Free-viewpoint Video Synthesis Technology for a New Video Viewing Experience

Kazuki Okami, Kouta Takeuchi, Megumi Isogai, and Hideaki Kimata

Abstract

There is a growing desire among viewers of sports programs to watch videos of matches with an even higher sense of presence. Free-viewpoint video synthesis has been attracting attention as one scheme for achieving a video viewing experience with an ultrahigh sense of presence. This technology reconstructs the match space targeted for viewing and provides video from the viewpoint of the user's choosing. This article describes the technologies for free-viewpoint video synthesis developed by the NTT laboratories and introduces application examples.

Keywords: free-viewpoint video, video viewing, image processing

1. Introduction

Free-viewpoint video synthesis enables the space targeted for a sports competition or event to be viewed from an arbitrary viewpoint as desired by the user. The viewpoint of synthesized video is not restricted by the arrangement of the cameras used for shooting. That is to say, the technology achieves video viewing from a *free* viewpoint; it does not simply switch cameras or interpolate between them. For example, in a soccer match, free-viewpoint video synthesis can achieve video viewing even from positions where no cameras are installed such as a viewpoint alongside a player standing in the field or a viewpoint tracking the ball. In short, free-viewpoint video synthesis can provide the user with a viewing experience that exceeds the experience of actually being at the stadium.

2. Flow of free-viewpoint video synthesis and technology trends

Various methods of free-viewpoint video synthesis have been developed [1]. We introduce a geometry-based technique that we have been working on. The process begins with capturing the scene desired for viewing and restoring three-dimensional shape and texture (image) information. Then, at viewing time, the system renders an image of such three-dimensional shapes and textures watched from the desired viewpoint.

To give some background, we describe Eye Vision, a well-known viewing system used in the 2001 Super Bowl telecast in the United States. The technology at that time involved shooting the target scene from various positions and directions by installing a large number of cameras on a scale of several tens of units and estimating three-dimensional information by performing image processing such as stereo matching [2] with the multi-view images obtained.

Today, however, high-definition cameras are becoming commonplace, so shooting a scene with many cameras and carrying out image processing using many multi-view images is considered to be prohibitive in time and cost.

However, a method developed recently can use both color images taken with ordinary cameras and depth images obtained with depth sensors [3]. Here, a depth

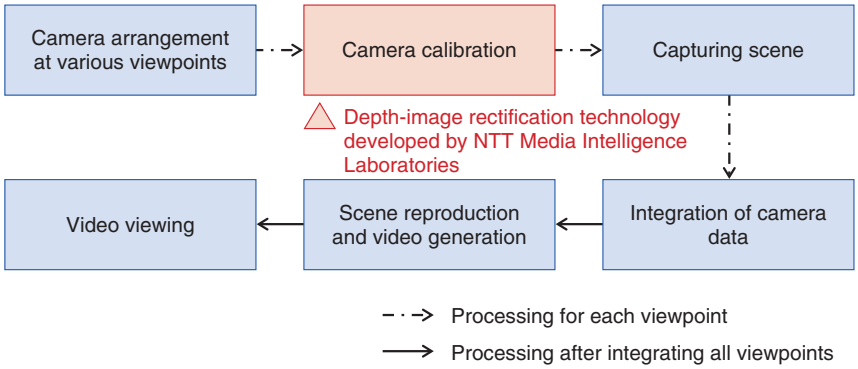


Fig. 1. Free-viewpoint video synthesis technology.

image represents the distance to an object measured by a depth sensor in the form of an image. It can directly represent the three-dimensional shape of an object from the shooting direction, which means that free-viewpoint video synthesis can be achieved with fewer cameras than the conventional technique of estimating three-dimensional information from multi-view images.

In addition, low-cost image-pickup devices equipped with a depth sensor have recently become available such as the Microsoft Kinect (Kinect). Such devices can be used to construct a low-cost system for shooting and synthesizing video from arbitrary viewpoints.

As described above, achieving free-viewpoint video synthesis in a wide-area space in a stadium or other venue is a near-future objective, but the first step here will be to use a depth sensor such as Kinect to establish free-viewpoint video synthesis technology targeting the relatively small space surrounding a person. The rest of this article describes the technology being researched and developed at NTT Media Intelligence Laboratories, presents a demonstration system, and introduces application examples.

3. Depth-image rectification technology for free-viewpoint video synthesis using depth sensors

We introduce here a small-scale demonstration system we constructed for researching free-viewpoint video synthesis using four Kinect units, each with a combined camera and depth sensor. The system flow and the shooting environment are respectively shown in **Figs. 1** and **2**. This system uses only a few compact Kinect devices as shooting cameras, making it possible to synthesize free-viewpoint video with high portability and at low cost. However, the mechanism and performance of this depth sensor and other factors limit the subject to a small space of about three square meters.

The four Kinect devices are each connected to a client personal computer (client PC). The system captures both color and depth images of the subject in the space and sends those images to the synthesis PC via a network switch. Then, by creating a correspondence between each camera and image, the system reproduces the captured subject as point-cloud data (a collection of points) or as mesh data that smoothly connect the point-cloud data.

We explain here the key elements of this technology that we are researching and developing. To obtain point-cloud data of the subject from color and depth images, it is necessary to calculate beforehand camera parameters such as the relative positions, orientations, and lens-related parameters of the cameras for both color images and depth images. This parameter-calculation process is called calibration. Specifically, there are two types of calibration for each camera: external and internal. External calibration is a process that determines the position of a camera. It must be performed to integrate the depth images received from each Kinect unit at their correct positions.

Internal calibration, meanwhile, corrects for lens distortion in the images captured by each camera. This calibration must be performed to obtain accurate images and point-cloud data.

Calibration using only color images has been used for a long time, and parameters have been obtained by performing the above external and internal calibration using a technique [4] featuring a structure with a fixed pattern (such as a black and white checkerboard).

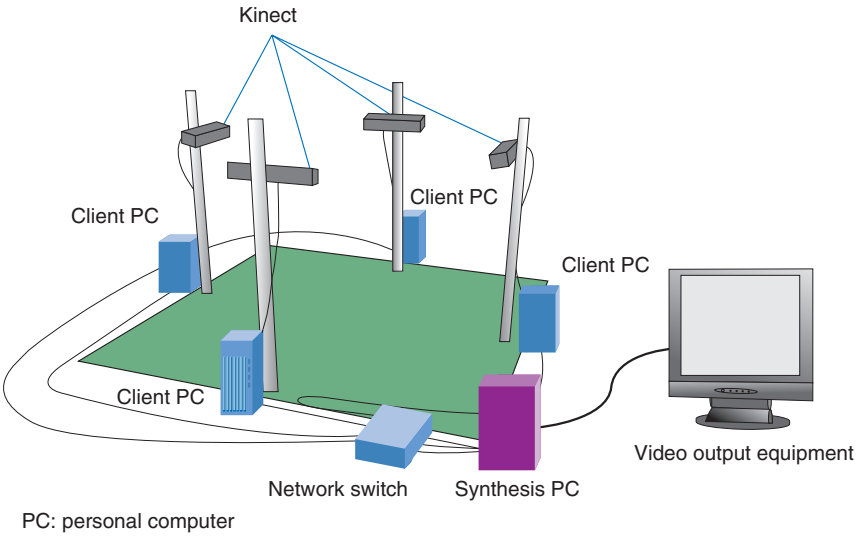


Fig. 2. Shooting environment.

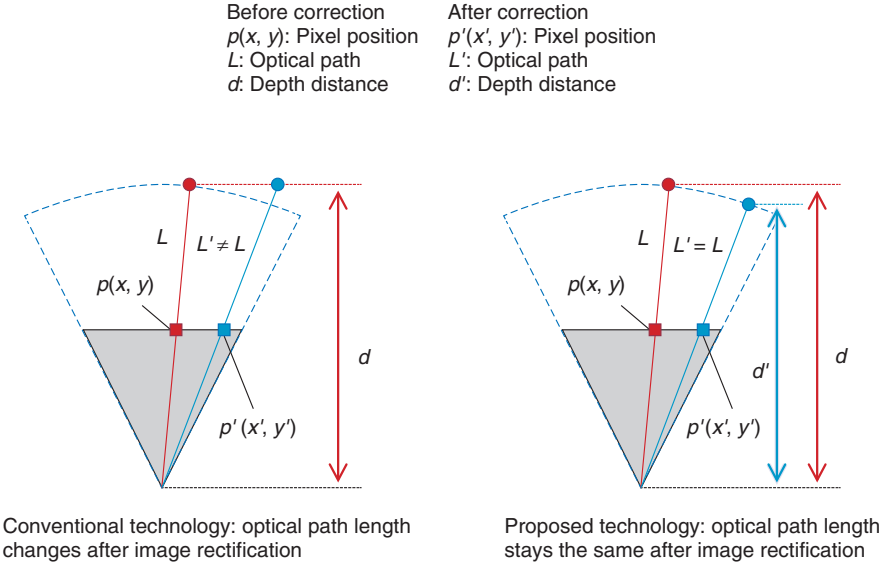


Fig. 3. Conceptual diagram of depth-image correction technology.

However, the depth images cannot be sufficiently corrected by conventional internal calibration, which means that accurate depth images cannot be obtained. This problem originates in the technique used to obtain the depth images. A Kinect device irradiates the target object with infrared light, obtains the optical path length by measuring the time taken for the light to be reflected and to return, and calculates the perpendicular distance to the object from the optical path length as the depth value. Consequently, if we use only camera lens distortion taken into account by conventional internal calibration, the result is a depth value calculated from an optical path different from the correct one, as shown in **Fig. 3**.

To solve this problem, we proposed a technique for depth-image rectification [5] that correctly calculates depth values after lens rectification by using mapping positions before and after lens distortion rectification

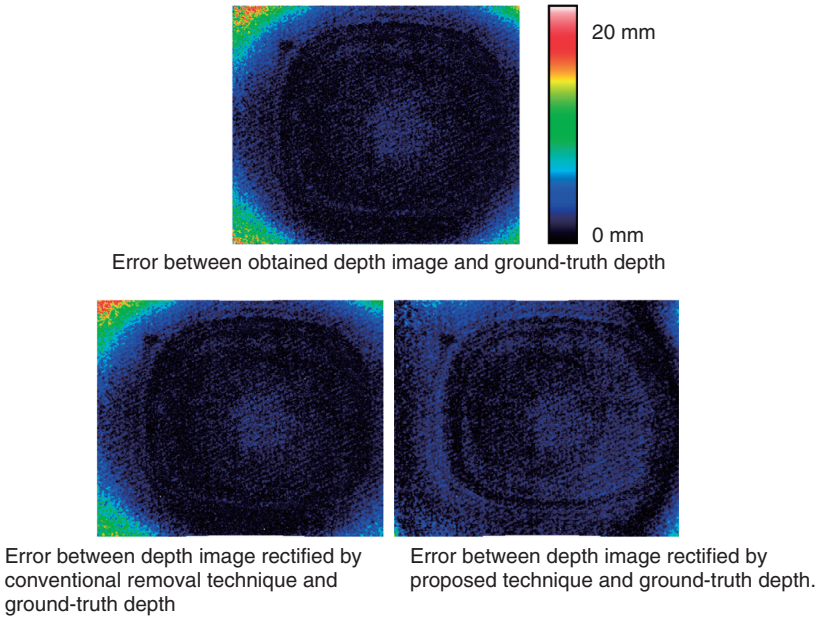


Fig. 4. Comparisons of error with ground-truth depth of depth image.

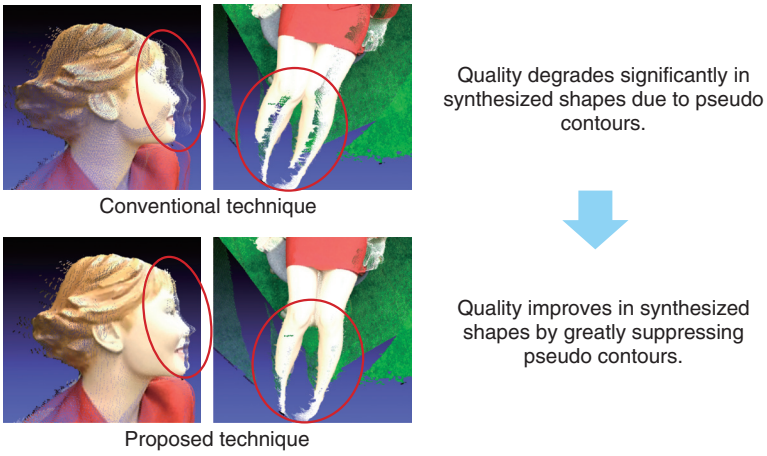


Fig. 5. Comparison of results applying depth-image rectification techniques.

and constraints that prevent the obtained optical-path values from changing. This technique makes it possible to obtain depth images with less error than that using the conventional calibration technique (**Fig. 4**).

A comparison of reconstruction results using the conventional technique and our proposed technique is shown in **Fig. 5**. The conventional technique results in quality degradation due to the appearance of pseudo contours of a subject. In contrast, the proposed technique greatly suppresses pseudo contours, thereby significantly contributing to improved quality in reproducing the three-dimensional shape of the subject.

4. Applications of free-viewpoint video synthesis in small spaces

We are considering a real-time system and an off-line one for free-viewpoint video synthesis application in small spaces. The real-time system displays

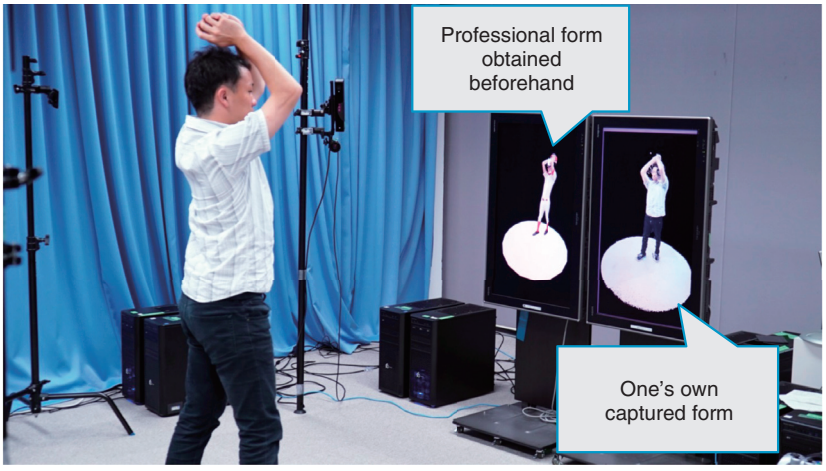


Fig. 6. Application to self-check one's form in real time.

the images obtained from the Kinect units as point-cloud data in real time. For example, we can imagine an application that would enable an athlete or dancer to check his or her form in real time. As shown in **Fig. 6**, lining up one's own actions in the present with those of a professional athlete obtained beforehand makes it possible to see where one might have poor form (e.g., dropping of the elbow compared with that of a professional baseball pitcher). Moreover, as this application would perform three-dimensional reconstruction different from simple images or video, it would make it possible to compare postures from an arbitrary viewpoint that would not normally be available. An application of this type should provide a more efficient means of correcting defects in one's athletic or artistic form.

The off-line system, on the other hand, would first integrate the images obtained from the Kinects as point-cloud data on the synthesis PC and then perform mesh processing to create surface data from the point-cloud data. Then the system would render the surface data instead of simply rendering the point-cloud data. This processing would enable the rendering of a three-dimensional model of even higher quality and the reconstruction of a high-density, smooth three-dimensional model of the subject (**Fig. 7**).

At present, however, the computational cost is too high to achieve real-time rendering. The off-line application could therefore be used to view three-dimensional shapes of a subject at a leisurely pace after shooting has been completed. One example of such an application would be self-checking of one's

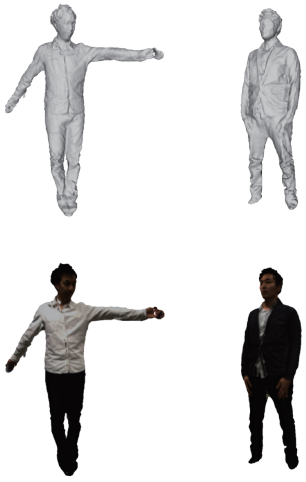


Fig. 7. Three-dimensional reproduction using proposed technology.

form just as in the real-time application. Here, however, instead of checking one's form instantly, the application could be used to save one's own model on a device such as a smartphone and review it whenever and wherever one likes.

It would also be possible to place a three-dimensional model of oneself in a virtual space and use a head-mounted display (HMD) to check one's life-size form. Another application differing from the real-time applications is to use it as a three-dimensional photo. This application could be used in many enjoyable ways such as lining up three-dimensional models of oneself and one's friends, viewing them

from any direction on a smartphone or HMD, and adding lighting effects.

5. Future development

As the first step to achieving free-viewpoint video synthesis, we researched and developed technologies for synthesizing small spaces with an emphasis on portability while envisioning applications for trouble-free capturing and rendering of the space surrounding a person. In this article, we outlined free-viewpoint video synthesis technology and a demonstration system and introduced promising application examples. As the next step, we plan to target larger spaces such as stadiums and provide a video viewing experience that can make the viewer feel like a user standing in any location even if no cameras are installed at that point. Achieving such an application will require free-viewpoint video synthesis oriented to larger spaces, and to this end, we are researching and developing technologies for capturing and synthesizing objects in such an environment.

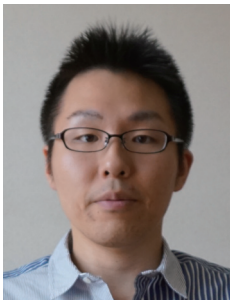
Furthermore, while the free-viewpoint video synthesis system that we have so far researched and developed performs synthesis processing on site, that is, at the capturing location, we are now developing techniques for transferring data to the cloud over the network and performing synthesis processing on the cloud. This will enable viewing of free-viewpoint video content from remote locations and provide a more enjoyable viewing experience.

References

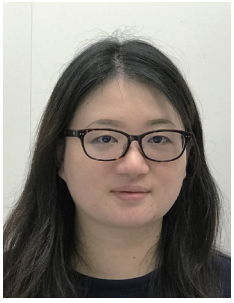
[1] A. Smolic, "3D Video and Free Viewpoint Video—From Capture to Display," Pattern Recognit., Vol. 44, No. 9, pp. 1958–1968, 2011.
[2] C. Zitnick and T. Kanade, "A Cooperative Algorithm for Stereo Matching and Occlusion Detection," IEEE Trans. Pattern Anal., Vol. 22, No. 7, pp. 675–684, 2000.
[3] D. Alexiadis, D. Zarpalas, and P. Daras, "Fast and Smooth 3D Reconstruction Using Multiple RGB-Depth Sensors," Proc. of the IEEE Visual Communications and Image Processing Conference 2014, pp. 173–176, Valletta, Malta, Dec. 2014.
[4] Z. Zhang, "A Flexible New Technique for Camera Calibration," IEEE Trans. Pattern Anal., Vol. 22, No. 11, pp. 1330–1334, 2000.
[5] A. Louie, K. Takeuchi, N. Ito, and A. Kojima, "Multiple Plane View Camera Calibration for RGB-D Sensor Rectification," IEICE Tech. Rep., Vol. 115, No. 350, IE2015-93, pp. 81–86, 2015.

Trademark notes

Kazuki Okami
Researcher, Visual Media Project, NTT Media Intelligence Laboratories.
He received a B.E. and M.E. from Waseda University, Tokyo, in 2012 and 2014. He joined NTT Media Intelligence Laboratories in 2014, where he has been engaged in research and development of free-viewpoint video synthesis.



Kouta Takeuchi
Engineer, NTT Plala Inc.
He received a B.E. and M.E. in engineering from Nagoya University, Aichi, in 2009 and 2011. He joined NTT in 2011 and conducted research on free-viewpoint video synthesis and super-resolution of a light field camera. Since July 2017, he has been with NTT Plala and working on a set-top box development project.



Megumi Isogai
Senior Research Engineer, Visual Media Project, NTT Media Intelligence Laboratories.
She received a B.E., M.E., and Ph.D. in communication network engineering from Okayama University in 2004, 2005, and 2010. She joined NTT Cyber Space Laboratories (now, NTT Media Intelligence Laboratories) in 2006 and has been studying three-dimensional video processing and high-reality communication technology.

Hideaki Kimata
Senior Research Engineer, Supervisor, High-Reality Visual Communication Group Leader, NTT Media Intelligence Laboratories.
He received a B.E. and M.E. in applied physics in 1993 and 1995 and a Ph.D. in electrical engineering in 2006 from Nagoya University, Aichi. He joined NTT in 1995 and has been researching and developing a video coding algorithm, visual communication systems, and machine learning. He is a Chief Examiner of the Information Processing Society of Japan (IPSJ) Special Interest Group on Audio Visual and Multimedia information processing.