Real-time Free-viewpoint Video Based on 3D Shape Reconstruction

Ueda, Megumu
Department of Intelligent Systems, Kyushu University

Nabeshima, Rui
Department of Intelligent Systems, Kyushu University

Arita, Daisaku
Department of Intelligent Systems, Kyushu University

Taniguchi, Rin-ichiro
Department of Intelligent Systems, Kyushu University

出版情報：Proceedings of the 1st Joint Workshop on Machine Perception and Robotics, 2005-09
バージョン：
権利関係：

Real-time Free-viewpoint Video Based on 3D Shape Reconstruction

M.Ueda R.Nabeshima D.Arita R.Taniguchi

Department of Intelligent Systems
Kyushu University
6-1, Kasuga-koen, Kasuga, Fukuoka 816-8580 Japan

Abstract

In this paper, we present a system generating free-viewpoint video in real-time using multiple cameras and a PC-cluster. Our system firstly reconstructs a shape model of objects by the visual cone intersection method, secondly transforms the shape model represented in terms of a voxel form into a triangular patch form, thirdly colors vertexes of triangular patches, and finally displays the shape-color model from the virtual viewpoint directed by a user. Here, we describe implementation details of our system and show some experimental results.

1. Introduction

Currently, televisions are used for real-time, or live, distribution of scenes of the world. In television, however, a video captured by a camera is displayed on a screen and the viewpoint is chosen only among camera positions specified by the program director, not among arbitrary positions. On the other hand, computer graphics techniques can generate a free-viewpoint video, in which a viewer can changed the viewpoint to arbitrary positions. However, computer graphics requires a structure and motion model of objects and it is time consuming to construct such a model in advance. Then, we aim to construct a computer graphics model by computer vision techniques in real-time for generating live free-viewpoint videos.

1.1. Previous Work

Several researches have been done for generating free-viewpoint videos using multiple cameras since Kanade et al.[1] had proposed the concept of "Virtualized Reality". We can classify the researches into two approaches. The first approach reconstructs 3D shapes of objects and the second one does not reconstruct them. We select the first approach because it can generate free-viewpoint videos with less cameras and less memories. As the first approach, Matsuyama et al.[2], Carranza et al.[3] and Davis et al.[4] have developed systems which generate a computer graphics model from multiple camera videos. Although they achieve relatively precise shape reconstruction and model coloring, it requires a lot of computation time, and, then, real-time processing can not be achieved. Gross et al.[5] and Grau et al.[6] have developed systems which generate a computer graphics model from multiple camera videos in real-time, sacrificing precision of 3D shape reconstruction. However, they color 3D shape model without judging a visibility from cameras, so model coloring makes error often. In contrast, our system can not only generate free-viewpoint videos in real-time by sacrificing precision of 3D shape reconstruction but also color 3D shape model with a new model coloring method.

2 Free-viewpoint Video Generation

It is quite time consuming to generate free-viewpoint videos, and, therefore, its online, or real-time, processing requires a high-performance computing system. We have employed a PC-cluster to implement a real-time free-viewpoint video system, because it provides quite a high cost performance based on parallel processing techniques. The key issue is programming methodology, especially for real-time algorithms, on a PC cluster. Here, we have implemented our system using RPV[7] on a PC-cluster, which is a programming environment for real-time image processing on a distributed parallel computer such as a PC-cluster.

If it is possible to divide a process into several packet-based processings, the process is divided and allotted to PCs, and the process is processed in parallel. Then we must consider process time and communication. We must make process time of each PC be almost the same and we must not make too many PCs since too many PCs make a large latency and the number of PCs is limited. So we have thought that following processes are best balance.

1. Reconstructing a shape model of objects by the visual cone intersection method[8].
2. Transforming the shape model represented in terms of a voxel form into a triangular patch form by the discrete marching cubes method[9].

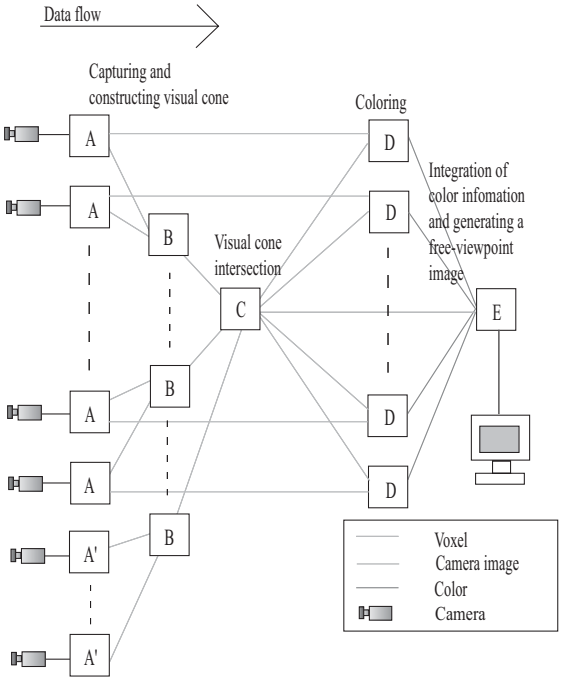


Figure 1: System configuration

3. Coloring vertexes of triangular patches varying with the position relation between the virtual viewpoint directed by a user and the viewpoints of cameras.

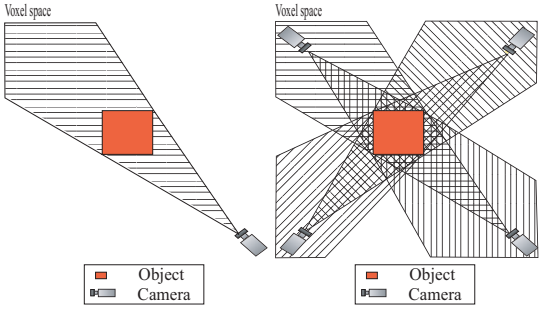
4. Displaying the shape-color model from the virtual viewpoint with painting triangular patches by interpolating among vertexes.

These processes are distributed to PCs shown in Fig. 1 and executed in pipeline parallel.

Node-A First, each node-A extracts object silhouettes from video frames captured by a camera by background subtraction and noise reduction. Secondly, each node-A constructs visual cones (Fig. 2(a)). A visual cone is defined as a cone whose apex is the viewpoint and whose cross section coincides with the silhouette of the object. Visual cones are represented in terms of a voxel space. Finally, each node-A sends the visual cones to a node-B and sends the colored silhouette image to a node-D.[1]

Node-B and Node-C Each node-B gathers and intersects visual cones from multiple viewpoints to construct a shape model of the objects represented in terms of a voxel space(Fig. 2(b)). Since receiving this large data at a time

[1]To reduce the computation time, some of node-As, indicated as node-A' in Fig. 1, do not send the colored silhouettes to reduce their redundancy. The selection is done in advance based on the camera arrangement.



(a) Visual cone construction

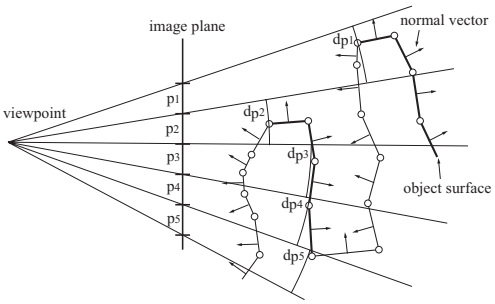
(b) Visual cone intersection

Figure 2: Visual cone intersection

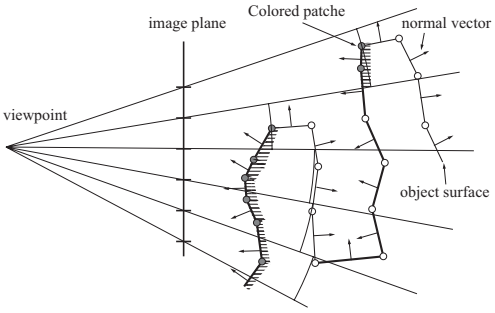
is time consuming, it is distributed to multiple node-Bs and Node-C hierarchically. Node-C transforms the finale shape model represented in terms of a voxel space into that in terms of triangular patches. However, node-C sends the voxel space and its corresponding patterns of the discrete marching cubes method instead of triangular patches since the triangular patch form is not efficient from the viewpoint of data size.

Node-D First each Node-D transforms the shape model represented in terms of a voxel space into those of triangular patches by the discrete marching cubes method using patterns sent from node-C. Then, each Node-D colors visible vertexes of the shape model based on one camera image. Finally, each Node-D sends color information of all vertexes of the shape model.

For coloring vertexes in real-time, it is necessary to quickly judge whether each vertex is visible from the camera or not. Conservative visibility check method has to check whether each vertex is occluded by each triangular patch. That computation amount is $O(N^2)$, where N is the number of vertexes. So we propose a new method based on the Z-buffer method, whose computation amount is $O(N)$. Our method consists of two steps (See Fig. 3). At the first step, Node-D searches for the object surface which faces against the viewpoint and which is nearest to the viewpoint in each pixel p. This step is realize by the Z-buffer method altered to taking account of not all surfaces but only surfaces facing against the viewpoint. Then, Node-D lets d_p be the distance between the viewpoint and the nearest surface. At the second step, Node-D colors all vertexes which faces toward the viewpoint and which is nearer to the viewpoint than d_p in each pixel p. The color of the vertexes is that of pixel p. At this time, each triangular patch is divided into six triangular patches as shown in Fig. 4 since increas-



(a) step 1



(b) step 2

Figure 3: Coloring vertexes

ing the number of vertexes makes coloring resolution higher without lengthening processing time for shape reconstruction.

Node-E First Node-E receives the position of the virtual viewpoint directed by a user.

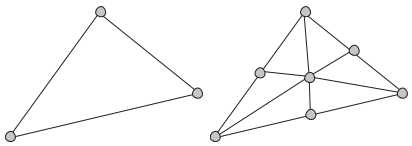
Secondly Node-E transforms the shape model represented in terms of a voxel space into those of triangular patches by the discrete marching cube method in the same way as Node-D. There are two reasons why shape model transformation is made on both Node-D and Node-E. The first one is because the data size of triangular patches is very large and the time to transport triangular patches is too long. The second one is because processing times of node-C, Node-D and Node-E are balanced best.

Thirdly, Node-E integrates color information of all cameras. The integrated color value for each vertex is weighted mean of color value from Node-D. The weight W_n of camera n is calculated by the following expression;

$$W_n = \frac{(cos\theta_n + 1)^\alpha}{\sum_{x=0}^{N}(cos\theta_x + 1)^\alpha} \quad (1)$$

where N is the number of cameras visible the vertex, θ_n is the angle between the vector from the virtual viewpoint to the vertex and that from the camera viewpoint to the vertex (See Fig. 5). By this equation, the more an angle between virtual viewpoint and a camera becomes small, the more the camera's weight becomes big. α is decided from balance between processing time of Node-E and other nodes. In this experiment, we let α be 5. Color value of a vertex visible from no camera is let be same as the mean value of neighbor vertexes.

Finally, Node-E generates an image from the directed viewpoint. Each triangular patch is painted by interpolating among vertexes.



(a) Original triangular patch (b) Divided triangular patch

Figure 4: Dividing triangular patch

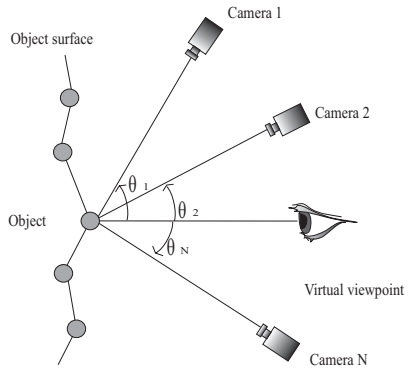


Figure 5: Angle between camera and virtual viewpoint

3 Experiments

Using our proposed system, we generate free-viewpoint video in real-time to evaluate the precision of generated images, processing time of each node, latency, and the amount

Table 1: Amount of data sending from each node

Node	Average (Kbyte)
A (Image)	93.5 (variable)
A and B (voxel)	256 (constant)
C	28.6 (variable)
D	92.0 (variable)

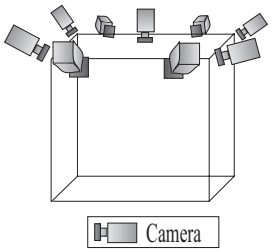


Figure 6: Camera arrangement

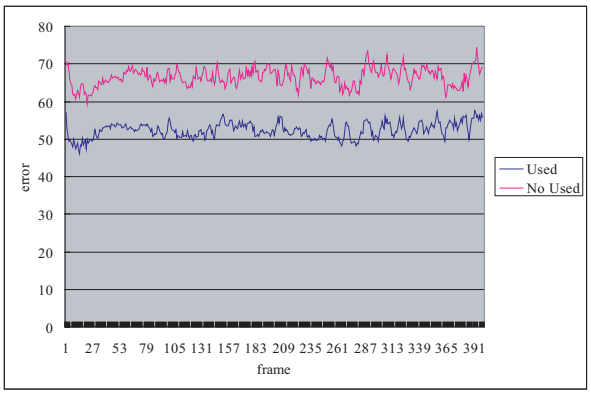


Figure 9: Error: ”Used” is cameras used for shape reconstruction and model coloring, ”No Used” is cameras used only for shape reconstruction

of data transfer. We have used seven IEEE-based cameras with 320×240 pixel resolution (see Fig. 6), and 20 PCs (six node-As, three node-A', three node-Bs, one node-C, six Node-Ds, one Node-E), each of which has an Intel Pentium4 (3GHz), 1GB memory and NVidia GeForce FX. PCs are connected by Myrinet. All the cameras are calibrated in advance by Tsai's method[10]. Camera resolution is 320×240. Voxel space resolution is $128 \times 128 \times 128$ and the size of a voxel is 2cm.

Fig. 7 shows two pairs of a camera image and a generated image whose viewpoints are same. Fig. 8 shows four generated images from virtual viewpoints. Each image is well-generated. Fig. 9 shows the sum of root mean square errors between a camera image and a generated image with a same viewpoint. This may be caused by

- shape reconstruction error,
- camera calibration error,
- individual difference between cameras,
- re-sampling error from image pixels to triangular patch vertexes, and
- color integration error.

Fig. 10 shows the mean of processing time of each node in case that there is one person in the experimental space. Fig. 11 shows the throughput of the system. The mean of the throughput is about 47fps and 38fps of one person and two persons respectively. The latency of the system is 150ms. We think balance of allotting processes is good from Fig. 10. Table. 1 shows the amount of data sending

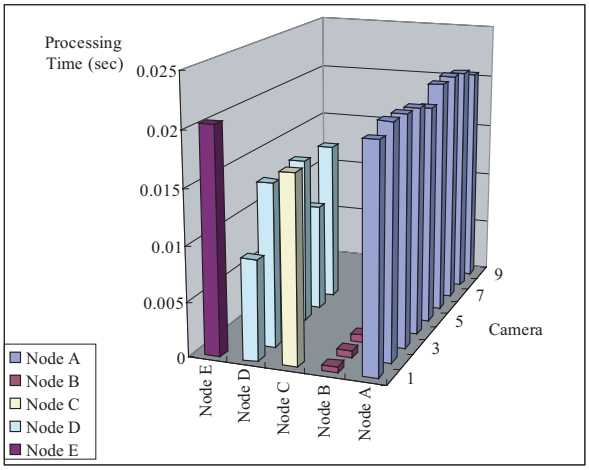
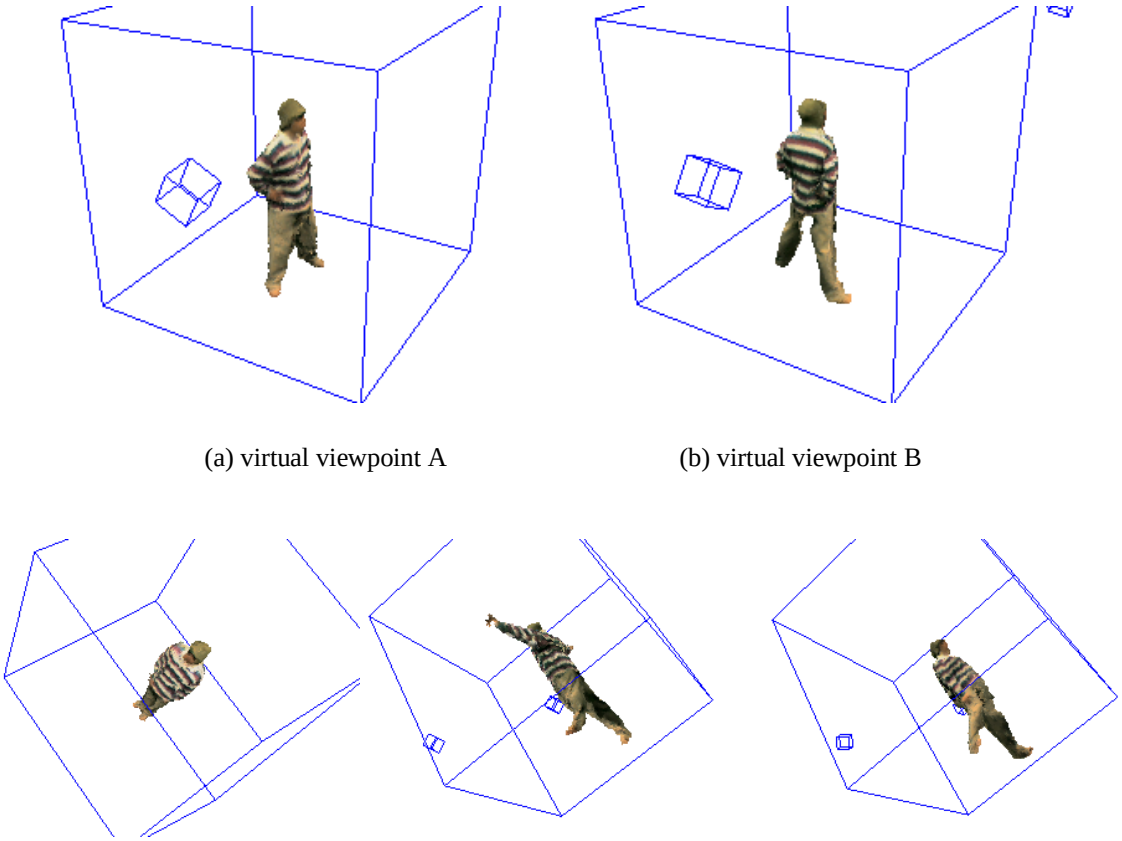


Figure 10: Processing time from each node

(a) viewpoint 1 (b) viewpoint 2

(c) viewpoint 3 (d) viewpoint 4

Figure 7: Camera images(upper) and generated images(lower)



(a) virtual viewpoint A (b) virtual viewpoint B

(c) virtual viewpoint C (d) virtual viewpoint D (e) virtual viewpoint E

Figure 8: Generated virtual-viewpoint images

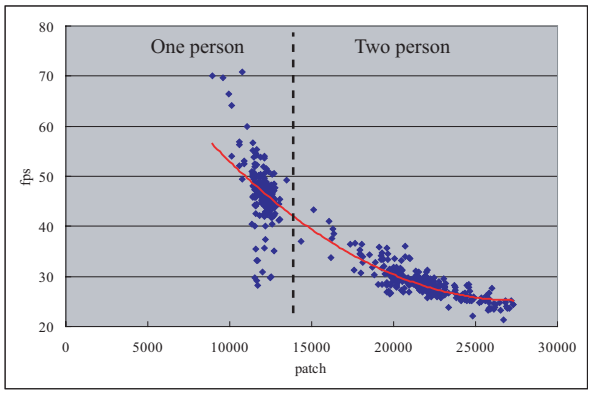


Figure 11: Relation between throughput and triangular patch

from each node. Node-B and Node-C receives the largest amount of data, 768KB/frame, of all nodes. This data size requires 6.1ms for receiving via Myrinet. We think that this data size is large because the sending time is equivalent to about 1/3 of the throughput. However, enough performance, 47fps or less is realized by using a PC-cluster and communication dose not influence the throughput so much.

4 Conclusion

In this paper, we propose a system generating free-viewpoint videos using multiple cameras and a PC-cluster in real-time. And we make some experiments to show the performance and the precision of our system.

Major future works especially from the view point of parallel/distributed processing are as follows.

Reduction of latency

The latency of the current system is about 200ms. That value is not small. We think that packet-based processing and data compression is effective for reduction of

latency. Currently, the packet-based processing is introduced only from node-A to node-C on the current system. Therefore, we will introduce it to all nodes.

Invariable throughput

The throughput of our system varies depending on the scenes since the number of voxels and the number of triangular patches depend on the size and the shape of objects. By introducing variable resolution of the voxel space we suppose we can make the throughput relatively unvaried.

Precise shape reconstruction

The more precise shape is reconstructed, the more photographic generated images become.

References

[1] P. J. Narayanan and T. Kanade and P. W. Rander: ”Concepts and early results”, Proc. IEEE Workshop on the Representation of Visual Scenes, pp. 69–76, jun, 1995.

[2] Takashi Matsuyama and Xiaojun Wu, Takeshi Takai and Shohei Nobuhara: ”Real-time generation and high fidelity visualization of 3d video”, Proc. of MIRAGE2003, pp. 1–10, Mar, 2003.

[3] Joel Carranza and Christian Theobalt and Marcus A. Magnor and Hans-Peter Seidel: ”Free-viewpoint video of human actors”, ACM Trans. on Graphics, pp. 569–577, Vol. 22, No. 3, Jul, 2003.

[4] E.Borovikov and L.Davis: ”A Distributed System for Real-time Volume Reconstruction”, Proc. 5th Int. Workshop on Computer Architecture for Machine Perception (CAMP2000), pp. 183–189, 2000.

[5] M. Gross, S. Wurmlin, M. Naef, E. Lamboray, C.Spagno, A.Kunz, E. Koller-Meier, T. Svoboda Luc Van Gool, S. Lang, K. Strehlke, A. Vande Moere, O, Staad, ”blue-c: A Spatially Immersive Display and 3D Video Portal for Telepresence”, Proceedings of ACM SIGGRAPH2003, pp. 819–827, 2003.

[6] O. Grau, T. Pullen, G. A. Thomas, ”A Combined Studio Production System for 3D Capturing of Live Action and Immersive Actor Feedback”, IEEE Transactions on Circuits and Systems for Video Technology, Volume 14, No. 3, pp. 370–380, March 2004.

[7] Daisaku Arita and Rin-ichiro Taniguchi: ”RPV-II: A stream-based real-time parallel vision system and its application to real-time volume reconstruction”, Proc. of Second International Workshop on Computer Vision System, pp. 174–189, Jul, 2001.

[8] W. N. Martin and J. K. Aggarwal: ”Volumetric description of objects from multiple views”, IEEE Trans. on Pattern Analysis and Machine Intelligence, pp. 150–158, Vol. 5, No. 2, 1983.

[9] Yukiko Kenmochi and Kazunori Kotani and Atsushi Imiya: ”Marching cubes method with connectivity”, Proc. on International Conference on Image Processing, pp. 361–365, Vol. 4, Oct, 1999.

[10] Roger Y. Tsai: ”A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses”, IEEE Trans. on Robotics and Automation, pp. 323–344, Vol. 3, No. 4, 1987.