Construction of an Immersive Mixed Environment Using an Omnidirectional Stereo Image Sensor

Jun Shimamura, Naokazu Yokoya, Haruo Takemura and Kazumasa Yamazawa
Graduate School of Information Science
Nara Institute of Science and Technology
8916-5, Takayama, Ikoma, Nara 630-0101, Japan
{jun-s, yokoya, takemura, yamazawa}@is.aist-nara.ac.jp

Abstract

Recently virtual reality (VR) systems have been incorporating rich information available in the real world into VR environments in order to improve their reality. This stream has created the field of mixed reality which seamlessly integrates real and virtual worlds. This paper describes a novel approach to the construction of a mixed environment. The approach is based on capturing the dynamic real world by using a video-rate omnidirectional stereo image sensor. The mixed environment is constructed of two different types of models: (1) texture-mapped cylindrical 3-D model of dynamic real scenes and (2) 3-D computer graphics (CG) model. The cylindrical 3-D model is generated from full panoramic stereo images obtained by the omnidirectional sensor that uses a pair of hexagonal pyramidal mirrors and twelve CCD cameras. A prototype system has been developed to confirm the feasibility of the proposed method, in which panoramic binocular stereo images of the mixed environment are projected on a cylindrical immersive display depending on user's viewpoint in real time.

1. Introduction

The recent progress in mixed reality [6, 10] has made it possible to construct virtual environments from real world scenes [1, 2, 4, 11]. However, there still remain some problems in capturing a large scale real environment such as urban or natural scenes. One of the most significant problems is to digitize a dynamic outdoor scene for creating an immersive virtualized environment. This requires to obtain 3-D models of real scenes containing dynamic events so that the following two can be realized: (1) interactive viewing of the environment with 3-D sensation and (2) computer graphics (CG) object synthesis maintaining correct occlusion between real and virtual objects.

Our primary objective is to develop a methodology for obtaining such models of dynamic real environments from their images. Our approach is based on acquiring full panoramic 3-D models of real worlds using a video-rate omnidirectional stereo image sensor [5]. The sensor can produce a pair of high-resolution omnidirectional binocular images that satisfy the single viewpoint constraint; that is, each omnidirectional image is obtained from a single effective viewpoint. A full panoramic 3-D model is made from a cylindrical panoramic image and depth to each pixel from the center the cylinder. The secondary objective is to construct a prototype system which presents a mixed environment consisting of real scene model and CG model. CG objects are merged into the full panoramic 3-D model of real scene maintaining consistent occlusion between real and virtual objects. Thus it is possible to yield rich 3-D sensation with binocular and motion parallax. Moreover, CG objects can be manipulated in the mixed environment. We have developed a prototype of immersive mixed reality system using a large cylindrical screen, in which a user can walk through the mixed environment and can manipulate CG objects in real time.

This paper is structured as follows. Section 2 describes the omnidirectional stereo image sensor using hexagonal pyramidal mirrors as well as brief reviews of omnidirectional sensors. Described in Section 3 is a new method of virtualizing a dynamic real environment as well as some examples. In Section 4, we present an immersive mixed reality system prototype realized by using the proposed method. Finally, Section 5 summarizes the present work.

2. Omnidirectional Stereo Imaging

Recently wide-angle or omnidirectional imaging has attracted much attention in several different fields. There exist a variety of video-rate ominidirectional sensors with different characteristics [14]. Important requirements for imaging sensors in virtual/mixed reality applications are as follows: (1) single viewpoint property [9], (2) high-resolution image acquisition, and (3) stereoscopic imaging. The single viewpoint property means that the image is obtained as a central projection of the scene onto an image plane. There exist some sensors [5, 7, 8, 15] that satisty the single viewpoint constraint. Sensors with a single standard camera, for example [8] and [15], usually have difficulty in obtaining high-resolution images. Omnidirectional stereo imaging systems can be constructed by appropriately aligning two omnidirectional sensor components [5, 8, 12]. We employ an omnidirectional sensor using pyramidal mirrors [5] which satisfies the three requirements above.

2.1. Omnidirectional stereo image sensor

At present, high-resolution omnidirectional image acquisition requires multiple cameras or a rotating camera [17]. The interesting idea for obtaining high-resolution omnidirectional images at video rate is the combination of multiple video cameras and a plane-sided pyramidal mirror that are aligned so as to satisfy the single viewpoint constraint. This idea was originally proposed by Yoshizawa [16] and later two sensors based on the same idea were actually developed [5, 7].

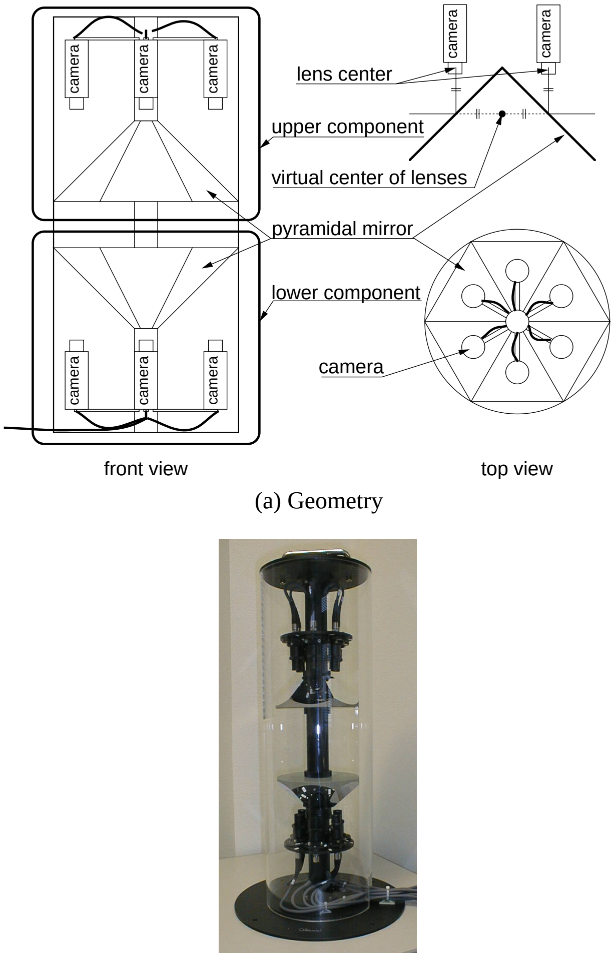
We use an omnidirectional stereo imaging sensor [5] that is composed of twelve CCD cameras and two hexagonal pyramidal mirrors. The sensor component is designed so that the virtual lens centers of six cameras are located at a fixed point as shown in Figure 1(a). The top of the mirror faces the six cameras and the base plane of the mirror is placed to be perpendicular to the line of sight of the cameras. The sensor, properly arranged, captures images of a real scene through the reflection on the pyramidal mirror. It can take a 360-degree omnidirectional image satisfying the single viewpoint constraint. In the total imaging system, two symmetrical sets of the component are aligned for omnidirectional stereo imaging. Each camera is a standard NTSC CCD camera with a wide-angle lens. Figure 1(b) shows an appearance of the sensor.

2.2. Generation of panoramic stereo images

The omnidirectional stereo image sensor described above produces synchronized twelve video streams. Figure 2 shows a set of captured images of “Heijo-kyo” (historical site in Nara) which contains the reconstructed “Suzakumon” gate. This data will be used as an example throughout the paper.

In order to generate panoramic stereo images from original twelve images captured by the sensor, the followings should be done:

1. elimination of geometric distortion in images,



(a) Geometry

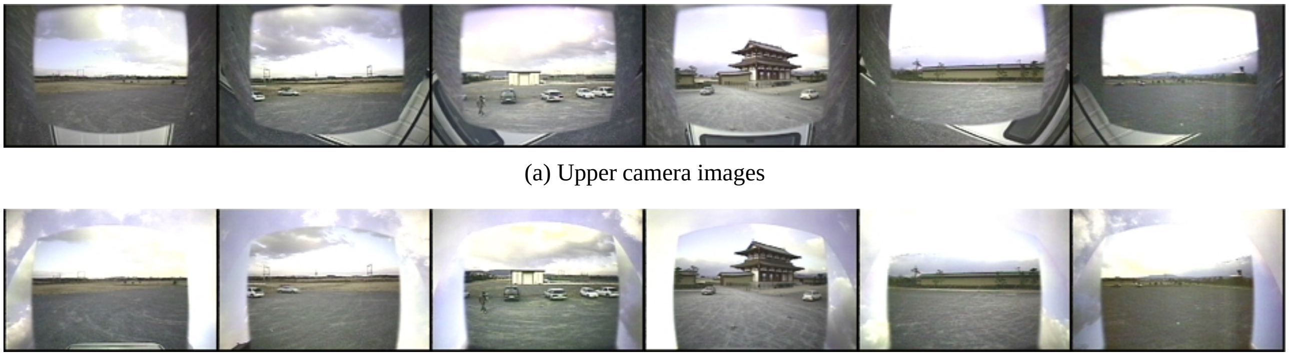
(b) Appearance

Figure 1. Omnidirectional stereo imaging sensor.

2. color adjustment of different camera images,
3. concatenation of six images for completing upper and lower omnidirectional images of stereo pair.

The image captured by each CCD camera of the sensor is geometrically distorted because of using a wide-angle lens. The Tsai’s calibration method [13] is applied to eliminate the distortion. Another problem is color difference among different camera images that is caused by the inequality of multiple CCD cameras. This problem is resolved by linearly transforming each color component of images imposing the continuity constraint on the boundaries of adjacent camera images.

Finally, by projecting each image generated from upper and lower cameras onto a cylindrical surface, twelve sheets of images are combined into a pair of full panoramic stereo images; upper and lower panoramic images. Consequently, a pair of panoramic stereo images satisfying the vertical



(a) Upper camera images

(b) Lower camera images

Figure 2. Twelve camera images of a real outdoor scene (historical site under reconstruction in Nara).

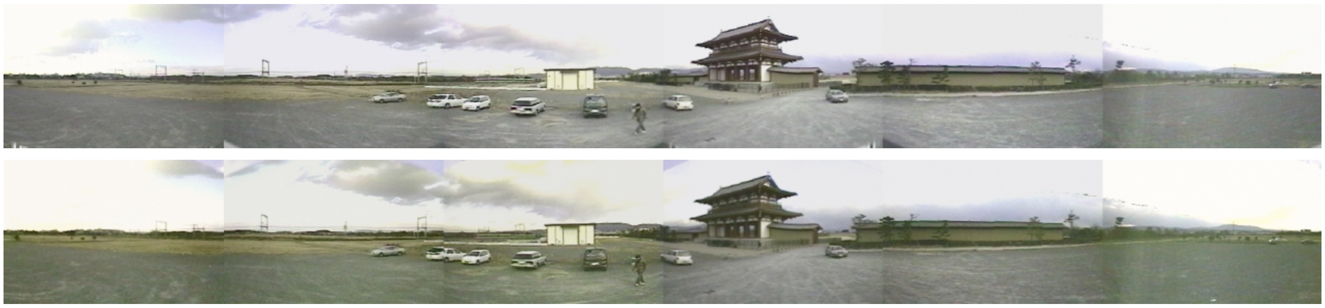


Figure 3. A pair of computed panoramic stereo images.

epipolar constraint is generated (see [5] for detail). It should be noted that the stereo images here have vertical disparities, in contrast with familiar binocular stereo images with horizontal disparities. Figure 3 shows a pair of panoramic stereo images computed from those in Figure 2. The size of each panoramic image is 3006×330 pixels.

3. Virtualizing a Dynamic Real Scene

To increase realistic sensation, real scene images often would be used in constructing virtual environments. The real images are simply mapped onto a planar or cylindrical plane in most extisting VR systems. Thus a user can not sense binocular nor motion parallax in constructed environment. In this section, we describe a novel method to construct a cylindrical 3-D model of a dynamic real scene that provides realistic 3-D sensation in virtual/mixed reality applications.

3.1. Layered representation of a dynamic scene

In our approach, a dynamic real scene model is constructed of two layers: (1) cylindrical (panoramic) 3-D model of a static scene and (2) 3-D model of dynamic event (moving object). The static scene image and moving object regions are extracted from panoramic images using existing techniques as follows.

1. Static scene image generation:
   A panoramic image of a static scene is generated by applying a temporal mode filter to a panoramic image sequence in a time interval. A stereo pair of panoramic static scene is obtained by applying this filter to both upper and lower images of omnidirectional stereo images. Figure 4 shows panoramic stereo images of a static scene generated from a sequence of dynamic panoramic stereo images including Figure 3. It can be clearly observed in Figure 4 that moving objects are eliminated from Figure 3.

2. Moving object extraction:
   Moving objects are extracted by subtracting consecutive image frames in time sequence. Figure 5 shows the result of extracting moving object regions from Figure 3 in the upper panoramic image.

Texture-mapped cylindrical 3-D models are then generated for both of two layers above. Figure 6 shows a flow diagram for the construction of a cylindrical model. After acquiring a pair of panoramic stereo images and extracting moving object regions (part A of Figure 6), Panoramic depth maps of the real world are computed for both a static

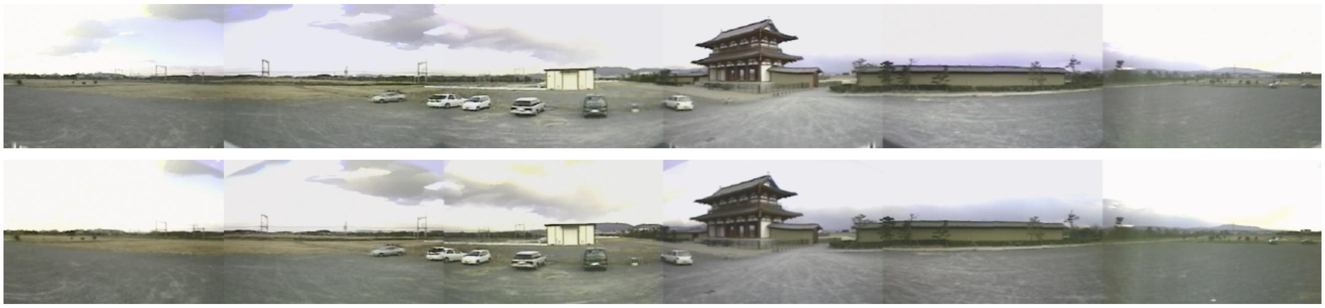


Figure 4. A pair of panoramic stereo images of a static scene without moving objects.

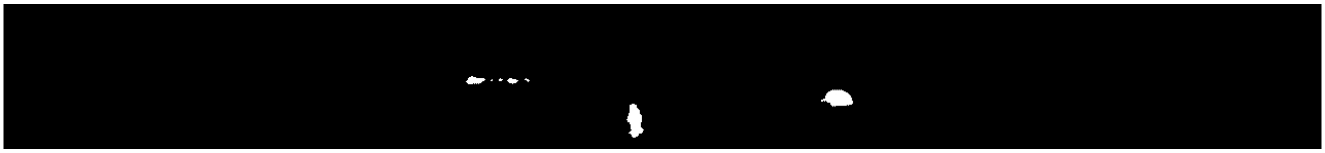


Figure 5. Extracted moving object regions in the upper panoramic image.

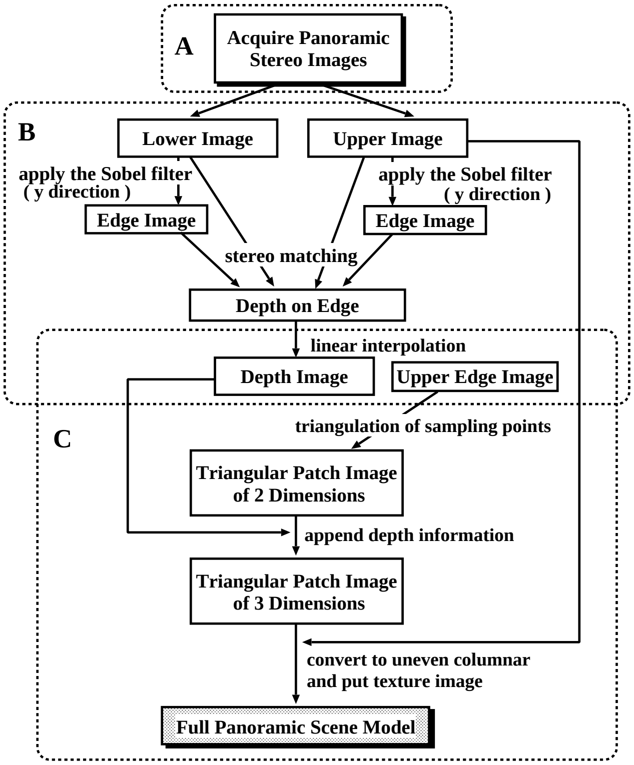

A
Acquire Panoramic Stereo Images
B
Lower Image
Upper Image
apply the Sobel filter (y direction)
apply the Sobel filter (y direction)
Edge Image
Edge Image
stereo matching
Depth on Edge
linear interpolation
Depth Image
Upper Edge Image
C
triangulation of sampling points
Triangular Patch Image of 2 Dimensions
append depth information
Triangular Patch Image of 3 Dimensions
convert to uneven columnar and put texture image
Full Panoramic Scene Model

Figure 6. Flow diagram of constructing a 3-D scene model.

scene and dyamic events by stereo matching (part B of Figure 6). Then cylindrical 3-D models are constructed from depth images and finally texture images are mapped on the models (part C of Figure 6). In the following sections, details of depth estimation and cylindrical model construction are described.

3.2. Depth estimation from panoramic stereo images

In this section, the depth estimation from panoramic stereo images is described. The depth of an existing real scene is the principal factor in representing depth relationship between virtual and real objects correctly. We acquire panoramic depth based on stereo matching. However, there is high possibility of false matching caused by noises in performing stereo matching on the whole image. In order to reduce incorrect correspondences, we estimate depth values only on edges, where the matching is thought to be reliable. Thereafter, intermediate data are approximated by linear interpolation. The following steps describe the method in more detail.

1. By adopting the vertical Sobel filter, non-vertical edges are detected in the upper and lower images as feature points.

2. Stereo matching is performed and the depth values are computed. Note that only pixels on detected edges in the upper image are matched to those in the lower image, matching window size is 9×9 pixels, and similarity measure is the normalized cross-correlation with a threshold (0.9 in experiments). In the same way, the

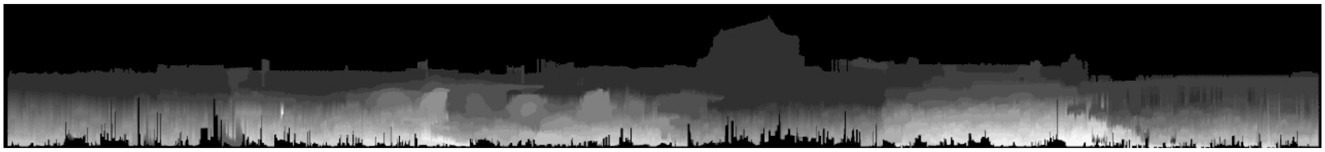


Figure 7. Panoramic depth map generated from panoramic stereo images.

lower image as a reference image is matched to the upper image.

3. Matching errors are excluded by considering the consistency between upper-to-lower and lower-to-upper matchings.

4. By adopting the median filter (5×3), noise and lacking values are revised at the upper edges.

5. The depth values at the pixels between the edges are linearly interpolated in vertical direction to complete a dense depth map.

Depth map computed from panoramic stereo images in Figure 4 is shown in Figure 7 in which depth values are coded in intensities. A brighter pixel is closer and a darker pixel is farther. A black pixel is a pixel of its depth is not computed from stereo images.

3.3. Generation of layered 3-D model

By using a panoramic depth map estimated in Section 3.2, cylindrical 3-D models of the real environment are constructed for static scene and dynamic events using the following steps.

1. Edges are detected from the upper image and points on edges with reliable depth are sampled at a fixed interval. Then non-edge points are sampled at a fixed interval over an entire region.

2. By applying the Delaunay's triangulation [3] to points extracted in Step 1, 2-D triangle patches are generated.

3. A 3-D triangular patch model is generated by assigning 3-D data of the depth image created in Section 3.2 to the vertices of 2-D triangles obtained in Step 2.

4. Finally, the upper panoramic texture image is mapped onto the constructed 3-D cylindrical model.

The steps above are for polygonizing panoramic texture and depth images for a static scene. Dynamic events can also be polygonized by applying similar steps to moving object regions in time sequence extracted in Section 3.1. Figure 8 illustrates a bird's-eye view of texture-mapped 3-D model. Distance to a pixel of which distance is not computed are set to infinity. The whole cylindrical 3-D model in Figure 8 consists of 13400 polygons.

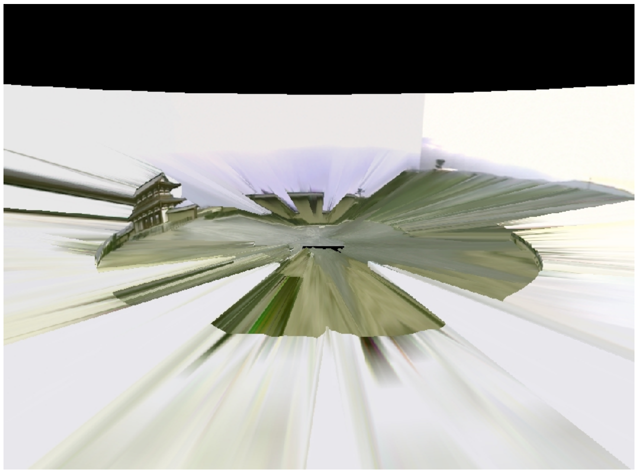


Figure 8. Bird's-eye view of texture-mapped 3-D scene model.

4. Immersive Mixed Reality System

This section describes a prototype system of immersively presenting a mixed environment in which virtual objects are merged into a virtualized dynamic and real scene.

4.1. Merging virtual objects into a virtualized real world scene

Virtual objects can be easily merged into a layered real scene model maintaining correct occlusion among real and virtual objects because the real scene model has depth information.

Figure 9 shows a mixed environment consisting of a static real scene and 3-D virtual objects (trees), in which virtual objects are created by using a computer graphics software (Alias/WaveFront). From nine different viewpoint images in Figure 9, it is celarly seen that motion parallax is clearly presented in the system. Figure 10 shows examples of superimposing dynamic event layers onto the static scene, in which images of two time instances are rendered assuming two different viewpoints. A walking person is rendered as a dynamic event in this scene.

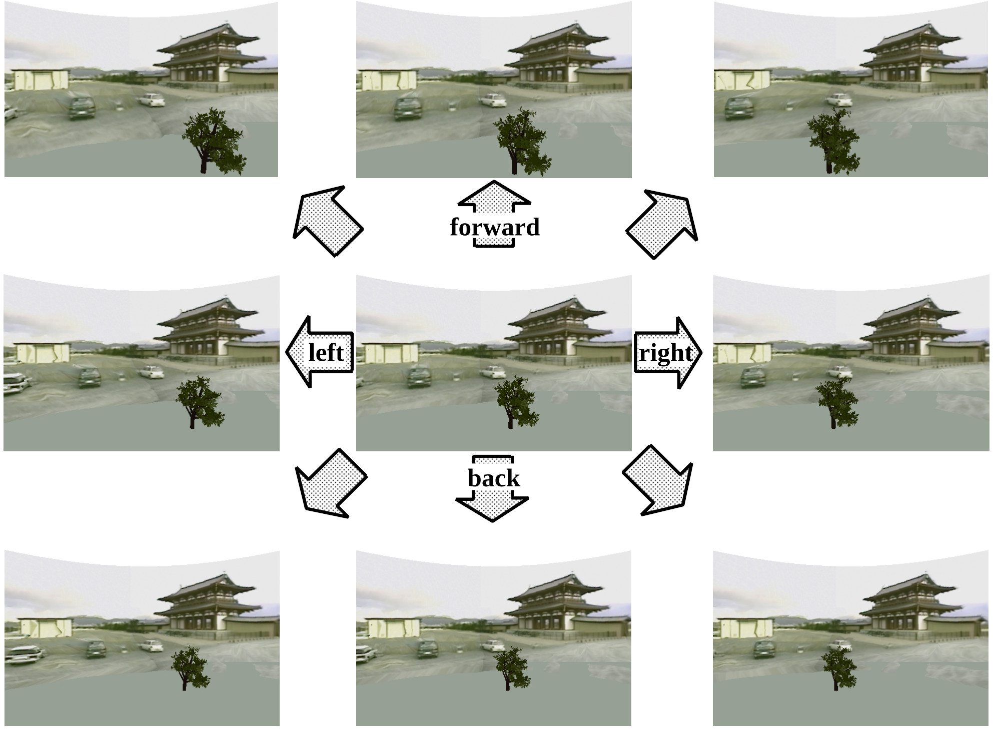


Figure 9. Mixed environment observed from different viewpoints (center: original viewpoint of sensor).

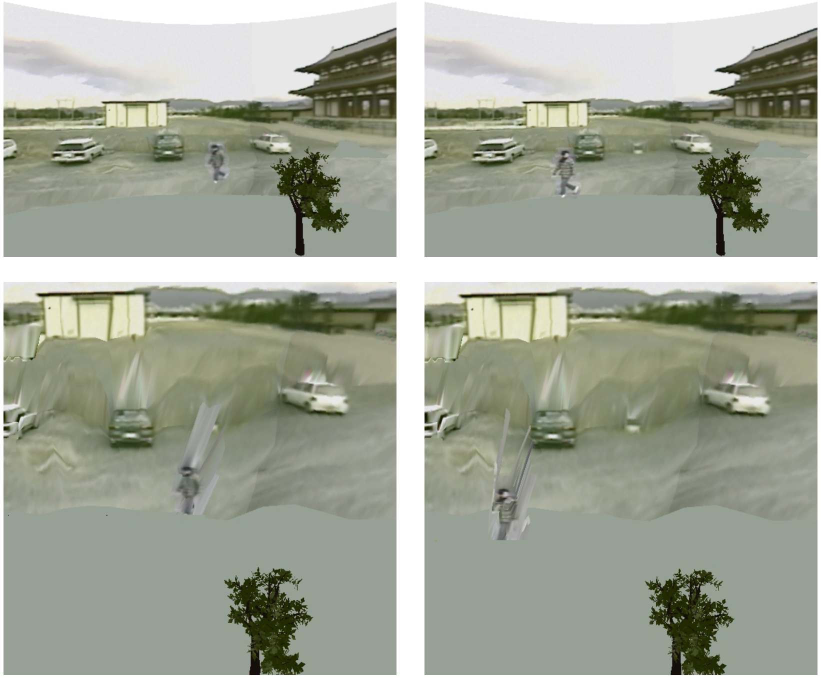


Figure 10. Superimposing dynamic event layers onto a static scene layer with virtual objects (trees) (top: original viewpoint; bottom: new higher viewpoint).

4.2. Prototype system

In order to confirm the feasibility of the proposed method, we have developed a prototype system for presenting a mixed environment of a real scene and CG objects. The hardware configuration of the system is illustrated in Figure 11. The cylindrical 3-D model is constructed on a graphics workstation, SGI Onyx2 (Infinite Reality2×2, 8CPUs (MIPS R10000, 250MHz)). Virtual objects are created by using a computer graphics tool (Alias/WaveFront) as described earlier. For presenting the mixed environment to a user in the present system, 3-D images are projected on a large cylindrical screen with the size of 6m in diameter and 2.4m in height of the CYLINDRA[1] system, which has a 330-degree view covered by six projectors. Note that the projected images are a pair of nearly-full panoramic stereo images with horizontal disparities. In the system, a user is able to change viewing position and orientation by a joystick device (SideWinder Precision Pro/Microsoft Inc.) and is able to experience stereoscopic vision as well as motion parallax through liquid crystal shutter-glasses (SB300/Solidray Inc.).

In the present system with the example used throughout this paper, image updating rate is about 13 frames/sec, when 2 CPUs are used to compute stereoscopic images of 6144×768 pixels. The total number of polygons in the static model is 54740 (13400 polygons for cylindrical 3-D real scene model and 41340 polygons for CG objects). Figure 12 shows a user performing the walk-through in the mixed environment using the CYLINDRA system.

It has been found that a user can feel real-time feedback and deep realistic sensation in the mixed environment constructed by the proposed method. In addition, we have confirmed that a user can handle virtual objects by using depth relationships among objects, and can sense binocular and motion parallax in the panoramic environment.

On the other hand, the current implementation of generating dynamic panoramic images takes considerable amount of time, because the system processes image from each CCD camera frame by frame. Multiple channel acquisition of a video stream in real time will make the system more efficient and usable. We also would like to point out that a user of the system have to stay relatively close to the center of cylindrical 3-D model in order to have better realistic sensations. When a user moves far away from the center, occluded area which is not observed from the panoramic image sensor appears in a scene and causes a sense of incompatibility. To solve this problem, a method of preparing multiple cylindrical 3-D models with different center position and switching among models based on a user's position will be effective.

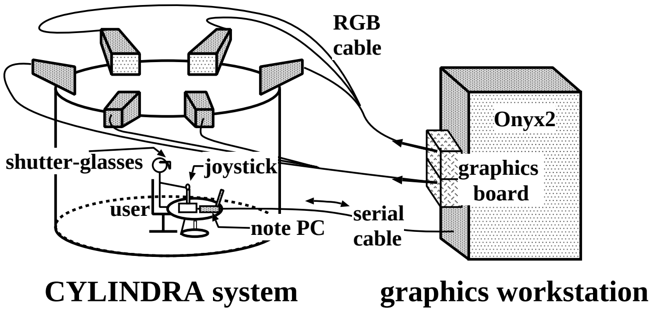


Figure 11. Hardware configuration of immersive mixed reality system.

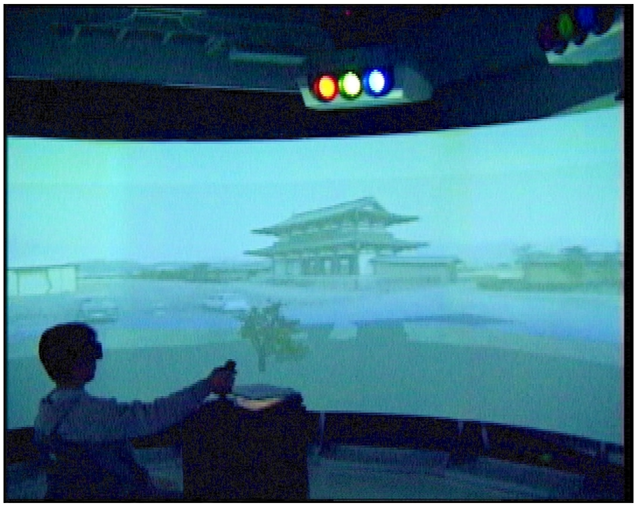


Figure 12. User's appearance in mixed environment using CYLINDRA system.

5. Conclusion

In this paper, we have proposed a novel method of constructing a large scale mixed environment. The constructed environment consists of cylindrical 3-D model of a real scene captured by an omnidirectional stereo image sensor and polygonal CG objects. The former is used in background of the scene, the latter are used to represent objects in user's vicinity. In order to represent approximate depth relationship among objects, depth values are appended to the cylindrical panoramic image of real scene. Consequently, the proposed method maintains real-time rendering because of constructing an approximate scene using real images and increases realistic sensation.

Using the prototype system, a user can virtually walk to arbitrary directions in real time in the mixed world of real and virtual objects as same as in the real world, and can handle virtual objects smoothly by using depth relationships

[1] CYLINDRA is an abbreviation for Cylindrical Yard with Large, Immersive and Novel Display for Reality Applications.

among objects. In the environment, a user can also feel deep realistic sensation of a mixed world.

As the future work, we will extend the model far larger by using multiple panoramic stereo images. We will also implement an algorithm that smoothly switch constructed models when user's viewpoint changes.

Acknowledgments

This work was supported in part by the Telecommunications Advancement Organization of Japan.

References

[1] G.U. Carraro, T. Edmark, and J.R. Ensor. Techniques for handling video in virtual environment. *Proc. SIGGRAPH'98*, pages 353–360, 1998.

[2] S.E. Chen. QuickTime VR – An image-based approach to virtual environment navigation. *Proc. SIGGRAPH'95*, pages 29–38, 1995.

[3] P. Heckbert, Ed. *Graphics Gems IV*. Academic Press Professional, Boston, 1994.

[4] T. Kanade, P. Rander, and P.J. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE MultiMedia*, 4(1):34–47, Jan. 1997.

[5] T. Kawanishi, K. Yamazawa, H. Iwasa, T. Takemura, and N. Yokoya. Generation of high-resolution stereo panoramic images by omnidirectional imaging sensor using hexagonal pyramidal mirrors. *Proc. 14th IAPR Int. Conf. on Pattern Recognition*, I, pages 485–489, Aug. 1998.

[6] P. Milgram and F. Kishino. A taxonomy of mixed reality visual display. *IEICE Trans. on Information and Systems*, E77-D(12):1321–1329, Dec. 1994.

[7] V. Nalwa. *A True Omnidirectional Viewer*. Tech. Report, Bell Laboratories, Holmdel, NJ, Feb. 1996.

[8] S.K. Nayer. Omnidirectional video camera. *Proc. DARPA Image Understanding Workshop*, 1, pages 235–241, May 1997.

[9] S.K. Nayer and S. Baker. Catadioptric image formation. *Proc. DARPA Image Understanding Workshop*, 2, pages 1431–1437, May 1997.

[10] Y. Ohta and H. Tamura, Eds. *Mixed Reality –Merging Real and Virtual Worlds*. Ohmsha & Springer-Verlag, Tokyo, 1999.

[11] 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*, 71(2):154–165, Aug. 1998.

[12] D. Southwell, A. Basu, M. Fiala, and J. Reyda. Panoramic stereo. *Proc. 13th IAPR Int. Conf. on Pattern Recognition*, I, pages 378–382, Aug. 1996.

[13] R.Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, RA-3(4):323–344, Aug. 1987.

[14] Y. Yagi. Omnidirectional sensing and its applications. *IEICE Trans. Information & Systems*, E82-D(3):568–579, Mar. 1999.

[15] K. Yamazawa, Y. Yagi, and M. Yachida. New real-time omnidirectional image sensor with hyperboloidal mirror. *Proc. 8th Scandinavian Conf. on Image Analysis*, 2, pages 1381–1387, May 1993.

[16] M. Yoshizawa. Omniazimuth photographing device and omniazimuth image synthesizer. Japanese Patent No.06260168 (pending), Oct. 1994.

[17] J.Y. Zheng and S. Tsuji. Panoramic representation of scenes for route understanding. *Proc. 10th IAPR Int. Conf. on Pattern Recognition*, I, pages 161–167, June 1990.