A Stereoscopic Video See-through Augmented Reality System Based on Real-time Vision-based Registration

Masayuki Kanbara[1], Takashi Okuma[2], Haruo Takemura[1] and Naokazu Yokoya[1 3]
[1]Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN
{masay-ka, takemura, yokoya } @is.aist-nara.ac.jp
[2] Electrotechnical Laboratory, MITI
1-1-4 Umezono, Tsukuba-shi, Ibaraki, 305-8568, JAPAN
okuma@etl.go.jp
[3] Nara Research Center, Telecommunications Advancement Organization of Japan
8916-19 Takayama-cho, Ikoma-shi, Nara, 630-0101, JAPAN

Abstract

In an augmented reality system, it is required to obtain the position and orientation of the user's viewpoint to display the composed image maintaining correct registration of real and virtual worlds. All the procedures must be done in real-time. This paper proposes a method for augmented reality with a stereo vision sensor and video see-through head mounted display. It can synchronize the display-timing between the virtual and real worlds, so that an alignment error is reduced. The method calculates camera parameters from three markers in image sequences captured by a pair of stereo cameras mounted on a HMD. In addition, it estimates depth of real world from a pair of stereo images to generate a composed image maintaining consistent occlusions between real and virtual objects. A region of depth estimation is efficiently limited by calculating a position of virtual object using camera parameters. Finally, we have developed a video see-through augmented reality system which mainly consists of a pair of stereo cameras mounted on the HMD and a standard graphics workstation. The feasibility of the system has been successfully demonstrated with experiments.

1. Introduction

This paper describes a method of merging real and virtual worlds for a video see-through augmented reality system. Augmented reality produces an environment in which virtual objects are superimposed on a user's view of the real environment. Augmented reality has received a great deal of attention as a new method for displaying information or increasing the reality of virtual environments. A number of applications have already been proposed and demonstrated [1]. For example, Feiner et al. proposed an augmented reality system that explains to an end-user how to maintain a printer using see-through HMD[2]. State et al. presented an augmented reality system applied to a medical procedure known as the ultrasound-guided needle biopsy of a breast[3].

To implement an augmented reality system, we must resolve some problems. A geometric registration is especially the most important problem because the problem is a principal factor which provides a user with a sense of incongruity. The registration includes a problem of geometric alignment of the real and virtual coordinates and a problem of resolving occlusion between real and virtual objects. The former problem is considered as one of acquiring the position and orientation of the user's viewpoint in terms of registering the real and virtual worlds geometrically[4]. The latter problem can be resolved by measuring the real world in advance when the real world is static. However, since the real world is usually dynamic, we must estimate a depth of the real scene in real-time.

In addition to the problems above, we need to discuss how to combine a method of visually merging the real and virtual worlds and a method of resolving geometric registration because there are some methods of composition that cannot present correct occlusion. The following briefly reviews the respective methods.

In general, the following two major methods are known for acquiring the user's viewpoint in geometric registration.

- A method that uses a 3-D tracker, such as electromagnetic, ultrasonic or mechanical trackers; for example, see [2, 3].
- A method that estimates the user's viewpoint by using camera images captured at the user's viewpoint; for example, see [5, 6, 7]. This is sometimes referred to as

a vision sensor.

The 3-D trackers used in the former method can directly acquire 3-D position and orientation of receivers. However, the drawbacks of the method are that the system requires a special equipment and its measurement range is limited to a relatively narrow area. On the other hand, the latter can estimate the position and orientation of the user's viewpoint from an acquired image and there is potentially no limitation in measurement range. When the relationship between a camera position and the user's viewpoint is known, the user's viewpoint can be obtained by calculating the camera parameters from captured images. This means that the traditional techniques studied in the field of computer vision can be used to measure the viewpoint [8].

There exist two methods of showing a user images in which the real and virtual environments are merged.

- An optical see-through shows a user the real environment through a half-transparent mirror and the virtual environment reflected on the half-transparent mirror.
- A video see-through shows a user image sequences of the real environment captured by a camera and the virtual environment overlaid upon the image sequences.

The optical see-through system can show a user the real environment without delay. However, virtual objects are not synchronized with the real world imagery. This may cause the displacement between real and virtual objects. This type of system has another drawback that the amount of light from the real environment is reduced, because it uses a half-transparent mirror. In contrast, the video see-through system makes the display-timing of the real environment and that of the virtual environment synchronous. Therefore, it reduces the alignment error. However, the augmented real environment is shown to the user with delay [9].

We focus on an augmented reality system that adopts the combination of vision sensor and video see-through system. Since the image captured by a camera is used both to estimate camera parameters and to show the user the real environment, the combination is able to synchronize the real and virtual environments and reduce alignment error between them. For implementation of an augmented reality system with highly realistic sensations, all the processes from acquisition of the user's viewpoint to displaying the composed image by using obtained viewpoint must be done in real-time.

This paper is structured as follows. Section 2 briefly reviews related work. Section 3 describes the algorithms to calculate camera parameters and to estimate the depth of real world as well as stereo image composition with examples. In Section 4, the experimental results with the proposed methods and the discussion about the prototype system are described. Finally, Section 5 summarizes the present work.

2. Related Work

The registration problem is considered as a problem of acquiring the position and orientation of the user's viewpoint. Using a vision sensor, if we assume that the optical axes of the cameras are set to be parallel to viewer's gaze direction, the acquisition of position and orientation of the user's viewpoint is equal to the estimation of external camera parameters. A number of methods of estimating camera parameters from tracked feature points in the image sequence have been proposed by several research groups [6, 10, 11, 12]. The authors have already built a prototype of a vision-based video see-through augmented reality system with a single camera that uses four colored markers as feature points [5]. In this system, the position and orientation of the user's viewpoint are calculated from the positions of the feature points in the captured image. However, the monocular vision system does not show the user the stereoscopic view. Moreover, virtual objects that should be behind the real objects occlude the real objects, so that the occlusion conflict occurs. State, et al. [13] proposed a video see-through augmented reality system that provides the user with the stereoscopic view using a pair of stereo cameras, but the system could not show the user the images of virtual objects with a dynamic real world.

This paper describes algorithms to realize a stereo-vision based augmented reality. The camera parameters can be estimated from captured stereo images containing three features in the scene. In addition, a depth of the real world can be estimated by a stereo vision technique. Static real objects can be properly treated by the state of the art of stereo vision techniques in video see-through augmented reality applications using standard workstations[14, 15]. However, dynamically moving arbitrary objects are still difficult to be mixed with CG objects in real-time without special hardwares [7, 16, 17], maintaining correct occlusions among them. The method proposed in the paper overcomes this problem by estimating depth only for regions where virtual objects are composed; that is, it is an augmented reality-oriented approach to stereo vision.

3. Algorithms for Stereoscopic Video See-through Augmented Reality

In order to evaluate the feasibility of the augmented reality environment that uses a pair of stereo cameras, the prototype system of augmented reality was implemented. The system estimates camera parameters using a pair of stereo images. The system also realizes the occlusion of virtual objects by real objects.

Figure 1 shows the flowchart of the prototype system. A pair of stereo image sequences captured by stereo cameras

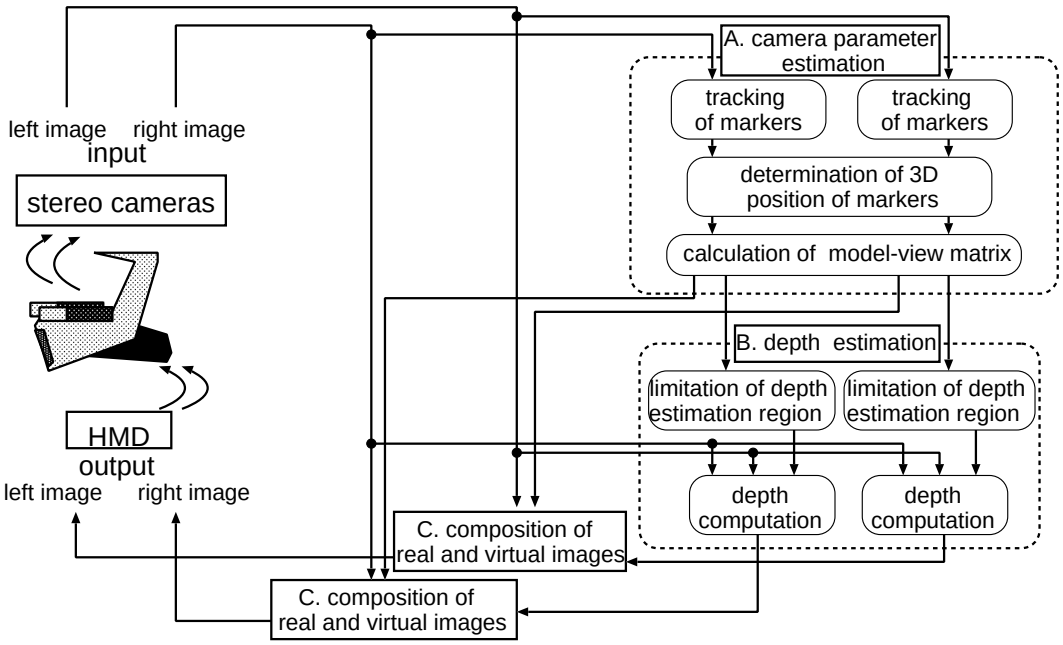

Figure 1. Flow diagram of stereoscopic video see-through augmented reality system.

are fed into a standard workstation. First, in order to capture the sight of user's viewing direction, the stereo cameras are mounted on a HMD. The camera parameter is estimated using these image sequences of the real world containing markers. Next, depth of real world is estimated using both these image sequences and the camera parameter. Finally, the stereo images are composed with the computer graphics (CG) images (images of virtual objects) and are outputted to the HMD that is worn by the user.

In the following sections, the details of the camera parameter estimation, depth estimation of the real world and the image composition are described.

3.1. Camera parameter estimation

In our prototype system, blue markers are used as feature points in a real environment. It should be noted that a pair of stereo cameras are calibrated in advance by using a standard technique[18]. The following three steps are executed to estimate camera parameters.

1. From the input images, blue regions are extracted and the positions of markers in the images are obtained.

2. The markers' 3-D positions in the camera coordinate system are estimated by using a stereo matching algorithm.

3. A projection matrix (model-view matrix) that represents the relationship between the real and virtual environments is calculated using 3-D coordinates of extracted markers.

Let us describe these steps in more detail in the following.

3.1.1. Extraction of markers. The extraction of blue marker regions from the entire image needs considerable amount of computation. Therefore, only in the first frame of the image sequence, the blue region extraction is performed over the entire image. In the subsequent frames, extracted regions are tracked by using the results obtained in the previous frame. The extracted and tracked regions are used to calculate the screen coordinate of each marker.

In the first frame of the image sequence, the following steps are used to extract blue marker regions.

1. The entire images in the first stereo pair are scanned to extract blue regions.

2. The center of gravity of each blue region is treated as the screen coordinate of the marker.

3. When three markers are found in both left and right images, stereo matching is performed based on the epipolar constraint. All markers are then labeled (Labels: 1,2,3).

When the processing speed is fast enough, the markers' positions in adjacent frames can be assumed to be close to each other. Thus the following steps are used in the subsequent frames in the image sequence.

1. A search area for each marker is determined in the current frame based on the position of the marker in the previous frame.

2. The center of gravity for all blue pixels in the search area is calculated and is treated as the screen coordinate of the marker. The matching between left and right images in the first frame is used to determine the stereo pair of markers between the current left and right images.

When the stereo matching of markers or tracking of markers fails, the system starts over the procedure for the first frame described above.

3.1.2. Determination of 3-D position of markers. In this phase of the process, the system determines the 3-D position of each marker. The relationship between two cameras and a marker is illustrated in Figure 2. The origin of the camera coordinate system is placed at the middle between the centers of projection of two cameras. The X-axis is set along the baseline of the cameras. The Z-axis is set to the direction parallel to the optical axes of the cameras. A marker at $P(X, Y, Z)$ in 3-D space is projected onto the left and right images at $P_l(x_l, y_l)$ and $P_r(x_r, y_r)$ in the screen coordinates, respectively. Then following equations stand:

$$\begin{aligned} X &= \frac{B(x_l + x_r)}{2(x_l - x_r)}, \\ Y &= \frac{B(y_l + y_r)}{2(x_l - x_r)}, \\ Z &= \frac{fB}{x_l - x_r}, \end{aligned}$$

where f is the focal length and B is the baseline length. Thus when f and B are known, $P(X, Y, Z)$ is calculated from the coordinates $P_l(x_l, y_l)$ and $P_r(x_r, y_r)$.

3.1.3. Calculation of model-view matrix. In order to compose the image of a real environment and CG objects, A model-view matrix that represents the transformation from the world coordinate, in which the shape of CG objects is defined, to the camera coordinate is required. Figure 3 illustrates the relationship between the world and camera coordinates as well as markers. Among a model-view matrix - **M**, a position of a point in the world coordinate system - **w**, and its position in the camera coordinate system - **c**, the following equation stands.

$$\mathbf{c} = \mathrm{M}\mathbf{w}.$$

The matrix **M** is a rigid transformation in homogenous coordinates consisting of a rotation **R** and a translation **T** as follows.

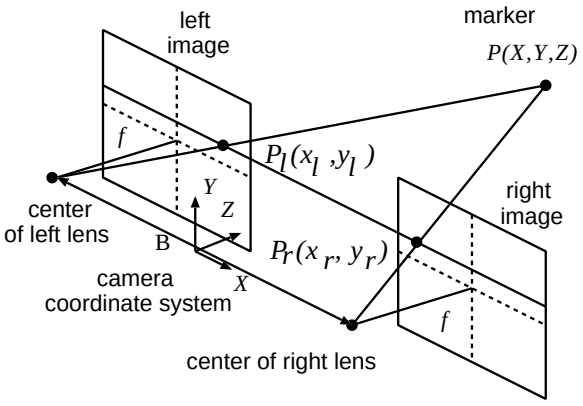

Figure 2. Geometry of stereoscopic projection of a marker.

$$\mathrm{M} = \left[\begin{array}{ccc|c} & \mathbf{R} & & \mathbf{T} \\ \hline 0 & 0 & 0 & 1 \end{array}\right].$$

In the prototype system, the world coordinate system is simply defined as follows:

- The origin is set at the marker labeled as No. 1 (M1 in Figure 3).
- The x-axis is set on the line that connects the markers labeled as No. 1 and No. 2 (M2 in Figure 3).
- The $x - y$ plane is set on the plane on which three markers reside.

According to the definition above, the translation component **T** can be given by the camera coordinate position of the marker No. 1. The rotation component **R** can be calculated with the following steps.

1. When the positions of the markers Nos. 1, 2, and 3 are given as $\mathbf{V_1}$, $\mathbf{V_2}$, and $\mathbf{V_3}$, respectively, the direction of the vector of each axis of the world coordinate ($\mathbf{x_n}$, $\mathbf{y_n}$, $\mathbf{z_n}$) can be defined as follows.

$$\begin{aligned} \mathbf{x_n} &= \mathbf{V_2} - \mathbf{V_1}, \\ \mathbf{y_n} &= (\mathbf{V_3} - \mathbf{V_1}) - \frac{\mathbf{x_n} \cdot (\mathbf{V_3} - \mathbf{V_1})}{\mathbf{x_n} \cdot \mathbf{x_n}} \mathbf{x_n}, \\ \mathbf{z_n} &= \mathbf{x_n} \times \mathbf{y_n}. \end{aligned}$$

2. Normalize $\mathbf{x_n}$, $\mathbf{y_n}$, and $\mathbf{z_n}$ into $\mathbf{x'_n}$, $\mathbf{y'_n}$, and $\mathbf{z'_n}$, respectively.

$$\mathbf{x'_n} = \frac{\mathbf{x_n}}{\|\mathbf{x_n}\|}, \quad \mathbf{y'_n} = \frac{\mathbf{y_n}}{\|\mathbf{y_n}\|}, \quad \mathbf{z'_n} = \frac{\mathbf{z_n}}{\|\mathbf{z_n}\|}.$$

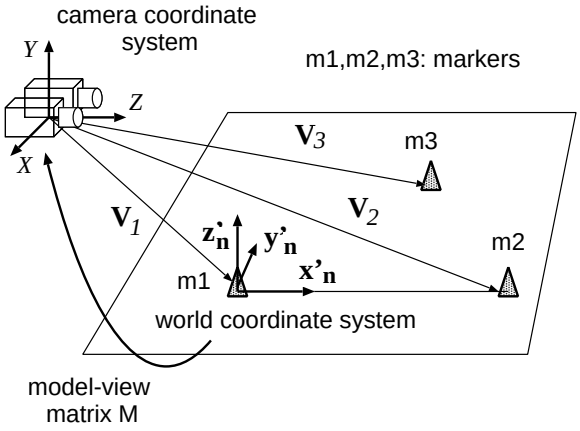

Figure 3. Relationship between the world and camera coordinates.

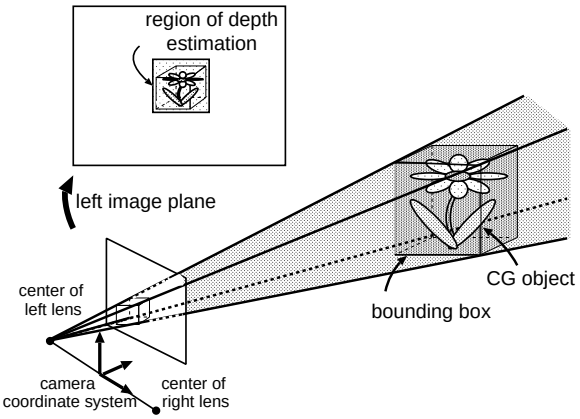

Figure 4. Determination of depth estimation region.

3. **R** can be determined by using $\mathbf{x}'_\mathbf{n}$, $\mathbf{y}'_\mathbf{n}$, and $\mathbf{z}'_\mathbf{n}$ as follows.

$$\mathbf{R} = [\mathbf{x}'_\mathbf{n}\ \mathbf{y}'_\mathbf{n}\ \mathbf{z}'_\mathbf{n}].$$

By using **T** and **R**, the model-view matrix **M** can be uniquely determined. Now it should be noted that the geometric alignment of the real and virtual world coordinates is achieved by the model-view matrix.

3.2. Depth estimation of real world

In this section, the depth estimation of real world is described. The depth estimation is used to compose the image of the virtual objects and the image of the real world without occlusion conflicts. In order to avoid occlusion conflicts, we need the depth information of only the region onto which virtual objects are projected. When the positions of virtual objects are known, the region for depth estimation can be limited as shown in Figure 4. Therefore, the depth estimation region of real world can be determined by projecting a CG object's bounding box using the model-view matrix estimated in Section 3.1. The following steps describe a method of depth estimation of real world.

1. A position of the CG object in the world coordinate system is transformed into its position in the camera coordinate system using the estimated model-view matrix **M**.

2. A bounding box of the virtual object is projected onto the left image as shown in Figure 4. A depth estimation region is obtained as a bounding rectangle of the projected box. The same process is done for the right image as well.

3. By adopting the Sobel filter, edges are detected in the region of depth estimation on the left and right images.

4. Stereo matching is performed and the depth value is computed. Note that only pixels on detected edges in the left image are matched to those in the right image, matching window size is 5×5 pixels, and similarity measure is the sum of absolute differences (SAD). In the same way, the right image as a reference image is matched to the left image.

5. Matching errors are excluded by considering the consistency between left-to-right and right-to-left matchings. The depth values at the pixels between the edges are interpolated.

3.3. Composition of real and CG images

By using the estimated model-view matrix and the depth map of the real world, CG images of virtual objects are mixed into the image of real world. At this stage, the depth of real world and each virtual object are compared. When the real objects are closer to the user's viewpoint, a transparent virtual object is drawn at the 3-D position where the real objects exist. By using a hardware z buffering algorithm, the virtual objects that are farther than the transparent objects are not actually drawn on the frame buffer. Therefore the composed image is looked as if the real objects are occluding the virtual objects. These rendering steps are illustrated with examples in Figure 5.

First, only the background image of a real scene is rendered on the frame buffer as in Figure 5 (a). Z-buffer value is also set to the farthest value through out the screen. Then, Z-buffer of the real objects are set to the depth of the real objects in Figure 5 (c). The regions of real objects are illustrated with gray levels of depth in Figure 5 (b). Finally, the virtual objects are rendered by using the model-view matrix in Figure 5 (d),(e). These steps are applied to both left

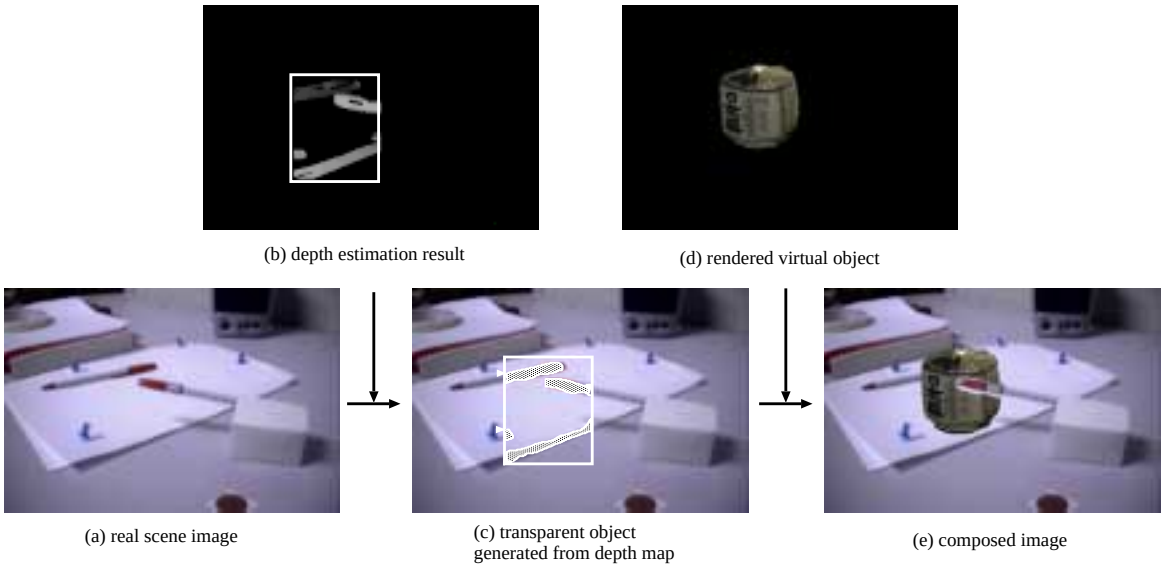

Figure 5. Process of image composition.

Figure 6. Appearance of stereoscopic video see-through HMD.

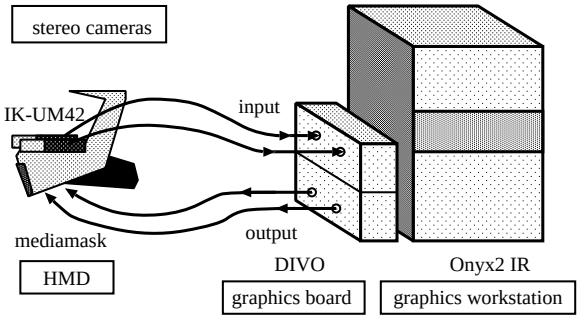

Figure 7. Hardware configuration of prototype system.

and right images for obtaining a stereo pair of composite images.

4. Experiments

4.1. Prototype system and results

In the present system, two small identical CCD cameras (Toshiba IK-UM42) are mounted on a HMD (Olympus Media Mask) as shown in Figure 6. An optical see-through function of the HMD is not used; that is, it works in closed-view mode. The baseline length between two cameras is set to 6.5 cm. The optical axes of the cameras are set to be parallel to the viewer's gaze direction (actually the head direction). The images captured by the cameras are fed into a graphics workstation (SGI Onyx2 IR: 16CPU MIPS R10000 195MHz) through the digital video interface (DIVO). The input real world images are merged with virtual objects and outputted from the DIVO interface to the HMD. The hardware configuration of the whole system is illustrated in Figure 7.

Figure 8 shows the result of the stereoscopic image composition. The images on the first and second rows (Figure 8(a),(b)) show the composition of a virtual sake cask into a real desktop scene by using three markers placed on the desktop. As is observed from the Figure 8, the stereoscopic composition of the virtual sake cask and the real environment are realized. In the third and fourth rows (Figure 8(c),(d)), real objects are placed in front of or back of the virtual sake cask. As we can see from these examples, the occlusion between the virtual and real objects are correctly established. Moreover, the stereoscopic video see-through HMD gives a good depth sensation to users and has improved the quality of augmented reality.

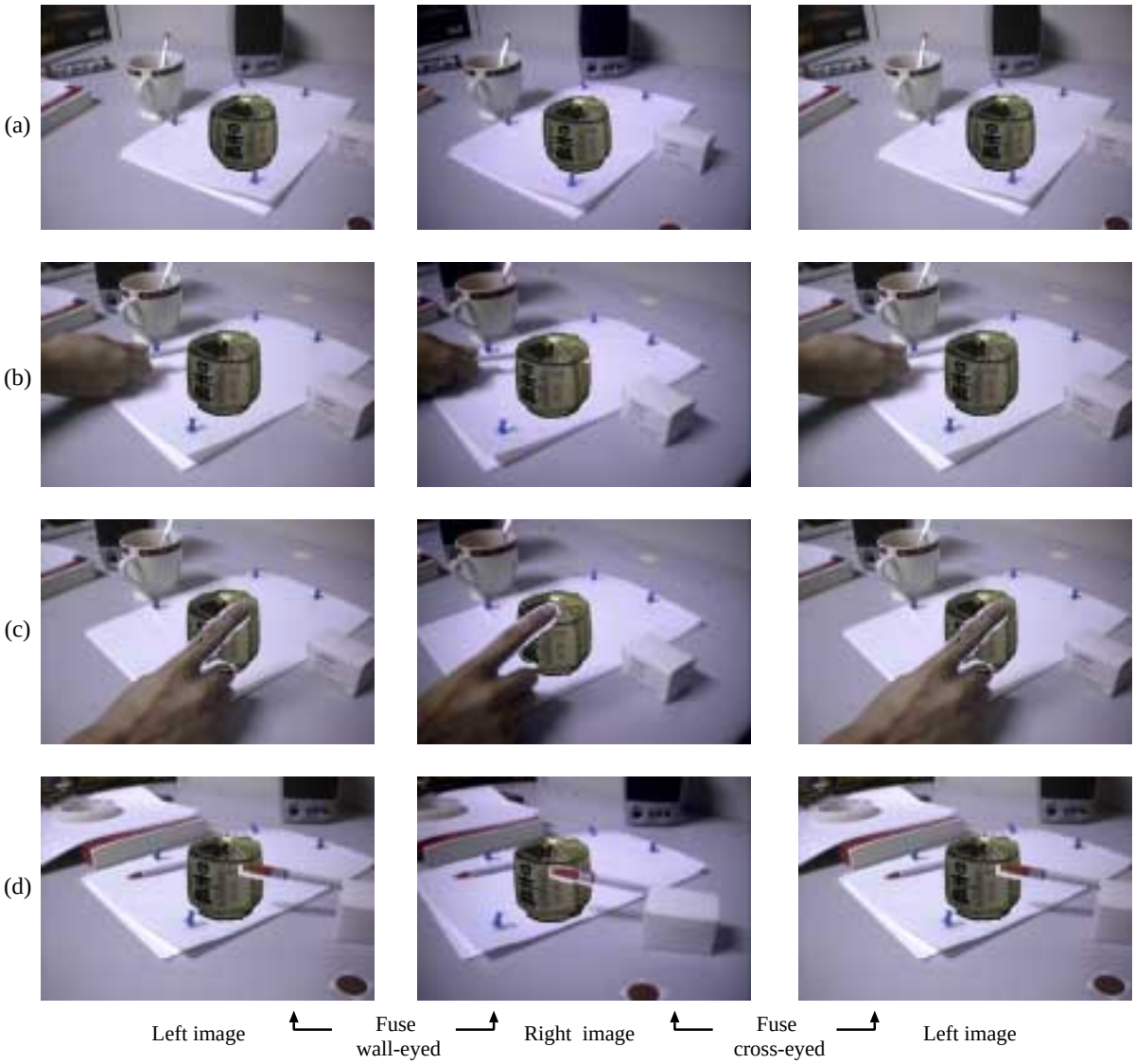

Figure 8. Composed stereo images.

4.2. Discussion

The average frame rate of the system is 10 frames per second, when the virtual object shown in Figure 8 (about 3700 polygons with a texture) is synthesized. This means that 100 ms are allocated for each frame. The approximate processing time at each stage is as follows: 20 ms for camera parameter estimation, 40 ms for depth estimation of real world, 30 ms for image composition with CG. Due to the nature of video see-through augmented reality, there is no synchronization error between the real scene and virtual objects. On the other hand, an entire scene has a time latency of 133 ms.

If the system cannot locate at least three markers in the real environment, it cannot estimate the model-view matrix. To overcome this problem, we should set more markers in the real environment, so that the system could stably estimate the model-view matrix.

The vision sensor essentially has no limitations in measurement range. However, the present prototype system actually has a limitation in measurement range because the system uses a limited number of predefined markers. If we could automatically detect and track feature points other than markers in a real scene to calculate the model-view matrix, a measurement range of the system would be extended.

5. Conclusion

In this paper, we have proposed a method of composing virtual (CG) objects with real world images in real-time

for video see-through augmented reality applications. As a pilot study, we also have developed a prototype of stereoscopic video see-through augmented reality system, which is implemented by using existing computer vision techniques on a graphics workstation and a HMD with a pair of CCD cameras. The prototype system can produce composite images of real and virtual objects nearly at video-rate, maintaining correct occlusions between CG objects and real objects. It should be noted that the system can provide a user with excellent 3-D depth sensation of augmented environments. We have proven that the stereo-vision based video see-through augmented reality system is feasible and has the possibility of building up actual applications.

In the future work, we will concentrate our attention on robust estimation of the model-view matrix in some ways; for example, (1) using more than three markers, (2) automatic detection and tracking of natural feature points in a real scene without predefined markers.

Acknowledgments

This work was supported in part by Grant-in-Aid for Scientific Research under Grant No. 09480068 from the Ministry of Education, Science, Sports, and Culture, and also by the Telecommunications Advancement Organization of Japan.

References

[1] R. T. Azuma: "A Survey of Augmented Reality," Presence, Vol. 6, No. 4, pp. 355–385, 1997.

[2] S. Feiner, B. Maclntyre and D. Seligmann: "Knowledge-based Augmented Reality," Commun. of the ACM, Vol. 36, No. 7, pp. 52–62, 1993.

[3] M. Bajura, H. Fuchs and R. Ohbuchi: "Merging Virtual Objects with the Real World: Seeing Ultrasound Imagery within the Patient," Proc. SIGGRAPH'92, pp. 203–210, 1992.

[4] M. Tuceryan, D. S. Greer, R. T. Whitaker, D. E. Breen, C. Crampton, E. Rose and K. Ahlers: "Calibration Requirements and Procedures for a Monitor-Based Augmented Reality System," Trans. on Visualization and computer graphics, Vol. 1, No. 3, pp. 255–273, 1995.

[5] T. Okuma, K. Kiyokawa, H. Takemura and N. Yokoya: "An Augmented Reality System Using a Real-time Vision Based Registration," Proc. ICPR'98, Vol. 2, pp. 1226–1229, 1998.

[6] M. Uenohara and T. Kanade: "Vision-Based Object Registration for Real-time Image Overlay," Proc. CVRMed'95, pp. 13–22, 1995.

[7] N. Yokoya, H. Takemura, T. Okuma and M. Kanbara: "Stereo Vision Based Video See-through Mixed Reality," in Mixed Reality - Merging Real and Virtual Worlds, Ohmsha-Springer Verlag, pp. 131–145, 1999.

[8] G. J. Klinker, K. H. Ahlers, D. E. Breen, P-Y. Chevalier, C. Crampton, D. S. Greer, D. Koller, A. Kramer, E. Rose, M. Tuceryan and R. T. Whitaker: "Confluence of Computer Vision and Interactive Graphics for Augmented Reality," Precense, Vol. 6, No. 4, pp. 433–451, 1997.

[9] E. K. Edwards, J. P. Rolland and K. P. Keller: "Video See-through Design for Merging of Real and Virtual Environments," Proc. VRAIS'93, pp. 197–204, 1993.

[10] M. Bujura and U. Neumann: "Dynamic Registration Correction in Video-Based Augmented Reality Systems," IEEE Computer Graphics and Applications, Vol. 15, No. 5, pp. 52–60, 1995.

[11] J. Rekimoto: "Matrix: A Realitime Object Identification and Registration Method for Augmented Reality," Proc. APCHI, pp. 63–68, 1998.

[12] A. State, G. Hirota, D. T. Chen, W. F. Garrett and A. Livingston: "Superior Augmented Reality Registration by Integrating Landmark Tracking and Magnetic Tracking," Proc. SIGGRAPH'96, pp. 429–438, 1996.

[13] A. State, A. Livingston, W. F. Garrett, G. Hirota and H. Fuchs: "Technologies for Augmented Reality Systems: Realizing Ultrasound-Guided Needle Biopsies," Proc. SIGGRAPH'96, pp. 439–446, 1996.

[14] M. M. Wloka and B. G. Anderson: "Resolving Occlusion in Augmented Reality," Proc. 1995 ACM Sympo. on Interactive 3D Graphics, pp. 5–12, 1995.

[15] B. Ross: "A Practical Stereo Vision System," Proc. CVPR'93, pp. 148–153, 1993.

[16] T. Kanade, A. Yoshida, K. Oda, H. Kano and M. Tanaka: "Stereo Machine for Video-rate Dense Depth Mapping and Its New Application," Proc. CVPR'96, pp. 196–202, 1996.

[17] N. Yokoya, T. Shakunaga and M. Kanbara: "Passive Range Sensing Techniques: Depth from Images," IEICE Trans. Information and Systems, Vol. E82-D, No. 3, pp. 523–533, 1999.

[18] 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, Vol. RA-3, No. 4, pp. 323–344, 1987.