Immersive Telepresence System Using High-resolution Omnidirectional Video with Locomotion Interface

Sei IKEDA, Tomokazu SATO, Masayuki KANBARA and Naokazu YOKOYA
Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara, 630-0192 Japan
sei-i@is.naist.jp, tomoka-s@is.naist.jp, kanbara@is.naist.jp, yokoya@is.naist.jp

Abstract

A telepresence system using real environment images is expected to be used in a number of fields such as entertainment, medicine and education. This paper describes a novel telepresence system which enables users to walk through a photorealistic virtualized environment by actual walking. To realize such a system, a wide-angle high-resolution video is projected on an immersive multi-screen display to present users the virtualized environments and a treadmill is controlled according to detected user's locomotion. In this study, we use an omnidirectional multi-camera system to acquire images of real outdoor scenes. The proposed system provides users with rich sense of walking in a remote site.

Key words: Telepresence, Omnidirectional Multi-camera System, Multi-screen Display, Treadmill

1. Introduction

Technology that enables users to experience a remote site virtually is called telepresence [1]. A telepresence system using real environment images is expected to be used in a number of fields such as entertainment, medicine and education. The telepresence system especially using an image-based technique attracts much attention because it can represent complex scenes such as outdoor environments. Our ultimate form of telepresence is an immersive system in which users can naturally move and look anywhere by their actions in a virtualized environment reproduced from a real environment faithfully. However such an ideal system does not exist today.

Conventional telepresence systems have two important problems. One is that high human cost is required to acquire images and to generate virtualized environments in the case of large-scale outdoor environments. The other is concerned with presentation of virtualized environments. Chen [2] has developed an image-based telepresence system: QuickTime VR that generates a virtualized environment as panoramic images. In his system, users can see any directions and move their view positions in the environment through a standard display. Panoramic images are generated from multiple standard still images by using a mosaicing technique. The image acquisition task takes much time and effort to enable users to move their view positions in a wide area. Moreover, standard displays are not suitable for giving the feeling of virtually walking in remote sites.

In some recent works such as [3], omnidirectional video camera systems are used to acquire panoramic images at various positions and to reduce human cost in acquisition of images. In a telepresence system developed by Onoe, et al. [3], multiple users can look around a scene of remote site in real time. They used an omnidirectional camera combining a standard single lens camera and a curved mirror. A part of omnidirectional video is displayed to users according to their view directions through a head mounted display.

More recent works [4-7] have improved the resolution of omnidirectional videos using multiple cameras. Kotake, et al. [4] developed a telepresence system using an immersive three-screen display. They used a multi-camera system radially arranged on a moving car to acquire high-resolution panoramic videos of an outdoor scene. In the immersive display, users can see a wide field of view direction to provide the feeling of high presence in remote sites. However, a game controller is used to move the view position. This system provides users with the sense of riding a carriage rather than the sense of walking in a virtualized environment of a real outdoor scene.

In this paper, we propose a novel telepresence system which enables a user to move by actual walking and change his view point in a photorealistic virtualized environment using a high-resolution omnidirectional video. For this system, first, videos of outdoor scenes are acquired by an omnidirectional multi-camera system (OMS) without much human cost in acquisition of images. Virtualized environment videos are then generated by using an image-based representation from the captured multiple videos. Calibration and motion estimation of the OMS are performed in advance. Finally, generated virtualized environment videos are projected on an immersive multi-screen display according to user's locomotion detected on a treadmill.

2. Immersive Telepresence System Using High-resolution Omnidirectional Videos

This section describes a method for realizing a telepresence system with a locomotion interface. Our system consists of three processes: acquisition of images, generation of a virtualized environment and presentation to users. For the first process, we use an OMS constructed of multiple camera units to reduce human cost in acquisition of images. In the second process, a virtualized environment is reconstructed as video streams from captured multiple image sequences. Geometric and photometric calibration of the OMS is required in order to generate a virtualized environment. Motion of the OMS is also estimated to reduce shake effect of video and to correct replay speed of the video streams. The final process is to present it to users. We use an immersive three-screen display with a treadmill to present a generated virtualized environment to users, as shown in Figure 1. The following sections describe details of these processes.

2.1 Acquisition of Images of a Real Dynamic Scene

An OMS has an important advantage that human cost in acquisition of images can be reduced because of the following two reasons. One is that the OMS has a wide field of view. The other is that the total resolution of images captured by the OMS is usually higher than an omnidirectional camera system using a single camera.

For the acquisition of images, videos are captured by an OMS Ladybug [8] mounted on a moving electric wheel chair in outdoor environments, as shown in Figure 2. This camera system obtains six 768x1024 images at 15 fps; five for horizontal views and one for a vertical view. In this work, the OMS is fixed on a tripod at the height of human eyes and the speed of the wheel chair is kept as constant as possible for simplification of the process of presentation to users because presentation of virtualized environments should be controlled according to user's locomotion regardless of variation of the car speed in the presentation process. Motion of OMS is estimated to compensate the shortage of the accuracy of chair control.

2.2 Generation of Virtualized Environment

In this section, virtualized environment videos are generated according to the shape of an immersive screen of a telepresence system by using the method in [6]. Geometric and photometric calibrations are required to generate videos of virtualized environments automatically. Motion estimation of OMS is also required to remove shaking the effect from the acquired videos and to correct replay the speed of virtualized environment videos. The following paragraphs describe the calibration and the motion estimation of OMS and the generation of virtualized environment videos.

Calibration of OMS

In the geometric calibration, intrinsic parameters (focal length, lens distortion parameters, center of distortion, aspect ratio) and extrinsic parameters (relative position, orientation) of each camera unit *i* should be estimated. First, 3D positions of many markers are measured by using a calibration board and a total station. 3D coordinates of three corners of the calibration board are measured by the total station, and all 3D positions of the markers on the board are calculated by linear interpolation among its corners. This method can virtually arrange markers around an OMS simultaneously. Intrinsic and extrinsic parameters of each camera are estimated using the obtained pairs of 3D and 2D positions of markers in the captured images [6]. Then, each relation $\mathbf{T}_i$ of position and orientation between one reference camera unit 0 and the rest ($i \neq 0$) is calculated from the estimated extrinsic camera parameter $\mathbf{E}_i$ represented by a 4x4 projection matrix in a homogeneous coordinate system, as follows.

$$\mathbf{T}_i = \mathbf{E}_0 \mathbf{E}_i^{-1}.$$

In the photometric calibration, the limb darkening of each camera and color balances among multiple cameras are corrected. The strength of limb darkening is calculated from estimated intrinsic parameters. The difference of color balances between camera units is measured by a histogram matching based on assuming a linear relation between radiance and irradiance.

Estimation of Camera Motion Parameters

In the proposed method, an electric wheel chair is used to acquire images. However, as mentioned earlier, it is difficult to control the position, orientation and speed of

Figure 1. Appearance of the system.

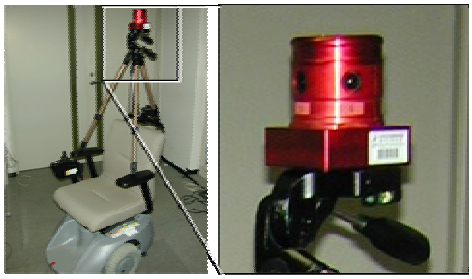


Figure 2. Image capturing System.

an OMS precisely. In this step, the motion of OMS is estimated from the acquired images to correct the shake effect of video and the replay speed of a virtualized video.

Motion of OMS is estimated by tracking both feature landmarks and image features in input video frames automatically [10]. The motion defined by a 4x4 matrix $\mathbf{M}_j$ means translation and rotation of OMS of j th frame. Feature landmarks mean image features whose 3-D positions are known. In this method, the position and orientation of all the cameras are calculated and optimized simultaneously, so that projection errors of the other feature landmarks and features points in images of all the cameras are minimized. This method makes it possible estimate camera motion more precisely than conventional methods using a single camera system since the motion of OMS is constrained by feature landmarks and natural features existing in all the directions. In this method, a small number of feature landmarks need to be visible and to be specified manually in key frames for minimizing accumulative errors.

Generation of Virtualized Environment Video

This step is based on re-projecting input images to a virtual image surface which corresponds to the shape of immersive screens for presentation. In advance, the limb darkening and the color balance of input images are corrected. Corrected images are then projected on a projection surface by using the parameters obtained by the geometric calibration and motion parameters of an OMS. This section first describes reduction of shake effect of acquired videos. The correction of replay speed is then described. Finally, the effect related to the violation of single view point constraint of the OMS is described.

Sake effect reduction: Shaking effect of video is caused by rotation and translation of a camera system. In this step, the rotation factor is canceled to reduce this effect by using the estimated motion parameters of an OMS. The projection method of image is formulated by a direction vector $\mathbf{p} = [p_x, p_y, p_z, 0]^t$ from the user's view point to a point on a display and the corresponding position (*u, v*) in an undistorted image, as follows,

$$[au, av, a, 0]^t = \mathbf{T}_i \mathbf{M}_j \mathbf{R}_j(\theta)\mathbf{p},$$

where the rotation matrix $\mathbf{R}_j(\theta)$ rotates a point by the angle θ around a vertical axis. The angle θ is also defined as shown in Figure 3. This matrix $\mathbf{R}_j(\theta)$ is used to replay videos so that a part corresponding to user's advance direction of the generated video is displayed in front of the user.

Replay speed correction: Frame indices of the input video are re-assigned so that the relation between user's locomotion and the displayed frame index become linear in the process of presentation to users. This correction can be easily performed by using translation parts of the estimated motion parameters.

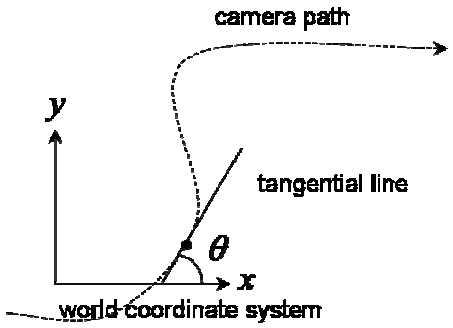


Figure 3. Correction of shake effect of OMS.

Effect of violation of single view point constraint: The single viewpoint perspective projection model is not applicable for this camera system since the centers of projection of six camera units of the OMS are different from each other. However, when the distance of a target from the system is sufficiently large, the centers of projection can be considered as the same. Therefore, we assume that the target scene is far enough from the OMS and set the projection surface far enough from the camera system. A frame of a virtualized environment video is generated by projecting all the pixels of all the six input images onto the projection surface. Note that a blending technique is used for generating a smooth image, when a point on the projection surface is projected from multiple images of different cameras. Although there is no exact definition of resolution for the above reason, the total horizontal resolution of an omnidirectional video acquired by Ladybug can be approximated as about 3340 pixels, assuming a horizontal 13% overlapping region between two adjacent cameras.

2.2 Presentation to Users

This section describes a method for presenting virtualized environment videos generated in the previous section. As shown in Figure 4, our system is composed of (a) a locomotion interface, (b) a graphics PC cluster and (c) an immersive three-screen display. The locomotion interface detects user's locomotion as input data, and sends calculated displacement information to the PC cluster. The PC cluster draws twelve images synchronized with the speed of user's walk because each screen image is generated by four projectors. As output data, these videos are displayed according to the user's motion. The scene in presented videos is appropriately changed according to the user's walking. The display system is described in some more detail below.

Locomotion Interface: This interface is composed of a treadmill, a couple of 3-D position sensors and a PC for

control as shown in Figure 4. and Table 1. User's locomotion is detected by two 3-D position sensors fixed on user's legs (Figure 4 (1)). The treadmill is controlled by the control PC based on position information from the sensors (Figure 4 (2)). The belt of the treadmill is automatically rotated so that the center of gravity of two sensors coincides with the center of the belt area [9]. This virtually realizes an infinite floor. Although a user can walk in any direction on this device, only the forward and the backward directions are used for the present system. The control PC calculates the displacement of user's position and sends it to the graphics PC cluster (Figure 4 (3)).

Graphics PC Cluster: The graphics PC cluster is composed of twelve PCs. Each graphics PC is networked through 100Mbps LAN and is controlled by the control PC. The control PC sends frame indexes to twelve PCs using the UDP protocol simultaneously (Figure 4 (3)). Each machine draws synchronized frame images according to the user's motion (Figure 4 (4)). Simultaneously the frames are interpolated by a blending technique between frames when a user walks slowly. Note that the images are accumulated in local hard disk in advance and only the frame index is carried via network.

Immersive Display: The immersive display is composed of three slanted rear-projection screens (Solidray VisualValley) and twelve projectors. To obtain a wide field of view, the screens are located in user's front, left and right sides. To realize high-resolution image projection, each screen image is made by four projectors. The resolution of each projector is 1024x768 (XGA) pixels. Because there are some overlapping areas projected by multiple projectors and some areas are not projected on the screen, the resolution of each screen is potentially about 2 million pixels.

3. Experiment

In the experiment, an omnidirectional video was actually obtained in an outdoor scene by Ladybug mounted on the electric wheel chair controlled manually. The speed of the chair is kept approximately constant 0.6m/sec. A telepresence system was also prototyped by using generated videos whose shake effect and replay speed was corrected. Figure 5 shows an example set of input images. Generated videos corresponding to the twelve projectors are also shown in Figure 6. The resolution of each generated video is set as 480x360 so as to be almost the same as the actual resolution of Ladybug assuming the 15% overlapping regions projected by two adjacent projectors.

Table 1. Components of the display system

instrument	specification
Control PC	CPU: Intel Pentium4 2.4GHz
Position Sensors	Polhemus Fastrak
Treadmill	WalkMaster
Graphics PC	CPU: Intel Pentium4 1.8GHz, Graphics Card: Geforce4 Ti4600
Immersive Display	Solidray VisualValley

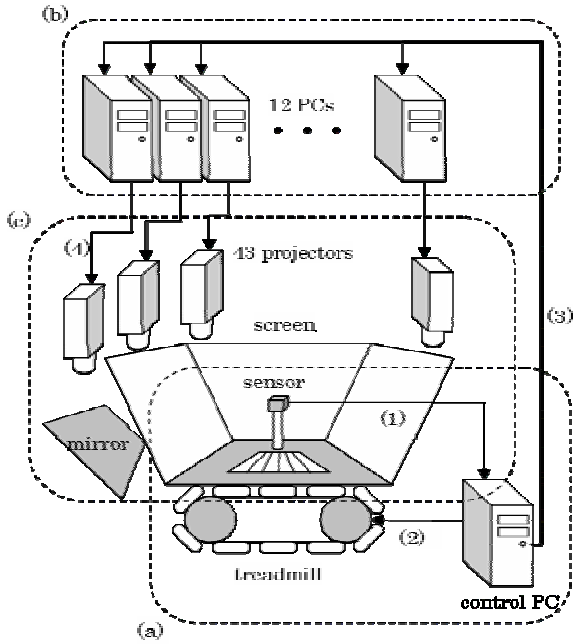


Figure 4. Display System.

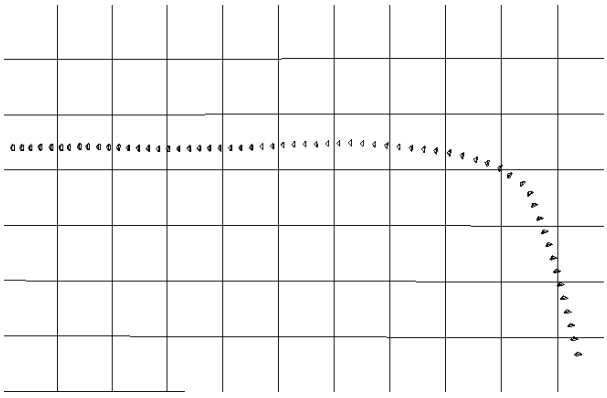
First, we have confirmed the result of the generated virtualized environment video. As shown in Figure 6, the geometric and photometric discontinuities among adjacent camera images could not be recognized except in areas that are very close to the camera system due to its violation of single view point constraint.

Secondly, we have also confirmed the estimated motion of Ladybug, which is drawn as view frustums of a representative camera unit every 10 frames, as shown in Figure 7. The motion of Ladybug is recovered very smoothly. However, Figure 8 indicates that the motion of the Ladybug includes the large shake effect defined as absolute rotation of the OMS. The maximum absolute rotation is 2.25×10^{-2} radian corresponding to about 26 pixels in a full resolution omnidirectional image of Ladybug. Although we kept the speed of the chair approximately constant, the change of the chair speed is large, as can be observed in Figure 9. These results indicate the importance of correcting camera shake and replay speed.

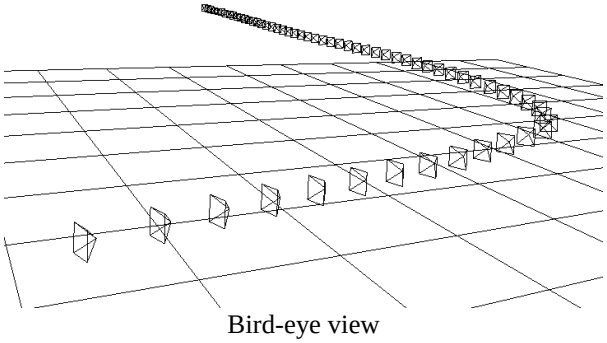
Next, the result of presented videos was confirmed. Figure 10 shows three pictures obtained from the user's view position in three directions. We have confirmed that the geometric discontinuities between regions projected by different projectors and synchronization errors could not be recognized except for the borderlines between two screens. The system could render the generated virtualized environment at 26 fps.

Figure 5. Sampled frame of captured videos acquired by six camera units of Ladybug.

Figure 6. Sampled frame of generated video accumulated in twelve graphics PCs.



Top view



Bird-eye view

Figure 7. Estimated camera motion.

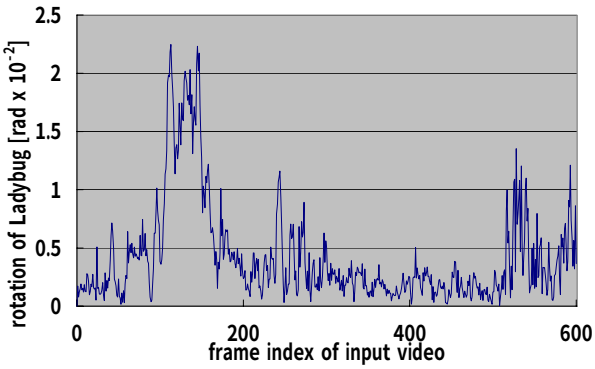


Figure 8. Magnitude of shake effect of Ladybug.

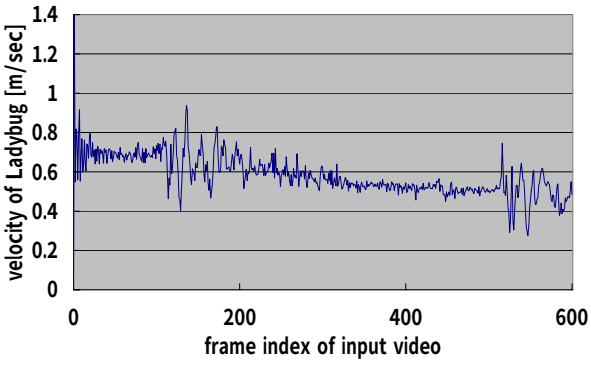


Figure 9. Change of camera speed.

left view front view right view

Figure 10. User's view.

Finally, we have confirmed that the proposed telepresence system provides us with the feeling of rich presence in remote sites in this experiment. However, poor presence was felt due to the limitation that the user's view position in a virtualized environment can not move in two dimensions. We also felt unnatural in the control of the treadmill when a user begins to walk, because the motion of upper part of the body is not considered in motion measurement; that is, the displayed image is not actually synchronized with head motion but with leg motion.

4. Conclusion

In this paper, a novel telepresence system using an immersive projection display and a locomotion interface has been proposed. This system can interactively present the feeling of walking in remote sites by showing a virtualized environment generated from real outdoor scene images. For construction of a virtualized environment, omnidirectional high-resolution videos are acquired by an omnidirectional multi-camera system. The camera system is calibrated geometrically and photometrically in advance. A virtualized environment is generated as multiple video streams whose shake effect and replay speed are corrected by using the motion of the camera system estimated from the acquired videos. The proposed display system presents the generated videos to users according to their locomotion by using the treadmill and 3D position sensors.

In experiments, a virtualized environment was generated from real images. We have confirmed the four processes were successfully achieved. The experiment has shown that the proposed telepresence system provides us with the feeling of rich presence in remote sites. In future work, we will relax the limitation in movement of user's view in virtualized environments using a new view synthesis [11].

References

1. "Special issue on immersive telepresence," IEEE Multimedia, vol. 4, no. 1, 1997.
2. S. Chen, "Quicktime VR: An image-based approach to virtual environment navigation," Proc. SIGGRAPH '95, pp. 29-38, 1995.
3. Y. Onoe, K. Yamazawa, H. Takemura, and N. Yokoya, "Telepresence by real-time view-dependent image generation from omnidirectional video streams," Computer Vision and Image Understanding, vol. 71, no. 2, pp. 154-165, 1998.
4. D. Kotake, T. Endo, F. Pighin, A. Katayama, H. Tamura, and M. Hirose, "Cybercity walker 2001 : Walking through and looking around a realistic cyberspace reconstructed from the physical world," Proc. 2nd IEEE and ACM Int. Symp. on Augmented Reality, pp. 205-206, 2001.
5. W. Tang, T. Wong, and P. Heng, "The immersive cockpit," Proc. Int. Workshop on Immersive Telepresence, pp. 36-39, 2002.
6. S. Ikeda, T. Sato, and N. Yokoya, "High-resolution panoramic movie generation from video streams acquired by an omnidirectional multi-camera system," Proc. IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent System, pp. 155-160, 2003.
7. M. Uyttendaele and A. Criminisi and S. B. Kang and S. Winder and R. Hartley and R. Szeliski, "High-quality Image-based Interactive Exploration of Real-World Environments", IEEE Computer Graphics and Applications, vol. 24, no. 3, pp. 52-63, 2003.
8. Point Grey Research Inc., http://www.ptgrey.com/.
9. H. Iwata, "Walking about virtual environments on an infinite floor," Proc. IEEE Virtual Reality '99, pp. 286-293, 1999.
10. T. Sato, S. Ikeda, and N. Yokoya: "Extrinsic camera parameter recovery from multiple image sequences captured by an omni-directional multi-camera system", Proc. 8th European Conf. on Computer Vision, vol. 2, pp. 326-340, 2004.
11. M. Irani, T. Hassner, and P. Anandan, "What does the scene look like from a scene point?" Proc. 7th European Conf. on Computer Vision, vol. 2, pp. 883-897, 2002.