An Optical Design for Avatar-User Co-axial Viewpoint Telepresence

Kei Tsuchiya*
The University of Electro-Communications

Naoya Koizumi†
The University of Electro-Communications
JST PRESTO

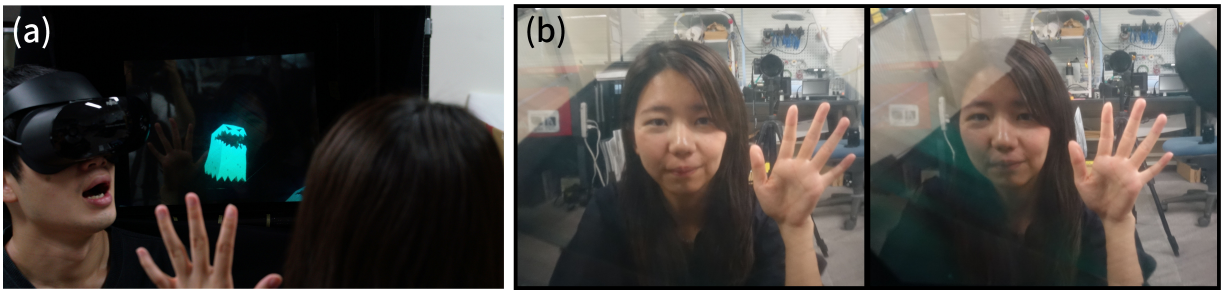

Figure 1: (a) User controls the motion of the mid-air CG avatars using a head-mounted display (HMD) tracking system and our proposed system, (b) he can see the video from the viewpoint of a mid-air image via the HMD.

ABSTRACT

We propose a mid-air image system for telepresence. Virtual reality (VR) social networks enable users to interact with each other through CG avatars and choose their appearances freely. However, this is only possible in VR space. We propose a system that takes the avatar from VR space to real space with the help of mid-air imaging technology. In this system, the micro-mirror array plates (MMAPs) display the mid-air image and optically transfer the camera viewpoint to capture users from the mid-air image position. Luminance measurement and modulation transfer function (MTF) measurement were performed to evaluate the image capturing performance of this system. As a result, we found that the MMAPs ccause a decrease in brightness and an increase in blur. In addition, the stray light generated by the MMAPs was in the captured video. We also confirmed that face detection works correctly on the captured video by adjusting the ISO sensitivity of the camera. Furthermore, we designed an application for telepresence called Levitar, which uses a dual camera to output the captured video to the HMD and controls the camera gaze direction.

Index Terms: Human-centered computing—Human computer interaction (HCI)— Interaction devices—Displays and imagers;

1 INTRODUCTION

In recent years, VR social networks such as VRChat [23], where users communicate through CG avatars, have been popular. One of the advantages of the CG avatar is that the appearance can be changed easily. Changing the appearance of the avatar is important because some studies have shown that its appearance influences to its communication [27] [24]. By using an HMD, we can transform ourselves into different avatars in VR space without any physical restrictions. However, this requires the user to immerse himself or herself in VR space. To realize this in real space instead of VR space, it is necessary to bring the CG avatar to real space. Although there is a system that display avatars in real space by using AR technologies, the avatar user remains in VR space. In addition, displaying the avatar using see-through HMDs requires all users to wear the devices. A system in which the users in different spaces share the same real space has not been proposed.

In this paper, we propose a system that takes the avatar from VR space to real space with the help of mid-air imaging technology. A mid-air image is an image formed in the air that can be displayed by the MMAPs. This technology enables us to display CG images in real space. In this system, we show a mid-air CG avatar of the user in the telepresence system. Therefore, the system presents the user with a video captured from the avatar's viewpoint. A previous study [2] has shown that misalignment of the camera viewpoint and avatar image makes eye contact difficult in video conferencing. Therefore, it is preferable that the viewpoint positions of the avatar and the camera are aligned. This system optically transfers the camera viewpoint with the MMAPs to superimpose the positions of the mid-air image and the camera. A beam splitter is used to place the camera and the display that is the light source for the mid-air image at a conjugate position, and these positions are transferred by the MMAPs. In this way, the mid-air image is displayed and the video from that viewpoint is captured.

The effects of camera viewpoint transfer using the MMAPs on capturing performance are stray light, brightness, and blur. Therefore, we confirmed the effect of stray light based on the previous research and evaluated the effects of brightness and blur by luminance measurement and MTF measurement, respectively. In luminance measurement, the luminance of the mid-air image and its light source were measured and compared to investigate the change in brightness due to the MMAPs. In MTF measurement, we calculated the MTF of images taken with a normal camera and the transferred camera to confirm the blur caused by the MMAPs. In addition, we conducted an experiment using face detection as an overall evaluation of the captured video. In this experiment, the face detection accuracy was measured when the ISO sensitivity of the camera was changed under two conditions: using the normal camera and using the transferred camera.

We designed Levitar, an application for telepresence based on the proposed system. A dual camera was used to output the captured video to the HMD, and two motors were added to control the camera gaze direction horizontally and vertically. The camera is rotated in an arc to control the camera gaze direction horizontally, and the beam splitter that reflects the camera is rotated to control the camera gaze direction vertically. Independent control of the camera and half mirror simplifies the mechanism and alleviates the interference between the camera and the display. The HMD measures the user's

*e-mail: tsuchiya@media.lab.uec.ac.jp
†e-mail: koizumi.naoya@uec.ac.jp

head movements and the camera gaze direction is controlled based on the measurement data. Levitar realizes interaction with other users via CG avatar in real space.

Our contributions in this paper are as follows.

1. We combine a mid-air image display and an optical transfer of the camera viewpoint to superimpose the viewpoint of the CG avatar and the camera.

2. We evaluate the basic capturing performance of the transferred camera and the face detection accuracy on the captured video to confirm that the optical system we designed is applicable to the telepresence system.

3. We design an application for telepresence called Levitar, which uses a dual camera to output the captured video to the HMD and controls the camera gaze direction. Levitar has a simple mechanism that rotates the mirror and the camera independently, enabling complex horizontal and vertical gaze control of the transferred camera.

2 Related work

2.1 Mid-air interaction

A mid-air image means an image formed in the air by reflection and refraction of light. One method for displaying mid-air images is to use a retro-transmissive optical system. A retro-transmissive optical system is an optical system that forms an image at a position that is symmetrical to the light source with respect to the optical element. Examples of retro-transmissive optical systems include dihedral corner reflector array (DCRA) [9], ASKA3D [14], and aerial imaging by retro-reflection (AIRR) [8]. DCRA is an optical element composed of a dihedral corner reflector array. ASKA3D is an optical element composed of two layer micro-mirror array and called micro-mirror array plates (MMAPs). DCRA and ASKA3D can form mid-air images with one optical element. AIRR is an optical system composed of a beam splitter and a retroreflector. In this study, we use MMAPs, which can display high-luminance mid-air images and are easily available.

Several interactive systems using mid-air images have been proposed. Yamamoto et al. [25] have developed a mid-air display that can be operated with a fingertip by combining mid-air image display using AIRR and gesture recognition using a high-speed camera. Systems that interact with mid-air CG images such as MARIO [5] and Scooplit [11] have also been proposed. In addition, Abe et al. [1] have developed a mid-air image system for behavioral biology experiments. Thus, the mid-air image is used for displays and interactions with CG images and is also used in systems that target not only humans but also creatures.

Furthermore, the retro-transmissive optical system enables us to display a 3D mid-air image. Terashima et al. [20] have developed a 3D mid-air imaging system that combines the mid-air image using AIRR and the autostereoscopic image using depth-fused 3D (DFD). Kurogi et al. [18] have developed a system that displays 3D mid-air images using time-division stereoscopic image. This system consists of the MMAPs, motion capture, transparent LCD, and two types of light sources. On the other hand, a stereoscopic mid-air imaging system based on motion parallax [19] and the one that uses a swept volume display as a light source [3] have been proposed. As above, various methods for displaying 3D mid-air images can be applied to the retro-transmissive optical system.

In this study, we apply the mid-air CG avatar to the telepresence system. HaptoClone [10] is a telepresence system that uses a mid-air image, but it displays the user as the mid-air image and does not use the CG avatar. We aim to realize a telepresence system in which the user becomes the mid-air CG avatar and interacts with other users. In particular, there are several systems that use an HMD to present video from an avatar viewpoint to the user.

2.2 Telepresence system

Various methods such as VR, AR, and robot methods are used for the telepresence system. Fig. 2 shows a comparison of telepresence systems in terms of the user interaction space and the type of avatar. VRChat [23] allows users to interact via CG avatar in VR space. In this system, all users are in VR space. Mini-Me [16] is a telepresence system combining VR and AR. This system displays a CG avatar in real space using a see-through HMD, and the user who plays the role of the avatar is in a VR space that reproduces the real space. In addition, there are many systems that use robots as avatars [17]. In these systems, the robots and users interact in real space. Our proposed system uses the CG avatar such as VR or VR and AR-based systems, and all users interact with each other in real space such as robot-based systems.

When designing a telepresence system, it is necessary to consider the position of the camera that functions as the avatar's eyes. According to Chen's research [2], eye contact perception decreases below 90% when the observer's viewing angle with respect to the camera and the avatar's eye increase by 1° or more in the horizontal direction and 5° or more in the vertical direction. Therefore, it is preferable to place the camera on the observer's line of sight. MMSpace [15] enables eye contact by installing cameras behind a transparent screen that displays the avatar. This system uses multiple cameras for one user and switches cameras according to the user's gaze direction. Since this method requires multiple cameras, it is not suitable for many users. Jones et al. [4] have developed a system that places a camera viewpoint near the avatar's eyes using a beam splitter. We use this design to place the camera and the light source of the mid-air image in a conjugate position.

2.3 Optical transfer of camera viewpoint

We use the optical transfer of the camera viewpoint to superimpose the mid-air image and the camera viewpoint. This is a method for transferring the camera viewpoint to a position that protrudes from the camera body by combining the camera and the optical system, and this method uses lenses or the retro-transmissive optical system.

Okumura et al. [12] have developed a saccade mirror for high-speed active vision. They have used the optical transfer method of the camera viewpoint with pupil shift lenses consisting of three lenses. The camera gaze direction is controlled by two small rotational mirrors placed at the transferred camera viewpoint. The camera viewpoint transfer in this system is used to control the gaze direction at high speed. Unlike our study, they do not aim to combine the mid-air image and the camera.

Yasui et al. [26] have developed an occlusion-robust sensing method toward interaction with a 3D image that combines a retro-

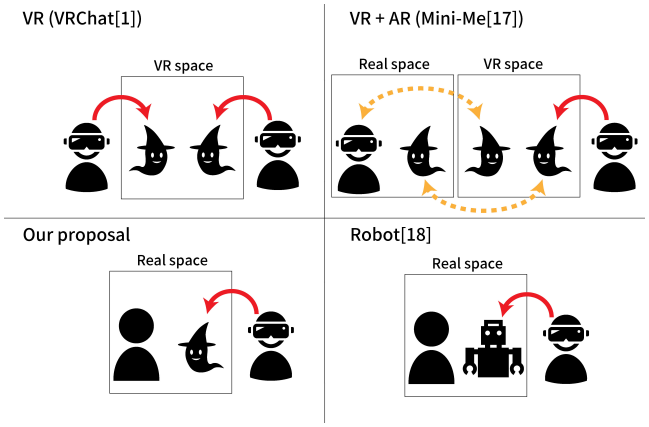

Figure 2: Comparison of telepresence systems

transmission optical system and a camera. In this system, the light is emitted onto the target through AIRR, and the reflected light is projected onto the diffuser screen through AIRR again. Capturing the projected light with a camera enables three-dimensional measurement of the object. This system was developed for occlusion-robust sensing and does not capture the image of the target directly. We aim to capture the image of the target with a camera for communication.

GoThro [6] uses camera viewpoint transfer with MMAPs to capture the image of the target behind the gap and make the target unaware of the presence of the camera. In this system, the MMAPs, which form a real image at a position away from the light source, are applied to the camera. Unlike lens viewpoint transfer, the MMAPs have the advantage of not causing distortion due to the optical axis. When transferring the camera viewpoint with MMAPs, the focus cannot be adjusted because the focus position is reversed. GoThro has solved this problem by attaching a concave lens to the camera lens. Our proposed optical system is based on GoThro's design and combines it with the mid-air imaging design.

3 Principle

3.1 Optical design

The optical system proposed in this paper is a combination of GoThro [6] design and the mid-air imaging design. The optical design is shown in Fig. 3. This design consists of a display, camera, concave lens, beam splitter, and the MMAPs. The concave lens is attached to the tip of the camera lens. The light emitted from the display is reflected by the beam splitter to form virtual image D'. The MMAPs form a mid-air image with the light from D' and transfer the camera viewpoint to the mid-air image position. By arranging the display, camera, and beam splitter so that D' and the camera overlap, the mid-air image position and the transferred camera viewpoint are superimposed. Thus, the mid-air image is displayed, and the video from that viewpoint is captured. The angles of the MMAPs and the beam splitter are both 45°. To prevent undesirable light, we shielded the MMAPs with light shielding material.

3.2 Experimental prototype

We made experimental prototypes based on the optical design. The display was a Samsung Galaxy Note 8 (OEL, 6.3 inches, 521 ppi), the camera was a Sony α7R II, and the camera lens was a Sony SEL2470GM (focal length: 24 mm to 70 mm). The concave lens was a Kenis H-07 (focal length: -250 mm), the beam splitter was an Edmund Optics plate beamsplitter (transmittance: 50%, reflectance: 50%, size: 127 mm × 178 mm), the MMAPs were ASUKANET ASKA3D-488 (size: 488 mm × 488 mm, pitch width: 0.5 mm). We used video capture (AverMedia's BU110) to capture the video taken with the camera and transferred it to a computer. The maximum resolution of the video is 1920 px × 1080 px, and the maximum frame rate is 60 fps.

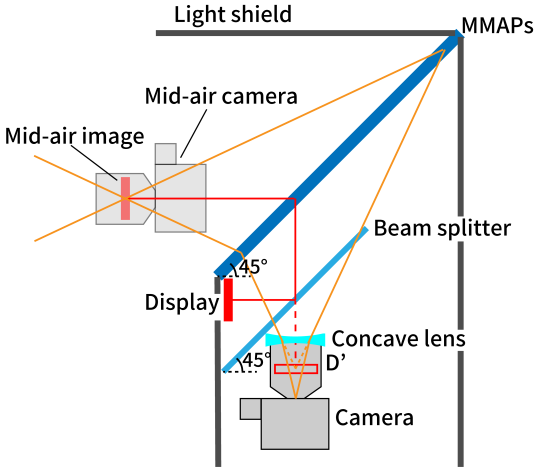

Figure 3: Optical design

3.3 Design of interaction area

We designed the interaction area based on the parameters of the camera, display, and MMAPs. In this paper, the interaction area is defined as the area where the camera is focusable and the mid-air image is viewable.

The maximum distance from the MMAPs to the mid-air image and the viewable area of the mid-air image are derived as follows. As shown in Fig. 4, θ is the vertical angle of view of the camera, L is the length of one side of the MMAPs, d is the distance from the camera viewpoint to the MMAPs, I is the vertical length of the mid-air image, ϕ is the viewable angle of the mid-air image, and α is the distance from the intersection of the MMAPs and the rays perpendicular to the mid-air image to the bottom edge of the MMAPs. ϕ is the angle between the line that connects the lower end of the MMAPs and the lower end of the mid-air image and the line that connects the upper end of the MMAPs and the upper end of the mid-air image, as shown by the blue line. θ is the angle of view when the camera lens and concave lens are compounded and is obtained from the composite focal length. It is assumed that the field of view is the widest possible without using zoom. When transferring the camera viewpoint, to reproduce the original camera angle of view, the following conditions must be satisfied so that the MMAPs are within the camera angle of view.

$$d \leq \frac{L}{\sqrt{2}\tan\theta} \tag{1}$$

Since L is 488 mm and θ is 65° in the experimental prototype, the maximum d is calculated to be 160 mm from Equation 1. Therefore, the experimental prototype can display the mid-air image at a position that protrudes up to 160 mm from the MMAPs. At this time, the viewable angle of the mid-air image can be obtained from the following equation.

$$\phi = \arctan\left(\frac{\sqrt{2}\alpha - I}{2d - \sqrt{2}\alpha}\right) + \arctan\left(\frac{\sqrt{2}(L-\alpha) - I}{2d + \sqrt{2}(L-\alpha)}\right) \tag{2}$$

In the experimental prototype, α is 89 mm and I is 75 mm. Hence, ϕ is calculated as 43° from Equation 2.

Fig. 5 shows the interaction area designed as above. The minimum distance at which the camera can be focus depending on the

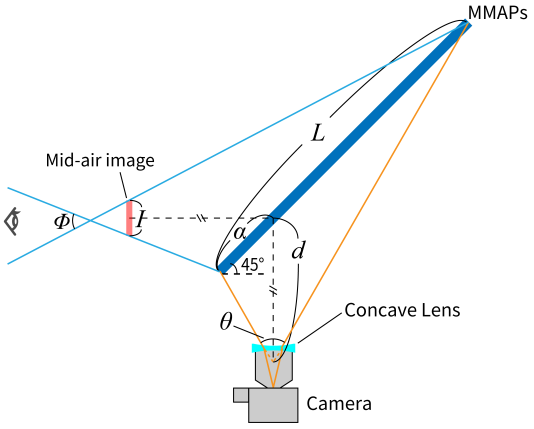

Figure 4: Parameters for design of interaction area

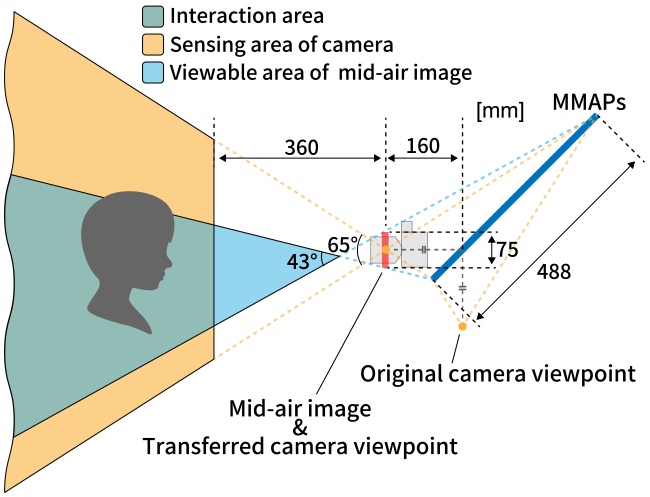

Figure 5: Interaction area

camera lens is determined. In the case of the experimental prototype, the distance is 360 mm from the camera viewpoint. Therefore, an image of the user is captured if he or she is at a position more than 360 mm away from the mid-air image and within the camera angle of view. Furthermore, if the user's eyes are within the viewable area of the mid-air image, the user can observe the image.

4 Evaluation

To confirm that this system can be applied to telepresence, we evaluated the capturing performance and the accuracy of face detection on the captured video. We assumed that the telepresence system needed a video of sufficient quality to recognize human faces. Therefore, we first evaluated three points: stray light, brightness, and blur to measure the basic capturing performance. As for the brightness, it is predicted that that problem can be solved by adjusting the ISO sensitivity of the camera. Hence, we conducted a face detection experiment to investigate the relationship between ISO sensitivity and face detection accuracy.

4.1 Evaluation of capturing performance

When transferring the camera viewpoint with the MMAPs, the three points of stray light, brightness, and blur may affect the quality of the captured video. Here, stray light is defined as light that does not form a mid-air image in an optical system using the MMAPs. We confirmed the effects of the above three points from previous research and evaluation experiments. According to the result of previous research [9] , MMAPs cause stray light. It does not reveal the effect of stray light to the brightness and blur of MMAPs. Therefore, two experiments were conducted to verify these two points.

4.1.1 Effect of stray light by MMAPs

When the mid-air image is displayed by the MMAPs, the light taht is reflected twice by the mirror array in the element forms the image. In this paper, we call this light imaging light. On the other hand, Maekawa et al. [9] have found that light reflected once or more than three times does not form an aerial image but is observed as stray light. In this research, the light is classified into one reflection and two reflections according to the incident angle of the light to the element. In camera viewpoint transfer with the MMAPs, the camera captures every incident light beam from outside. Therefore, it is expected that the captured light contains not only imaging light but also stray light.

The image captured in the experimental prototype is shown in Fig. 6. There is stray light in the upper left and upper right parts of

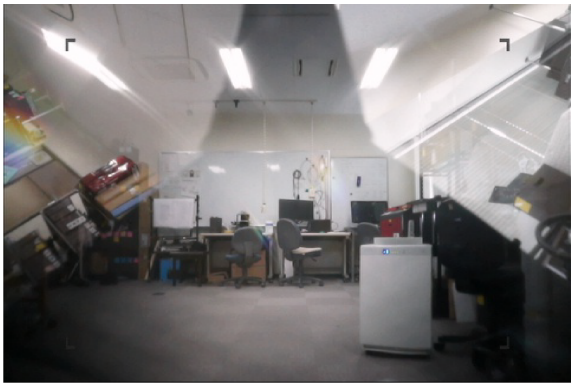

Figure 6: Image captured by the experimental prototype

the figure. Since the target subject cannot be captured in this part, the effective angle of view of the camera is decreased. This may also affect the overall brightness ratio of the image.

4.1.2 Brightness decay caused by MMAPs: Luminance measurement

Purpose To evaluate the brightness decay due to MMAPs, we measured the luminance of the light source and the mid-air image and calculated the luminance ratio.

Procedure The luminance of the mid-air image and its light source were measured as shown in Fig. 7 (a) and (b). We used an Apple iPad 5 (IPS LCD, 9.7 inches, 264 ppi) with a white circle as the light source, as shown in Fig. 7 (c). This was placed horizontally, and the luminance of the central part of the display and the mid-air image was measured from several angles using a luminance meter. The angle was set to 11 points at 5° intervals of 10° to 60°, with 0° being perpendicular to the display. The measurements were conducted under two conditions: a vertical display and a horizontal display with respect to the luminance meter. The luminance ratio of the mid-air image to the light source was calculated and averaged between the vertical display and the horizontal display.

Results The measurement results are shown in Fig. 8. The horizontal axis is the angle, and the vertical axis is the luminance ratio. These results show that the luminance ratio is the highest at an angle of 45° – 55° and decreases as the angle becomes larger or smaller.

Discussion In this system, the MMAPs are arranged at an angle of 45° to the camera. Thus, it is expected that the MMAPs reduce the light reaching the camera to about 40% or less.

Figure 7: Overview of luminance measurement (a) and (b) and display (c)

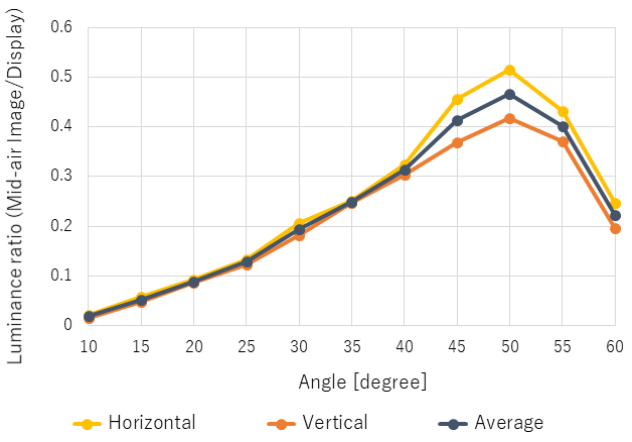

Figure 8: Results of luminance measurement

4.1.3 Blur increase caused by MMAPs: MTF measurement

Purpose To confirm the increase in blur of the captured video in this system, we measured the modulation transfer function (MTF) under two conditions: using a normal camera and a transferred camera with MMAPs.

Procedure In this experiment, the MTF was measured using the edge method. We photographed a display that displayed the edge image shown in Fig. 9 in a sufficiently dark room to measure the MTF. The display used as the subject was the same as that used in the luminance measurement described in Sect. 4.1.2, and the pieces of equipment were described in Sect. 3.2. The video via video capture was saved as an image using OpenCV. We assumed the actual operation of the system and decided the camera settings to photograph with as much brightness as possible. The F-number was set to 4.0, with the aperture fully open. The shutter speed was set to 1/30s for processing at 30fps in real time. To suppress the effects of noise, we set the ISO sensitivity to 100, which has the least noise. The distance from the camera viewpoint to the display was 360 mm, which is the shortest distance for focusing the camera.

The two conditions in this experiment are shown in Fig. 10 and below.

(i). Normal: Photographing with the camera alone

(ii). MMAPs: Photographing with the transferred camera by the MMAPs

The luminance of the display was 135.7 cd/m^2 in (i) and 342.9 cd/m^2 in (ii). These values were obtained by measuring the central part of the white image on the display from the camera position in (i) with a luminance meter. We adjusted the luminance of the display so that the light intensity reaching the camera was equal to that of (i) and (ii) with reference to the luminance ratio obtained in Sect. 4.1.2.

Figure 9: Edge image

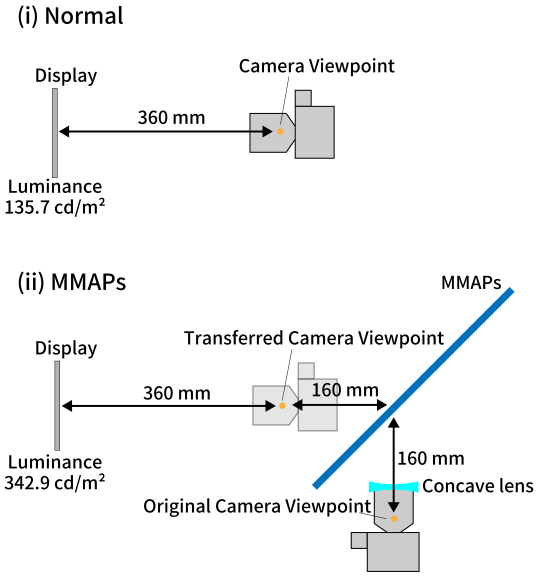

Figure 10: Conditions of MTF measurement

The distance from the camera viewpoint to the MMAPs in (ii) was 160 mm, which was the setting when the camera viewpoint was protruded to the maximum described in Sect. 3.2.

It is necessary to find the theoretical resolution limitation of the camera to calculate the MTF. The theoretical resolution limitation is the maximum number of line pairs per mm. A line pair consists of two pixels that display white and black. The derivation of the theoretical resolution limitation in this experiment is shown below. The resolution of the image captured by video capture is 1920 px × 1080 px, but the aspect ratio of the image photographed by the camera is 2: 3. Hence, 1620 px × 1080 px out of 1920 px × 1080 px is valid. Since the size of the image sensor of the camera used for this measurement is 36 mm × 24 mm, the length per px of the image sensor is obtained as follows.

$$\frac{36}{1620} = \frac{1}{45}[\mathrm{mm/px}] \tag{3}$$

In addition, since one line pair consists of 2 px each of white and black pixels, the number of line pairs per mm is calculated as follows.

$$\frac{1}{\frac{2}{45}} = 22.5[\mathrm{LP/mm}] \tag{4}$$

Therefore, the theoretical resolution limitation of the camera in this experiment is 22.5 LP/mm.

The calculation procedure for MTF is shown below. First, we converted the captured image to grayscale and obtained pixel values for the center 100 rows of the image. The edge part of the image was determined from the gradient of the pixel values. The pixel values for 50 columns before and after this edge part were differentiated and Fourier transformed. We performed this procedure for all 100 lines and then averaged them. Finally, the MTF curve was obtained by normalizing these values.

Results Fig. 11 and Fig. 12 show the edge images used for the measurement and the results. These results show that the MTF of (ii) is lower than that of (i) in the range of 0.0 LP/mm to 15.0 LP/mm.

Discussion From the measurement results, it was confirmed that blur increased when photographing with the transferred camera with MMAPs compared with photographing with the camera alone.

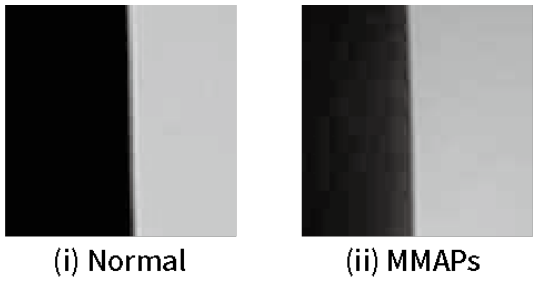

Figure 11: Photographed edge image

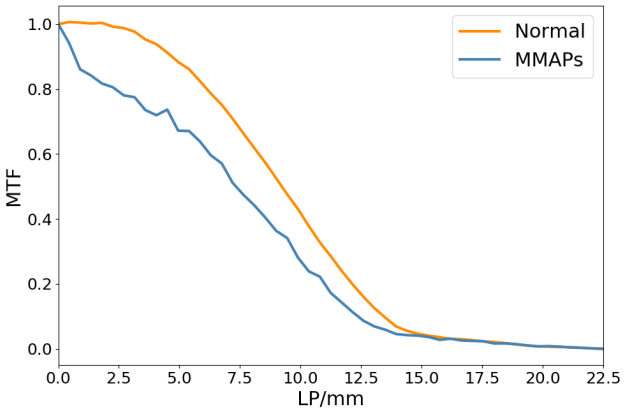

Figure 12: Results of MTF measurement

4.2 Face detection experiment

4.2.1 Purpose

To verify that the quality of the captured video is sufficient for the telepresence system, we confirm that face detection works on the captured video. To communicate with others, users need a quality image to recognize the face. In addition, the face detector is not robust enough under the environment we tested but humans are good enough to recognize the face in such an environment. Therefore, we assume if face detection works on the quality of the captured video, people can recognize faces and communicate using the same video. From Sect. 4.1, it was found that the three points of stray light, brightness, and blur deteriorate the image, but it is not known how much they affect the communication. To improve the brightness of the captured image, it is possible to adjust camera settings such as shutter speed, F-number, and ISO sensitivity. Setting a slow shutter speed can improve brightness, but it is limited because the video is processed in real-time. As for F-number, some cameras such as compact cameras do not have a function to adjust it. On the other hand, the ISO sensitivity is more accessible in a variety of cameras, and it is easy to apply to the brightness adjustment based on the exposure value. We have also confirmed that setting the ISO sensitivity to auto did not work on face detection. This is probably because the change in brightness was different from that of a general camera due to the influence of stray light. Therefore, we investigated the relationship between ISO sensitivity and face detection accuracy.

4.2.2 Procedure

The experimental procedure is shown below. The face detection accuracy was measured by photographing one subject and measuring the number of detection frames in 300 frames on the face detection program. Between the beeps played at intervals of 0.6 seconds, the subject faced up, down, left or right and moved her mouth to imitate the movement of human communication. We changed the ISO sensitivity setting and measured face detection accuracy at each setting. This measurement was conducted three times, and the detection rate was calculated from the average value. The ISO sensitivity was set to 11 conditions from the minimum value of 100 to the maximum value of 102,400. We selected the OpenCV 3.4 [13] for face detection as the most common and available software. The frame rate was 30 fps. The pieces of equipment were described in Sect. 4.1.3. Similar to the MTF measurement described in Sect. 4.1.3, the F-number was 4.0, the shutter speed was 1/30 seconds, and the distance from the camera to the subject was 360 mm. We used the Haar Cascade classifier [7] [22] in OpenCV 3.4 [13] for face detection and a green cloth as the background during the experiment to prevent false detection.

The two conditions of this experiment are shown in Fig. 13 and below.

(a). Normal: Photographing with the camera alone

(b). Proposed system: Photographing with the transferred camera with the MMAPs using the beam splitter

In (b), the transmittance of the beam splitter was 50%, and the distance from the MMAPs to the camera viewpoint was 160 mm. The luminance of the subject's face measured from the camera position in (a) was 9.46 cd/m^2. This value was the average of the luminance values of the nose, forehead and cheeks measured with a luminance meter.

4.2.3 Results

The results of the experiment are shown in Fig. 14. The horizontal axis is the ISO sensitivity setting and the exposure in (a), and the vertical axis is the detection rate. The results of three measurements were plotted, and the average value is represented by a solid line. The exposure was expressed as the difference in the exposure values when the appropriate exposure was set to 0. In other words, 0 means appropriate exposure, a positive value means overexposure, and a negative value means underexposure. When the F-number is 4.0, the shutter speed is 1/30s, and the subject luminance is 9.46 cd/m^2; the ISO sensitivity for appropriate exposure calculated based on the exposure value is about 710. Therefore, an ISO sensitivity of

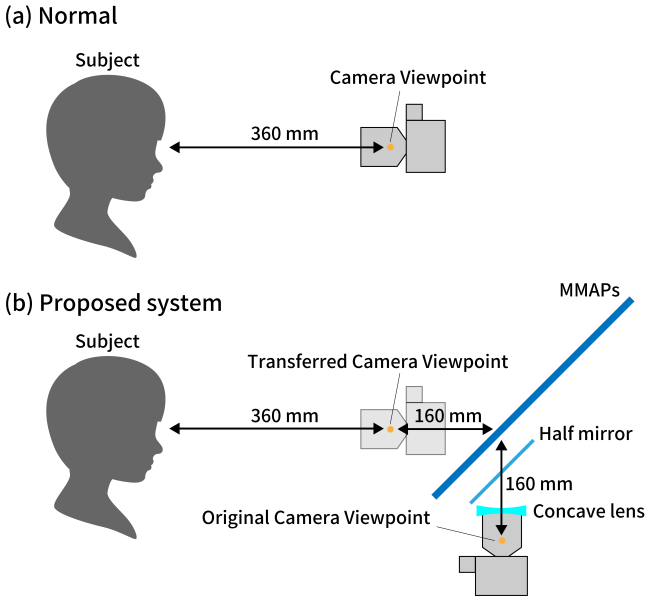

Figure 13: Conditions in face detection experiment

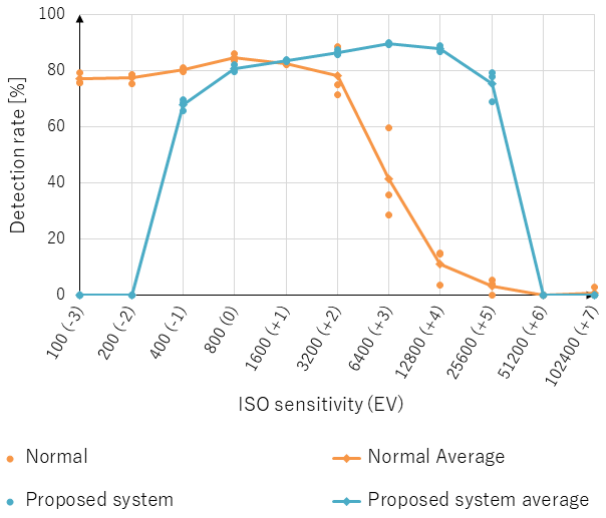

Figure 14: Results of face detection experiment

800, which is the closest to 710, is assumed to be the appropriate exposure.

The results show that the ISO sensitivity setting is optimal when the exposure state is +2 to +4 in the experimental prototype. These results suggest that these settings enable face detection with the same degree of accuracy as capturing with the camera alone. In (a), the detection rate is 75% or more in the range of ISO sensitivity of 100 (exposure: -3) to 3200 (exposure: +2). This indicates that face detection works even if the exposure is somewhat under or over the normal exposure. The highest detection rate in (a) is an ISO sensitivity of 800 (exposure: 0), and the detection rate is about 85%. It is a reasonable result that the detection accuracy is the highest in the appropriate exposure. On the other hand, in (b), the detection rate is highest at ISO sensitivity of 6400 (exposure: +3), which is about 90%. This is consistent with the fact that the brightness decreases to less than 40% with MMAPs and 50% with the beam splitter. In addition, when the ISO sensitivity is 3200 (exposure: +2) and 12800 (exposure: +4), the detection rates are about 87% and about 88%, respectively, which are close to the maximum value of (a). From the above, in the experimental prototype, when setting the ISO sensitivity to 4 to 16 times the appropriate exposure, face detection works with the same accuracy as capturing with the camera alone.

4.2.4 Discussion

We confirmed that face detection worked correctly on the captured video by adjusting the ISO sensitivity from the experimental result. Hence, this system can be applied to a telepresence system because it presents a video with sufficient quality to recognize the face.

4.3 Deign guidelines

We found that adjusting the ISO sensitivity is effective to use this system for telepresence from the experiments. The results show that the MMAPs cause the generation of the stray light, the brightness decay and the blur increase, and the brightness decay is solved by adjusting the ISO sensitivity. Specifically, the adjustment is to set the ISO sensitivity to 4 to 16 times the appropriate exposure. It is optimal for our system that uses a beam splitter having 50% of transmittance. In addition, the maximum detection rates were comparable between the two conditions in the face detection experiment. Therefore, adjusting the ISO sensitivity enables face detection with the same accuracy as the normal condition without considering the influence of blur. This shows that the blur increase does not influence face detection. As for the stray light, it is generated in the upper left and upper right parts of the captured video as shown in Fig. 6. Therefore, this system is suitable for use in an environment where there is no subject on the left and right sides of the apparatus. In such an environment, adjusting the ISO sensitivity realizes the telecommunication that requires the user to see the audience's face.

5 Application

As an application example of the optical system described above, we applied a dual camera to the optical system and developed a mechanism to control the camera gaze direction. It is called Levitar, and it takes the avatar from VR space to real space with the help of mid-air imaging technology (Fig. 1 (a)). A video captured from the mid-air image position is presented to the user via the HMD (Fig. 1 (b)). We control the camera gaze direction according to the user's head movements. In other words, this system provides the user with the experience of becoming a mid-air CG avatar and interacting with other users in real space.

5.1 System design

Fig. 15 shows the optical design and system flow of Levitar. We used a dual camera to display the captured image on the HMD and added two motors to control the camera gaze direction. The position of the camera and the display was switched for the gaze control. In other words, the virtual image of the camera reflected by the beam splitter overlaps the display. This is because the pan/tilt camera cannot be applied because the anteroposterior relationship of the camera gaze direction is reversed by the MMAPs. Furthermore, we control the camera gaze direction with two motors synchronized with the head movements of the HMD wearer. At the same time, we can change the facing direction of the CG avatar on the display.

An overview of the camera gaze control is shown in Fig. 16. The gaze is controlled horizontally and vertically by two motors. The camera is moved to control the gaze in the horizontal direction. We control the gaze direction of the mid-air camera by moving it in an arc. The beam splitter is rotated to control the gaze in the vertical direction. We control the gaze direction of the mid-air camera by rotating the beam splitter around the intersection of the camera gaze and the beam splitter. The radius of gyration control is 90 mm, which is close to the radius of the human head. In this way, it is possible to avoid interference between the camera and the display by independently controlling the gaze direction horizontally and

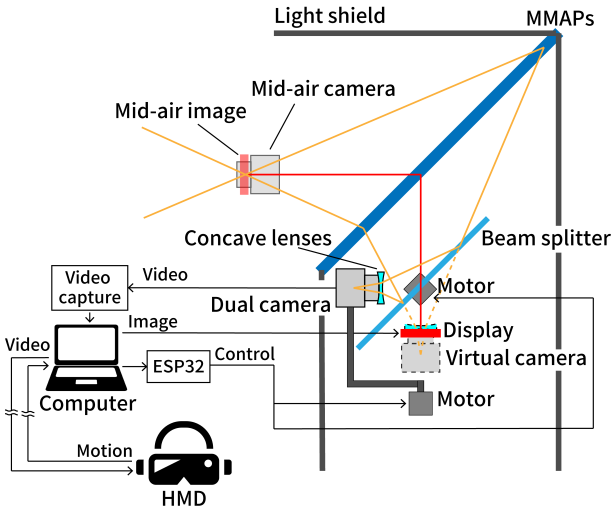

Figure 15: Optical design & System flow of Levitar

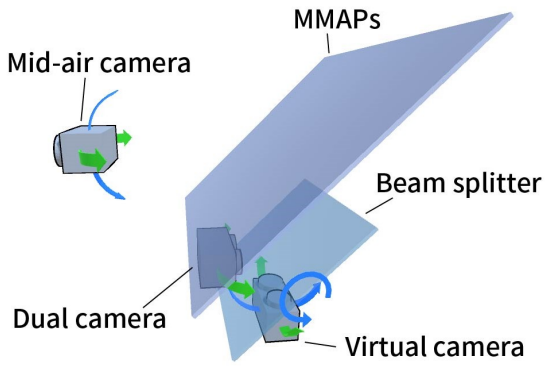

Figure 16: Overview of camera gaze control

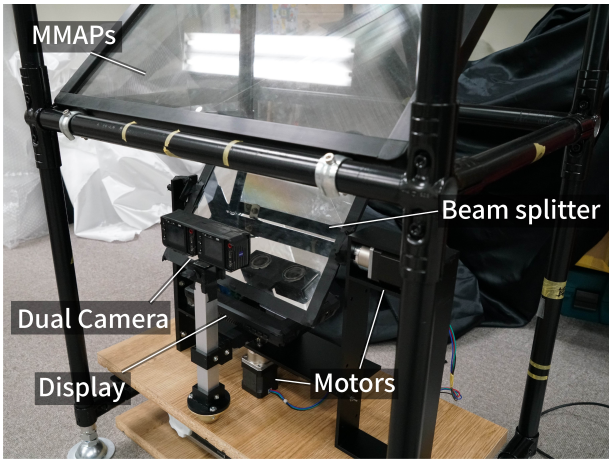

Figure 17: Implementation

vertically.

5.2 Implementation

The implementation of the system is shown in Fig. 17: We used two Sony RX0 cameras for the dual camera. The focal length of the concave lens was -250 mm, and the reflectivity of the beam splitter was 70%. The MMAPs were ASUKANET ASKA3D-Plate (size: 488 mm × 488 mm, pitch: 0.5 mm), and the display was Feelworld FW279S (IPS, 7 inch, 323 ppi, brightness: 2200 cd/m^2). We covered this equipment with black cloths to avoid the effects of ambient light.

5.3 Experience

The system enables users to transform themselves into CG avatars and interact with other users in real space. The users can communicate with other users through various CG avatars without revealing their own appearances. From the perspective of a user who observes the mid-air image, it makes human-to-human interaction in a real space such as an amusement park more enjoyable. The appearance of a customer service agent at the park, for example, can be changed according to the visitor's preferences.

We demonstrated Levitar at SIGGRAPH Asia 2019 [21]. In this demonstration, we adjusted the ISO sensitivity, and the user recognized the audience's face. Hence, we confirmed that this system can present a video with sufficient quality to recognize the face by adjusting the ISO sensitivity.

6 General Discussion

The limitations of this system are that the capturing performance deteriorates and the communication method, available environment, and activity range are limited. First, as described in Sect. 4.1, stray light is generated, brightness is decreased, and blur is increased by the transfer of the camera viewpoint in this system. These always occur due to the structure of the optical element. However, adjusting the ISO sensitivity is expected to solve the decrease in the brightness as described in Sect. 4.2. Furthermore, this may give the user extra visual capabilities, such as good visibility in dark environments. Secondly, since the mid-air image does not have a physical body, this system alone cannot achieve physical communication. Third, the environment needs to be dark to some extent to observe the mid-air image. Therefore, this system has a limitation regarding the available environment. Finally, the field of view and movable range of the user who plays the role of a mid-air CG avatar are limited. However, it is anticipated that these problems can be reduced by moving the entire apparatus.

For future work, we need to focus on the specific uses for Levitar and its applicability in other optical systems, and further evaluation is also needed. The specific uses for Levitar are the talk show type attraction and the teleconference. This system can be applied to attractions where actors can talk to guests as CG characters. In addition, if we prepare our own face model, this system can display not only CG images but also our face image. Therefore, users can participate in teleconferences with their real appearances. Furthermore, this design is expected to be usable with other optical systems such as AIRR. Since stray light as described in Sect. 4.1.1 is a characteristic of MMAPs, it may be possible to remove stray light by using AIRR. In addition, it is necessary to evaluate the communication factors other than face detection, such as facial expression recognition and audio design.

7 Conclusion

In this paper, we proposed a telepresence system using a mid-air image. We designed an optical system that places the camera viewpoint at the mid-air image position by combining a mid-air image display and optical transfer of the camera viewpoint with MMAPs. From the evaluation, the three points of stray light, brightness, and blur were confirmed as affecting the capturing performance by the MMAPs. It was also found that face detection works well on the captured video. In addition, we designed Levitar as an application for telepresence.

Acknowledgments

This research was supported by PRESTO, JST (JPMJPR16D5).

References

[1] E. Abe, M. Yasugi, H. Takeuchi, E. Watanabe, Y. Kamei, and H. Yamamoto. Development of omnidirectional aerial display with aerial imaging by retro-reflection (airr) for behavioral biology experiments. *Optical Review*, 26:221–229, Feb. 2019. doi: 10.1007/s10043-019-00502-w

[2] M. Chen. Leveraging the asymmetric sensitivity of eye contact for videoconference. In *Proc. SPIE 6392, Three-Dimensional TV, Video, and Display V, 63920E*. SPIE, 2002. doi: 10.1117/12.690574

[3] S. Hunter, R. Azuma, J. Moisant-Thompson, D. MacLeod, and D. Disanjh. Mid-air interaction with a 3d aerial display. In *Proceedings of ACM SIGGRAPH 2017 Emerging Technologies*. ACM, New York, NY, USA, 2017. doi: 10.1145/3084822.3084827

[4] A. Jones, M. Lang, G. Fyffe, X. Yu, J. Busch, I. McDowall, M. Bolas, and P. Debevec. Achieving eye contact in a one-to-many 3d video teleconferencing system. *ACM Transactions on Graphics (TOG)*, 28(3), Aug. 2009. doi: 10.1145/1531326.1531370

[5] H. Kim, I. Takahashi, H. Yamamoto, S. Maekawa, and T. Naemura. Mario: Mid-air augmented reality interaction with objects. *Entertain-*

ment Computing, 5(4):233–241, Dec. 2014. doi: 10.1016/j.entcom.2014.10.008

[6] N. Koizumi, Y. Niwa, H. Kajita, and T. Naemura. Optical design for transfer of camera viewpoint using retrotransmissive optical system. *Optical Review*, Jan. 2020. doi: 10.1007/s10043-019-00575-7

[7] R. Lienhart and J. Maydt. An extended set of haar-like features for rapid object detection. In *Proceedings. International Conference on Image Processing*, vol. 1, pp. I–I, Sep. 2002. doi: 10.1109/ICIP.2002.1038171

[8] Y. Maeda, D. Miyazaki, and S. Maekawa. Aerial imaging display based on a heterogeneous imaging system consisting of roof mirror arrays. In *Proceedings of 2014 IEEE 3rd Global Conference on Consumer Electronics (GCCE)*. IEEE, 2014. doi: 10.1109/GCCE.2014.7031217

[9] S. Maekawa, K. Nitta, and O. Matoba. Transmissive optical imaging device with micromirror array. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 49–56. ACM, New York, NY, USA, 2006. doi: 10.1145/503376.503386

[10] Y. Makino, Y. Furuyama, S. Inoue, and H. Shinoda. Haptoclone (haptic-optical clone) for mutual tele-environment by real-time 3d image transfer with midair force feedback. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2016. doi: 10.1145/2858036.2858481

[11] Y. Matsuura and N. Koizumi. Scoopirit: A method of scooping mid-air images on water surface. In *Proceedings of the 2018 ACM International Conference on Interactive Surfaces and Spaces*, pp. 227–235. ACM, New York, NY, USA, 2018. doi: 10.1145/3279778.3279796

[12] K. Okumura, H. Oku, and M. Ishikawa. High-speed gaze controller for millisecond-order pan/tilt camera. In *Proceedings of 2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011. doi: 10.1109/ICRA.2011.5980080

[13] OpenCV. Cascade classifier, 2019. `https://docs.opencv.org/trunk/db/d28/tutorial_cascade_classifier.html`.

[14] M. Otsubo. U.s. patent no. 8,702,252., Filed January 30, 2012, issued April 22.

[15] K. Otsuka. Behavioral analysis of kinetic telepresence for small symmetric group-to-group meetings. *IEEE Transactions on Multimedia*, 20(6):1432–1447, June 2018. doi: 10.1109/TMM.2017.2771396

[16] T. Piumsomboon, G. A. Lee, J. D. Hart, B. Ens, R. W. Lindeman, B. H. Thomas, and M. Billinghurst. Mini-me: An adaptive avatar for mixed reality remote collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2018. doi: 10.1145/3173574.3173620

[17] S. Tachi. *Teleexistance*. World Scientific, 2nd ed., 2015. doi: 10.1142/9248

[18] K. Tadatoshi, N. Hideaki, P. R. Lalintha, and M. Kouta. Scalable autostereoscopic display with temporal division method. In *ICAT-EGVE 2018 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*. The Eurographics Association, 2018. doi: 10.2312/egve.20181323

[19] M. Takasaki, K. Ohashi, and S. Mizuno. Interaction of a stereoscopic 3dcg image with motion parallax displayed in mid-air. In *Proceedings of SIGGRAPH Asia 2018 Posters*. ACM, New York, NY, USA, 2018. doi: 10.1145/3283289.3283343

[20] Y. Terashima, S. Suyama, and H. Yamamoto. Aerial depth-fused 3d image formed with aerial imaging by retro-reflection (airr). *Optical Review*, 26:179–186, Feb. 2019. doi: 10.1007/s10043-018-0473-9

[21] K. Tsuhciya and N. Koizumi. Levitar: Real space interaction through mid-air cg avatar. In *Proceedings of SIGGRAPH Asia 2019 Emerging Technologies*. ACM, New York, NY, USA, 2019. doi: 10.1145/3355049.3360539

[22] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, vol. 1, pp. I–I, Dec 2001. doi: 10.1109/CVPR.2001.990517

[23] VRChatNet. Vrchat, 2019. `https://www.vrchat.com/`.

[24] M. E. Walker, D. Szafir, and I. Rae. The influence of size in augmented reality telepresence avatars. In *Proceedings of 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019. doi: 10.1109/VR.2019.8798152

[25] H. Yamamoto, M. Yasui, M. S. Alvissalim, M. Takahashi, Y. Tomiyama, S. Suyama, and M. Ishikawa. Floating display screen formed by airr (aerial imaging by retro-reflection) for interaction in 3d space. In *Proceedings of 2014 International Conference on 3D Imaging (IC3D)*. IEEE, 2014. doi: 10.1109/IC3D.2014.7032590

[26] M. Yasui, Y. Watanabe, and M. Ishikawa. Occlusion-robust sensing method by using the light-field of a 3d display system toward interaction with a 3d image. *Applied Optics*, 58(5):A209–A227, Feb. 2019. doi: 10.1364/AO.58.00A209

[27] B. Yoon, H. il Kim, G. A. Lee, M. Billinghurst, and W. Woo. The effect of avatar appearance on social presence in an augmented reality remote collaboration. In *Proceedings of 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2019. doi: 10.1109/VR.2019.8797719