IEEE and ATR Workshop on Computer Vision for Virtual
Reality Based Human Communications (CVVRHC'98)
Indian Institute of Technology, Bombay, India, Jan.3, 1998

Vision and Graphics in Producing Mixed Reality Worlds

Hideyuki Tamura and Hiroyuki Yamamoto
Mixed Reality Systems Laboratory Inc.
6-145, Hanasaki-cho, Nishi-ku, Yokohama 220-0022, Japan
e-mail : {tamura, ymmt}@mr-system.com

Abstract

This paper introduces prominent topics of our Mixed Reality (MR) project and states what kind of role computer vision and graphics play in the project. MR is a part of VR in broader sense, but it treats the physical space as well as the virtual space created in computers. Here we use the term "mixed reality" instead of often used "augmented reality (AR)." This is because MR is not only the mixture of real world and virtual world but also the mixture of AR and "augmented virtuality (AV)." In that sense, we first describe that there is no clear distinction between AR and AV, then introduce some research results at the early stage of our MR project.

Keywords: *Mixed reality, augmented reality, augmented virtuality*

1. Introduction

There has been nearly ten years since we started to use a phrase "Virtual Reality (VR)." Most of VR systems we have experienced in this decade made it possible for participants to interact with virtual environments which are totally synthesized in computers. As everyone knows, the reality in synthetic world is limited in its nature. Because of this limitation, people tend to positively utilize rich information in the real world.

We have been participating in the Research Project on Mixed Reality whose target is to develop the technology merging the real world and the virtual world seamlessly. Our Mixed Reality Systems Laboratory Inc. was established to conduct the project in January 1997 on the investment by the Japanese government and Canon Inc. This national project is relatively in short term and is planned to be extended to March 2001 collaborating with three universities, Univ. of Tokyo (Prof. M. Hirose), Univ. of Tsukuba (Prof. Y. Ohta), and Hokkaido Univ. (Prof. T. Ifukube), in Japan.

"Mixed Reality (MR)" is a kind of VR in broader sense but the term is not widely recognized yet. People often use the term "Augmented Reality (AR)," instead of MR, which augments the real world with synthetic information. In other words, the system adds electronic data from a cyberspace on the physical space as a base. On the opposite side, there is a term "Augmented Virtuality (AV)." That enhances or augments the virtual environment with data from the real world. Our intent for the term "MR" is in mixing AR and AV.

This concept inherits the definition of MR stated in [1] by Paul Milgram. His taxonomy said that there is no clear distinction between AR and AV, and MR is "virtual continuum." He has pointed out six classes of display for MR. We are now feeling that there is no border between AR and AV by considering the following four classes.

Class A : Optical see-through MR
The most prominent AR method is seeing the real world through optical see-through glasses on which images are superimposed. In this system, the two worlds are merged on the retinas of an observer.

Class B : Video see-through MR
This method uses TV camera attached on a head mounted display (HMD) to take scenes in front of an observer, which one usually see through his/her naked eyes, into the system and electrically merges the video images of the real world and virtual images. Observers feel images from the two different worlds merged evenly.

Class C : On-line tele-presence
You can think of a system, by extending the concept of video see-through, which provides video images transmitted from a remote site into the HMD. This system can produce a MR world by merging computer generated images into the video images from a remote stereoscopic camera which is synchronized with tracked head action of an observer.

Class D : Off-line tele-presence

With a little modification of the idea of Class C, you can think of a system in which images are reconstructed from an image database having various prerecorded images instead of realtime capturing of remote scene. The system of this class is developed by only adding a method to provide images according to the viewpoint of an observer. This kind of method must not be classified to AR but AV. The system, however, which merges computer generated images into reconstructed realistic images can be thought as a MR.

The virtual world is “synthesized” or “created” or “generated.” On the other hand, the real world is already presents or exists physically. Therefore, we dared to use a word “produce” for the creation of MR worlds. In the MR worlds implemented by using computers, HMDs and various sensor devices, there can be some displays which stimulate auditory or tactile sense. However, original and innovative portion of our research project mainly focuses on the visual mixture. Conventional computer graphics techniques had been playing the main role when people had dealt only with the completely synthetic worlds. On the contrary, methods of computer vision and image processing are now expected to play some important roles in the production of MR worlds. This is because it is required to recognize outer spaces and to register the physical space and the virtual space both geometrically and radiometrically in order to seamlessly merge these two spaces.

This paper introduces prominent topics of our MR project and states what kind of role vision and graphics play in our project.

2. Topics on Augmented Virtuality Research

An example of most elementary AV is the texture mapping used in the computer graphics. This method maps image data captured from the physical space on to a geometric model in order to enhance reality of the model. However, since this method requires a certain geometric model of an object, it becomes difficult to represent complex shaped objects. For this reason, people now focusing on a method which reconstruct an arbitrary view directly form captured images.

2.1 Photorealistic presentation of complicated objects

A few years before starting research of MR, we had already studied about AV. In that study we had been seeking a way to handle objects and their backgrounds having complex shapes which could not be drawn using conventional computer graphics techniques in a virtual space. We tried to reconstruct a scene which coincides with the viewpoint of observers from various real images without expressing a virtual space with data based on the geometric models. At that time, Apple Computer's QuickTimeVR [2] had already been introduced which could reproduce a scene as a photographic panorama seen from a limited viewpoint. Note that, in QuickTimeVR, the scene did not follow the motion when an observer moved.

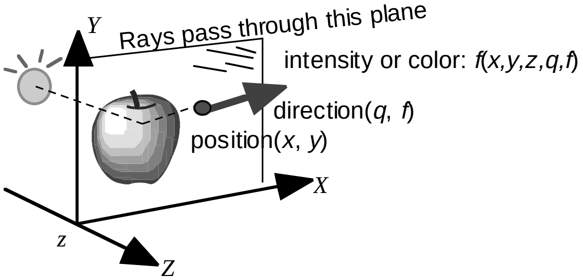

Y
Rays pass through this plane
intensity or color: f(x,y,z,q,f)
direction(q, f)
position(x, y)
X
z
Z

Figure 1: Ray space description

Our goal was to find out a method to reconstruct an image which produced motion parallax when an observer moved around it. The method had to reconstruct a required image from images captured by multiple cameras placed evenly on a line by interpolating images from those cameras. The problem eventually became a simpler one to find out a straight line from an epipolar plane image (EPI) []. That was really a technique of computer vision or image processing. Applying this theory, we have developed a Holomedia system [4] that gives an observer stereoscopic images through liquid crystal shutter glasses with a head tracker. No geometric data was used in this method at all. Now, such a method is called “image-based rendering (IBR).”

By generalizing the method based on the EPI, we have advanced to image-based rendering based on the “Ray Space” method. This method advocated by H. Harashima and others [5] is the one to produce radiometric representation of an object as a bundle of rays which go through a certain point on a screen at a certain time. Figure 1 illustrates this. The theory has the same basis as the Lumigraph by Cohen [6] or the Light Field by Levoy [7]. All these methods perform image-based rendering from a lot of pictures captured from the real world.

We have then tried to draw an image by merging geometric model-based data and ray-based data. Finally, we have completed a system in which an observer can walk through MR space which is constructed by complex objects represented by textual image-based data placed in circumstances of polygon-represented graphic data. Figure 2 shows an example of this type of data structure. The system expanded from the VRML Viewer is called CyberMirage [8].

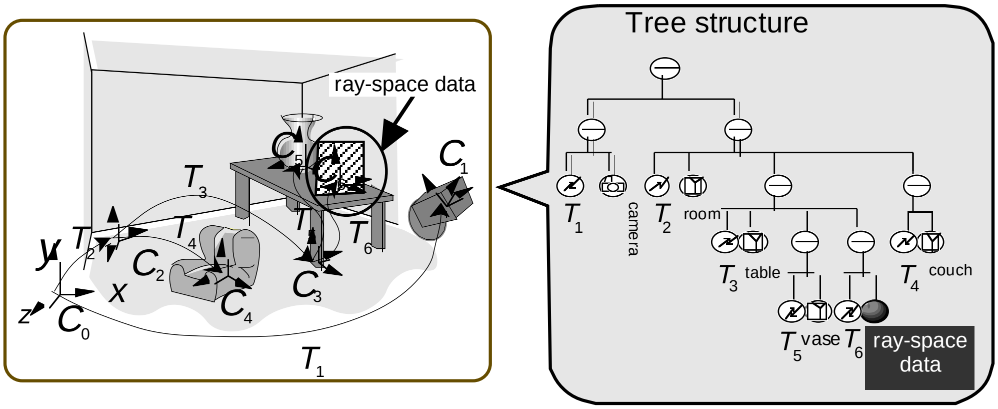


Figure 2: Ray space data embedded in VRML data structure

We are now publishing CyberMirage Viewer for a single user onto the WWW [9]. Collaborative CyberMirage [10] is an expanded inheritance of the CyberMirage in which multiple remote participants can visit a cyberspace built on a network and communicate with each other in realtime while recognizing other participants as avatars. Figure 3 shows a user's view when three participants collaborate in a MR world. Our Collaborative CyberMirage taking a cybershopping in a virtual mall a model was tested by linking multiple points apart several dozen kilometers with each other with lines of 6 Mbps. The research is mainly reviewed from the point of telecommunication system such as how to compress and transmit huge image-based data such as some megabytes per object.

In our MR project, we have to merge the image-based data in any class of implementation from the Class A to Class D. We have already achieved a certain extent for the photoreality of a single object not affected by circumstances. Therefore, the next problem to be solved is the shading of an object placed under some lighting. The shading becomes fixed when we reproduce an object from images captured under fixed lighting. In order to achieve seamless merge, we have to find out some measure to alter shading according to the lighting condition in the real world in which the object exists.

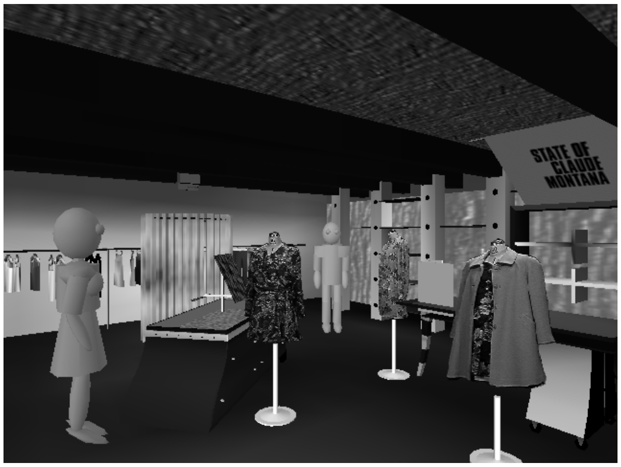


Figure 3: A user's view in Collaborative CyberMirage

In order to realize this, we must be able to estimate lighting in the real world. This is an inverse problem of the shape from shading in computer vision.

2.2 Building image-based cybercities

We are now thinking about to render circumstances or backgrounds with real images not only solid objects as in the virtual mall. This is a plan to take some hundreds square of a city into a cyberspace where participants can walk through [11].

In the famous Aspen Movie Map developed by MIT Machine Architecture Group [12], participants can see the city in a certain angle in which the images are taken. Our aim is to make it possible for participants who walking through the city to see everything in it from their point of views. This is a problem of Class D.

Figure 4 shows our approach to building image-based cybercities.

Sampling of images

In this part, a huge number of images from a wide area are sampled. For this purpose, we need a transducer with physical mobility. We adopted an automobile with plural video cameras on board. To know the position and posture of cameras, we use three kinds of sensors, a GPS sensor, gyro sensors, and geo-magnetic sensors. Geo-magnetic sensors are used to compensate drifts of gyro sensor. Fig.5 shows outside and inside of the automobile. Both image database and position-posture database are indexed by time codes from which these two databases can be integrated easily.

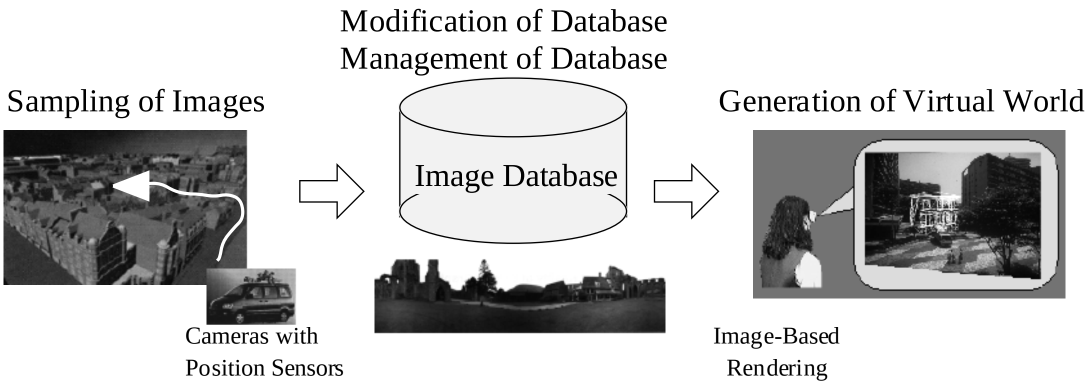


Figure 4: Process flow of building cybercities

Modification of database

In this part, errors involved in the position and image database are modified. For position data, sensor-fusion with plural position sensors and modification of errors with respect to data output timing must be operated. And for image data, modification of lens distortion, adjustment of white balance and brightness, and removal of obstructions such as people and cars in images should be enabled. Since the shadings or shadows can change as time goes by (such that the sun is covered by a cloud), these shadings have also to be compensated. This is a problem of the computer vision.

Management of database

In this part, an image database by indexing position data to images is produced. In our prototype, seven images captured at nearly same locations are integrated into a panoramic image. In order to access to the database effectively, we have to examine image formats to store data into the database.

Compression technology for the database is also necessary because this image database is estimated to be a massive amount. This compression technology requires a high compression rate, allowance of random access, and real time decompression. Vector quantization [13] is one candidate of this kind of method.

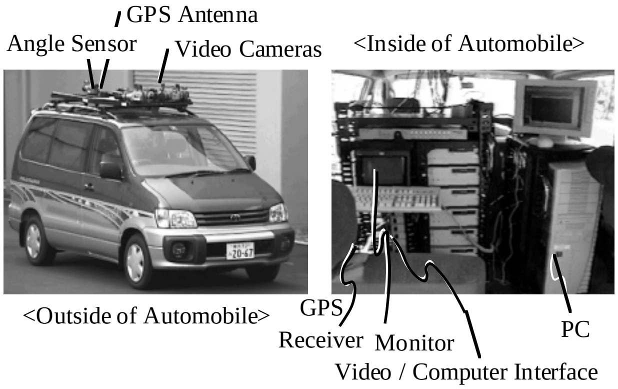


Figure 5: Data sampling equipment

Generation of virtual spaces

In this part, 3-D virtual space from an image database is generated. To generate images from arbitrary view points, image interpolation techniques can be applied. Methods described in [5], [14], and [15] are considerable to be applied to this system. These techniques are also useful to reduce the quantity of image sources.

In our system, the view morphing method is now used to reconstruct images and makes it possible for a participant to walk-through a limited area of the cybercity. In this method, corresponding points of captured images must be known in order to generate images from arbitrary viewpoint. Therefore auto-detection algorithm of corresponding points is required in view of an amount of data to be processed. This is also a problem of computer vision.

3. Topics on Augmented Reality

The AR technology that seamlessly enhances a physical space with computer generated information has a wider range of applications than conventional VR. Instructing repair or maintenance procedures of a complex machinery [16], manufacturing, in-patient visualization of medical data [17], annotation [18] are some of the application fields in which AR can be utilized. These AR researches, however, have been made mainly on single-user applications so far. New application fields, especially in the field of human communication, will appear if multiple operators can share a physical space and if we can seamlessly offer a virtual space into the shared physical space [19]. For example, it becomes possible for multiple people to collaborate to design something while exchanging their ideas through virtual objects [20].

We have developed the AR AiR Hockey (AR^{2}Hockey) system [21] as a case study of the collaborative AR for human communications. In this study, the collaborative

AR is a method to establish an environment in which participants get together and make collaboration while sharing a physical space and a cyberspace simultaneously.

3.1 Collaborative AR

Air hockey is a game in which two players hit a puck with mallets on a table and shoot it into goals. In our AR^{2}Hockey, a puck is in a virtual space. Each player wears an optical see-through HMD and hits a virtual puck placed on a real table with a physical hand. Figure 6 (a) shows the scene of playing AR^{2}Hockey and (b) is an image seen through the HMD when the system is operated.

Figure 7 (a) shows the typical coordinate systems used in simple AR. The registration is the process that transforms the viewing matrix C_C . In collaborative AR, the physical space and virtual space are shared by all the participants. Thus the coordinate system C_R and C_V exist in the system and shared by the participants. On the other hand, the coordinate system C_C and C_D that relate to the viewing transformations exist for each participant. Figure 6 (b) illustrates this situation. Thus the registration described in the following subsection is implemented independently for each participant.

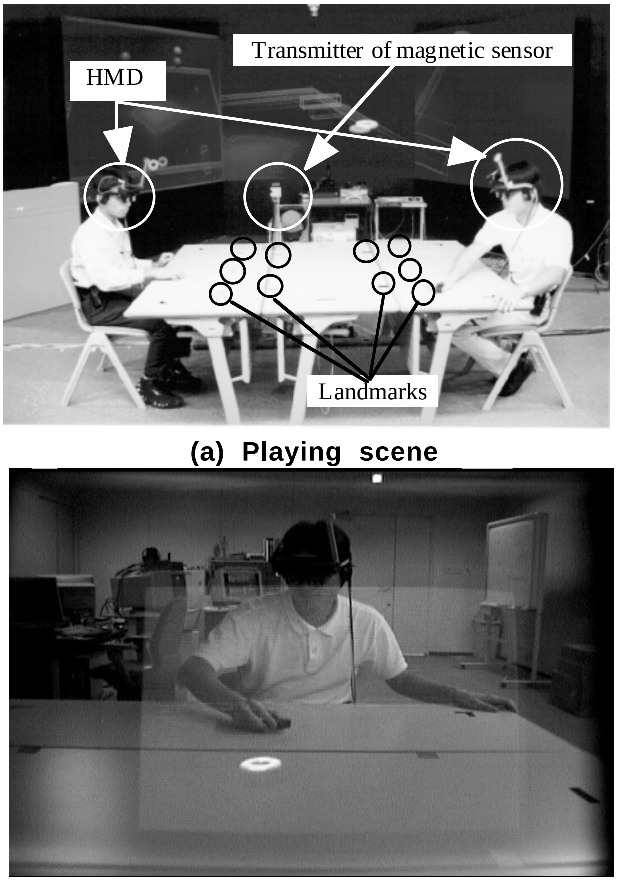
3.2 Registration

Registration is a technique to seamlessly combine virtual objects according to the condition of the physical space. It includes recognition of a physical scene and compensation process of the virtual space according to the result of the recognition. Therefore, the technology of the computer vision, which makes it possible for a computer to recognize or understand images of the physical space, becomes quite important to realize smart registration system.

In our AR^{2}Hockey, it is required to minimize positional misalignment and time lag between the physical space and the virtual space to allow participants to hit a virtual puck with physical hands.

Positional misalignment

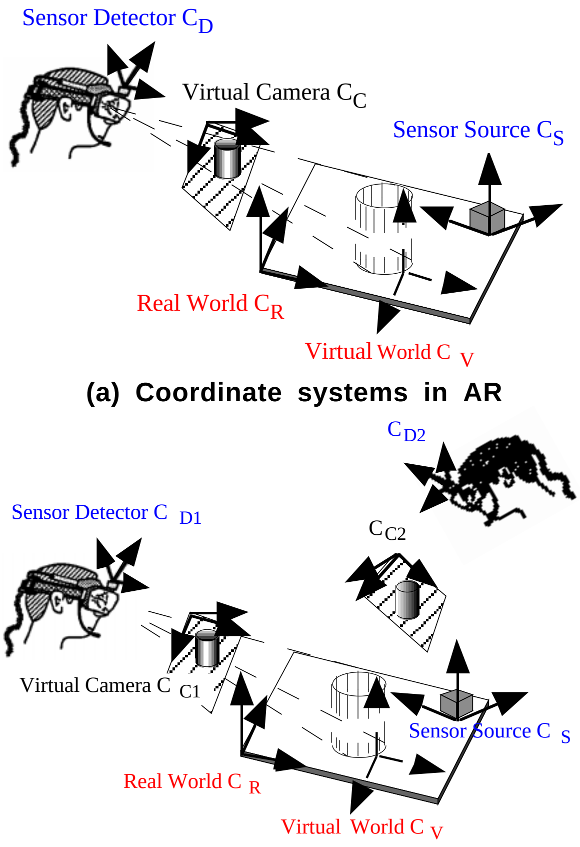
Registration of positional misalignment can be translated into the problem to determine the 3D position of a user's viewpoint. To decide the viewpoint, it is generally measured by 3D sensors such as magnetic



(a) Playing scene

(b) Player's view

Figure 6: Playing scene of AR^{2}Hockey



(a) Coordinate systems in AR

(b) Coordinate systems in collaborative AR

Figure 7: Coordinate systems in AR

sensors, ultrasonic sensors or gyroscopes. Since these sensors can not always give us enough accuracy required, the error in these sensors causes positional misalignment. The image captured by the camera attached to HMD can be used to register such positional misalignment. One approach is to utilize a camera calibration method developed in computer vision research and discard the sensory output. The other is to correct the sensory output based on the image information. Our AR^{2}Hockey adopts the latter method considering speed and reliability of processing. The following explains this method.

Figure 8 shows the basic theory to correct positioning error using one landmark. The discussion below assumes that all the inner camera parameters are already known and an image is captured by an ideal capturing system without any distortion.

In the figure, let C, I, and Q be the camera position, the image plane, and the landmark position in the physical 3D space, respectively. For those C, I, and Q, the projected landmark position on the image Q' is determined as the point at which the line l_Q connecting C and Q intersects the image plane I. On the other hand, the landmark position P in the camera coordinate system and the corresponding position on the image P' are calculated based on the 3D sensor data. These P and P' can be thought as the landmark in the virtual space and its corresponding image position. Ideally, point Q and P coincide in the 3D space. That is , the projected position Q' and P' coincide on the image plane. This is, however, usually not true because of the 3D sensor error.

The correction is done by translating the virtual space coordinates so that the corrected predicted observed coordinate of the landmark P' coincides with the point Q'. This is done by translating objects in the virtual space by

$$v = n(v_1 - v_2) \qquad (1)$$

where n is a scale factor derived by

$$n = \frac{|CP|}{|CP'|} \quad . \qquad (2)$$

Since this method only registers the positions on 2D image plane, the three dimensional positions may not coincide correctly even after the registration. This method, however, is still effective if the sensor error is not so much (see Fig. 8).

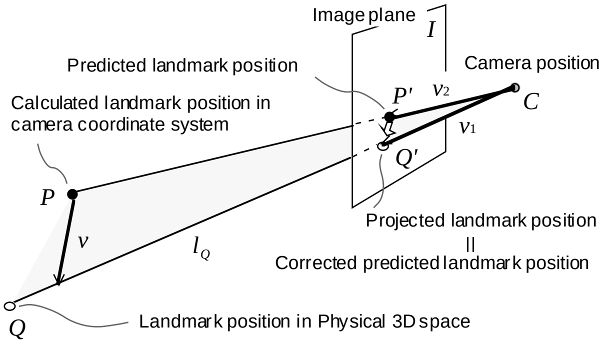
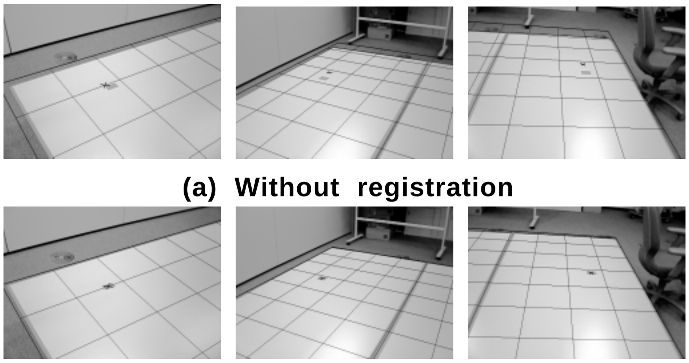


Image plane
I
Predicted landmark position
Camera position
Calculated landmark position in camera coordinate system
P'
v_2
C
v_1
Q'
P
v
l_Q
Projected landmark position
||
Corrected predicted landmark position
Q
Landmark position in Physical 3D space

Figure 8: Registration with one landmark



(a) Without registration

(b) With registration

Figure 8: Effects of registration

Time lag

The physical space follows the viewpoint of an observer without any time lag. On the other hand, it requires some amount of time to sense the change of the viewpoint and render a virtual space accordingly. If this duration becomes larger, observers feel virtual objects floating in the physical world.

To minimize this floating feeling, we have to speed up every component of the system as fast as possible. In our AR^{2}Hockey, we have adopted a high-speed graphic workstation to reduce rendering time. On the software side, we have applied simple algorithm to predict the head position, orientation, and marker position at the time the rendering is finished. This is done by the second order prediction algorithm described in Eq. (3).

$$\hat{p} = p_t + \frac{1}{2} a_t \Delta t^2 + v_t \Delta t \qquad (3)$$

where $\hat{p}$ is the predicted value,

p_t is the latest recorded data,

v_t, is the velocity at the time p_t is recorded,

a_t is the acceleration at the time p_t is recorded,

Δt is the elapsed time from the moment p_t is recorded.

The configuration of processes has also been considered. As shown in the Fig. 9, the process is composed of six sub processes and one master process. Four subprocesses, head tracking, marker tracking, registration and rendering are invoked for each player. The system is asynchronous.

That means that three tracking processes that drive input devices proceed independently from the master process and parallel to each other. On the other hand, the registration, the space management and the rendering processes are synchronous to the master process. By configuring the system in this way, it becomes possible to reduce the effect caused by the difference of the sampling rates of 3D sensors, video capturing rate and the rendering rate.

High-speed computer vision method is an important theme. Since research in computer vision so far has focused mainly on algorithms and theories, the processing speed has not been examined well enough. Since the update rate of 60Hz may be not enough for the AR application, we have to speed up the processing time of computer vision method.

Shadowing

In the AR system, some measure is required to reduce or completely remove incongruity from the real world by adjusting contrast and hue of the rendered virtual world. Observers also feel incongruity if the lighting condition is different between the real and virtual worlds. In our AR^{2}Hockey environment, we have applied the virtual space almost the same lighting condition as the physical space since it is easy to control the lighting of the physical space. This reduces the feeling of the puck floating on the table. In the future, we have to estimate the lighting condition of the physical space, which is dynamically changing, using the technique of the computer vision, and apply the results of the estimation onto the virtual space.

Depth-keying

In order to realize seamless mixed reality space, it is required to represent mutual occlusion between the real and virtual worlds, which has not been implemented on the AR^{2}Hockey. For this purpose, we require some method to measure instantly and accurately the distance from the observer to the objects and circumstances of the real world [22]. The stereo matching method in computer vision can be applied to this field.

4. Concluding Remarks and Future Work

We perceived the concept of “Mixed Reality” that Paul Milgram proposed as the concrete research target of our project. This paper described its early results.

Mixed reality has to deal with the mixed problems of computer vision and graphics. Image-based rendering is in the spotlight and becomes an attractive research topic in computer graphics. This is because the image processing is recognized as very useful in order to overcome the limit of conventional computer graphics techniques. Computer vision is necessary to make the construction process of IBR data (semi-)automatic and to make it practical.

In section 2, we presented topics of rendering objects on a table and circumstances such as streets without any explicit geometric data. We now recognize that the middle

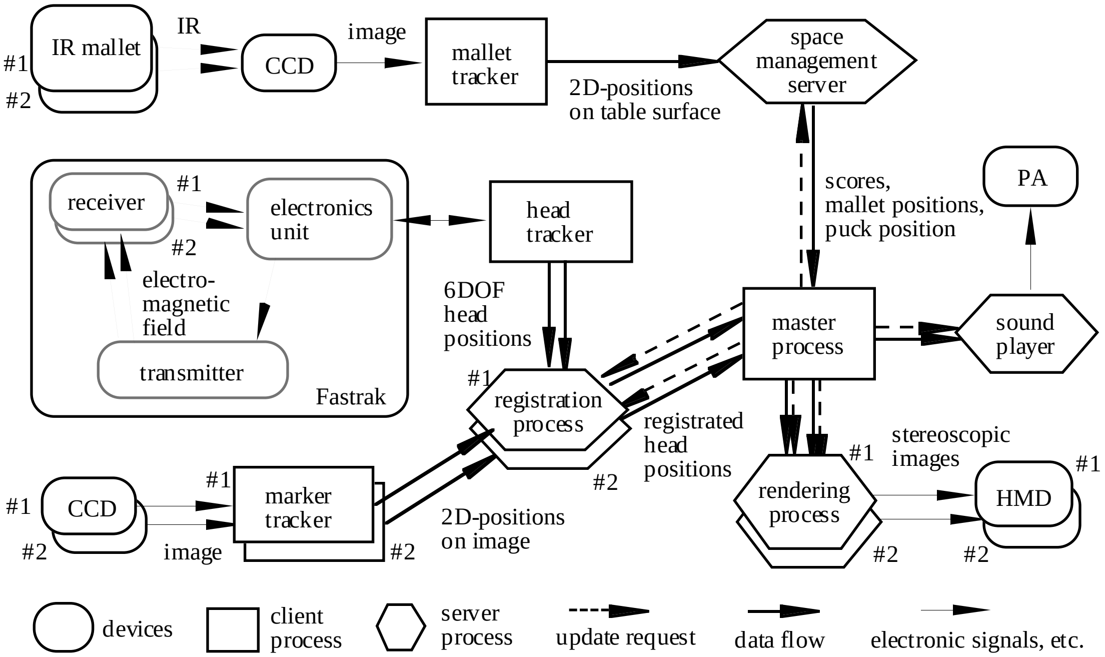


Figure 9: Software architecture in AR^{2}Hockey

range space, that is, from a few meters to around ten meters away, is not a easy target. Considering the required resolution and quality of reconstructed images, building sparse geometric data from range images is a better approach. We can make the most use of computer vision technique for implementing a range and image scanner to get such data.

Many problems should be solved to generalize and make AR of practical use. In the AR^{2}Hockey, the movement of mullet is constrained on top of the table, that is a 2-D plane. We actually did not choose the table tennis or squash but the air-hockey in order to apply this constraint. This is because suitable 3-D tracking sensors are not currently available.

Sensor fusion with physical 3-D sensors and tracking landmarks on the captured images is an optimal solution for the smart registration for the time. Prediction of the head movement might be effective to decrease the time lag. Smart researchers in computer vision area could be aware of that AR is a treasure house of problems that they can make the most use of computer vision.

Acknowledgments

The authors would like to thank Mr. Akihiro Katayama, Dr. Toshikazu Ohshima, Mr. Shinji Uchiyama, and Dr. Kiyohide Sato for their cooperative works.

References

[1] P. Milgram and F. Kishino, "A taxonomy of mixed reality visual display," *IEICE Trans. Inf. & Sys.*, Vol. E77-D, No.12, pp.1321-1329 (1994).

[2] S. E. Chen, "QuickTime VR - An imaged-based approach to virtual environment navigation," *Proc. SIGGRAPH '95*, pp.29-38 (1995).

[3] R. C. Bolles, H. H. Baker, and D. H. Marimont, "Epipolar-plane image analysis: An approach to determining structure from motion," *Int. J. Computer Vision*, Vol.1, No.1, pp.7-55 (1987).

[4] A. Katayama, K. Tanaka, T. Oshino, and H. Tamura, "A viewpoint dependent stereoscopic display using interpolation of multi-viewpoint images," *Proc. SPIE*, Vol. 2409A, pp.11-20 (1995).

[5] T. Naemura, T. Takano, M. Kaneko, and H. Harashima, "Ray-based creation of photo-realistic virtual world," *Proc. VSMM '97*, pp.59-68 (1997).

[6] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The Lumigraph," *Proc. SIGGRAPH '96*, pp.43-54 (1996).

[7] M. Levoy and P. Hanrahan, "Light field rendering," *Proc. SIGGRAPH '96*, pp.31-42 (1996).

[8] S. Uchiyama, A. Katayama, H. Tamura, T. Naemura, M. Kaneko, and H. Harashima, "CyberMirage: Embedding ray based data in VRML world," *Video Proc. VRAIS '97* (1997).

[9] http://www.x-zone.canon.co.jp/CyberMirage/

[10] S. Uchiyama, A. Katayama, A. Kumagai, H. Tamura, T. Naemura, M. Kaneko, and H. Harashima, "Collaborative CyberMirage: A shared cyberspace with mixed reality," *Proc. VSMM '97*, pp.9-18(1997).

[11] M. Hirose, S. Watanabe, and T. Endo, "Generation of wide-range virtual spaces using photographic images," *Proc.VRAIS '98* (to appear).

[12] A. Lippmann, "Movie-Maps/An application of the optical videodisc to computer graphics," *Computer Graphics*, Vol.14, No.3, pp.32-42 (1980).

[13] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, Vol. 28, No. 1, pp.84-95 (1980).

[14] S. M. Seitz and C. R. Dyer, "View morphing," *Proc. SIGGRAPH '96*, pp.21-30 (1996).

[15] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," *Proc. SIGGRAPH '95*, pp39-46 (1995).

[16] S. Feiner, B. MacIntyre, and D. Seligmann, "Knowledge-based augmented reality," *Comm. ACM*, Vol.36, No.7, pp.52-62 (1993).

[17] M. Bajura, H. Fuchs, and R. Ohbuchi, "Merging virtual reality with the real world: Seeing ultrasound imagery within the patient," *Computer Graphics*, Vol. 26, No.2, pp.203-210 (1992).

[18] E. Rose et al., "Annotating real world objects using augmented reality," *Proc. Computer Graphics International '95*, pp.357-370 (1995).

[19] M. Billinghurst, S. Baldis, E. Miller, and S. Weghorst, "Shared space: Collaborative information spaces," *Proc. HCI International '97*, (1997).

[20] J. Rekimoto, "TransVision: A hand-held augmented reality system for collaborative design," *Proc. VSMM '96*, pp.85-90 (1996).

[21] T. Ohshima, K. Sato, H. Yamamoto, and H. Tamura, "AR^{2}Hockey: A case study of collaborative augmented reality," *Proc. VRAIS '98* (to appear).

[22] T. Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka, "A stereo machine for video-rate dense depth mapping and its new applications," *Proc. CVPR '96*, pp.196-202 (1996).