Human-Like Telepresence System Using Dummy Head Projection for Real-Time Conversation with the Presence of a Remote Participant*

Takayoshi Yamada[1], Hiiro Okano[1], Akito Fukuda[1], Vibol Yem[2,3] and Keiichi Zempo[2,4,*]

[1]Graduate School of Science and Technology, University of Tsukuba, Ibaraki, Japan
[2]Institute of Systems and Information Engineering, University of Tsukuba, Tsukuba, Japan
[3]Organization of Advanced Teaching and Learning, University of Tsukuba, Tsukuba, Japan
[4]School of Transdisciplinary Science and Design, University of Tsukuba, Kuala Lumpur, Malaysia

Abstract

With the increasing demand for video conferencing, effectively reproducing the presence of remote participants has become a significant challenge. In addition, in environments where communication latency occurs, it is essential to mitigate the impact of such latencys to enable smooth communication. To address these issues, this study develops a telepresence system that projects remote participants onto a dummy head and combines it with an application that enables low-latency audio transmission. Furthermore, the system employs a binaural microphone to provide immersive audio communication, enabling high-quality interaction. The effectiveness of this system was evaluated by analyzing the interaction using object recognition technology. Additionally, through a fieldwork study connecting Japan and Malaysia, it was confirmed that local persons and remote participants could engage in enjoyable interactions. These findings suggest that this system contributes to the realization of more effective communication as a new human-like telepresence method for remote video conferencing that can be shared by multiple participants.

Keywords

Telepresence, Communication, Video conference

1. Introduction

With the increasing demand for video conferencing, it has become possible to communicate with anyone around the world while seeing their face, despite being physically distant. In traditional video conferencing, it is common to display the face of the conversational partner on a 2D screen, which has also been utilized as a telepresence system [1, 2, 3]. Telepresence refers to technology that provides a sense of presence to remote participants, making it seem as though they are physically present in the same space, while also creating the illusion of their presence in the environment [4, 5, 6, 7]. Video conferencing, by enabling communication through visual and auditory channels, is one means of realizing this telepresence.

In traditional video conferencing systems, dynamic displays [8, 9] and methods that add mobility by incorporating displays onto robots [10, 11] have been explored to enhance the satisfaction and presence of the conversational partner. However, 2D displays still present the challenge of insufficiently conveying non-verbal cues, such as eye contact [9]. On the other hand, displaying the face of the remote participant on a 3D display has been suggested as a way to facilitate the sensation of eye contact [12]. Furthermore, representing remote participants in 3D may improve the accuracy of non-verbal cues [13].

The presence of the conversational partner plays a significant role in influencing the immersion, intimacy, and trust in the conversation [14]. When the presence of the remote participant is not adequately established, important tasks such as decision-making are often completed by the local

APMAR'25: The 17th Asia-Pacific Workshop on Mixed and Augmented Reality, Sep. 26-27, 2025, Busan, South Korea
*Corresponding author.
zempo@iit.tsukuba.ac.jp (K. Zempo)
0000-0002-4012-4417 (T. Yamada); 0000-0002-8539-5843 (H. Okano); 0009-0007-4205-4858 (A. Fukuda); 0000-0002-7254-6614 (V. Yem); 0000-0003-2339-5298 (K. Zempo)

persons, leaving the remote participant in a supplementary role [15]. Furthermore, even during video conferences, local persons may naturally form groups, making it difficult for remote participants to join and possibly resulting in neglect [16]. On the other hand, enhancing the presence of the remote participant can improve task performance in remote collaboration [17]. Additionally, beyond the presence of the conversational partner, in the field of medical training, using interactive mannequins with dynamic facial expressions to enhance presence has been shown to improve the concentration and learning outcomes of trainees [18]. These studies suggest that enhancing the presence is crucial in various scenarios such as conversation, learning, and collaboration.

One of the challenges in video conferencing is technical issues related to communication, such as latency, audio/video delays, and freezes. These technical challenges have been shown to interfere with conversation and may degrade the quality of communication, potentially hindering the development of rapport with the conversation partner [19, 20, 14]. Moreover, in collaborative tasks involving multiple participants, it has been suggested that as latency increases, the team's performance declines non-linearly [21]. To address these issues, it is essential not only to enhance the presence of remote participants but also to create a communication environment that minimizes latency.

Therefore, this study proposes a telepresence system that projects the face of a remote participant onto a dummy head for 3D representation. In this study, we aim not only to verify whether this system actually enhances the presence of remote participants but also to combine it with a low-latency audio transmission application, thereby enabling high real-time video conferencing. Additionally, the effectiveness of the system is evaluated under the conditions of international remote conferencing between Japan and Malaysia, a scenario that has posed significant challenges in previous studies. This study investigates whether, when connecting Japan and Malaysia, this telepresence system can enhance the presence of remote participants while maintaining low-latency communication. The research question of this study is as follows:

RQ: Does 3D projection improve the presence of remote participants compared to traditional 2D displays?

This study contributes to enhancing the presence of remote participants in telepresence systems and facilitating smooth communication. Improving the presence of remote participants is essential for enabling local persons to interpret non-verbal cues more easily and immerse themselves in the conversation, which in turn enables effective remote video conferencing and collaboration.

2. Related Work

Telepresence [4, 5, 6, 7] systems enable remote participants to feel as though they are sharing the same space with local persons. One of the advantages of these systems is that they allow remote meetings to be conducted with a sense of presence that is close to face-to-face communication. This section focuses on telepresence systems that recreate the presence of remote participants in the local environment.

Misawa et al. introduced an approach in which a 2D display is worn by local persons, and remote participants are projected onto it to perform substitutional actions [1, 2]. Faridan et al. applied this approach to the field of education [22]. Beck et al. implemented a 3D telepresence system using two coupled stereoscopic multi-viewer systems, which improved the quality of non-verbal communication and user satisfaction in collaborative tasks [23]. Kim et al. used a cylindrical 3D display in their telepresence system to enhance the accuracy of non-verbal cues, suggesting that both motion parallax and stereoscopic vision improve the sense of presence and embodiment [13], and Gotsch et al. further enhanced this system [24]. Pejsa et al. successfully improved presence and reduced task completion time by projecting life-size virtual copies of remote participants into a room using a projector, compared to traditional 2D video conferencing [17].

Additionally, research has been conducted to further enhance the presence of remote participants by using technologies such as AR (Augmented Reality), VR (Virtual Reality), and MR (Mixed Reality) through the use of HMD (Head-Mounted Displays). Orts et al. enabled a near face-to-face experience

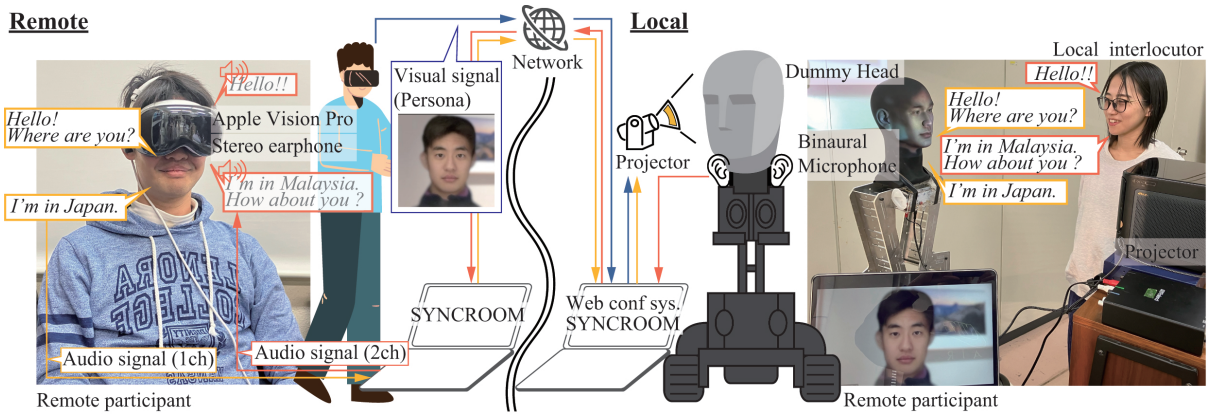

Figure 1: Overview of the proposed system. The remote participant wears the Apple Vision Pro and communicates with the local person. The local person interacts with the remote participant's persona, which is projected onto a dummy head. Integrating this system with a low-latency audio transmission application supports smooth communication.

by projecting a remotely captured person onto an HMD [25]. Piumsomboon et al. improved the experience of remote collaboration tasks by presenting small avatars [26], and Kim et al. proposed a method to generate avatar placement and deictic gestures in VR space, which enhanced engagement and social presence [27]. Schlagowski et al. demonstrated that projecting mutual presence through HMDs enhanced the sense of co-presence during jam sessions [28]. However, while these HMD-based approaches can enhance presence, a limitation remains in that the improvement of the remote participant's presence is restricted to the HMD-wearer.

Moreover, telepresence robots are another form of telepresence system. Telepresence robots have attracted considerable research and development due to their ability to recreate the presence of remote participants without the need for a human actor [29, 30]. Rae et al. investigated the relationship between control and trust in telepresence robots during collaborative tasks, as well as the impact of robot mobility on user presence and task performance [31, 11]. Lee et al. demonstrated, through a long-term study, that using a mobile telepresence robot allowed remote workers to experience living and working in the same space as their local colleagues [32]. Liu et al. combined telepresence robots with the chameleon effect [33] to enhance the user experience [34]. Sakashita et al. proposed a method to recreate the neck movements of remote participants with a telepresence robot, which not only enhanced the sense of presence but also facilitated shared attention with local persons [10]. As such, telepresence robots have garnered significant interest from researchers as a technology with the potential to improve the quality of remote meetings.

Building on the existing research on telepresence systems, this study adopts an approach that projects remote participants onto a dummy head, proposing a telepresence system that does not rely on human actors and is non-wearable. Unlike traditional systems limited to one-on-one conversations, this approach can also be applied to one-to-many interactions, thereby expanding its potential for use in a wider range of scenarios.

3. Proposed Method

3.1. Overview

This study aims to enhance the presence of remote participants in remote meetings by proposing a novel telepresence system that combines projection onto a dummy head using a projector with low-latency audio communication. The proposed system is illustrated in Figure 1.

It has been suggested that displaying the remote participant's face in 3D can facilitate the interpretation of non-verbal cues [12]. Therefore, this study projects the remote participant's face onto a dummy head, representing the remote participant's face in 3D. This approach is expected to enhance the

presence of remote participants compared to traditional video conferencing systems using 2D displays, thereby improving both the sense of immersion and the smoothness of interaction. Additionally, by using a high-quality audio communication application with minimal transmission latency, near-real-time conversation becomes possible.

The proposed system consists of the following components. First, the face image of the remote participant is directly projected onto a dummy head using a projector, enhancing the visual sense of immersion and presence. Second, a binaural microphone is used to collect and transmit the local sound information as immersive stereo signals, enabling natural sound field reproduction on the remote side. Additionally, by combining this system with a low-latency audio transmission application, smooth communication can be achieved while maintaining real-time interaction. By integrating visual enhancement through 3D projection, real-time communication, and the transmission of high-quality audio information, the system aims to provide superior presence and communication efficiency in interactions with remote participants, surpassing conventional methods.

3.2. Projection

In this system, a laptop and a projector are connected, and the face image of the remote participant is projected onto a dummy head. A laser-type projector with excellent brightness and resolution is used for the projection equipment. This allows for enhanced visibility of the image on the surface of the dummy head. Furthermore, the direction of the projector's light is focused solely on the dummy head, and the configuration ensures that the light does not directly shine on the local persons, thus reducing unwanted glare and interference for the local persons.

The remote participants join the video conference using the Apple Vision Pro (Apple Inc.). The Apple Vision Pro has the capability to scan the wearer's face and transmit a generated persona, allowing the wearer to communicate their face and expressions even while wearing a Head-Mounted Display. Since the persona accurately replicates the wearer's face and expressions, it enables communication that feels natural even from a remote location.

3.3. Acoustic System

To enable remote participants to immerse themselves in the local environment, this system adopts an audio collection and transmission method using a binaural microphone. The binaural microphone used is equipped with artificial ears that replicate the human ear, allowing it to capture audio information that closely resembles human auditory characteristics. This enables the transmission of spatial information, including the left-right sound pressure differences and head-related transfer functions (HRTF) generated in the local environment, allowing remote participants to experience a heightened sense of presence. The binaural microphone is connected to a laptop via an audio interface.

The remote participants engage in conversation with local persons, similar to conventional video conferencing systems. By using stereo headphones, the remote participants can listen to the local environment's audio captured by the binaural microphone.

Additionally, to achieve real-time interaction, this system introduces a low-latency audio transmission application separate from video transmission. By using SYNCROOM (Yamaha Inc.) for audio transmission, the system minimizes audio communication latency and reduces the time lag in bidirectional communication with remote participants. In this way, by combining spatial audio via the binaural microphone with low-latency audio transmission, the system creates an environment where remote participants can experience a heightened sense of presence, similar to being on-site.

4. Deployment

In this study, the latency performance of video calls during the system's deployment phase was evaluated by measuring the latency time for several applications. The video conferencing applications evaluated included Zoom with the Musicians' original sound (hereafter, Zoom (Musician)) and Zoom

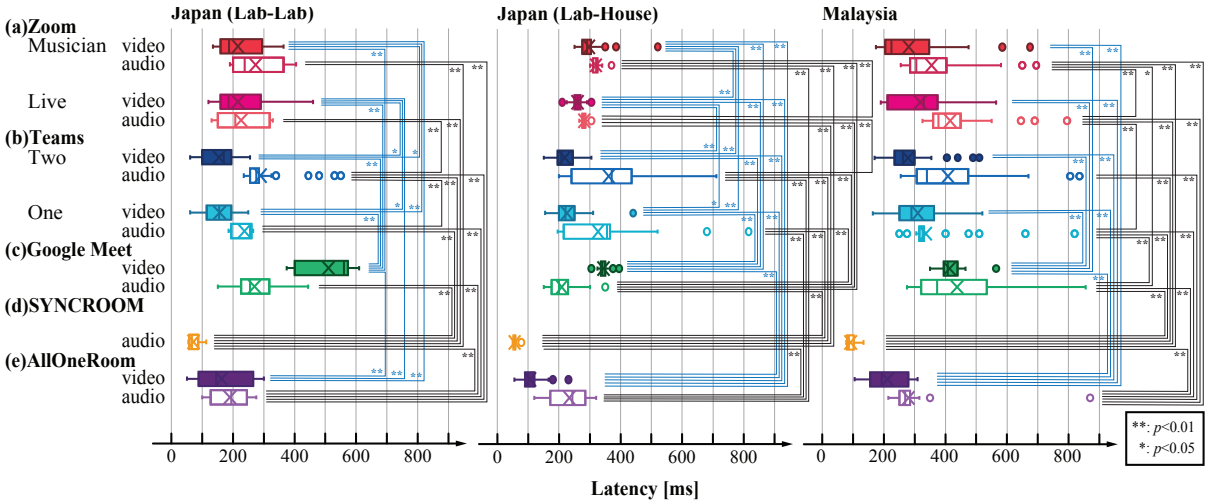

Figure 2: Results of the latency measurement. The comparison was conducted across three connection locations, from left to right: Japan (Lab) - Japan (Lab), Japan (Lab) - Japan (House), and Japan (Lab) - Malaysia. In this study, the analysis was divided into video latency and audio latency.

with the Live Performance Audio (hereafter, Zoom (Live)), Teams with high-quality music mode for both local and remote participants (hereafter, Teams (Two)) and Teams with high-quality music mode for remote participants only (hereafter, Teams (One)), Google Meet, SYNCROOM (audio-only), and the AllOneRoom system developed by the Virtual Reality Laboratory in University of Tsukuba using the WebRTC Sora SFU communication API (hereafter referred to as AllOneRoom) [35]. It should be noted that in Zoom, enabling both Musicians' original sound and Live Performance Audio for both local and remote participants simultaneously causes feedback, so for the experiment, the settings were only enabled for the remote participants. Additionally, since SYNCROOM is a system dedicated to audio calls, video latency evaluation was not included in the analysis.

The latency for video was measured by connecting the local and remote locations through a video conferencing system, where the local side shared a stopwatch it started with the remote side. The time taken for round-trip communication was then measured. The remote participant used the virtual camera function of OBS Studio to re-broadcast the received video, displaying the same stopwatch screen on the video conferencing system. These interactions were recorded, and the timestamp of the stopwatch values were compared. The video latency was calculated by dividing the round-trip communication latency by two. For audio latency, the audio sent from the local side was looped back using a virtual microphone on the remote side and sent back to the local side. These interactions were recorded, and the time difference between feature points on the waveform was measured using the audio analysis software Audacity. The audio latency was calculated by dividing the round-trip communication time by two. This measurement was conducted three times for each application, with 20 samples taken per trial, resulting in a total of 60 samples being collected.

In this study, the latency measurements were conducted by dividing the connection locations into three categories: Japan (Lab) - Japan (Lab), Japan (Lab) - Japan (House), and Japan (Lab) - Malaysia. The network speeds at each location (download/upload [Mbps]) were as follows: 375.73 Mbps / 526.41 Mbps for Japan (Lab) - Japan (Lab), 62.04 Mbps / 63.2 Mbps for Japan (Lab) - Japan (House), and 46.94 Mbps / 45.84 Mbps for Japan (Lab) - Malaysia.

The latency measurement results for video calls are shown in Fig. 2. To analyze the results, the Shapiro-Wilk test was first applied to check the normality of the data. The null hypothesis of normality was rejected. As a result, the Kruskal-Wallis test was conducted, and significant differences were observed at all locations ($p < 0.01$). Therefore, as a post-hoc test, Dunn's test with Bonferroni correction was performed.

At Japan (Lab) to Japan (Lab) connection, for video latency, significant differences were observed

between Google Meet and all other tools, as well as between Zoom (Musician) and Teams (One), Zoom (Musician) and AllOneRoom, and Zoom (Live) and AllOneRoom ($p < 0.01$). Additionally, significant differences were observed between Teams (Two) and Zoom (Musician), Teams (Two) and Zoom (Live), and Teams (One) and Zoom (Live) ($p < 0.05$). For audio latency, significant differences were observed between SYNCROOM and all other tools, as well as between AllOneRoom and all tools except Zoom (Live), Teams (Two) and Zoom (Live), and Teams (Two) and Teams (One) ($p < 0.01$).

At Japan (Lab) to Japan (House) connection, for video latency, significant differences were found between Google Meet and all other tools, AllOneRoom and all other tools, Zoom (Musician) and all other tools, and Teams (Two) and Zoom (Live) ($p < 0.01$). Furthermore, significant differences were observed between Teams (One) and Zoom (Live) ($p < 0.05$). For audio latency, significant differences were found between SYNCROOM and all other tools, all tools except Google Meet and AllOneRoom, Teams (Two) and Zoom (Live), Teams (Two) and AllOneRoom, Teams (One) and AllOneRoom, Zoom (Musician) and Zoom (Live), and Zoom (Musician) and AllOneRoom ($p < 0.01$).

At Japan (Lab) to Malaysia connection, for video latency, significant differences were observed between Google Meet and all other tools, and between AllOneRoom and all other tools ($p < 0.01$). For audio latency, significant differences were observed between SYNCROOM and all other tools, AllOneRoom and all other tools, Teams (One) and Google Meet, Teams (One) and Zoom (Live), and Zoom (Musician) and Zoom (Live) ($p < 0.01$). Furthermore, significant differences were observed between Google Meet and Zoom (Musician) ($p < 0.05$).

5. Evaluation

5.1. Set up

This study investigated whether the human-like presence of remote participants can be enhanced by projecting their faces onto a dummy head, compared to traditional video conferencing systems using 2D displays. In this evaluation, the human-like presence was assessed by recording the interactions and analyzing the videos using object recognition technology. The setup of the evaluation conducted in this study is shown in Fig. 3.

In this study, three conditions were established for evaluating the human-like presence: a projection system using a dummy head (the proposed system), a system using a 2D display (the traditional system), and a condition where a human participant is physically present (the face-to-face condition). The effectiveness of the proposed projection system was verified by comparing these three conditions. Additionally, to ensure fairness in the experimental conditions, the projection surface and the area below the neck of the human figure were covered with dark curtains in all conditions, preventing any external information from influencing the object recognition results. This measure minimized the impact of elements such as mechanical parts or the human body on the object recognition outcomes. Furthermore, rather than using fixed-point shooting, the shooting was conducted with a 180-degree range of motion around the projection target. This approach aimed to reduce biases caused by fixed locations or perspectives, thereby enhancing the reliability of the evaluation by dispersing the shooting conditions.

In this evaluation, recorded videos were used as the projected images. This choice was made to reduce the variability in results caused by differences in facial movements and expressions, thereby minimizing biases related to the projected content. Since using recorded videos was not feasible in the face-to-face condition, the expressions of the person set up in the condition were minimized to reduce variability between conditions. Additionally, the evaluation considered not only differences in the projection surface but also variations in the type of projected video. The two types of projected videos were: a recording using the Apple Vision Pro's Persona and a recording captured using a typical video conferencing tool. The impact of different combinations of projection systems and video types on human-like presence was evaluated. Therefore, the experimental conditions consisted of five scenarios: two types of videos in the dummy head-based projection system, two types of videos in the 2D display system, and the face-to-face condition. In all conditions, the projected person, speech content, and facial

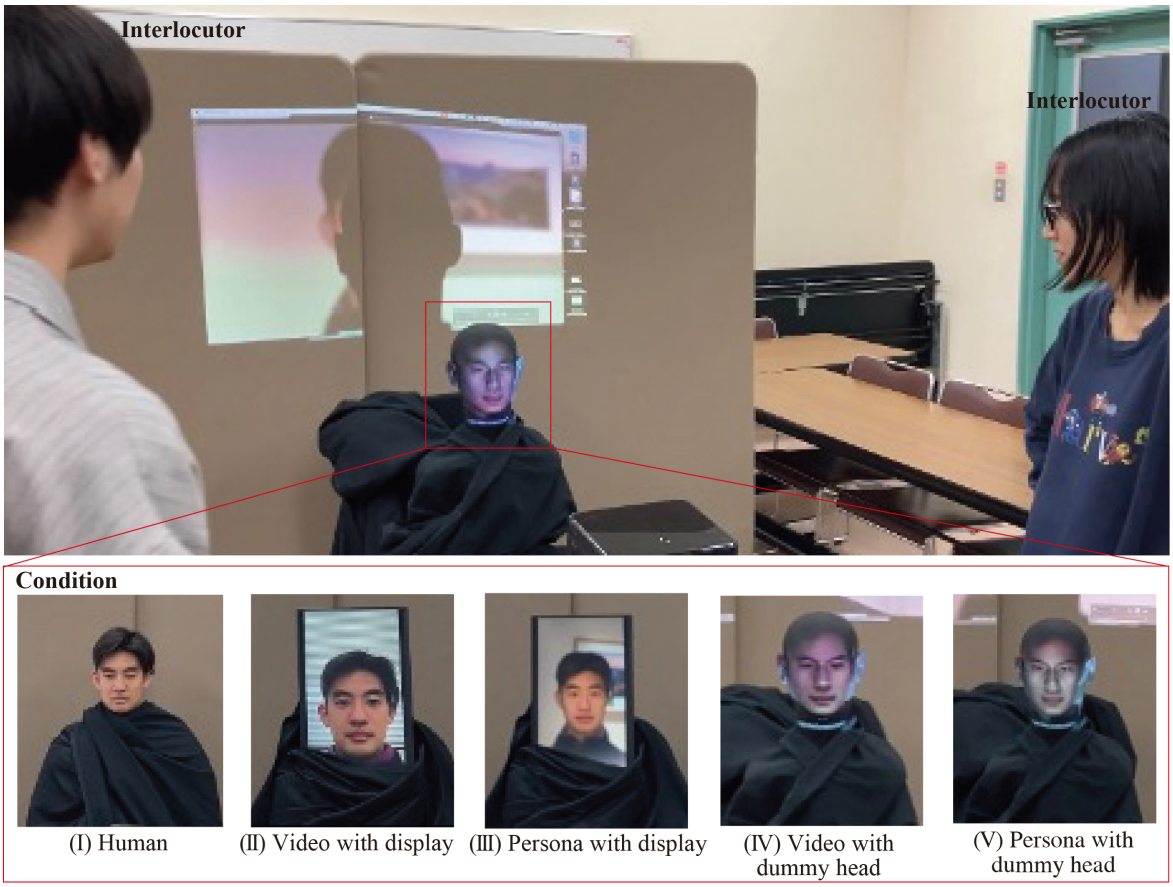

Figure 3: Observation of interactions recorded to evaluate the system. Five conditions, formed by different combinations of the projection surface and projected video, were recorded and analyzed using object recognition to assess the system's sense of presence. The recording was conducted while moving around, rather than from a fixed point.

expressions were standardized to minimize any bias arising from content or expression differences. The interaction time for each trial was approximately 160 seconds. To record the actual interaction, two actors were assigned to interact with the projected person. These actors interacted only with the projected person and were not involved in subjective assessments through surveys or interviews.

Object recognition analysis in this study utilized the YOLO11 detector. For the object detection models, three different versions of the pre-trained YOLO11 model were selected, each with varying speed and accuracy. The chosen versions were the high-speed, low-accuracy YOLO11n, the mid-speed, mid-accuracy YOLO11m, and the low-speed, high-accuracy YOLO11x. By using models with different detection accuracies, the study aimed to investigate the factors contributing to the variations in the detection results. Therefore, the analysis of five conditions in this study was conducted using different object detection models.

5.2. Results

This study evaluated the impact of different projection surfaces on the human-like presence of remote participants by recording interactions with the projected subject and using object recognition technology for assessment. In the object recognition process, still images were extracted at one-second intervals from the recorded videos for each condition. As a result, the total number of still images was 159 samples.

The results of the object recognition were categorized into four patterns: when the projection surface was recognized as "Person", when it was recognized as both "Person" and "Display" with a higher

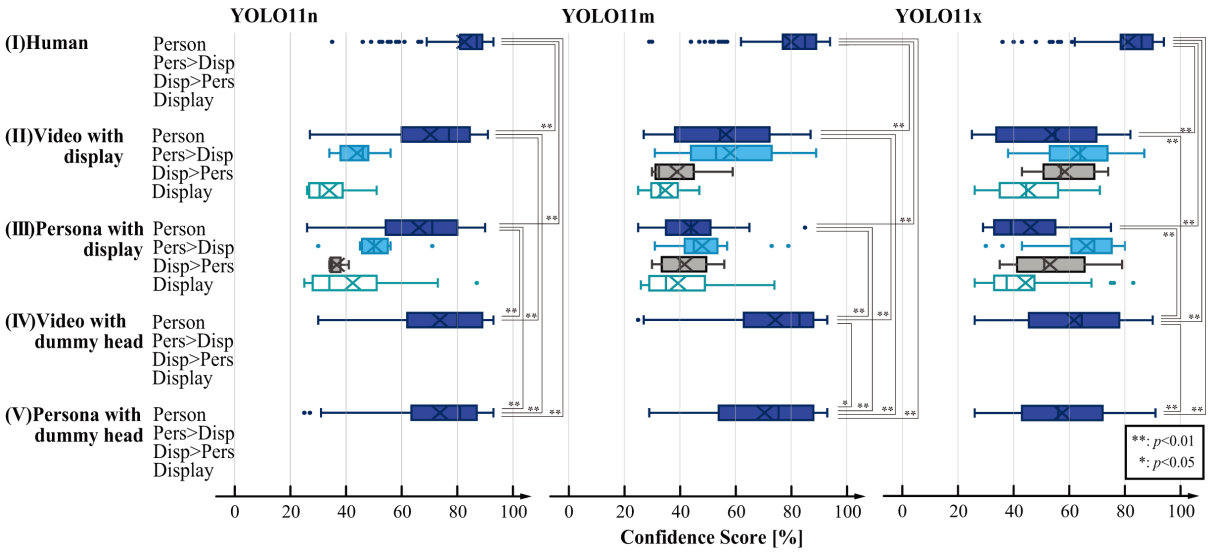

Figure 4: Results of object recognition. The effectiveness of the system was evaluated using three different object recognition tools. "Person" indicates that only a person was detected, while "Pers > Disp" signifies that both a person and a display were detected but the confidence score for the person was higher. "Disp > Pers" similarly indicates that both were detected, yet the display's confidence score was higher, and "Display" indicates that only a display was detected.

confidence score for "Person", when it was recognized as both "Person" and "Display" with a higher confidence score for "Display", and when it was recognized as "Display". Additionally, during object recognition, the projection surface was occasionally recognized as a TV or Laptop, and these were grouped together under the category "Display". The confidence scores for object recognition obtained in this experiment are shown in Fig. 4.

In this study, the data analysis focused on the confidence score when the projection surface was recognized as "Person" during object recognition of the recorded interaction. This focus was intended to promote the evaluation that the remote participant appears to be present in the local environment by recognizing the projection surface as a person.

data analysis was conducted using three different object detection models for five conditions. To check the distribution characteristics of the confidence scores, the Shapiro-Wilk test was first performed to assess the normality of the data. As a result, the null hypothesis of normality was rejected for all conditions. Consequently, the Kruskal-Wallis test, which is suitable for non-normal distributions, was applied to the data. Significant differences were observed for all models ($p < 0.01$). Therefore, as a post-hoc analysis, Dunn's test with Bonferroni correction was conducted. When using the YOLO11n object detector, significant differences were found between all conditions, except for the face-to-face condition and the condition where a video recorded using a video conferencing system was projected onto the dummy head. Significant differences were also found between the condition where Persona video was projected onto the dummy head and all conditions where video was projected onto a 2D display, and between the condition where a video recorded using a video conferencing system was projected onto the dummy head and all conditions where video was projected onto a 2D display ($p < 0.01$). When using the YOLO11m object detector, significant differences were observed between all conditions except for the face-to-face condition and the condition where a video recorded using a video conferencing system was projected onto the dummy head. Significant differences were also found between the condition where a video recorded using a video conferencing system was projected onto the dummy head and all conditions where video was projected onto a 2D display, and between the condition where Persona video was projected onto the dummy head and all conditions where video was projected onto a 2D display ($p < 0.01$). Additionally, significant differences were observed between the condition where a video recorded using a video conferencing system was projected onto the dummy head and

the condition where Persona video was projected onto the dummy head ($p < 0.05$). When using the YOLO11x object detector, significant differences were observed between the face-to-face condition and all other conditions, as well as between the condition where a video recorded using a video conferencing system was projected onto the dummy head and all other conditions ($p < 0.01$).

6. Field work and Discussion

In this study, a fieldwork was conducted by connecting Japan and Malaysia using the proposed telepresence system, through which interactions were facilitated. The remote participants wore the Apple Vision Pro and participated in the meeting via the system. The goal of this fieldwork was to observe how remote participants interact with local persons and to identify the challenges and advantages of using the system in practical applications. The setup of the fieldwork conducted in this study is shown in Fig. 5. During the fieldwork, a 360-degree camera was connected to the system to broadcast the local environment to the remote participants.

As a result of installing the system in a public space, many local persons showed interest and actively enjoyed conversations with the remote participant projected onto the dummy head. The local persons were engaging in conversation while looking at the eyes of the remote participant displayed on the dummy head, and it was frequently observed that this provided an immersive experience that was not achievable with a 2D display. To further investigate this, a laptop displaying the same image was placed next to the system to see if local persons would interact with it. Despite the laptop being available, local persons continued to interact primarily with the system, showing little attention to the laptop. This result suggests that interactions accompanied by 3D presence, as provided by the proposed system, may be preferred over those using a 2D display. Moreover, it is known that in remote meetings, local persons tend to value the presence of remote participants [36], and this study suggests that the proposed system could facilitate advanced interactions, including non-verbal cues. This was one of the key findings from the fieldwork conducted. Furthermore, existing robotic systems that project video onto 3D face models, similar to the proposed system, have reported cases of negative reactions from elderly participants in studies [37]. However, in this study, which targeted a diverse group of people rather than just the elderly, the reactions from local persons were generally positive. This may indicate that the dynamic interaction with an actual remote participant reduced the robotic elements that are often associated with the uncanny valley effect, leading to a more natural acceptance.

Additionally, by combining low-latency audio communication using SYNCROOM and a binaural microphone, this system enables remote participants to accurately perceive the local sound environment and the position of the speaker. This setup has been confirmed to create conditions where remote participants can more easily experience the local environment's sense of presence through sound source localization. The results of the fieldwork suggest that remote participants were able to determine who was speaking from which position, alleviating the lack of spatial information in remote meetings. Furthermore, the use of low-latency audio communication ensured that the conversation proceeded smoothly without disrupting the rhythm of the dialogue. These findings suggest that the system has the potential to address the spatial and temporal delays in remote video communication, which were identified as challenges in previous studies [20, 14].

7. Limitation

This study has several limitations that need to be considered when interpreting the results. First, in the measurement of communication latency, it was difficult to completely eliminate the inherent network jitter. In this study, the sample size was set to 60, and the impact of jitter was partially mitigated by introducing variance, but it was not fully eliminated. Therefore, more detailed studies focusing on latency performance should consider increasing the sample size over a longer period and explore further experimental designs. Additionally, in the analysis of latency, there is the potential for mechanical and human errors. Consequently, the results must be interpreted while considering these factors. Moreover,

Figure 5: General visitors enjoying interactions with the system during the exhibition. The projector is connected to a laptop, the laptop shown in the photo displays the exact same facial visuals as those projected by the projector. The local visuals were captured using a 360-degree camera and transmitted to participants in remote locations.

the measurements in this study were conducted in December 2024. It is expected that the performance of video conferencing applications will change over time due to system updates and other factors. Therefore, the latency results obtained in this study should be considered with the understanding that they may vary as the systems evolve. Furthermore, in the measurement of audio latency, a loopback method was used, and conditions were selected that allowed for relatively high-quality audio extraction. As a result, there are audio conditions that were not compared in this study, and these conditions should be further examined in future research.

In addition, numerous studies have utilized questionnaire surveys as a method for evaluating the sense of presence [17, 13] and anthropomorphism [38, 39]. However, since this study was a preliminary investigation to examine whether projecting the remote participant's face onto a dummy head affects the perception of presence, no surveys involving users were conducted. In the future, it will be necessary to quantitatively verify the effectiveness of the system in enhancing presence through surveys or other means involving users.

8. Conclusion

In this study, we proposed a telepresence system that projects a remote participant's face onto a dummy head and integrates a low-latency audio transmission application to enhance the human-like presence of remote participants. We evaluated the human-like presence by conducting machine-based object recognition. As a result, our approach was found to provide a more human-like telepresence system compared with a conventional two-dimensional display. Furthermore, in an international field study connecting Japan and Malaysia, we confirmed that local and remote participants could enjoy seamless interactions using the proposed method.

The telepresence approach presented in this study opens up new possibilities for achieving low-latency, immersive remote communication. By delivering value beyond traditional video conferencing systems, this method holds promise for novel approaches to remote collaborative tasks and meetings. It enables more effective and immersive interaction, and could potentially be extended in the future to large-scale telepresence systems for multiple simultaneous users.

Acknowledgments

This work was supported by JST, PRESTO Grant Number JPMJPR2269, Japan. It was also supported by JST SPRING, Grant Number JPMJSP2124, Japan.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Google Gemini in order to: Grammar and spelling check, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] K. Misawa, J. Rekimoto, Wearing another's personality: A human-surrogate system with a telepresence face, in: Proceedings of the 2015 ACM International Symposium on Wearable Computers, 2015, pp. 125–132.

[2] K. Misawa, J. Rekimoto, Chameleonmask: a human-surrogate system with a telepresence face, in: SIGGRAPH Asia 2015 Emerging Technologies, 2015, pp. 1–3.

[3] H. Tobita, S. Numanoi, Inteach: Enhanced personal e-learning with tabletop telepresence and real-world objects, in: Proceedings of the Seventh International Conference on the Internet of Things, 2017, pp. 1–8.

[4] M. Minsky, Telepresence (1980).

[5] R. Held, Telepresence, The Journal of the Acoustical Society of America 92 (1992) 2458–2458.

[6] J. V. Draper, D. B. Kaber, J. M. Usher, Telepresence, Human factors 40 (1998) 354–375.

[7] K. Youssef, S. Said, S. Al Kork, T. Beyrouthy, Telepresence in the recent literature with a focus on robotic platforms, applications and challenges, Robotics 12 (2023) 111.

[8] N. Yankelovich, N. Simpson, J. Kaplan, J. Provino, Porta-person: Telepresence for the connected conference room, in: CHI'07 extended abstracts on Human factors in computing systems, 2007, pp. 2789–2794.

[9] D. Sirkin, G. Venolia, J. Tang, G. Robertson, T. Kim, K. Inkpen, M. Sedlins, B. Lee, M. Sinclair, Motion and attention in a kinetic videoconferencing proxy, in: Human-Computer Interaction–INTERACT 2011: 13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part I 13, Springer, 2011, pp. 162–180.

[10] M. Sakashita, R. Zhang, X. Li, H. Kim, M. Russo, C. Zhang, M. F. Jung, F. Guimbretière, Remotion: Supporting remote collaboration in open space with automatic robotic embodiment, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, 2023, pp. 1–14.

[11] I. Rae, B. Mutlu, L. Takayama, Bodies in motion: mobility, presence, and task awareness in telepresence, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2014, pp. 2153–2162.

[12] A. Jones, M. Lang, G. Fyffe, X. Yu, J. Busch, I. McDowall, M. Bolas, P. Debevec, Achieving eye contact in a one-to-many 3d video teleconferencing system, ACM Transactions on Graphics (TOG) 28 (2009) 1–8.

[13] K. Kim, J. Bolton, A. Girouard, J. Cooperstock, R. Vertegaal, Telehuman: effects of 3d perspective on gaze and pose estimation with a life-size cylindrical telepresence pod, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2012, pp. 2531–2540.

[14] B. Saatçi, K. Akyüz, S. Rintel, C. N. Klokmose, (re) configuring hybrid meetings: Moving from user-centered design to meeting-centered design, Computer Supported Cooperative Work (CSCW) 29 (2020) 769–794.

[15] N. D. Bos, A. Buyuktur, J. S. Olson, G. M. Olson, A. Voida, Shared identity helps partially distributed teams, but distance still matters, in: Proceedings of the 2010 ACM International Conference on Supporting Group Work, 2010, pp. 89–96.

[16] N. Bos, N. S. Shami, J. S. Olson, A. Cheshin, N. Nan, In-group/out-group effects in distributed teams: an experimental simulation, in: Proceedings of the 2004 ACM conference on Computer supported cooperative work, 2004, pp. 429–436.
[17] T. Pejsa, J. Kantor, H. Benko, E. Ofek, A. Wilson, Room2room: Enabling life-size telepresence in a projected augmented reality environment, in: Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing, 2016, pp. 1716–1725.
[18] G. Zhou, A. Nagle, G. Takahashi, T. Hornbeck, A. Loomis, B. Smith, B. Duerstock, D. Yu, Bringing patient mannequins to life: 3d projection enhances nursing simulation, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–15.
[19] S. Uhrig, T. Michael, S. Möller, P. E. Keller, J.-N. Voigt-Antons, Effects of delay on perceived quality, behavior and oscillatory brain activity in dyadic telephone conversations, in: 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX), IEEE, 2018, pp. 1–6.
[20] C.-L. Yang, X. Li, T. Narumi, H. Kuzuoka, Understanding the impact of technical issues on people's perception and attribution of responsibility in videoconferencing, in: CHI Conference on Human Factors in Computing Systems Extended Abstracts, 2022, pp. 1–6.
[21] A. Armstead, R. Henning, Effects of long audio communication delays on team performance, Proceedings of the Human Factors and Ergonomics Society Annual Meeting 51 (2007) 136–140. doi:10.1177/154193120705100304.
[22] M. Faridan, B. Kumari, R. Suzuki, Chameleoncontrol: Teleoperating real human surrogates through mixed reality gestural guidance for remote hands-on classrooms, in: Proceedings of the 2023 CHI conference on human factors in computing systems, 2023, pp. 1–13.
[23] S. Beck, A. Kunert, A. Kulik, B. Froehlich, Immersive group-to-group telepresence, IEEE transactions on visualization and computer graphics 19 (2013) 616–625.
[24] D. Gotsch, X. Zhang, T. Merritt, R. Vertegaal, Telehuman2: A cylindrical light field teleconferencing system for life-size 3d human telepresence., in: CHI, volume 18, 2018, p. 552.
[25] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, et al., Holoportation: Virtual 3d teleportation in real-time, in: Proceedings of the 29th annual symposium on user interface software and technology, 2016, pp. 741–754.
[26] T. Piumsomboon, G. A. Lee, J. D. Hart, B. Ens, R. W. Lindeman, B. H. Thomas, 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, 2018, pp. 1–13.
[27] M. Kim, S.-H. Lee, Deictic gesture retargeting for telepresence avatars in dissimilar object and user arrangements, in: Proceedings of the 25th International Conference on 3D Web Technology, 2020, pp. 1–6.
[28] R. Schlagowski, D. Nazarenko, Y. Can, K. Gupta, S. Mertes, M. Billinghurst, E. André, Wish you were here: Mental and physiological effects of remote music collaboration in mixed reality, in: Proceedings of the 2023 CHI conference on human factors in computing systems, 2023, pp. 1–16.
[29] A. Kristoffersson, S. Coradeschi, A. Loutfi, A review of mobile robotic telepresence, Advances in Human-Computer Interaction 2013 (2013) 902316.
[30] G. Zhang, J. P. Hansen, Telepresence robots for people with special needs: a systematic review, International Journal of Human–Computer Interaction 38 (2022) 1651–1667.
[31] I. Rae, L. Takayama, B. Mutlu, In-body experiences: embodiment, control, and trust in robot-mediated communication, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2013, pp. 1921–1930.
[32] M. K. Lee, L. Takayama, " now, i have a body" uses and social norms for mobile remote presence in the workplace, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2011, pp. 33–42.
[33] T. L. Chartrand, J. A. Bargh, The chameleon effect: The perception–behavior link and social interaction., Journal of personality and social psychology 76 (1999) 893.
[34] Z. Liu, M. Imai, Telepresence chameleon: Improve user experience of telepresence robot with chameleon effect, in: Proceedings of the 11th International Conference on Human-Agent Interac-

tion, 2023, pp. 55–62.
[35] University of Tsukuba, Virtual Reality Laboratory, Alloneroom system, 2025. URL: https://alloneroom.com/.
[36] C. N. Gunawardena, F. J. Zittle, Social presence as a predictor of satisfaction within a computer-mediated conferencing environment, American Journal of Distance Education 11 (1997) 8–26. URL: https://doi.org/10.1080/08923649709526970. doi:10.1080/08923649709526970. arXiv:https://doi.org/10.1080/08923649709526970.
[37] S. Thunberg, M. Arnelid, T. Ziemke, Older adults' perception of the furhat robot, in: Proceedings of the 10th International Conference on Human-Agent Interaction, 2022, pp. 4–12.
[38] C. Bartneck, D. Kulić, E. Croft, S. Zoghbi, Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots, International journal of social robotics 1 (2009) 71–81.
[39] N. Yun, S. Yamada, Investigation of factors that influence human presence and robot anthropomorphism in telepresence robot, IEEE Access (2024).