Improvement of Depth Images for Space-sharing Content Viewing System

Kazuma Yoshino†, Toshiyuki Niida, Takuya Handa and Kensuke Hisatomi

1NHK Engineering System, Inc., Tokyo, Japan
(Tel: +81-3-5494-2475; E-mail: yoshino.kazuma@nes.or.jp)
2NHK Osaka Station, Osaka, Japan
(Tel: +81-6-6941-3424; E-mail: niida.t-jm@nhk.or.jp)
3NHK Science & Technology Research Laboratories, Tokyo, Japan
(Tel: +81-3-5494-3269, 3396; E-mail: {handa.t-es, hisatomi.k-ko}@nhk.or.jp)

Abstract: We developed a space-sharing content viewing system based on the assumption that augmented reality (AR) glasses will be used extensively in the future. Our vision is to assist people in remote locations to enjoy AR and virtual reality programs together. To realize this, we developed a system that can display life-size persons virtually and communicate with them in real-time. In this study, we used the system for another application: for creating a television program that allows family members in remote locations to meet again virtually. Communication with three-dimensional (3D) figures demonstrated the possibility of creating new programs, such as family interactions. However, a technical problem was observed pertaining to the inability to obtain depth data for black-colored hair with infrared signals used for 3D measuring. Accordingly, herein we combined two different sensors, namely the time-of-flight (ToF) and stereoscopic vision with infrared projection, to capture 3D images without missing parts, thereby capitalizing on both the accuracy of ToF and stability of stereoscopic vision. The evaluation results showed that the proposed method filled in 70% of the initial holes in the depth data and provided depth to 90% of the color pixels in the area around the face.

Keywords: Augmented reality, Virtual reality, 3D reconstruction, Social VR

1. INTRODUCTION

We studied a new viewing style for future media in which augmented reality (AR) glasses will be used extensively. In this study, we focus on the ability of AR glasses to display life-size people. This will enable them to share and experience AR and virtual reality (VR) content with people in remote locations. We have developed and demonstrated this concept as a space-sharing content viewing system [1]. We confirmed that by displaying a three-dimensional (3D) virtual person in real-time, it is possible to share an immersive experience with people in remote locations.

Here, we report on a project that used this developed system for broadcasting programs. During the COVID-19 pandemic, the importance of face-to-face meetings has increased. By using the system, we created a program in which families living far away from each other could meet again virtually. This was achieved by capturing and transmitting live-action 3D images of the family in real-time and synthesizing them in conjunction with real-space environments. The meeting scene became a heartwarming scene in the program. However, there are challenges associated with the system concerning the capture and display of a person in a virtual environment. When an infrared (IR) time-of-flight (ToF) sensor is used to capture 3D images of people, the lack of depth data for black-colored hair resulted in a problem. In this study, we propose a method that uses multiple sensors to solve the problem and discuss its effectiveness.

† Kazuma Yoshino is the presenter of this paper.
transmit in real time. Volumetric live-action images have the potential to provide new value to remote communication that is different from those provided by avatars. Orts-Escolano et al. realized a remote communication system with AR glasses such that remote participants appeared to almost be present in the same physical space [5]. Gunkel et al. developed a system using Kinect v2 to transmit 3D data and conducted meetings remotely [6]. They investigated the possibility of real image transmission and showed that the conferencing and education applications are highly expected use case. Communication systems with live-action people are in high demand, and are becoming more functional and sophisticated. In 2021 Google released “Project Starline” [7], and Microsoft Research presented “VirtualCube” at IEEE VR in 2022 [8]. All these studies are aimed at creating a communication system that makes it seem as if a real person is right in front of you, allowing improved communication for better work outcomes. Hence, we expect there is a considerable value in communication with live-action figures. At this time, we focused on the fact that the meetings of people would be beneficial not only to people who meet but also to those who are watching. Given the dramatic changes caused by COVID-19, we planned to create a broadcasting program wherein a family could meet again virtually for a heart-warming experience. There are no examples where the focus is on the power of watching the scene of a meeting with a virtual person, nor are there any examples of the use of these AR and VR techniques in broadcasting. We describe the on-air realization and solutions to problems experienced during the process.

3. APPLICATION IN BROADCASTING PROGRAMS

One of the core technologies of the space-sharing content viewing system is the transmission of 3D live-action images of people and composing them to real and virtual spaces at life-sizes in real-time. We attempted to use a new type of expression and applied this technology to a singing-contest program for children in the Kansai area produced by NHK Osaka Station. The family that appeared on the program included the mother and daughter who were living in Osaka and the father who was living in Tokyo. COVID-19 restricted the long-distance travel of people, which made it difficult for the father to return home. Therefore, we planned to visit the father in Tokyo such that he could join his family in Osaka virtually to sing and dance with his daughter for the program.

3.1. System Overview

To realize singing and dancing from remote locations, images and audio need to be transmitted. We constructed the system for visual images as shown in Fig. 1. The audio was connected by mobile phones. In Tokyo, we set up a depth and color camera, which acquired 3D images of the father and transmitted shot images via a network to Osaka in real-time. These images were then combined with the images of the actual space in Osaka. We used the Azure Kinect Developers Kit (AKDK) from Microsoft as the depth and color camera, and HOLO-COMMUNICATION from HoloLab Inc. was used for image transmission. The color camera was used in the 720p mode (1280 x 720 pixels progressive mode), and the depth camera was set in the NFOV 2x2 Binned mode (narrow field-of-view and long-distance mode by combining adjacent sensor pixels into a bin). In Osaka, we used a head-mounted display (HMD; Rift S from Meta), to which a stereo camera [ZED Mini (ZED M)] was attached. The HMD, which has a stereo camera, is often called a video see-through HMD and was used as an AR glass. ZED M can capture real spaces and their depth from the camera by stereo vision. HMD can obtain its spatial position, which is the same as the position of ZED M. The transmitted 3D person was composited with real spaces captured by the ZED M according to the position of the HMD. The composited images were rendered by calculating the visibility or hiddenness of real objects, called occlusion, using the depth information acquired by ZED M. Thus, we created a composite image that emulated the presence of a virtual person in a real space. We used Unity from Unity Technologies for the system development. The produced image is like the one shown in Fig. 2. While this image is presented as an illustration, it is actually a composite of live images. The composited images were displayed both on HMD and monitor screens. The HMD was used as an AR glass as well as in the form of a camera for program production. When users put on the HMD, they could see the composited images in stereoscopic view. Monitor screens displayed images of the HMD’s left eye, and the child in Osaka experienced the composited images with the monitor. These images are also used for broadcasting. The specifications of the PC used for transmission and composition are listed in Table 1.

3.2. Results and subsequent challenges

We connected Tokyo and Osaka as shown in Fig. 1 and measured the line bandwidth and latency. Line bandwidth was checked by task managers on both the transmission and reception sides. The delay was measured by comparing the images of the radio clock transmitted from Tokyo with that in Osaka. The results showed that the line bandwidth ranged from 30 to 50 Mbps, and the delay was...
approximately 1 s. The transmission rates at the transmission and reception locations were almost the same, and a minor loss occurred during the transmission. We also identified the number of staff members needed to set up and test the system. We estimated that two technical staff members could complete the setup in approximately 2 h. Given that this was the first case of using the system for a TV program setup and broadcast, we discussed the system limitations with directors and decided on the type of images that we could acquire. The filming day progressed as planned. This was a recorded (not a live) program. During the recording, we could observe heart-warming communication expressed in the form of touching, which was enabled by combining real-time images with occlusion and by localizing live-action images like Fig. 2. After the broadcast, there were comments, such as “amazed by VR technology” and “impressed by the family connection” on social networking sites, which demonstrated that this program added new value to technology in the field of program production. In the future, we need to solve the technical issues that we identified. Specifically, the IR distance measurement by AKDK could not generate depth data from black-colored hair. Depth image taken with AKDK creates person images as shown in Fig. 3 (left). This problem is attributed to the low reflection of IR signals that are used for depth measurements. To solve this problem, we temporarily introduced morphological transformations, as shown in Fig. 3, to expand the cropping area and achieved good quality images. However, this approach has a disadvantage associated with incorporation of the surrounding space as well as including the extent of the hair. In the next section, we propose another approach for improvement.

3. RESULTS

Fig. 3 Results obtained following the expansion of the clipping area.

4. IMPROVEMENT OF DEPTH DATA

To improve depth data, we combined two different sensors. We added the RealSense D435i (RS) from Intel Corporation as the second sensor. The AKDK measures distance based on ToF, while the RS achieves this in accordance with stereoscopic vision. The differences in these two outputs are shown in Fig. 4. Depth data were captured more precisely with the AKDK. However, some black areas, such as that of black-colored hair, caused depth data losses. The AKDK emits IR light and measures the time of its reflection. Therefore, objects with low IR reflectance are not captured sufficiently, thereby making measurements impossible. Conversely, RS has fewer missing parts and can stably acquire data. The sensor detects a relatively large area by combining stereoscopic vision with marker projection. Therefore, we attempted to fill the areas where AKDK lacked depth data with that obtained by the RS. The color texture was used from the AKDK because its depth data were filled in.

Fig. 4 Output images and correspondence of each camera parameter used in this study.

4.1. Camera placements

First, we considered the placements of the cameras. Given that the color and depth cameras in AKDK were spatially separated from each other, parallax inconsistencies were observed. To minimize the visual effects of these inconsistencies, we placed the color and depth cameras along the vertical direction. This can make inconsistencies appear on the upper or lower sides of objects. Fig. 5 shows the images acquired with the color camera (top) and depth camera (bottom). The two images on the left are color and depth data, which are converted to the coordinates of the color camera. The functions provided in the AKDK software development kit (SDK) were used for coordinate transformation. The image on the right is a combination of the two data, where the depth images are transparently overlaid. In this figure, the depth data contain three notable areas.

1. area missing IR reflection;
2. upper boundaries of the objects;
3. lower boundaries of the objects.

Our method improved areas 1 and 2 and did not target area 3. Area 1 was caused by low IR reflectance, for which black-colored hair is a typical example (glasses are also included herein). Areas 2 and 3 were caused by parallax between the color and depth cameras. In area 2, the visible area of the color camera was wider than that of the depth camera. Almost all color pixels of the front object in this area have depth data. In area 3, the visible area of the color camera was narrower than that of the depth camera. In this area, all color points had corresponding depths. The mismatch observed there was not because of the lack of depth data but was because there were two depths for each color. In the case of Fig. 6, the color at point X1 corresponds to two depths, namely X1 and X2. The selection of one over the other is based on the SDK of the AKDK. Because the purpose of the second camera was to compensate for the part that was not obtained from the main depth, mainly in the upper direction, we did not...
consider area 3 for correction. To achieve this, the second sensor was positioned above the main sensor, as shown in Fig. 7. The tilt angle was adjusted according to the distance from the target object and height of the sensor. This ensures changes in the reflective properties of IR radiation and mitigates parallax errors by setting and tilting the sensor separately. We refer to the combined sensor as "AKpRS".

Fig. 5 Correspondence between color and depth in AKDK.

Fig. 6 Principle of arising two corresponding points.

Fig. 7 An image of the combined sensors.

4.2. Process

All images acquired by each camera were converted to the viewpoint of the AKDK color camera. For conversion, the internal and external parameters of all cameras were needed. The correspondence is summarized in Fig. 4. The internal parameters of the cameras and the external parameters between color and depth included in the same sensor used the factory default values. The external parameters of the other sensors (AKDK and RS) were manually calculated from several corresponding 3D positions in the software, as demonstrated in Fig. 8. The corresponding points used as inputs were identified by markers attached to mannequins. Using the calculated camera parameters, all images were transformed to the coordinate color camera in AKDK. The missing areas of the IR data in AKDK were represented as zero depths. If the same point in the RS depth was not zero, it was replaced with that non-zero value. Based on these processes, we acquired AKDK depth data filled with RS data. The image processing program was created in C++, and the calculated 3D data were shared with Unity for display.

4.3. Evaluation

We evaluated the effects of the combined sensors on depth improvement. We captured images of five subjects (four males and one female) in our laboratory where fluorescent lights were set on the ceiling above the subjects’ heads, and analyzed the data. This procedure is shown in Fig. 9. The color and depth data were shot from a distance of 1.2 m, with the AKDK positioned directly in front of the face. We evaluated the face area because the lack of data around the face significantly impacts impression. We obtained the face area in the color image by removing the background with the image processing tool in Microsoft PowerPoint. This enabled us to extract the outline of the face for evaluation in the depth image, as shown by the dotted line in Fig. 9. We counted the pixels inside the selected region and calculated two parameters, which we named depth coverage rate (DCR) and filling rate (FR). DCR refers to the percentage of the color image that contains depth data and is expressed as

$$ DCR(\%) = \frac{N_{\text{non-zero}}}{N_{\text{total}}} \times 100 $$

(1)

where $N_{\text{non-zero}}$ represents the numbers of pixels in the face area with non-zero depth and $N_{\text{total}}$ is the numbers of pixels in the face image. FR refers to the percentage of the image that is filled in with data from the RS and is expressed as

$$ FR(\%) = \frac{N_{\text{zero,AKDK}} - N_{\text{zero,AKpRS}}}{N_{\text{zero,AKDK}}} \times 100 $$

(2)

where $N_{\text{zero,AKDK}}$ and $N_{\text{zero,AKpRS}}$ represent the numbers of pixels in the face area with zero depth for the AKDK and AKpRS, respectively.
4.4. Results

The results are shown in Table 2. It shows our method can fill in 70% of the missing holes and improve the fill rate to over 90%. Fig. 9 shows the results of our study on subject A. Most of the missing IR data areas, particularly at the top of the head, are filled in. In addition, transparent objects, such as glasses are improved. In this figure, some areas are classified as zero points and background. This mismatch may be attributed to multiple factors, such as calibration inaccuracies and camera distortion.

Fig. 10 is displayed as a 3D image and is viewed from above. The hair area was filled in, and the image was volumetric. The data output rate of the sensor was 30 frames per second, and the processing was completed within the frame rate.

Table 2 Results for each subject (%)

<table>
<thead>
<tr>
<th>Subjects ID</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCR (AKDK)</td>
<td>73</td>
<td>72</td>
<td>80</td>
<td>73</td>
<td>78</td>
<td>75.2</td>
</tr>
<tr>
<td>DCR (AKpRS)</td>
<td>91</td>
<td>91</td>
<td>95</td>
<td>93</td>
<td>95</td>
<td>93.0</td>
</tr>
<tr>
<td>FR</td>
<td>66</td>
<td>69</td>
<td>73</td>
<td>74</td>
<td>79</td>
<td>72.3</td>
</tr>
</tbody>
</table>

Fig. 10 Results of filling in based on the top view.

5. CONCLUSIONS

This study described a new application of a space-sharing content viewing system in the field of broadcasting. Moreover, a solution to the problem of missing data for black-colored hair is proposed by applying a combination of different types of sensors. The constructed system was able to achieve the filling-in process without reduction in the frame rate and with improvement in the depth images. These results enabled us to produce programs with better quality. In the future, we plan to further study the applications of this system, including the design of an environment to watch AR and VR content.

REFERENCES