The Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality 2007
November 13-16, 2007, Nara-Ken New Public Hall, Japan

A Two-by-Two Mixed Reality System That Merges Real and Virtual Worlds in Both Audio and Visual Senses

Kyota Higa[*], Takanobu Nishiura[*], Asako Kimura[†,‡], Fumihisa Shibata[*], and Hideyuki Tamura[*]

[*]Graduate School of Science and Engineering, Ritsumeikan University
[†]PRESTO, Japan Science and Technology Agency
[‡]Research Organization of Science and Engineering, Ritsumeikan University

ABSTRACT

There have been many implementations of virtual reality, using audio and visual senses. However, implementations of mixed reality (MR) have thus far only dealt with the visual sense. We have developed an MR system that merges real and virtual worlds in both the audio and visual senses, wherein the geometric consistency of the audio sense was fully coordinated with the visual sense. We tried two approaches for merging real and virtual worlds in the audio sense, using open-air and closed-air headphones.

CR Categories: H.5.1 [Multimedia Information Systems]: Artificial, augmented, and virtual realities—

Keywords: Mixed Reality, Audio and Visual Senses, Geometric Consistency, Open-Air Headphones, Closed-Air Headphones

1 INTRODUCTION

This paper relates to a novel audio–visual mixed reality (MR) system. We have succeeded in developing a system that can simultaneously and consistently manage visual and auditory MR in real time.

Augmented reality (AR) and MR are significant extensions of the conventional virtual reality (VR) that handle virtual environments constructed by a computer. Originally, VR was advocated to treat various types of sensory information such as visual, auditory, tactile, and olfactory senses. Auditory VR [1]-[3] was achieved by generating a 3-D sound field. Until data, there have been many systems that simultaneously manage visual and auditory VR, for example [4]-[6].

On the other hand, most of the earlier AR/MR systems were primarily designed to cover a visual mixture of real and virtual worlds [7][8]. Sometimes audio and tactile sensations were added to visual MR in the form of stage effects [9][10]. More recently, some AR/MR systems have been developed, which also incorporates 3-D sound [11]. However, all these systems simply added auditory VR to visual MR. They never developed any auditory MR that compatibly presents both real sound (originating from the surrounding environment) and artificial sound (generated through HRTFs) in real time, and did not discuss about the auditory MR (for example, how to mix real and artificial sounds, etc.).

In this study, we focused on the implementation and development of an MR system that could present visual MR and auditory MR, both simultaneously and compatibly. The system we present realizes the "merging of real and virtual worlds" and the "joining of visual and auditory sensations"; thus, we call it "two-by-two mixed reality."

In order to achieve its purpose, this system should be implemented with

(1) A combination of real and virtual worlds in the visual sense
(2) A combination of real and artificial sounds processed by a computer and presented in the audio sense
(3) Geometrical and real-time coordination of visual and audio MR

There have already been many studies covering (1), and we have a track record for using the existing system. Therefore, in this study, we implemented (2) and (3). In other words, we developed an MR system using both audio and visual senses.

The users of (1) generally wear a see-through head-mounted display (HMD) to experience the MR space. In addition, the combined sound for (2) is not presented from speakers but from headphones; there are two possible ways to hear the combination of real and artificial sounds—open-air or closed-air headphones. These are similar to the "optical see-through" and "video see-through" methods employed in visual MR. In this study, we adopted these two audio MR methods and discuss the results.

2 MIXED REALITY IN BOTH AUDIO AND VISUAL SENSES

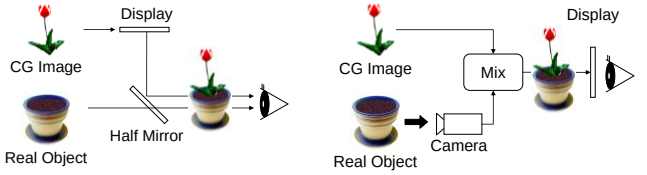
2.1. Mixed Reality of Visual Sensation

There are two well-known display methods that can be used for achieving a mixed reality of visual sensation—optical see-through and video see-through. Fig. 1 illustrates the differences between these two display methods. An optical see-through display can be implemented by a half-mirror mechanism (Fig. 1(a)). In other words, the real-world scene and the computer-generated image (CGI) are physically merged at the retina in the human eye. On the other hand, in the video see-through method (Fig. 1(b)), a pair of video cameras, mostly attached to or built inside a closed-type HMD, capture the real scene instead of the human eye and transfer the image to a computer. Then, the CGI is superimposed upon them and the resultant composite image is shown. Although each has its merits and demerits, the video see-through method is widely used because of its superior optical consistency. Therefore, the video see-through HMD was used in this study.

2.2. Mixed Reality of Auditory Sensation

It is quite difficult to present 3-D sounds to multiple users in an MR space using the transaural method. Consequently, in this study, the binaural method was preferred for presenting a 3-D sound field to each subject.

[*, ‡] 1-1-1 Noji-Higashi, Kusatsu 525-8577, Shiga, Japan
[†] 4-1-8 Honcho, Kawaguchi 332-0012, Saitama, Japan
{higa, asa, shibata, tamura}@mclab.ics.ritsumei.ac.jp
{nishiura}@is.ritsumei.ac.jp



(a) Optical see-through (b) Video see-through

Figure 1. Two see-through methods

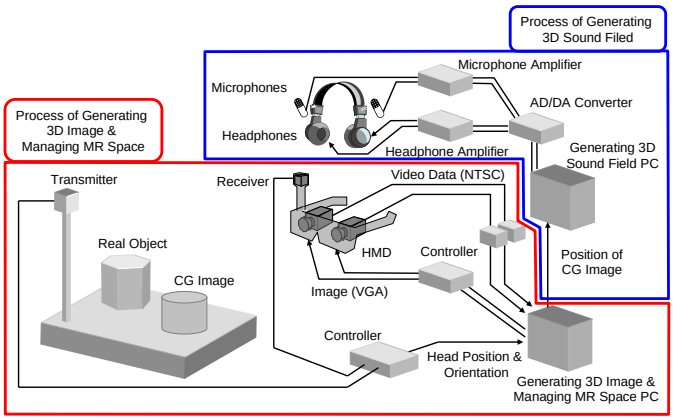


Figure 3. System configuration

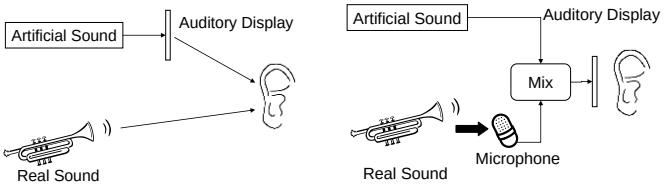
When thinking about an auditory mixture of real and virtual worlds, two methods are possible similar to that applicable to a visual mixture. Fig. 2 illustrates the differences between two auditory mixing displays. The "physical mixing" in Fig. 2(a) can be experienced using open-air headphones. Here, artificial sounds through the headphones, either pre-recorded or synthesized, and real sounds heard through a pair of gaps between each ear and the headphone's ear covers, are physically mixed. This is similar to the optical see-through method used in the visual sense. "Electronic mixing," as shown in Fig. 2(b), can be achieved using closed-air headphones. Natural sounds in the real world are gathered by a pair of microphones attached to the outside of the headphones and then mixed with the artificial sound using a computer. The resultant mixed sound reaches the user through closed-air headphones. This auditory mixing method corresponds to the video see-through method.

One of the main purposes of our study was to evaluate and determine which method was the most appropriate for the desired MR system.

2.3. System Configuration of the Two-by-Two MR System

Using the above-mentioned methods, we tried to construct an MR system that merges real and virtual worlds in both the audio and visual senses. To avoid any possible confusion with a VR system handling a completely synthesized AV world or a simple visual MR system, we call it a "two-by-two mixed reality" system. This means that we can present any combination of data from real and virtual images through audio and visual means.

To achieve total geometric consistency of a two-by-two MR world in real time while the user is moving and interacting, we adopted the system shown in Fig. 3. This system is composed of two subsystems. One manages the total MR space and displays it visually. The other only displays the MR space in the audio sense. The former subsystem superimposes CGIs onto the real world image captured by a pair of video cameras built into the HMD and shows the resultant stereoscopic images on HMD-LCD panels. The user's head position and orientation are detected and tracked



(a) Physical mixing method (b) Electronic mixing method

Figure 2. Two methods of sound image presentation

by magnetic sensors. The latter subsystem generates a 3-D sound field and displays sounds associated with the virtual objects.

3 Acoustic Consistency Experiment for Real and Artificial Sounds

3.1. Purpose and Preparation

We assessed the adequacy of these two mixing methods for audio and visual MR from the viewpoint of acoustic consistency based on two types of experiments.

In the experiments, we defined the sound played from a speaker (MITSUBISHI DIATONE DS-7) as the sound of the real world. The artificial sound is processed by a computer and reproduced through the headphones. The open-air headphones (SONY MDR-F1) are used for physical mixing and the closed-air headphones (PELTOR HTM79B-S) are used for electronic mixing. In the electronic mixing method, nondirectional microphones are connected to the closed-air headphones to obtain real-world sounds. The influence of reflected sounds, which result from the properties of the real world, is not considered in this system.

The audio sampling frequency was 16 kHz, the frame length for sound processing was 64 ms per channel, and the image processing speed was approximately 12.5 fps. The experiments were undertaken in a regular office environment with a background noise of 48 dBA. Ten university students participated in these experiments.

3.2. Experiment 1: Evaluation Experiment of Mixing Sound by Static Sound Source

3.2.1. Description

The real and artificial sounds were set as a static sound source. The test subjects heard both sounds simultaneously and answered questions based on sound source localization and sound quality. The sound quality evaluation used a subjective appraisal scheme based on five-levels, from 1 (bad) to 5 (excellent). For artificial sound, a helicopter propeller's sound was chosen. The real sound (classical music) was provided by one of three speakers surrounding the subjects (Fig. 4).

During the experiment, the subjects wore HMD. The CGI helicopter was superimposed at the position of the virtual sound.

3.2.2. Procedures

(1) The real and artificial sounds were presented for 3 s from directions randomly selected from the 12 possible directions.
(2) The subject was asked to locate the direction of each sound.
(3) The sounds were presented for an additional 3 s.
(4) An evaluation of the sound quality for each sound was obtained using the 5-grade evaluation process.
(5) Comments were collected from the subject.
(6) The above steps were repeated until the subject located the source of the sound twice for any of the 12 possible directions.

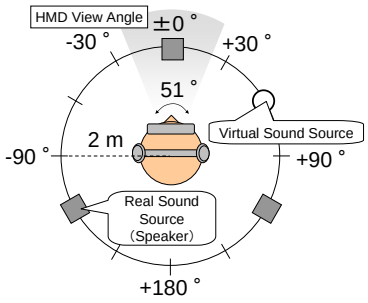


Figure 4. Reception of real and artificial sounds in Ex. 1

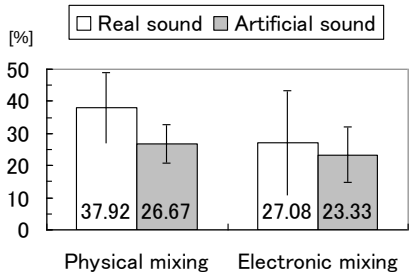


Figure 5. Evaluation of sound localization accuracy in Ex. 1

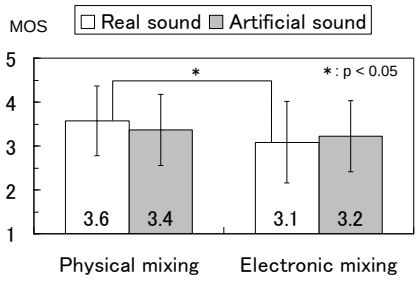


Figure 6. Evaluation of sound quality in Ex. 1

3.2.3. Results

The experimental results regarding localization of the real and artificial sounds are shown in Fig. 5, where the vertical axes show the percentage for each correct answer and its standard deviation. Fig. 6 shows the evaluation of the sound quality, where the vertical axes show the values of the mean opinion score (MOS). In the evaluation of sound localization accuracy, no significant difference was found between physical and electronic mixing of real/artificial sound, according to the t-test. The results themselves seem to be valid, considering the simultaneous presentation of two sounds from two directions.

In the sound quality evaluation results, a significant difference was found ($p < 0.05$) between physical mixing and electronic mixing of real sound; however, there was no significant difference between either method in the mixing of artificial sound.

Eight out of the ten subjects, who evaluated the electronic mixing method, made the comment, "I feel uncomfortable when I say something." This could be caused by hearing their own voice being directly transmitted through their body as well as via the headphones. In addition, most of the subjects also made the comment, "I am very bothered by environmental sounds (such as noise from PCs)." In particular, rustling of the cables from the headphones and the microphones produced harmful effects.

3.3. Experiment 2: Evaluation Experiment of Mixing Sound by Mobile Sound Source

3.3.1. Description

In this experiment, we evaluated whether the subjects could correctly localize the direction of sound from a mobile sound source. The sound localization of the mobile artificial sound source and the position fixed real sound was evaluated when subjects heard them simultaneously (Fig. 7).

The sounds presented were the same as those used in experiment 1. The velocity of the helicopter was 0.5, 1, 5, 10, 15, and 20 m/s. Subjects wore headphones and an HMD, and the CGI helicopter was superimposed with the virtual sound source position.

3.3.2. Procedure

(1) The real and artificial sounds were presented simultaneously.
(2) Evaluation of the sense of sound localization of the virtual sound was obtained using a 5-grade evaluation.
(3) Comments were collected from the subjects.
(4) The above steps were repeated with a change in the velocity of the virtual sound source.

3.3.3. Results

Fig. 8 shows the results from experiment 2. The highest value occurs when the helicopter's velocity is 1 m/s; as the helicopter speeds up, the value of MOS decreases. No significant difference was found between physical and electronic mixing at each propeller velocity. When the velocity is high, the movement of the helicopter's image appears discontinuous due to the frame rate restriction of the CGI.

3.4. Synthetic Judgment of Method Selection

Prior to the above-mentioned experiments, we presumed that the electronic mixing method was only suitable for presenting mixed sounds, when compared with the physical mixing method. This supposition was based upon the superiority of the "video see-through method" with regard to the visual sense. However, in experiment 1, the following major drawbacks of the electronic mixing method become clear:

- Duplication in hearing the voice, which is directly transmitted through the subject's body, as well as via the headphones
- Uncomfortable feelings associated with the negative effects of rustling sounds from the headphones and the microphones

The former is an unexpected problem, which cannot be overcome despite the superiority of the electronic mixing method. This drawback is caused by the delay time of 64 ms + alpha (sound processing time) in our system. The latter drawback can be overcome by changing the microphones to wireless ones. However, these requirements restrict implementation of the electronic mixing method.

On the other hand, the results from experiment 2 indicate that there are only small differences between the two methods, and there is no actual reason to restrict adoption of the physical mixing method.

Thus, we decided to adopt the physical mixing method in the

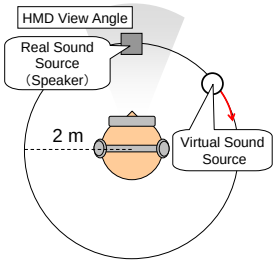


Figure 7. Reception of real and artificial sounds in Ex. 2

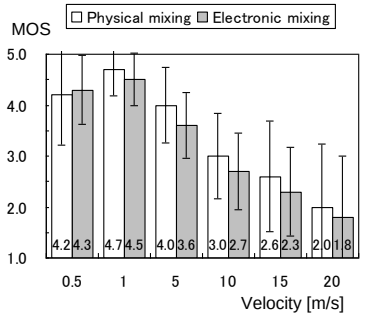


Figure 8. Evaluation of sound source localization

two-by-two audio–visual MR system.

4 Achievement of Geometric Consistency by Two-by-Two Method

To achieve geometric consistency between real and virtual worlds in the visual sense, it is necessary to achieve geometric consistency with the coordinate systems shown in Fig. 9. The relative positions of the three coordinate systems *W*, *M*, and *S*, are set constant, and the relative positions of the coordinate systems *H* and *C* are kept constant. The consistency of these coordinate systems is manually calibrated prior to performing the experiment. The position and orientation of the magnetic sensor attached to the HMD is constantly measured on *S*, and it could relate *H* with *S*. Consequently, geometrical consistency can be achieved among the three coordinate systems, *W*, *M*, and *C*.

Likewise, it is also necessary to aurally achieve geometric consistency among the three coordinate systems *W*, *M*, and *A*, as shown in Fig. 9. *A* has its origin at the center of the user's head. Thus, *A* could be related to the origin of *H* if the user's head size is known in advance. Since the consistency among *H*, *W*, and *M* has already been given, the geometrical consistency of *W*, *M*, and *A* could be easily achieved.

Fig. 10 shows the reproduction method for the artificial sound heard by a user.

5 Implementation of an Audio–visual Two-by-Two Method Mixed Reality System

According to the results and discussions in Chapter 3, we adopted a "physical mixing method" to produce an audio–visual MR system using the two-by-two method by implementing the contents of Fig. 3. The hardware shown in Table 1 is used to implement the prototype system shown in Fig. 3. In this system, we tried to superimpose a CGI helicopter onto the real space (W, 8.8 m; D, 6.2 m; and H, 2.7 m) as a sample of MR space. We also implemented the system for multiple users and multiple sound sources.

As shown in the results from experiment 1 (Fig. 5, 6), the users could hear real and artificial sounds, but without noticing that there were any significant differences. Therefore, we conclude

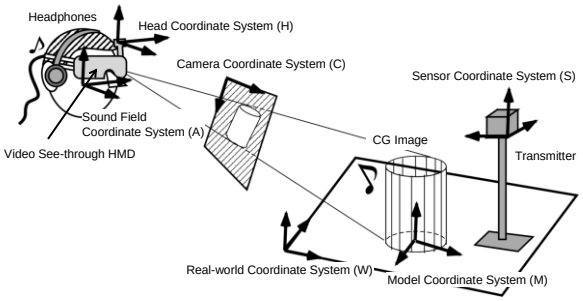


Figure 9. Coordinating systems for geometric registration

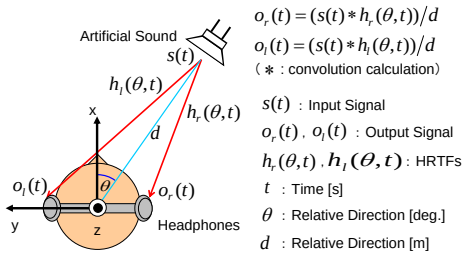


Figure 10. Reception of artificial sound

Table1. Hardware configuration

Constituent Elements	Model Numbers/Names
PC Generating 3-Dl Image and Managing MR Space	Canon MR Platform System
PC Generating 3-D Sound Field	Dell Precision 670
Magnetic Sensor	Polhemus 3SPACE FASTRAK
Head Mounted Display	Canon VH-2002
DA Converter	Thinknet DF-2032D
Open-Air Headphones	SONY MDR-F1

that this prototype system has achieved audio and visual MR using the two-by-two method.

6 Conclusion

In this paper, we have described the development of a system that can simultaneously and consistently manage both visual and auditory MR, in real time, as the first stage of an ongoing project. In order to achieve an auditory MR display, there are two alternatives—the physical mixing method using open-air headphones and electronic mixing method using closed-air headphones. Initially, we had assumed that the electronic mixing method would be as appropriate as the video see-through method is in visual MR. However, during our experiments, it was found that open-air headphones are superior to closed-air headphones. Hence, we decided to use the physical mixing method in our development of, namely, the two-by-two audio–visual MR system.

As mentioned in the latter half of this paper, we were able to implement the co-existence of visual and auditory MR with a satisfactory level of geometric consistency.

This research was supported by the Japan Society for the Promotion of Science through Grants-in-aid for Scientific Research (A), "A mixed reality system that merges real and virtual worlds with three senses."

References

[1] P. Zahorik: Auditory display of sound source distance, *Proc. Int. Conf. on Auditory Display*, pp. 326 - 332, 2002.

[2] K. Lyons, *et al.*: Guided by voice: An audio augmented reality system, *ibid.*, pp. 57 - 62, 2000.

[3] A. Härmä, *et al.*: Techniques and applications of wearable augmented reality audio, *Proc. 114th CAES*, 2003.

[4] Y. Tamura, *et al.*: Virtual reality system to visualize and auralize numerical simulation data, *Proc. Conf. on Comput. Physics*, pp. 227 - 230, 2000.

[5] W. W. Gaver, *et al.*: Effective sounds in complex systems: The ARKOLA simulation, *Proc. CHI*, pp. 85 - 90, 1991.

[6] P. Flanagan, *et al.*: Aurally and visually guided visual search in a virtual environment, *Human Factors*, Vol. 40, pp. 461, 1998.

[7] Y. Ohta and H. Tamura (eds.): Mixed Reality―Merging Real and Virtual Worlds, *Ohm-sha & Springer*, 1999.

[8] R. T. Azuma, *et al.*: Recent advances in augmented reality, *IEEE Compt. Graph. & App.*, Vol. 21, No. 6, pp. 34 - 47, 2001.

[9] T. Ohshima, *et al.*: AR^{2}Hockey: A case study of collaborative augmented reality, *Proc. VRAIS*, pp. 268 - 275, 1998.

[10] C. E. Hughes, *et al.*: Mixed reality in education, entertainment, and training, *IEEE Compt. Graph. & App.*, Vol. 25, No. 6, pp. 24 - 30, 2005.

[11] Z. Zhou, *et al.*: The role of 3-D sound in human reaction and performance in augmented reality environments, *IEEE Trans. Syst., Man & Cybern.*, Part A , Vol. 37, No. 2, pp. 262 - 272, 2007.