_j} of adjacent voxels beyond the region boundary. The minimization of this energy function is known to be solved based on the graph-cut algorithm.[5] In the proposed scheme, the binarization process for the voxel space is conducted in a similar way while assigning label 0 and 1 to the background and the foreground, respectively.

C. Refinement based on 3D model projections

1) Removal of unwanted regions for the visual hull: We introduce a removal process of unwanted regions which are misclassified as foreground. Since a voxel size of a real object is usually larger than that of unwanted regions, the voxel size is used to distinguish them. The proposed process eliminates small regions whose voxel size is not ranked in the top R-order of all objects. Prior to rank the regions, the visual hull is divided into the closed regions.

2) Removal of shadow regions for silhouette images: Shadow regions cannot be extracted as closed regions since the shadow is connected to the foreground regions. Therefore, the above removal process does not work well. However, it is possible to eliminate shadow regions based on the constraint that they exist on the floor in 3D space. At first, we represent each pixel value of viewpoints as vector $\mathbf{I}(p)$ in an appropriate color space, and calculate the difference between foreground image $\mathbf{I}_f(p)$ and base image $\mathbf{I}_b(p)$ captured without any objects. The pixel that satisfies the condition of equation (7) is regarded as an unwanted region candidate. In the equation, I_d indicates the threshold parameter.

$$\left|\mathbf{I}_f(p)-\mathbf{I}_b(p)\right|<I_d. \tag{7}$$

Considering that the difference of chroma signals between foreground images and base images is small in shadow regions, the color space UV is effective for detecting shadows.

Then, we calculate the intersection point where the light rays of each pixel and the visual hull cross in 3D space. When there is an intersection point whose height from the floor is nearly equal to 0, the pixel can be eliminated as a shadow region. When p_{n_i} indicates the pixel index from viewpoint *n*, corresponding light rays are expressed by equation (8) where $\mathbf{r}_{n_i}=(x_{n_i},y_{n_i},z_{n_i})$ and $\mathbf{M}_n=(X_n,Y_n,Z_n)$ indicate the gradient and the camera position, respectively.

$$(X,Y,Z)^T=\mathbf{M}_n+t\mathbf{r}_{n_i}. \tag{8}$$

For each pixel in an unwanted region candidate, we calculate the intersection point of the visual hull and the corresponding light rays as (X_v,Y_v,Z_v), and when the condition below is satisfied the pixel is eliminated. In the equation, Y_d indicates the threshold parameter which stands for the height in 3D space.

$$Y_v<Y_d. \tag{9}$$

3) Removal of unwanted regions for silhouette images: We introduce the removal process for unwanted regions that neighbor the contour of the foreground object. Each silhouette image is binarized by minimizing the energy function with graph-cut algorithm, and the pixels regarded as background after minimization are removed. Energy function $E(\mathbf{p};\boldsymbol{\alpha}_\mathbf{p})$ is defined by equation (10) where $\mathbf{p}=(p_1,...,p_i,...)$ and $\boldsymbol{\alpha}_\mathbf{p}=(\alpha_{p_1},...,\alpha_{p_i},...)$ indicate pixel indices and labels identifying whether the corresponding pixel belongs to the background or the foreground. $U(\mathbf{p};\boldsymbol{\alpha}_\mathbf{p})=\sum_i U(p_i;\alpha_{p_i})$ is a data term that depends on only likelihood $f(\mathbf{x}_i)$ calculated by pixel value $\mathbf{x}_i$, and gives the energy value related to α_{p_i} by equation (11). Here, b_p is a positive constant. $V(\mathbf{p};\boldsymbol{\alpha}_\mathbf{p})=\sum_{(i,j)\in N_p} V(p_i,p_j;\alpha_{p_i},\alpha_{p_j})$ is a smoothing term defined with the difference between $\mathbf{x}_i$ and $\mathbf{x}_j((i,j)\in N_p)$ by equation (12). In the equation, $dis_p(\cdot)$ is the Eucidean distance of adjacent pixels, and N_p indicates all the combinations of a pair of adjacent pixels p_i and p_j. Variables λ_p and κ_p are positive constants, and

κ_p is calculated by equation (13) where $<\cdot>$ indicates the expectation operation related to N_p.[5] [6]

$$E(\mathbf{p};\boldsymbol{\alpha}_\mathbf{p}) = U(\mathbf{p};\boldsymbol{\alpha}_\mathbf{p}) + V(\mathbf{p};\boldsymbol{\alpha}_\mathbf{p}). \tag{10}$$

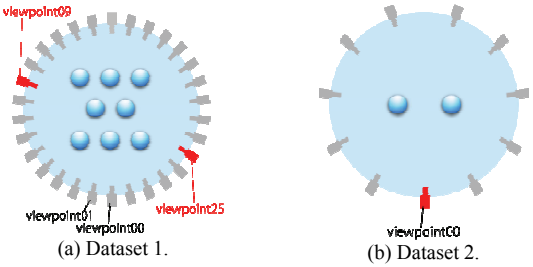
$$U(p_i;\alpha_{p_i}) = \begin{cases} (-\log_{b_p}(f(\mathbf{x}_i))) & (\alpha_{p_i}=0) \\ (-\log_{b_p}(1-f(\mathbf{x}_i))) & (\alpha_{p_i}=1). \end{cases} \tag{11}$$

$$V(p_i,p_j;\alpha_{p_i},\alpha_{p_j}) = \begin{cases} \dfrac{\lambda_p \exp(-\kappa_p(\mathbf{x}_i-\mathbf{x}_j)^2)}{dis_p(i,j)} & (\alpha_{p_i}\neq\alpha_{p_j}) \\ 0 & (\alpha_{p_i}=\alpha_{p_j}). \end{cases} \tag{12}$$

$$\kappa_p = (2<(\mathbf{x}_i-\mathbf{x}_j)^2>)^{-1}. \tag{13}$$

IV. Experimental Results

In order to evaluate the effectiveness of the proposed method, we conducted three experiments for multi-view video sequences. In experiment 1, we evaluated the reconstruction accuracy of visual hull compared with conventional works. In experiment 2, we analysed the contributions of each process in the proposed method. Finally, in experiment 3, we generated the free viewpoint video using the visual hull reconstructed by the proposed method. In these evaluations, we used multi-view images which were captured in a 360-degree circle with the spatial resolution of 640x360, and temporally aligned frames in each viewpoint. We prepared two kinds of sequences. Sequence A was captured with 30-cameras as in Fig. 2 (a), and sequence B was produced with 11-cameras as in Fig. 2 (b). Figures 3 and 4 show test images, which also include base images captured without any objects.

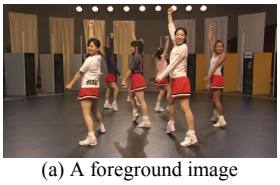


(a) Dataset 1. (b) Dataset 2.

Fig. 2. Camera Configuration

A. Experiment 1: reconstruction accuracy of the visual hull

In order to assess the accuracy of the visual hull reconstructed by the proposed method, we conducted an experiment for sequences A and B. The resolution of voxel space was 2cm^3 in both sequences, and the number of voxels were 256x128x256 and 160x100x100 in x-y-z coordinate system in sequence A and B, respectively. The likelihood function of each pixel was defined according to the equation (1), and base images of 60-frames in each viewpoint ware used to calculate the likelihood function. The parameters ware set as shown in TABLE I from preliminary experimental results. As indicated in TABLE I, the unwanted region removal process for silhouette images was not applied to sequence A, and the shadow removal process was not applied to sequence B. Therefore, in equation (1) each pixel value is represented as 3-demensional vector in RGB color space in sequence A, while it was represented as 2-demensional vector in UV space in sequence B.


(a) A foreground image

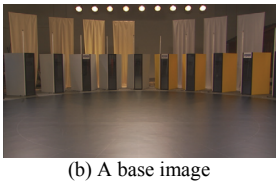

(b) A base image

Fig. 3. Dataset 1 (dancing)


(a) A foreground image

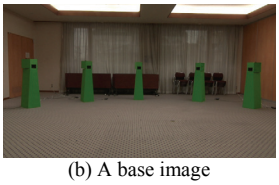

(b) A base image

Fig. 4. Dataset 2 (wrestling)

TABLE I
Parameters of The Proposed Method for Experiments

(a) Binarization process

	b_v	th_p	λ_v
Exp. 1	10	8.0	1.0
Exp. 2	10	3.4	1.0

(b) Unwanted region removal for the visual hull

	R
Exp. 1	8
Exp. 2	2

(c) Shadow region removal

	I_d	Y_d
Exp. 1	3.0	2.0
Exp. 2	–	–

(d) Unwanted region removal for the silhouette image

	b_p	λ_p
Exp. 1	–	–
Exp. 2	10	250.0

We evaluated the accuracy of the finally reconstructed visual hull by comparing the projections of the visual hull and the ground-truth images. The quantitative performance was assessed as follows. First, we prepare ground-truth images which were manually segmented, and then compared them pixel-wise with the projections of the visual hull. Finally, three values Recall, Precision, and F-measure ware calculated based on true positives, false positives and false negatives by equation (14), (15), and (16) as in the case of related work.[8]

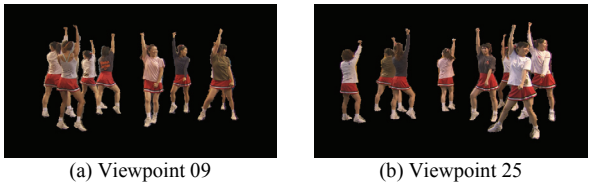
$$\text{Recall} = \frac{\#\text{true positives}}{\#\text{true positives} + \#\text{false neegatives}} \tag{14}$$

$$\text{Precision} = \frac{\#\text{true positives}}{\#\text{true positives} + \#\text{false positives}} \tag{15}$$

$$\text{F}-\text{measure} = \frac{2\times\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}} \tag{16}$$

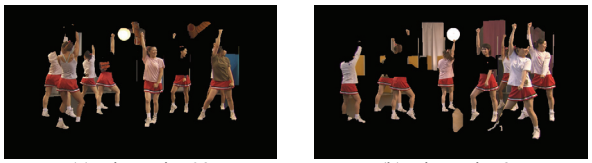
As conventional schemes, the GrabCut method[6] which uses only single image information, and Zeng's method[9] which uses the information of multi-view images, were also evaluated for comparison. Regarding the GrabCut method, we measured pixel-wise differences between the ground-truth and extracted results in each viewpoint. With regard to Zeng's method, we used projections of the visual hull to evaluate with the same criteria as used in the proposed method.

The projection images of the visual hull reconstructed by proposed method are shown in Fig. 5. The resultant images are shown with textures to help understanding. The results of the GrabCut method and Zeng's method are shown in Fig. 6 and Fig. 7, respectively. As shown in these results, it is obvious that the projection image of the visual hull generated by the proposed method almost corresponds to the object region in each viewpoint, except that there are a little false negatives neighboring the contour of the arms. Such false negatives are assumed to be caused by the estimated error of each projection matrix. The estimated error of the projection matrix directly affects the accuracy, since our proposed method employs projections between voxel space and each viewpoint. On the other hand, the GrabCut and Zeng's method resulted in a lot of false negatives and false positives. The accuracy of Zeng's method depends on the results of region segmentations that are applied first, and the method does not work well especially for objects with similar textures to background.



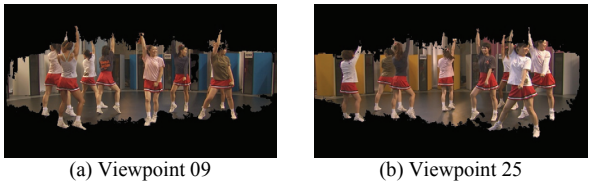
(a) Viewpoint 09 (b) Viewpoint 25

Fig. 5. Results of the proposed method



(a) Viewpoint 09 (b) Viewpoint 25

Fig. 6. Results of the GrabCut method



(a) Viewpoint 09 (b) Viewpoint 25

Fig. 7. Results of Zeng's method

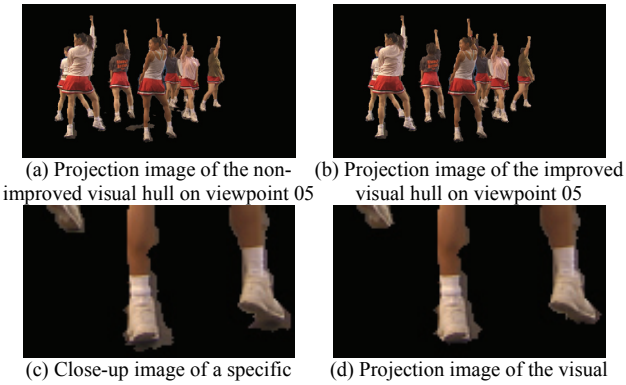
Additionally, we evaluated all the silhouette images from every camera viewpoint. For the proposed method, the GrabCut method, and Zeng's method, comparative results are shown in TABLE II. The difference in Recall values among the three methods is not large since all the methods are controlled to avoid false negatives in the early stage of processing to the extent possible. However, the difference in Precision values is significant, which proves that the proposed scheme is more effective than conventional methods.

TABLE II
COMPARISON OF QUANTITATIVE MEASUREMENT

	Recall	Precision	F-measure
Proposed method	0.951	0.924	0.937
GrabCut method	0.831	0.674	0.738
Zeng's method	0.965	0.376	0.541

B. Experiment 2:analyzation of contributions for refinement processes

The proposed scheme is assumed to be comprised by two key functions that are the binarization process of the voxel space with likelihood and the refinement process considering the 3D geometry. In order to clarify the contribution of the respective function, we analysed results of each process for proposed method. An example of projection images of the visual hull reconstructed by binarizing the voxel space is shown in Fig. 8 (a). This result corresponds to the performance without the refinement process. We can confirm that the reconstruction accuracy is comparatively high, though there remain a little false positives in the floor. On the other hand, Fig. 8 (b) shows the result when refinement process was additionally employed. The difference between Fig. 8 (a) and Fig. 8 (b) shows the improvement by refinement process of visual hull. For the specific region in Fig. 8 (b), the close-up image is shown in Fig. 8 (c). It shows that there remain shadow regions near objects of legs. We applied shadow removal process to all viewpoints, and reconstructed the visual hull using improved silhouettes. A projection image of the improved visual hull is shown in Fig. 8 (d). The difference between Fig. 8 (c) and Fig. 8 (d) indicates the improvement by shadow removal process.



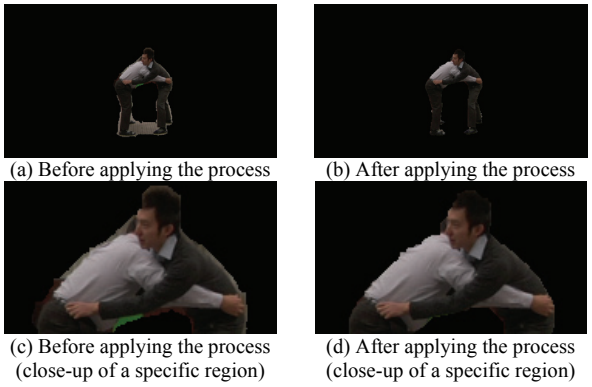
(a) Projection image of the non-improved visual hull on viewpoint 05 (b) Projection image of the improved visual hull on viewpoint 05

(c) Close-up image of a specific region in Fig.8(b) (d) Projection image of the visual hull whose shadow regions were removed on viewpoint 05

Fig. 8. Results of the refinement process

To evaluate the removal process for unwanted regions for silhouette images, we analysed the results for sequences B. An example of projection images of the visual hull reconstructed by binarizing the voxel space is shown in Fig. 9 (a). It is confirmed that false positives are remaining near object contours. The removal process was then applied for every

viewpoint, and the visual hull was reconstructed using improved silhouettes. A projection image of the improved visual hull to the corresponding viewpoint is shown in Fig. 9 (b). For the specific region in Fig. 9 (a) and Fig. 9 (b), close-up images are also shown in Fig. 9 (c) and Fig. 9 (d), respectively. Figure 9 shows that unwanted regions on the floor as well as neighboring contours have been removed appropriately.



(a) Before applying the process (b) After applying the process

(c) Before applying the process (close-up of a specific region) (d) After applying the process (close-up of a specific region)

Fig. 9. Results of the removal process for unwanted regions near contours

Furthermore, we quantitatively evaluated the projection images for both the non-improved visual hull and the improved visual hull in a similar way as in experiment 1, and the results are shown in TABLE III. The improvement of Precision values shows that false positives were successfully suppressed, and the flatness of Recall values suggests robustness for false negatives.

TABLE III
RESULTS OF THE REMOVAL PROCESS FOR UNWANTED REGIONS NEAR CONTOURS

	Recall	Precision	F-measure
Removal : OFF	0.996	0.805	0.890
Removal : ON	0.963	0.971	0.967

C. *Experiment 3:generation of free viewpoint images*

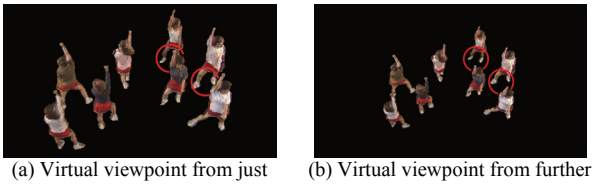
We generated a free viewpoint video for sequence A using the visual hull reconstructed by the proposed method. We made a polygon model by applying the marching cubes method,[11] and generated virtual viewpoint images by appropriate texture extraction from camera images and corresponding polygon mapping.

Two examples of generated virtual viewpoint images are shown in Fig. 10. We assumed two virtual viewpoints located above actual camera viewpoint 25. The results for the near viewpoint and the far viewpoint are shown in Fig. 10 (a) and Fig. 10 (b), respectively. Although there is a little error in the texture mapping, we can confirm that the occlusion regions of right-hand-side objects in Fig. 5 (b) are successfully generated in Fig. 10 (indicated by circles).

V. CONCLUSION

To realize highly precise 3D model reconstruction, we proposed a robust background subtraction method using the integrated information of multi-view images. As an inherent problem of the conventional schemes for 3D model reconstruction, the precision of the visual hull is highly dependent on the background subtraction result for the specific viewpoint. In order to overcome this problem, the proposed scheme employs two main features. One is determination of the background region based on the likelihood in a voxel space, and the other is the refinement of both visual hull and projection images considering 3D space geometry as well as visual information. From experimental results using actual multi-view images, it was confirmed that both key features greatly contributed to significant improvement compared with the conventional methods. Furthermore, it was also confirmed that the virtual viewpoint images were generated precisely while the occluded regions were reconstructed successfully.

As future works, we need to introduce a process that reduces the influence of estimated error of projection matrixes.



(a) Virtual viewpoint from just above viewpoint 25 (b) Virtual viewpoint from further above viewpoint 25

Fig. 10. Generated virtual viewpoint images

REFERENCES

[1] A. Ishikawa, M. P. Tehrani, S. Naito, S. Sakazawa, and A. Koike, "Free Viewpoint Video Generation for Walk-through Experience using Image-based Rendering", In Proc of the 16th ACM international conference on Multimedia, pp.1007-1008 (2008).

[2] T. Kanade, P. W. Rander, and P. J. Narayanan: "Virtualized Reality: Constructing Virtual Worlds from Real Scenes", IEEE Multimedia, vol. 4, no. 1, pp. 34-37 (1997).

[3] N. Inamoto and H. Saito: "Virtual Viewpoint Replay for a Soccer Match by View Interpolation from Multiple Cameras", IEEE trans. Multimedia, vol.9, no.6, pp.1155-1166 (2007).

[4] W. N. Martin and J. K. Aggarwal: "Volumetric Description of Objects from Multiple Views", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 5, no. 2, pp. 150-158 (1983).

[5] Y. Y. Boykov, M. P. Jolly, "Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images", IEEE conference on Computer Vision, CD-ROM.

[6] C. Rother, V. Kolmogorov, A. Blake: "GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts", In Proc of ACM SIGGRAPH, vol. 23, pp. 309-314 (2004).

[7] G. K. M. Cheung, T. Kanade, J. Y. Bouguet, and M. Holler, "A Real Time System for Robust 3d Voxel Reconstruction of Human Motions", IEEE conference on Computer Vision and Pattern Recognition, vol. 2, pp. 714-720 (2000).

[8] J. L. Landabaso, M. Pardas, and J. R. Casas, "Shape from Inconsistent Silhouette", Computer Vision and Image Understanding, vol. 112, no. 2, pp. 210-224 (2008).

[9] G. Zeng and L. Quan, "Silhouette Extraction from Multiple Images of an Unknown Background", Asian Computer on Computer Vision, pp.628- 633 (2004).

[10] K. N. Kutulakos and S. M. Seitz: "A Theory of Shape by Space Carving", Int. J. Comput. Vis., vol. 38, no. 3, pp. 192-218 (2000).

[11] W. E. Lorensen and H. E. Cline, "Marching Cubes: A High Resolution 3d Surface Construction Algorithm", In Proc of ACM SIGGRAPH, vol.21, no.4, pp.163-169 (1987).