International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2024)
S. Hasegawa, N. Sakata and V. Sundstedt (Editors)

XR remote dialogue system presenting speaker's expression using a real-space avatar robot

Yuto Yoneda[1] Yukiya Ojima[1] Yosuke Fukuchi[1] Vibol Yem[2] Yasushi Ikei[3] Nobuyuki Nishiuchi[1]

[1]Tokyo Metropolitan University, Japan
[2]University of Tsukuba, Japan
[3]The University of Tokyo, Japan

Abstract

This paper proposes a remote dialogue system that utilizes XR technology with an avatar robot. The proposed system is designed to facilitate on-site interaction between a worker wearing mixed reality glasses (HoloLens 2) and a remote interlocutor wearing a head-mounted display, including spatial objects. The remote interlocutor's facial expressions, head movements, mouth movements, and gaze direction were reproduced on a 3D avatar. The effectiveness of the proposed system was evaluated qualitatively. In addition, the effect of the facial expression generation of the real avatar was assessed.

CCS Concepts

• ***Computing methodologies*** → ***Mixed / augmented reality;***

1. Introduction

Recently, with the advancements in Beyond 5th Generation technology and the spread of telework, real-time communication among remote locations has become increasingly important. However, current common flat panel displays lack spatial information, thereby making nonverbal communication inadequate. In addition, the lack of a sense of presence reduces the quality of sharing the status of the remote locations, which can become an obstacle to effective communication [TNN*20, YY20]. Conventional commercial telepresence robots have primarily focused on presenting avatars on screens, with limited task performance capabilities. Recent studies have introduced a robotic avatar with a realistic operator's avatar model that can be seen by the on-site collaborator with MR glasses [JZWR21, LWH*23]. However, eye contact and realistic facial expressions have not been sufficiently achieved such that the remote collaborators can share the attention to spatial objects in question with the non-verbal spatial awareness based on facial expressions. The nuanced expression of emotions through avatars, beyond simple eye or mouth movements [KBL*19, LBW20], appears to be an unexplored aspect of telepresence systems.

Additionally, while 360° imaging of remote spaces has been implemented, the generation of binocular disparity for enhanced stereoscopic viewing and improved communication with 3D objects has not been extensively investigated in the context of telepresence robotics. An extended reality (XR) system, which can fully utilize 3D spatial information, is an effective tool to realize smooth communication between remote physical spaces. The application area of an XR system has been expanded with the emergence of lightweight mixed reality (MR) glasses with enhanced functionality [NLSB19b, NLSB19a]. In this study, we attempted to create a system in which a realistic CG avatar of the speaker at the remote site is presented to the place of the camera on a telepresence robot to provide a sensation that is equivalent to that when talking with the speaker in a real-world environment.

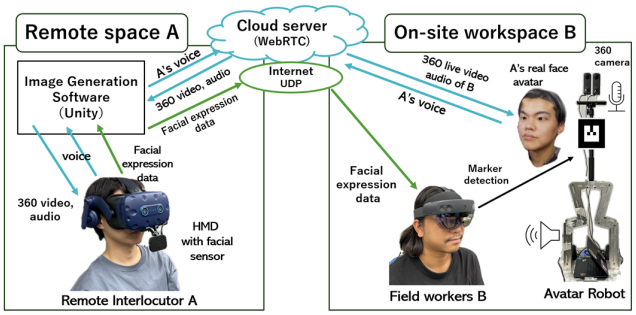

Figure 1: *Configuration of the XR remote dialogue system*

2. The XR remote dialogue system

The structure of our XR dialogue system is shown in Figure 1. On the left is remote space A, where the remote interlocutor A moves an avatar robot at point B without going to the on-site workspace B. The avatar robot is equipped with two omnidirectional cameras (THETA Z1, Ricoh) by which a 360° real-time 3D video (4K, 30 fps) is captured and presented to the remote person A's head-mounted display (HMD; VIVE Pro Eye, HTC). This stereoscopic

DOI: 10.2312/egve.20241362

image is viewed with no apparent delay to head turn. The remote worker B wearing MR glasses (HoloLens 2, Microsoft) can interact with the remote interlocutor A's CG avatar (Figure 2) at the position of the cameras of the avatar robot [KYNI20].

In this system, head rotation angle, eye movement, and mouth area movement data measured by the sensor in remote interlocutor A's HMD are transmitted to the MR glasses at the on-site workspace B, which reproduces the movements on the CG avatar.

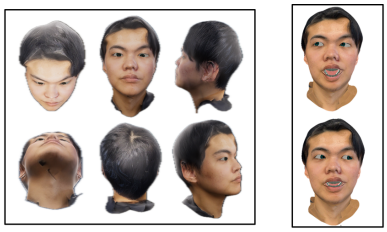

Figure 2: *A 3D CG avatar model. Model appearances (left), and blendshapes with eye and mouth displacements applied (right).)*

2.1. Real CG avatar

The 3D head model of the avatar of the remote interlocutor A was constructed using an optical capture process. Here, an iPhone 13 (LiDAR) and Scaniverse (Niantic, Inc.) were used to acquire the shape and color of the head. From this model, blendshapes for facial expressions shown in Figure 2 were created [KKYI21a, KKYI21b, KKYI22].

2.2. Facial expression data of remote interlocutor A

To reproduce the facial expressions of the remote interlocutor A on the CG avatar, the eye, head, and mouth movements (displacement) were measured using the eye movement sensor in the HMD and a lip tracker device (VIVE Facial tracker, HTC) added to the HMD. The acquired data comprised 4 degrees-of-freedom of the eyes, 6 degrees-of-freedom of the head, and 11 degrees-of-freedom of the mouth [App24].

2.3. Communication processing

Two communication protocols were employed to implement the system. Two WebRTC media channels were used to transmit Real-time 360° L/R videos from the on-site workspace B to the remote space A via a cloud server. The bidirectional voice channel was installed on the WebRTC. In addition, the facial expression data of remote interlocutor A was sent to the on-site workspace B via User Datagram Protocol.

2.4. Displaying CG avatar in the on-site work place

A model of a realistic CG avatar of the remote interloper A in remote space A was embedded in the on-site worker B's HoloLens 2 device. The CG avatar face was presented to the camera position of the avatar robot [KYNI20] following the motion of the robot. The position of the camera was located by HoloLens 2 using an AR marker at the bottom of the camera through a recognition library (Vuforia Augmented Reality SDK, PTC Inc.).

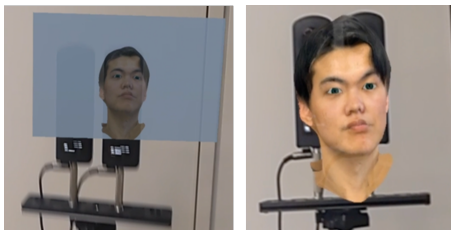

Figure 3: *Presentation of the remote interlocutor A's avatar. 2D (left) and 3D (right) rendering.*

3. Evaluation experiment

3.1. Objective

We evaluated the effectiveness of the XR remote dialogue system focusing the visibility and the effect of presenting the CG model (i.e., the head) of the speaker, qualitatively. The test work scenario was a model work where the user interactively instructs regarding real objects set in an indoor space and shared virtual objects floating before the camera. Printed photos of these objects were randomly pinned on a partition board to be searched from the remote site. This experiment partially assumes that maintenance work would be performed jointly at a job site or factory in the same manner.

3.2. Participants and presentation conditions

In this study, twelve healthy university (graduate school) students (eight males and four females; aged 21–24) participated in the experiment. The participants were divided into two groups, i.e., groups of remote interlocutors A and field workers B. There were six pairs comprising A and B roles. The avatar of remote interlocutor A displayed on the MR glasses (HoloLens 2) worn by remote worker B included 2D and 3D presentations, as shown in Figure 3. Here, the 2D condition assumed a conventional telepresence robot operator display.

3.3. Experimental tasks

In workspace B, we prepared real-space objects, as shown in Figure 4 (a), including paper cups, pens, books, scissors, staples, and tape, and we printed black and white photographs of these objects in three different sizes (9–25 cm). The printed photographs were then attached to a whiteboard panel (180 cm × 180 cm), as shown in Figure 4 (b). On the front side of the board, 17 photos of the objects were placed at random angles. On the back side, the black and white photos of three types of virtual paper cup objects and four types of virtual cylindrical objects were placed at random angles. Note that remote interlocutor A and on-site worker B can see the same virtual objects; thus, the cup and cylinder objects in Figure 4 (c,d) were placed approximately 30 cm in front of the avatar robot.

The experimental task required the worker B participant to hold one of the real objects in the hand in front of the camera (i.e., the face) of the avatar robot, to have the interloper A memorize it, remove it from their field of view, and then ask the interloper A to find a corresponding image among the pictures on the whiteboard.

(a) (b) (c) (d)

Figure 4: *Experimental environment. (a) Real objects used in the experiment. (b) Pieces of photos pasted on a whiteboard (Upper: front side; Lower: backside). (c) Virtual objects in the field of view of the interlocutor's HMD. (d) Field of view of HoloLens 2.*

3.4. Experimental procedure

The following procedure was used to perform the experimental task. Here, the real objects were targeted first, followed by the virtual objects.

1. Remote worker B picks up any real object and presents it in front of the camera.
2. Remote interlocutor A views the real object in 360° real-time stereoscopic view using the HMD.
3. Remote interlocutor A looks for how many pictures of the viewed real object are on the whiteboard by moving the robot's viewpoint, and answers by explaining the location of the object verbally. Worker B observes the avatar of remote interlocutor A searching with HoloLens 2.

The experimental task was performed on both sides of the whiteboard. For the virtual objects, Remote worker B specified one of the shared virtual objects and searched how many corresponding pictures were present on the whiteboard.

The viewpoint of the avatar robot was manipulated by remote control using a VIVE controller. However, due to insufficient positional accuracy during the experiment, the experimenter moved the avatar robot according to verbal instructions given by the interlocutor.

3.5. Assessment items

The following questionnaire was administered to each group of participants to evaluate the proposed system.

The questionnaire for the remote interlocutor A group covered the following aspects: reality (Q1), clarity of real-space objects (Q2), clarity of virtual objects (Q3), clarity of printed alternatives (Q4), and air-self-location (Q5). Here, the visual analog scale (VAS) was employed to answer the questions. The question text and anchors of the VAS are in the following.

Q1. *Did you feel a sense of reality and presence when you had a dialogue with the on-site participant B? Left VAS anchor: I had no idea what was going on with on-site worker B; right VAS anchor: it felt as if on-site worker B was right in front of me.*
Q2. *Did you clearly understand the real-space objects indicated by your interlocutor? Left VAS anchor: I had no idea what he was indicating; right VAS anchor: it was the same as if I had actually seen the object in the field.*
Q3. *Did you clearly understand the virtual object that your interlocutor showed you? Left VAS anchor: I had no idea what they were indicating; right VAS anchor: I could clearly understand which object it was.*
Q4. *Did you clearly understand the answer choices in the other person's space from the viewed video (including viewpoint shift)? Left VAS anchor: I could not see at all and could not answer; right VAS anchor: I could understand as if I were watching at the site.*
Q5. *When you changed the location of the dialogue, did you understand your position well? Left VAS anchor: I had no idea of direction and position; right VAS anchor: I had a perfect idea about the position and direction where I was working.*

A different questionnaire was administered to the field worker B group. This questionnaire covered the following aspects: presence (Q1), clarity of the avatar's eye movement (Q2), clarity of the real avatar's head movement (Q3), and virtual object localization (Q4). The question text and VAS anchors are shown in the following. Note that the respondents were asked to answer the questions for both the 2D and 3D views of the avatar.

Q1. *How did you feel when you saw the real avatar of the remote interlocutor? Left VAS anchor: it did not feel like a human being; right VAS anchor: it felt like a real person was right in front of me.*
Q2. *How clear were the eye movements of the real avatar when indicating the interaction target? Left VAS anchor: unclear and not understandable at all; right VAS anchor: clearly and unquestionably understandable.*
Q3. *How clear were the real avatar's head movements when indicating the interaction target? Left VAS anchor: unclear and not understood at all; right VAS anchor: clearly understood without doubt.*
Q4. *Were the virtual objects clearly localized and visible? Left VAS anchor: unclear and not understandable at all; right VAS anchor: it felt like a real object was right at the place.*

3.6. Experimental results and discussion

3.6.1. Remote A's responses to the questionnaire

The answers to the questions administered to the remote interlocutor A group (HMD side) are shown in Figure 5, where the error bars represent standard errors. Remote interlocutor A has a higher evaluation value than the middle value for the sense of reality (Q1) of the on-site worker B. He has a fairly good understanding of B's situation in the on-site space. In addition, the indicated objects (Q2 and Q3) were rated higher than the clarity of worker B because the real objects were not seen by the camera lens. This may be because the real objects were displayed close to the camera lens, and the virtual objects were CG; thus, the indicated object's clarity was sufficiently high. In particular, the fact that the real objects could be seen with nearly the same clarity as the CG objects suggests

that the real-time video of this XR remote dialogue system is sufficiently effective to observe remote locations. Although the system utilizes full-dome surround images, the work target was only on the whiteboard. Thus, the self-location (Q5) was considered to be highly recognized even though it was incidental information.

The clarity of the printed material (Q4) of the answer choice on the whiteboard was lower than the other results. This may be due to the low contrast of the black and white printed materials, the lower resolution of the camera and HMD (1440 x 1600 pixels per eye) than the human eye, and the insufficient lighting at the site.

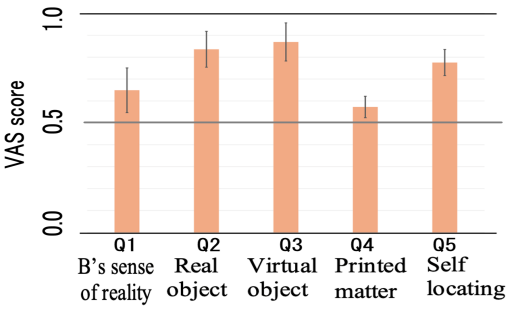

Figure 5: *Remote interlocutor A answers (n=6, SE)*

3.6.2. Responses to questionnaire for on-site worker B (2D and 3D views)

The answers from the field worker B group (HoloLens 2 side) are shown in Figure 6, with error bars representing standard errors. A two-level (2D/3D) t-test (MS Excel, Mac ver. 16.86) was conducted for Q1 to Q4. Results show that the presence of remote interlocutor A's avatar (Q1) was significantly higher in 3D at 1% level ($p = 0.00036$), indicating the effect of stereoscopic avatar display. The effectiveness of the 3D display was demonstrated at a significance level of 5% ($p = 0.0274$) for the clarity of whether A was aware of the topic of discussion (Q3) by the head rotation of the remote interlocutor A's avatar. The effectiveness of the 3D display was also demonstrated at a significance level of 5% ($p = 0.0274$). Furthermore, localization of the virtual object under discussion (Q4) exhibited a significant difference trend ($p = 0.077$) and was considered to be clearer in the 3D case. The scores for Q1, Q3, and Q4 were nearly equal to 0.8, which suggests that the level of clarity was quite high. The fact that the perceived reality of the remote participants was greatly enhanced compared to the 2D display suggests the effectiveness of the method.

The clarity of the avatar's eye movement to indicate the object under discussion (Q2) was slightly lower, and there was no significant difference between the 2D and 3D cases. Possible reasons for this observation are summarized as follows. Remote interlocutor B did not face the avatar of remote interlocutor A from the front. Instead, they observed it from an oblique angle; thus, they did not obtain a detailed image of its gaze because the remote interlocutor was facing the board to observe the details of the attached pictures on the white-board, and the field of view of the HMD (Max. 110 deg.) was not that wide. Thus, the eye movements were small. Remote interlocutor A's behavior was likely due to a tendency to observe head movements rather than make an assessment according to the eye movements.

The 2D display assumes that a flat display is mounted on the avatar robot and that the display is difficult to view from the side. Thus, it is considered that the dispersion of the visibility of the 2D display became larger.

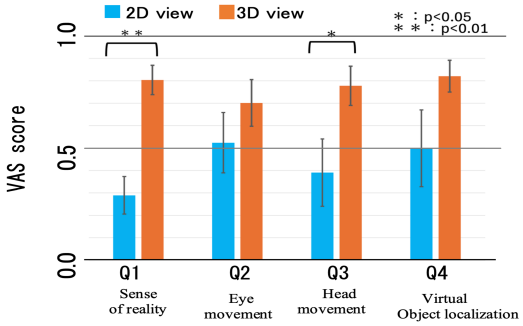

Figure 6: *Field worker B answers (n=6, SE)*

4. Conclusion

In this study, to facilitate collaborative work with a person on-site using an avatar robot sent to the remote work site, a CG avatar of a person operating the avatar robot was displayed at the camera position of the robot using an XR device. An environment was constructed to enable dialogue about shared spatial objects. To create highly realistic dialogue, the proposed system was designed such that the remote operator wearing MR glasses can observe the gaze, head movement, and facial expressions of the realistic 3D CG avatar of the remote interlocutor wearing the HMD.

Generally, both the remote interlocutor wearing the HMD and the on-site worker wearing MR glasses were able to interact and understand the environment using the 3D display, which suggests that the proposed system is effective for remote interactions. The clarity of the spatial objects was sufficient for the remote interlocutors by providing stereoscopic images. However, the clarity of the black-and-white printed images of the site workers and the search target in the indoor environment was somewhat lower due to the camera resolution and insufficient illumination. In the case of a site where the environment can be manipulated, it is possible to improve clarity with better conditions. With the MR glasses of the field workers, the images were easy to observe under indoor conditions, and the presentation of the 3D avatar and virtual objects was sufficient. However, under outdoor conditions, it would depend on the performance of the XR device. The eye, head, and mouth movements may have contributed to the avatar's sense of realism; however, more detailed control of the facial model is required to achieve a natural and expressive reproduction.

5. Acknowledgments

This work was partially supported by JSPS KAKENHI JP19K20325, 21H04883, 18H04118, 21K19785 in addition to Local-5G Project at TMU and the Telecommunications Advancement Foundation.

References

[App24] APPLE INC.: ARFaceAnchor.BlendShapeLocation, 2024. Accessed: 2024-10-25. URL: https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation. 2

[JZWR21] JONES B., ZHANG Y., WONG P. N. Y., RINTEL S.: Belonging there: VROOM-ing into the uncanny valley of XR telepresence. *Proceedings of the ACM on Human-Computer Interaction 5*, CSCW1 (Apr. 2021), 1–31. doi:10.1145/3449133. 1

[KBL*19] KIM S., BILLINGHURST M., LEE G., NORMAN M., HUANG W., HE J.: Sharing emotion by displaying a partner near the gaze point in a telepresence system. pp. 86–91. doi:10.1109/IV-2.2019.00026. 1

[KKYI21a] KATO A., KIKUCHI Y., YEM V., IKEI Y.: A study on facial expression generation method for VR avatar. In *Proceedings of the 26th Annual Conference of the Virtual Reality Society of Japan* (2021), pp. 3C1–4. 2

[KKYI21b] KATO R., KIKUCHI Y., YEM V., IKEI Y.: Real-time facial animation of a reality avatar based on Japanese vowels in a speech audio stream. In *Proceedings of the International Display Workshops* (2021), vol. 28, ITE and SID, pp. 526–529. doi:https://doi.org/10.36463/idw.2021.0526. 2

[KKYI22] KATO R., KIKUCHI Y., YEM V., IKEI Y.: Reality avatar for customer conversation in the metaverse. In *Human Interface and the Management of Information: Applications in Complex Technological Environments* (Cham, 2022), Yamamoto S., Mori H., (Eds.), Springer International Publishing, pp. 131–145. 2

[KYNI20] KIKUCHI Y., YEM V., NAGAI C., IKEI Y.: A study on MR dialogue-assisted telepresence system. In *Proceedings of the 25th Annual Conference of the Virtual Reality Society of Japan* (2020), pp. 3B1–5. 2

[LBW20] LEE M., BRUDER G., WELCH G. F.: The effect of vibrotactile feedback on teleport-based reorientation. *Journal on Multimodal User Interfaces 14* (Dec. 2020), 373–383. doi:10.1007/s12193-020-00343-x. 1

[LWH*23] LUO L., WENG D., HAO J., TU Z., JIANG H.: Viewpoint-controllable telepresence: A robotic-arm-based mixed-reality telecollaboration system. *Sensors 23*, 8 (2023), 4113. URL: https://www.mdpi.com/1424-8220/23/8/4113, doi:10.3390/s23084113. 1

[NLSB19a] NORMAN M., LEE G., SMITH R. T., BILLINGHURST M.: A mixed presence collaborative mixed reality system. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)* (2019), IEEE, pp. 1158–1159. doi:10.1109/VR.2019.8797966. 1

[NLSB19b] NORMAN M., LEE G. A., SMITH R. T., BILLINGURST M.: The impact of remote user's role in a mixed reality mixed presence system. In *Proceedings of the 17th International Conference on Virtual-Reality Continuum and its Applications in Industry* (New York, NY, USA, 2019), VRCAI '19, Association for Computing Machinery. URL: https://doi.org/10.1145/3359997.3365691, doi:10.1145/3359997.3365691. 1

[TNN*20] TAGUCHI W., NAKANO Y., NIHEI F., FUKASAWA S., AKATSU H.: Supporting video-mediated communication based on nonverbal signals by speech and facial expressions. In *Proceedings of the 34th Annual Conference of the Japanese Society for Artificial Intelligence* (2020), pp. 4E2–OS–19a–05. doi:10.11517/pjsai.JSAI2020.0_4E2OS19a05. 1

[YY20] YAMADA R., YUIZONO T.: A proposal of facial expression direction to activate remote meeting. In *Proceedings of the Information Processing Society of Japan* (2020), vol. 2020-DCC-24 of *Digital Content Creation*, pp. 1–6. URL: https://ipsj.ixsq.nii.ac.jp/ej/?action=pages_view_main&active_action=repository_view_main_item_detail&item_id=202936&item_no=1&page_id=13&block_id=8. 1