A User Study on Mixed Reality Remote Collaboration with Eye Gaze and Hand Gesture Sharing

Huidong Bai[1], **Prasanth Sasikumar**[1], **Jing Yang**[2], **Mark Billinghurst**[1]

[1]Auckland Bioengineering Institute, The University of Auckland, Auckland, New Zealand
[2]Department of Computer Science, ETH Zurich, Zurich, Switzerland

ABSTRACT

Supporting natural communication cues is critical for people to work together remotely and face-to-face. In this paper we present a Mixed Reality (MR) remote collaboration system that enables a local worker to share a live 3D panorama of his/her surroundings with a remote expert. The remote expert can also share task instructions back to the local worker using visual cues in addition to verbal communication. We conducted a user study to investigate how sharing augmented gaze and gesture cues from the remote expert to the local worker could affect the overall collaboration performance and user experience. We found that by combing gaze and gesture cues, our remote collaboration system could provide a significantly stronger sense of co-presence for both the local and remote users than using the gaze cue alone. The combined cues were also rated significantly higher than the gaze in terms of ease of conveying spatial actions.

Author Keywords

Mixed Reality; Augmented Reality; Virtual Reality; remote collaboration; 3D panorama; scene reconstruction; eye gaze; hand gesture.

CCS Concepts

•**Human-centered computing** → **Mixed / augmented reality; Collaborative interaction;** *Computer supported cooperative work;*

INTRODUCTION

A wide variety of communication cues are used in face-to-face collaboration, such as audio (e.g., speech, paralinguistic, intonation), visual (e.g., gaze, gesture, facial expression, body posture), and environmental information (e.g., object manipulation, writing, drawing, spatial layout). These different cues are combined to create more efficient communication and better mutual understanding between remote collaborators. Advances in telecommunication technology have enabled rapid developments in methods for remote collaboration, such as real-time audio and video streaming, or mobile conferencing. However, most of these technologies cannot convey all of the same communication cues as those which are presented in face-to-face collaboration. For example, on a video conferencing link, some subtleties of the gaze or gestures might be lost. It could also be tricky to share the same spatial cues present in the local conversation or to share environmental information around.

The use of Head-Mounted Displays (HMDs) with Mixed Reality (MR) technology creates the possibility for a more intuitive and immersive collaborative experience than with conventional 2D video-based systems. For example, by capturing and streaming 360° video of the local worker's view into a Virtual Reality (VR) scene viewed by a remote expert, the remote expert can feel like he/she is sharing the local user's workspace, being able to inspect the local environment in a 360 view [29]. Similarly, Augmented Reality (AR) technology enables the remote expert to overlay virtual content onto the local worker's view, such as showing 3D virtual annotations on top of real objects to demonstrate how to manipulate them [22].

We would like to study remote collaboration with natural communication cues in dynamically changing room-scale environments, such as remote maintenance of large machines or control rooms, crime scene forensics, emergency response, and remote teaching of dance or acting performance, etc. There are many diverse application areas where this could be valuable. MR remote collaboration has often been studied before from two perspectives: 1) Capturing more dimensional information about the local scene and building an unconstrained viewpoint for the remote expert [27]; 2) Adding and improving communication cues exchanged between local and remote users for more efficient and easier collaboration [32].

In this paper, we present a novel MR remote collaboration system that supports live streaming of an immersive 3D view of the local worker's environment at room-scale. It also supports live sharing of the remote user's eye gaze and hand gestures back to the local user to convey spatial task instructions. This system combines the advantages of capture technologies and communication interfaces mentioned above while minimizing their limitations. Compared to prior work, the primary novel contributions of this paper include: 1) A MR remote collaboration system that enables sharing of hand gesture and eye gaze communication cues within a live 3D panorama; 2) A formal user study that compares hand gestures and eye gaze as visual cues (standalone and combined) with the conven-

CHI'20, April 25–30, 2020, Honolulu, HI, USA

 ISBN 978-1-4503-6708-0/20/04…$15.00

DOI: https://doi.org/10.1145/3313831.3376550

tional verbal communication in a live 3D panorama-based MR remote collaboration task.

In the rest of the paper, we review related work and compare our approach to this prior work, and then describe the design and implementation of our prototype system. We report on a full user study with the platform focusing on the usability of proposed MR collaboration cues. We finally discuss the results we have found, conclusions, and directions for future work.

RELATED WORK

In this section, we discuss two perspectives of related work in MR remote collaboration: 1) Local scene capturing and sharing for the remote expert; 2) Remote communication cues shared between the local worker and the remote expert. We study these two aspects with our novel system and experiment.

Local Scene Capturing and Sharing

Conventional remote collaboration systems share 2D video feeds from local to remote users to help them work together [2, 22, 23]. These systems mainly use a fixed-view camera, which limits the monitoring angle and operating volume for the remote user. To overcome this, researchers started to explore alternative approaches such as using head-mounted cameras [7, 22], hand-held cameras [3, 39], or cutting between shots of multiple cameras [14] to support dynamic views from different positions and poses. With this changeable point of view, the remote expert can observe the local workspace from more perspectives, which improves the understanding of the local worker's situation.

Many prior studies focused on evaluating the usefulness of sharing the first-person view with head-mounted cameras for remote collaboration [10, 15, 21]. These systems can automatically follow the local worker's actions and effectively reduce the remote expert's cognitive load. However, the captured view from the head-mounted camera is controlled by the local worker and only gives a limited view of the local workspace. Use of a static wide-angle camera [7] or even 360° cameras [20, 27, 28, 29, 38] have been studied to enlarge the captured area of the local environment. In this case, the remote expert can independently control their viewpoint, but cannot change their position in the local space. Some researchers have tried to overcome this limitation. For example, the Giant-Miniature Collaboration system [37] explored MR collaboration through a tracked tangible 360-camera interface, in which both dependent and independent viewing of the 360° video feeds were possible at multiple scales and positions.

Although 2D video sharing can provide a real-time high-resolution view for detailed inspection and manipulation, it does not provide depth information for the remote user, which makes it challenging to understand the layout of the captured scene and also decreases the sense of co-presence [9]. To overcome this limitation, depth sensors have been used to reconstruct 3D geometrical information of the local scene into a static mesh online [11] or offline [44] for the remote expert to view in a VR environment. Researchers have experimented with capturing the local workspace as a 3D geometrical model [1, 11, 39, 43, 44], providing independent third-person 6 Degrees-of-Freedom (DoF) viewpoint control, and increasing the remote expert's depth perception and spatial awareness of the local surroundings. Two possible ways are used to present the captured 3D geometrical model. The first is to display the 3D model on a traditional 2D screen, such as a computer monitor or a handheld device; The second is to use a HMD to show the scene in a more immersive VR environment. A study by Johnson et al. [19] showed that using a HMD is better for giving frequent instructions during dynamic tasks.

Due to hardware limitations, previous research on 3D scene capturing and sharing either restricted the local workspace to a small area to ensure real-time streaming, or reconstructed a large local workspace as a static 3D model without updates at all. For example, Gao et al. [9] combined a pre-recorded low-resolution 3D mesh of the user's surroundings with a wide field-of-view high-resolution 2D video feed streaming from the local to the remote user. This hybrid method can be used to solve the update restriction while offering depth information at the same time. Stotko et al. [42] recently presented the first practical client-server system for real-time capture and many-user exploration of large-scale static 3D scenes. Teo et al. [45] mixed 360° live video and 3D offline reconstruction to enhance collaborative search tasks in 360 with 3D geometrical information while providing updates in 2D, which created a higher sense of social presence in the collaboration without affecting performance.

Several systems have tried to reconstruct and share a large workspace as a dynamic 3D model for remote collaboration by merging data from multiple sensors [1, 4, 5]. These systems were mainly implemented in an outside-in structure, in which all sensors were deployed around the scene, facing towards a center target. For example, the Remote Fusion system [1] was one of the first remote guidance systems to support an independent 3D viewpoint for the remote expert. Multiple depth sensors were used to capture the local work environment, which was then rendered as a 3D model in the VR world and streamed to a remote site. Dou and Fuchs [5] also merged pre-scanned static parts of a room by tracking them online with live data from commodity depth cameras to achieve a noise-free and complete 3D capture of the room. Maimone and Fuchs [31] further developed a telepresence system offering room-sized, fully real-time volumetric scene capture, and continuous-viewpoint head-tracked display. However, these works mainly focused on sharing of the entire environment and user's full body without sharing accurate communication cues such as gaze or gesture data.

Remote Communication Cues

One type of remote collaboration aims to guide a local worker to complete a real-world task with help from a remote expert using various communication methods. The most common way is using speech to talk with each other.

However, there is a lot of research showing that visual cues provide further communication details for collaborators to ground their utterances and improve their performance [6]. Device-centric input like moving a cursor pointer or writing virtual annotations is commonly used to provide spatial cues for MR remote collaboration. For example, Fussell et al. [8]

showed that using virtual annotations could increase guidance efficiency to where it was nearly identical to working side by side. Kim et al. [21] compared using annotations for remote guiding tasks to a cursor pointer, and found that the pointer was the most preferred additional cue by users for a parallel experience where both users had the same information. However, drawing annotations over 2D videos requires a static camera or the ability to freeze the live video; Otherwise virtual annotations may lose their referents easily when the camera viewpoint is changed [16, 24, 26]. To solve this issue, some researchers have used Simultaneous Localization and Mapping (SLAM) tracking techniques to track and map the 3D movements of the camera [12, 13]. With this technology, world-stabilized annotations can be supported with an independent view enabled at the same time.

With the development of behavior detection technology like eye tracking or hand detection, natural user-centric cues, such as hand gesture, eye gaze, or body posture, can be used in collaborative MR interfaces to match the face-to-face collaboration experience. There are many examples of remote collaborative MR systems that support hand tracking and gesture sharing. Alem et al. [2] indicated that hand gestures were richer than using a cursor in terms of representation of rotation and orientation in collaborative tasks. Tecchia et al. [44] investigated providing gesture guidance in a 3D VR world to simulate a more natural guidance scenario and found that experts could reach places that they were not able to point to before. Researchers have also introduced the sharing of head pointing or eye gaze in remote collaboration [25, 32, 35, 40] or attached these behaviors onto avatars in a collaborative virtual environment [41] to enhance the understanding of the local worker's intentions. Gupta et al. [15] used eye gaze to show the local worker's real-time focus during tasks. They found that sharing of gaze cues could significantly improve task performance even without supporting cursor pointing from the remote expert. However, gaze transfer also resulted in longer task time [34], and the interpretation of communicative intention could be difficult [33]. Other types of communication cues, such as a virtual view frustum, can also be shared to provide accurate information as to where each user is looking [36].

Most of these studies were conducted with no support for room-scale 3D live scene scanning or no use of both gaze and hand gestures at the same time, leaving questions about the usability of these cues in a large scale collaborative experience. By mixing a 3D live panorama and natural visual cues in our system, the local worker is able to capture and share a reconstructed 3D virtual replica of their physical workspace with a remote expert. The remote expert can walk through the VR scene independently of the local user's view, while sharing their eye gaze and hand gestures back to the local worker for efficient communication. Being visualized using a virtual gaze line and a hand mesh, these cues can be augmented in the real scene through the local user's AR display to help them with real-world tasks.

SYSTEM OVERVIEW

Inspired by the 360° video camera, we developed a live 3D panorama sensor cluster that supports instant 3D reconstruction with real-time updating (Figure 2a). This enables the remote expert to watch the entire local user's environment in 3D and receive all actions and changes without delay. The cluster has a large scanning volume (a semi-sphere shaped space with a $10m$ radius) with a reasonably small cluster size. The remote expert's gaze and gestures can be detected by the VR headset's built-in eye tracker and external gesture sensor, and can be seamlessly shared with the local worker for task guiding and attention focus, offering rich natural communication cues.

Figure 1 shows an overview of our system. A local worker wears a see-through AR display (Magic Leap One) with a set of position-fixed depth-sensing units for mapping out their workspace. The remote expert wears a VR headset (HTC Vive Pro Eye) with a built-in eye tracker and an external gesture sensor (Leap Motion). We connect all devices to the same private network for fast data exchange, and then align both the AR and VR systems to the same shared virtual coordinate frame, the origin of which is located in the center of an image marker in the workspace. We use the image tracking in the Magic Leap One to detect the image marker in the AR physical world, and use the Vive Lighthouse tracker to get the same position in the VR mapped representation. So the gaze and gesture data from the remote VR system can be directly visualized in the local AR system without any further pose transformation needed. This allows both the local and remote users to feel that they are co-located in a shared MR space, the same as in face-to-face communication.

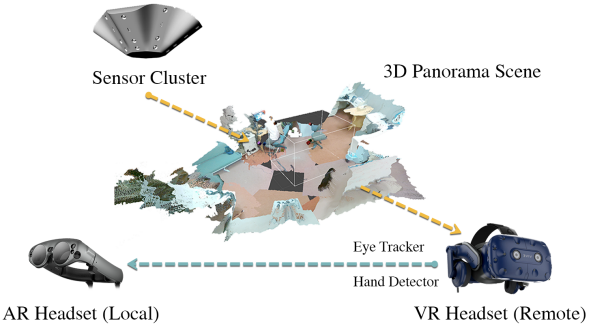

Figure 1. The 3D panorama unit reconstructs the local environment, and then streams the stitched point-cloud data to the remote VR expert via the network. The eye gaze and hand gesture information are shared back to the local AR worker from the remote VR expert synchronously.

In the following sections, we further explain the design and implementation of main features of our MR remote collaboration system: 1) The live volumetric fusion and sharing of point-cloud data from the depth sensor cluster; 2) The hand gesture and eye gaze guidance as natural visual cues; 3) The avatar and its pointing arrow for auxiliary awareness cues.

3D Live Panorama (Local to Remote)

To capture the local workspace, we assembled eight off-the-shelf RGB-Depth cameras (Intel RealSense D415) into a sensor cluster, as shown in Figure 2a. We designed a frame based on a field-of-view calculation, with which all point clouds from eight sensors can be correctly merged into a semi-sphere shape with no gaps left. To fuse all sensors accurately in space, we performed a one-time calibration using the Multiple-camera System Calibration Toolbox for MATLAB [30]. The calibrated result is shown in Figure 2b with the red frames. A single local computer retrieves all volumetric frames from the cluster via extended USB connections, stitches them in real time, and then streams the fused output to a remote VR computer via a 10Gb Ethernet connection. The final 3D data set is rendered as a dense panorama scene in the VR environment, which the remote user can freely navigate himself/herself through.

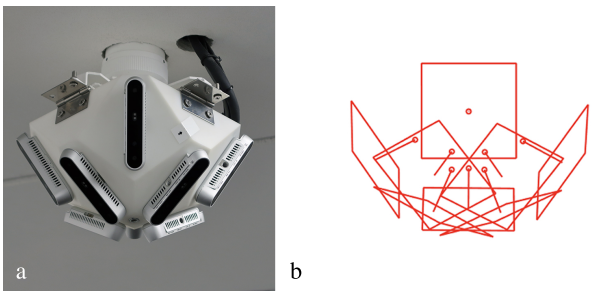

Figure 2. a) The sensor cluster with eight depth camera units; b) The calibration result of the sensor cluster.

Natural Visual Cues (Remote to Local)

We explored two types of natural communication cues as well as their combination (Figure 3) that could be shared from the remote VR mode to the local AR mode:

- *Eye Gaze* A virtual raycast line of the remote user's eye gaze overlaid onto the local user's AR view from a third-person perspective. This provides an explicit visual representation of the remote expert's gaze direction and what they are looking at in 3D space.
- *Hand Gesture* A virtual 3D mesh of the remote user's hand is overlaid onto the local user's view from a third-person perspective. This works as a proactive visual guide from the remote expert in the searching, placement, or rotation of target objects.
- *Eye Gaze and Hand Gesture* By combing two cues above, both the remote user's gaze raycast line and 3D hand mesh are augmented in the local AR scene, providing much richer interaction with the local user to help them with their tasks.

All visual cues from the remote user are rendered in both the AR and VR scenes with accurate depth perception and occlusion, which provides realistic physical interaction for better spatial understanding and more precise communication. For example, the 3D hand models from the remote user can be visually blocked by a real box in the local AR scene because of the spatial relationship and its real-time occlusion.

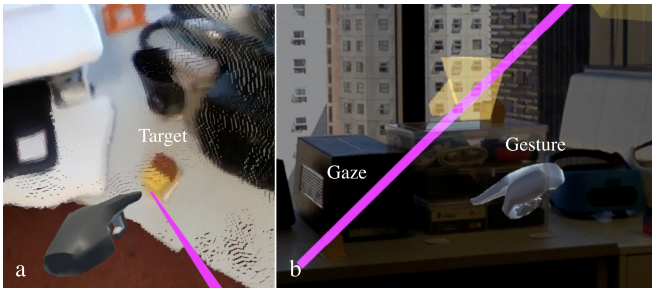

Figure 3. Natural visual cues shared from the remote to the local user: the purple gaze raycast line, and the grey hand mesh in a) the remote VR mode and b) the local AR mode.

Avatar and 360 Awareness Cues

Since our system shares a room-size 3D panorama scene, when the two users are located in different positions and looking in different directions, it is hard for one user to track the gaze or gesture cues of the other. To give both users a more unambiguous indication of their partner's location and viewing direction, we introduce a simple avatar made from a virtual head frustum and a 3D arrow cue that points to the location of the other user (Figure 4). These features are enabled on both sides, although the local worker can be watched through the shared panorama by the remote VR expert.

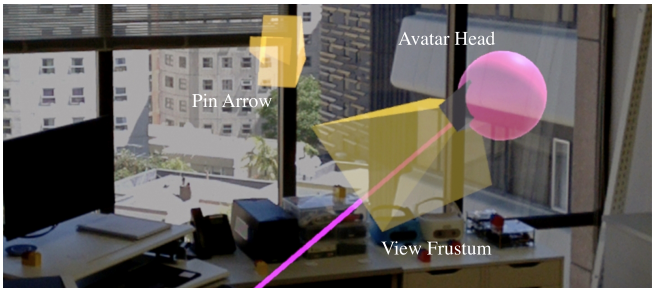

Figure 4. Auxiliary awareness cues for the local worker: the simple avatar has the purple sphere head and half-transparent yellow view frustum, and the pin arrow in orange always points to the head sphere.

- *Head Frustum* For both sides, the virtual view frustum represents the other user's head position and the general direction of where that user is looking in his/her workspace, as well as the field of view of the AR/VR headsets.
- *Pin Arrow* In case one of the user's avatars is out of view and cannot be seen by the partner, the system will always show a virtual 3D arrow pointing to where the other user's avatar head is located.

Implementation

The prototype system was built with AR and VR HMDs for the local worker and remote expert, respectively. The remote expert used the HTC VIVE Pro Eye, tethered to a desktop computer (Intel Core i7-8700 3.2GHz CPU with 6 Cores, 32GB DDR4 RAM, NVIDIA GeForce RTX 2070 GPU) running Microsoft Windows 10. A Leap Motion hand tracker was mounted on the front plate of the VR HMD for capturing the remote expert's hand motions that were then shared with the

local worker. The local worker used the Magic Leap One AR headset to watch the augmented cues. We used the Magic Leap One in Size 1 (the smaller fitting) and only recruited users with normal vision or corrected to normal vision with contact lenses. Some people could not participate in the study because their heads were too big or their eyes were too far apart for the Magic Leap build-in eye-tracking to work appropriately.

For capturing the live 3D panorama of the local workspace, eight Intel RealSense D415 cameras were connected to another desktop computer (the same hardware configurations as above) via a PCIe USB 3.0 expansion board and USB 3.0 extension cables. We used the native depth resolution of 848×480 of the D415 camera at 30 Frames Per Second (FPS) to reduce the computing workload and network delay while maintaining the smooth frame acquisition. We converted the color and depth frames into point-cloud data using the Intel RealSense library[1]. We then projected the spatial data of eight cameras into one unified virtual coordinate system with our calibration procedure. These mapped data could be stitched together to reconstruct the workspace as a dense live panorama. The average FPS of our system dropped to 20 FPS (at the depth resolution of 848×480) from native 30 FPS with this step.

The 3D panorama was then shared with the remote computer through the wired connection by using the Draco5 library[2] with socket APIs to combine encoding and decoding of point clouds into real-time streaming, which improves the storage and transmission of 3D geometric meshes and point-clouds. The data exchange delay between the two computers averages less than $300ms$ over our local area network (up to $10Gbps$). The AR and VR devices and all computers were wirelessly connected to the same local Wi-Fi network, so that the spatial information of all natural cues can be transmitted in two ways with no noticeable delay ($< 10ms$). Since the Magic Leap One and HTC Vive Pro Eye can track their pose in the environment with 6 DoF accuracy, the headset spatial location and orientation were also shared with each user in real time and visualized as a simple avatar. With the image tracking calibration method, the alignment of two coordinate systems was accurate to around 1–$2cm$ in the working range ($< 3m$).

The software was developed using the Unity 3D game engine (2017.4.18f1), and an image processing Unity plugin was coded in C++ for processing the point-cloud data. This framework allows us to rapidly prototype the MR remote collaboration system that supports various communication cues in the shared live 3D dense scene. However, one limitation is that the depth camera's depth frame is slightly noisy with a wavy surface observed in the visualized point-cloud where flat planes should be.

USER STUDY

We were interested in evaluating the impact of sharing gaze and gesture cues separately and together for spatial guiding in MR remote collaboration tasks. We conducted a formal user study to explore the usability of our visual cues.

[1]https://github.com/IntelRealSense/librealsense

[2]https://github.com/google/draco

Experimental Design

In our study, we used speech communication without any visual cues as our control condition and the verbal plus visual cues as the comparison. In this case, our primary independent variable was the type of natural visual cues that were shared from the remote to the local user, with four communication conditions:

1. *Verbal Only (Control Condition)*
2. *Eye Gaze*
3. *Hand Gesture*
4. *Eye Gaze + Hand Gesture (Combined Condition)*

In the user study, we mainly investigated the following two research questions: 1) How does the sharing of eye gaze or hand gestures from the remote user affect collaboration in a MR remote collaboration interface? 2) What are the benefits of mixing both gaze and gesture cues for MR remote collaboration compared with using each cue alone?

Our research hypotheses were:

- H1. *Sharing natural visual communication cues (eye gaze and hand gesture) rather than only verbal cues from a remote user in a MR remote collaborative interface would improve collaboration (as measured by task performance, social presence, etc.).*
- H2. *Mixing both gaze and gesture cues for MR remote collaboration would benefit remote communication compared with using each cue alone.*

It is important to note that augmented visual cues have different properties. Gaze and head frustum communication cues are continuously shown and do not require intentional action from the user, and so work as implicit cues. In contrast, hand gestures are proactive cues and require conscious effort from the user. For example, the remote user has to explicitly decide to reach out and point at an object. By mixing both types, we assume better usability on collaborative tasks than using either cue alone. The effect of these communication cues have not been studied before in a shared live room-scale 3D panorama collaborative workspace.

Experiment Set-up

We set up our experiment in a big open space office with two cubicles. The local user performed tasks in a $(7m \times 5m)$ cubicle, and the remote expert was located in another cubicle of the same size in the same office (Figure 5b). To avoid both users seeing each other, we placed them back to back and put a whiteboard between them. The panorama capture unit was installed in the center of the local cubicle's ceiling $2.4m$ from the floor, as indicated in Figure 5a. The reconstructed live 3D panorama example is shown in Figure 5c.

Experimental Task

The local workspace consisted of four L-shaped tables placed around one desk in the middle, three cabinets on the side, office supplies, and many devices lying around like a typical working environment. On four tables, we placed 20 Lego bricks of the same size and shape, each on top of a white square tag. We put the bricks into five groups, each including

Figure 5. Experiment environment: a) The local section with Lego bricks placed around desks and the live 3D panorama capture unit installed on the ceiling; b) The remote part with the VR headset installed; c) A screenshot of the reconstructed live 3D panorama in VR.

four bricks of the same color, and we randomly placed another 20 white square tags among the Lego bricks (Figure 6a).

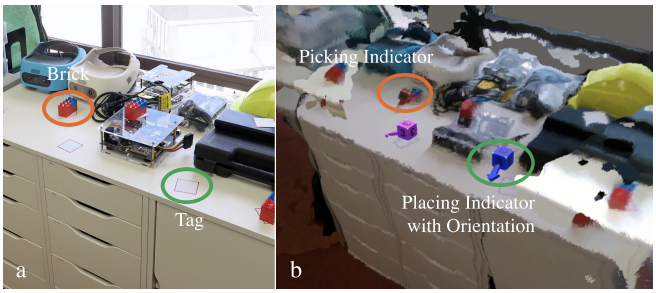

Figure 6. a) Lego bricks on the tags; b) The overlaid cubes with numbers to indicate the brick for picking up, and the overlaid cubes with numbers and arrows to indicate the target position and orientation of the Lego brick with the same pickup number.

The experimental task in each trial was to search and pick up four arbitrary Lego bricks from all 20 pieces, and place them with correct orientation on top of their corresponding numbered tags using different types of communication cues. The bricks to be taken, their destination tags, and their orientations were shown to the remote expert in the VR scene by overlaying visual cues next to the bricks and over the destination tags (Figure 6b). The local worker, wearing the AR headset, was able to walk between the L-shaped tables and the middle table to complete the task. He/she first needed to follow the remote expert's instructions (speech description, gesture pointing, and/or gaze indicating) to pick up a designated Lego brick, and then walk to the target location to place and align the picked brick on the target tag until it was correctly positioned and confirmed by the remote expert.

To avoid the shadow cast by users from the depth sensors while standing in front of the L-shaped tables, we asked participants to walk around sideways. In each trial, we defined four bricks to be picked up, one each from four different groups, and their destinations and orientations were also shuffled and assigned arbitrarily in each condition to avoid memorizing their positions. During the whole process, the participants were free to talk to each other, describing the brick direction, color, and surrounding reference objects. To avoid collaborators easily describing the color or orientation to identify the target, we grouped the same-color Lego bricks together and placed them with the same initial orientation to increase task difficulties, as shown in Figure 5a. Depending on the experiment condition, participants could communicate only verbally or together with augmented eye gaze and/or hand gestures.

Experimental Procedure

Before the experiment started, participants were assigned to their partner and their roles as the local worker or the remote expert with no specific preference. The experiment began with the participants signing a consent form and answering demographic questions and describing their VR/AR experience. The participants were then shown the remote collaboration system and the experiment tasks. The VR user's eye gaze was calibrated at the beginning, followed by a training session where paired local AR and remote VR participants tried each condition and got used to the collaboration system and interfaces. Each experiment session included four trials with different cue conditions shared from the remote to the local side. The order of the cue conditions was counterbalanced between participants. In each trial, participants performed four "pick-and-place" subtasks under a given shared cue. Based on pre-defined virtual tags, the remote user gave instructions to the local users to pick up the target bricks and put them with correct orientation on top of the designated tags. After finishing each trial and all four trials, they evaluated their experience and provided qualitative feedback about the communication cues and system in general. The study for each pair took about one hour on average to complete.

Measurements

We used a within-subject design between four trials of different cue conditions, as described above. For each pair of participants, one was the local worker and the other was the remote expert, without swapping for each condition as a between-group design. We chose this design because it reduced the time for the study, participants felt less tired or bored, and it alleviated the learning effect to some extent.

We collected both objective and subjective measures from each condition. The time for completing the tasks was recorded in a system log file to objectively measure task performance. At the end of each trial, the participants were asked to complete several subjective questionnaires. We used the NMM Social Presence Questionnaire [17] for measuring Social Presence, the MEC Spatial Presence Questionnaire [46] for measuring the sense of being together, and the NASA Task Load Index

Questionnaire [18] for measuring mental and physical load. We also measured the usability of the system using the System Usability Scale (SUS) [4]. After completing all four trials, participants were asked to rank the four conditions, in terms of advantages and disadvantages of each condition, and they provided qualitative feedback from open questions in a post-experiment questionnaire.

RESULTS

In this section, we report on the results of the user study regarding the performance and usability of all communication cue conditions, and summarize the subjective feedback collected from the participants. The mean difference was significant at the .05 level, and adjustment for multiple comparisons was automatically made with the Bonferroni correction unless noted otherwise.

Participants

We recruited 24 participants (12 male and 12 female) in 12 pairs from the local campus community with their ages ranging from 20 to 47 years old (M = 29.6, SD = 6.6). Most participant pairs knew each other. Four participants used video conferencing daily, and the rest a few times a month. Four participants were familiar with AR or VR interfaces, with ratings of four or higher on a 7-point Likert item (1: novice~7: expert).

Task Completion Time

There was a significant difference in average performance time across each of the four conditions. The Shapiro-Wilk test indicated that all task completion time data of Control (p = .162), Gaze (p = .059), Gesture (p = .367) and Combined (p = .138) were normally distributed. Mauchly's test ($\chi^2(5) = 9.41$, p = .095) did not indicate any violation of sphericity. A repeated measure ANOVA with Sphericity Assumed ($F(3, 33) = 6.073$, p = .002) determined that there was a statistically significant difference in performance time across the four conditions. A Post Hoc analysis using the Bonferroni correction revealed that the time (in seconds) to complete the tasks with Combined cues (M = 139.08, SD = 44.99) was statistically significantly faster than the Control condition (M = 227.75, SD = 88.76, p = .005). There was no significant difference found in time between all other conditions (Gaze (M = 179.08, SD = 59.51), Gesture (M = 173.00, SD = 70.11)).

Subjective Questionnaires

Social Presence

To investigate if the type of communication cues affected the participants' presence and attention, we used three sub-scales, Co-presence (CP), Attention Allocation (AA) and Perceived Message Understanding (PMU) of the Networked Mind Measure of Social Presence Questionnaire [17]. This survey consists of 18 rating items on a 7-point Likert scale (1: Strongly Disagree~7: Strongly Agree). Regarding CP, a Shapiro-Wilk test found some of the conditions were not following a normal distribution, so we applied an Align Rank Transform (ART) [47] before using a Two-Way Mixed ANOVA. The result showed that there were significant difference between communication cues ($F(3, 44) = 22.99$, $p < .001$), the user roles ($F(3, 44) = 4.15$, p = .044), as well as the interaction effect ($F(3, 44) = 4.69$, p = .004). Figure 7 shows the average CP rating.

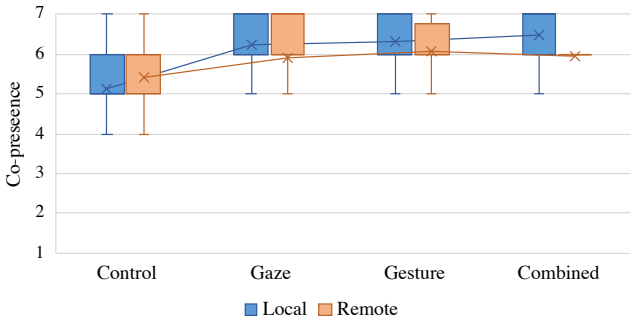

Figure 7. Results of Co-presence questionnaires (7-point Likert scale from 1 to 7, the higher the better).

Participants in the local user role gave a significantly higher rating in the CP scale overall (Local (M = 6.05, SE = 0.07)) than those in the remote user role (Remote (M = 5.84, SE = 0.07)), except in the Control condition (Local (M = 5.13, SE = 0.17), Remote (M = 5.42, SE = 0.17)). Based on pairwise comparisons, we found that paired participants gave a significantly higher rating to the three visual cues, Gaze (M = 6.08, SD = 0.89, $p < .001$), Gesture (M = 6.19, SD = 0.66, $p < .001$), Combined (M = 6.22, SD = 0.55, $p < .001$), than the Control condition (M = 5.27, SD = 1.46). Meanwhile, the difference between Gaze and Combined conditions was also statistically significant (p = .047) in terms of the user's co-presence experience. However, we found no significant difference in AA or PMU sub-scales.

Spatial Presence (Remote Only)

To study if the different types of communication cues affected the remote user's sense of being in a remote location or not, we used three sub-scales, the Spatial Situation Model (SSM), the Spatial Presence: Self Location (SPSL), and the Spatial Presence: Possible Actions (SPPA) of the MEC Spatial Presence Questionnaire [46]. These consisted of 18 rating items on a 5-point Likert scale (1: Fully Disagree~5: Fully Agree). The participants were asked to answer these questions only when they were the remote user. We ran a Friedman test on the collected data, and results showed that all sub-scales, SSM ($\chi^2(3) = 11.471$, p = .009), SPSL ($\chi^2(3) = 18.228$, $p <$ .001), SPPA ($\chi^2(3) = 9.878$, p = .020), had a significant difference between cue conditions. Gaze was rated statistically significantly higher for the remote expert's spatial layout and self-location awareness than the Combined condition in SSM (Z = -3.000, p = .003; Gaze (M = 4.17, SD = 0.444), Combined (M = 4.00, SD = 0.001)) and SPSL (Z = -2.236, p = .025, Gaze (M = 4.13, SD = 0.502), Combined (M = 3.99, SD = 0.118)). In contrast, the Combined cues gave significantly stronger feeling of spatial action than Gaze in SPPA (Z = -2.175, p = .030; Combined (M = 3.96, SD = 0.426), Gaze (M = 3.78, SD = 0.809)). The strong spatial action feeling suggested that the participants had the impression that he/she could act that same as in real life, like moving around among the objects

and having some effect on things in the environment of the presentation. No other significant difference was found. It is noticeable that all conditions have an average SSM score of higher than 4, which indicates that people in these conditions are feeling a high degree of understanding about the spatial environment.

Workload

To compare the participants' mental and physical effort in each condition, we used the NASA Task Load Index Questionnaire (TLX) [18], which consists of six rating items within a 100-points range with 5-point steps (0: very low~100: very high, the lower, the better). A Two-Way Mixed ANOVA method with ART showed that there was a significant difference in workload between the communication cues ($F(3, 44) = 7.85$, $p < .001$) and the roles ($F(3, 44) = 9.80$, $p = .002$), but there was no significant interaction effect between the two factors ($F(3, 44) = 1.43$, $p = .236$)). Figure 8 shows the average rating results of each condition for TLX. Participants in the local user role had a significantly lower workload rating in all cue conditions (Local ($M = 42.19$, $SE = 3.25$) compared to the remote user role (Remote ($M = 56.58$, $SE = 3.25$)). Based on the Pairwise Comparisons, we found that paired participants gave a significant lower rating to all visual cues, Gaze ($M = 48.68$, $SD = 30.34$, $p = .001$), Gesture ($M = 47.88$, $SD = 31.15$, $p < .001$), Combined ($M = 46.11$, $SD = 31.23$, $p < .001$), compared to the Control cue ($M = 54.86$, $SD = 29.44$). There was no difference between the visual cue conditions though.

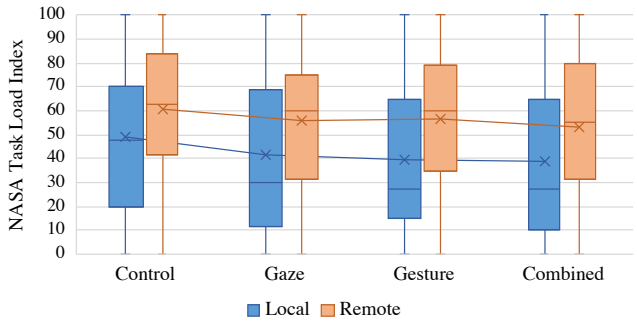

Figure 8. Results of the TLX questionnaire (100-points range with 5-point steps, 0: very low~100: very high, the lower the better).

System Usability

To evaluate the usability of our system, we used the SUS [4], which consists of 10 rating items with five response options for respondents (from Strongly Disagree to Strongly Agree). A SUS score of 68 or above is viewed as above average system usability. Although we investigated the system usability of natural visual cues for both sides, the 3D live panorama had a more important influence on the remote side. Therefore, we avoided a mixed design testing method but reviewed the result of each side separately.

The Shapiro-Wilk test on the local side indicated that the system usability scores for all conditions (Control ($p = .021$), Gaze ($p = .012$), Gesture ($p = .001$) and Combined ($p = .039$)) have marked deviations from normality, so we used a Friedman test to test for difference. The result ($\chi^2(3) = 8.119$, $p = .044$) showed that there was a statistically significant difference in the system usability between the four cue conditions for the local worker. A Post Hoc analysis with Wilcoxon Signed-rank tests was conducted with a Bonferroni correction applied, resulting in a significance difference between Gesture and Gaze ($Z = -2.147$, $p = .032$; Gesture ($M = 70.00$, $SD = 14.381$), Gaze ($M = 65.42$, $SD = 10.271$)) as well as Gesture and Control ($Z = -2.280$, $p = .023$; Gesture ($M = 70.00$, $SD = 14.381$), Control ($M = 58.75$, $SD = 15.429$)). The Shapiro-Wilk test on the remote side indicated that the system usability scores of some conditions were not following a normal distribution (Control ($p = .974$), Gaze ($p < .001$), Gesture ($p = .109$) and Combined ($p = .003$)). We used a Friedman test, and found that there was no statistically significant difference between all condition pairs.

Preference

At the end of all of the trials, we also asked participants to rank the four cue conditions in their preference for the remote collaboration task. Overall, participants mostly preferred Combined cues (16 out of 24) as their first choice, followed by Gesture, Gaze and Control options in sequence (Figure 9). Six local workers commented that they felt more like being face-to-face with the remote expert in Combined condition. Four local users regarded the eye gaze as redundant or distracting since they felt that hand gestures were sufficient for most tasks. There was a significant difference in the ranking results between four cue conditions in pairwise comparison tested through the Friedman test ($\chi^2(3) = 31.50$, $p < .001$). We ran a Wilcoxon Signed-rank test, and the results showed significant differences between all cue pairs except Gaze and Gesture ($Z = -1.622$, $p = .195$), or Gesture and Combined ($Z = -1.429$, $p = .153$). This shows that Combined cues were ranked and preferred on both local and remote sides in our tasks (by more than 50% of the participants).

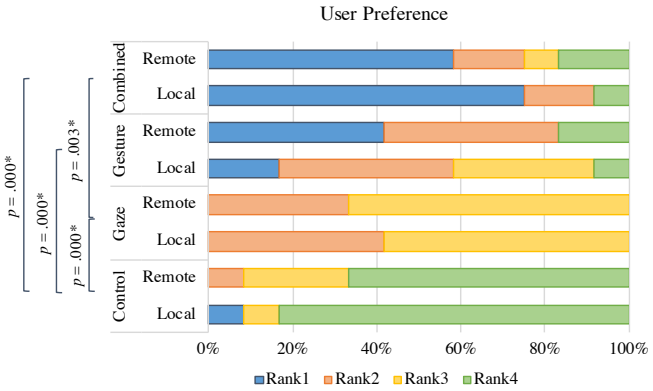

Figure 9. User preference based ranking results (Rank1 is the most preferred, *: statistically significant).

DISCUSSION

The user study results show that the use of gaze and gesture communication cues separately or together improves the local

worker and remote expert collaboration in a MR remote collaboration system, compared to using only speech cues. By combing gaze and gesture cues, our system can provide a significantly stronger feeling of co-presence for both the local and remote users than using a Gaze cue by itself. The Combined cues with the hand gesture included were rated significantly higher than gaze alone in terms of ease of performing spatial actions. Examining our research hypotheses, we found that H1 (sharing visual communication cues) was confirmed as the visual cues affected not only the performance, but also the social presence, spatial presence, and mental workload of the remote collaboration experience. H2 (mixing gaze and gesture cues) was partially verified since the combined cues provide significantly stronger feelings of co-presence and spatial action, but not the case for other terms.

In the following sections, we discuss the research results, some experiment observations, and the possible reasons for some of these results in more detail. We also compare our findings with previous related work, especially including gesture sharing in the 360 and 3D MR remote collaboration studies. Finally, we discuss the current limitations of our system and research, and present implications for collaborative MR interface design and development.

Visual vs. Verbal Cues

We found that combined visual cues significantly improved collaboration efficiency by reducing the communication time compared to using only verbal cues. This could be because when no visual cues were enabled, local workers had to act exclusively based on the remote experts' verbal instructions. Some remote users provided constructive descriptions which included spatial information such as "*at the edge of the table*" and "*in front of the blue headset*", which helped the local user locate the Lego bricks easier. However, other remote users often failed to convey useful spatial awareness messages, using more phrases like "*next one*", "*keep moving*", and "*turn a little bit*" with no detailed reference information, which prolonged the task completion time. Overall, participant pairs using hybrid gaze and gesture visual cues completed the tasks 39% faster on average than using verbal cues alone. Paired participants gave a significantly higher rating to all visual conditions than the verbal-only condition. The natural visual cues greatly enhanced the feeling of co-presence for the local users in the AR environment. This could be because the gaze and gesture cues shared by the remote expert conveyed different social information (gaze shows awareness, gesture shows interaction). The visual cues that greatly enhanced the co-presence for the local worker also had a similar influence on the remote expert. They created significantly stronger co-presence, especially with Combined cues for the remote expert. As one participant described, "*The condition with all the cues active in it provided me with the maximum flexibility. It was especially true when asking my partner to re-orient the bricks by being able to demonstrate it to him via hand gestures*".

Combined vs. Standalone Cues

There was no significant difference among three visual cues for the performance time, showing that combined visual cues were not as effective in reducing the task completion time compared with the two standalone visual cues. However, there was a difference in the benefit of using visual cues regarding usability. The usability results imply that gaze can be complemented by adding gesture cues to provide a significantly better co-presence experience. One of the participants commented that "*... the mix of all cues was much easier to use, and it feels more like the person is there with you*", and another participant mentioned that "*... hand plus gaze makes it more like a real person there*". Combined cues gave a significantly stronger feeling of spatial action than gaze alone. Most participants felt that they could interact with target objects better and easier as in real life when gaze worked with hand gestures, as they commented, "*Eye gaze was informing my partner about the exact object that was needed to be picked. Hand gestures helped me in telling him about the orientation in which the object needed to be placed. So, I think a combination of all would be pretty helpful in completing the task quickly and efficiently*".

Surprisingly the Gaze condition was rated statistically significantly higher by the remote expert for spatial layout and self-location awareness than the Combined condition in the shared MR environment. The gaze raycast supports real-time occlusion with the reconstructed point cloud, and the shadow collision on the 3D virtual replica created more accurate depth perception for the remote user to estimate the environment distance without moving around too much, which gestures alone could not provide. However, adding gestures on top of the gaze may reduce the spatial awareness especially self-location terms within our system based on some participants' feedback. Since our designed avatar representation is not human-like at all, the gaze raycast line and the hand gesture mesh will float around in mid-air with no direct visual connection with each other during the communication. Mixing the augmented gaze line (long working range) with the gesture model (short working range) would confuse people on the distance perception to some extent. In terms of the spatial action, gaze could only be used to indicate the user's point of interest but cannot support direct interaction with the environment or objects, and was rated lower than the Combined cue. For example, gaze alone cannot be used to position objects with 6 DoF in space.

The SUS scores indicated a significant difference in usability between the Gesture and Gaze conditions and the Gesture and Control conditions. The Gesture condition had higher usability scores in our remote collaboration system than the Gaze and Control conditions. This could be because remote experts usually guided the orientation of Lego bricks with fingers pointing to the corresponding direction, which efficiently assisted the local worker in identifying the target and placing it correctly. As one participant said, "*I can point in directions of things, and show which angle it needed to be turned to with hand gestures*". In comparison, the standalone gaze sharing helped with localization and showing the target objects but could not show object-orientation cues. This result is quite different from Lee et al. [29] for a similar experimental design, where they found that the view awareness cues were very useful, while the hand gestures were not used as much. This might be because our task was more 3D based with detailed depth data, where their system was not.

Local vs. Remote Users

In terms of the collaboration roles, there was a significant difference in co-presence for both sides. This finding is in line with Lee and Teo's research results [29, 45], that sharing live panorama enhances social presence in collaboration. Our remote collaboration system provides the remote user with a real-time 3D panorama scene to create an immersive experience with intuitive spatial awareness. Remote users gave much higher Spatial Presence ratings with all cues (> 3.9 out of 5) than the average value. One remote user reported that with the shared scene, he was "*more aware of the partner*", and could "*walk around*" and "*see everything*", so he felt "*more present*". However, the local worker generally felt more co-presence than the remote user, and they also had a significantly lower workload rating in all cue conditions compared to the remote user. As measured by the NASA TLX survey, the remote user had a much higher mental workload than the local one, which is expected since the local user just had to follow the lead of the remote expert. In contrast, the remote user was surrounded by the live 3D scene during the task, and had to search for targets as well as provide communication cues. One remote user commented that "*... too much stuff going on, better to have either eye gaze or hand, with both it gets too complicated*". One local user did mention that "*following all cues was little tiring and overwhelming. I had a neckache. But the voice and hand were easy to use and less tiring*", which indicates that combing the verbal and visual cues could reduce the communication overload. Although using spatial visual cues can improve remote performance, it is critical to limit the number of cues to avoid information overload.

Limitations

Our remote collaboration system works well with non-verbal communication cues in a simple controlled environment. However, there are still many aspects that need to be further improved. For example, the studied task was simplified for the usability evaluation purpose, and was different from typical real world tasks. More human factor elements need to be considered if we would like to apply our system in practical working scenarios in industry, such as the physical ergonomics of the headset, how to convey audio and visual cues, and how to reliably capture gaze and gesture cues.

Although our 3D capture unit is a small size considering its large scanning volume, the remote expert may suffer from depth shadows cast by the local worker when he/she stands between the unit and its surroundings. In our study, the local participant could easily walk sideways to avoid this problem, but this would not be ideal in practice. This might be fixed by adding another depth sensor on the AR headset worn by the local worker to fill in the shadowed area by stitching its data correctly in real time. During the user study, several participants expressed some concerns about aspects of the visual cue display. For example, four local users reported that the gaze did not accurately point to the target ("*... the gaze was not accurate enough*"). Two participants explicitly stated that gaze sharing did not feel comfortable because it moved too quickly when "*the remote person shifts their gaze around a lot*". Part of the problems was caused by the incorrect depth perception of the raycast and inconsistent eye-tracking. One solution would be to use different methods to display and smooth gaze movement. For example, it would be better to visualize gaze on top of the target with a collision-based indicator to provide more accurate depth perception.

We only focused on the comparison between the verbal and visual cues in this study but did not conduct any deeper conversational analysis to investigate their correlation. Moreover, only one-way eye gaze and/or hand gesture were shared, which could be enhanced by studying mutual sharing in further investigation. We used simple avatars to represent collaborators, which might affect the usability in our case. A more human-like or realistic avatar could be applied to the system to possibly improve the embodiment and collaboration.

Design Implications

By sharing natural cues from a remote expert to a local worker and visualizing these cues on both sides, we learned several design implications for future MR remote collaboration systems:

1. Combining gaze and gesture cues can provide a better co-presence experience and feeling of spatial action in shared MR remote collaboration.
2. Gaze cues can provide better spatial layout and self-location awareness in a shared MR environment as they enhance depth and distance perception.
3. By immersing themselves in the 3D 360 scene, remote users can get a higher spatial presence rating with all cues.
4. By receiving augmented visual cues, the local user can have a lower workload than the remote user.

CONCLUSIONS AND FUTURE WORK

In this paper, we present a MR remote collaboration system with gaze and gesture visual cues for real-time task assistance. The system supports capturing and sharing of a 3D live panorama in point-cloud format from the local to the remote side for better spatial awareness and immersion. We conducted a user study to investigate the benefits of providing natural gaze and gesture cues during the collaboration task. We concluded that sharing gaze and gestures from the remote user to the local can significantly reduce the task completion time. The local worker felt significantly higher system usability with the gesture cue than gaze, as well as significantly less mental workload than the remote expert. By combing gaze and gesture cues, the system can provide a significantly stronger feeling of co-presence for both local and remote sides than when using a gaze cue by itself. The combined cues with the hand gesture included were rated significantly higher than gaze alone in terms of the spatial action, but the gaze cue gave a significantly higher level of spatial layout and self-location awareness than the combined cues. The feedback also showed that both local and remote users preferred using combined cues over gaze or gestures alone.

In the future, we would like to explore MR collaboration with mutual natural communication cues from both sides. We would also plan to explore the usability difference between the 2D and 3D live panorama views, as well as single and multiple 360° live panorama views.

REFERENCES

[1] Matt Adcock, Stuart Anderson, and Bruce Thomas. 2013. RemoteFusion: Real Time Depth Camera Fusion for Remote Collaboration on Physical Tasks. In *Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry (VRCAI '13)*. Association for Computing Machinery, New York, NY, USA, 235–242. DOI:http://dx.doi.org/10.1145/2534329.2534331

[2] Leila Alem and Jane Li. 2011. A Study of Gestures in a Video-mediated Collaborative Assembly Task. *Advances in Human-Computer Interaction* 2011 (Jan 2011), 1:1–1:7.

[3] Martin Bauer, Gerd Kortuem, and Zary Segall. 1999. "Where are you pointing at?" A Study of Remote Collaboration in A Wearable Videoconference System. In *Digest of Papers. Third International Symposium on Wearable Computers*. IEEE, 151–158.

[4] John Brooke. 1996. SUS: A "quick and dirty" Usability Scale. 189, 194 (1996), 4–7.

[5] Henry Fuchs, Andrei State, and Jean-Charles Bazin. 2014. Immersive 3D Telepresence. 47, 7 (Jul 2014), 46–52.

[6] Susan R. Fussell, Robert E. Kraut, and Jane Siegel. 2000. Coordination of Communication: Effects of Shared Visual Context on Collaborative Work. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00)*. Association for Computing Machinery, New York, NY, USA, 21–30. DOI:http://dx.doi.org/10.1145/358916.358947

[7] Susan R. Fussell, Leslie D. Setlock, and Robert E. Kraut. 2003. Effects of Head-Mounted and Scene-Oriented Video Systems on Remote Collaboration on Physical Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03)*. Association for Computing Machinery, New York, NY, USA, 513–520. DOI: http://dx.doi.org/10.1145/642611.642701

[8] Susan R Fussell, Leslie D Setlock, Jie Yang, Jiazhi Ou, Elizabeth Mauer, and Adam DI Kramer. 2004. Gestures over Video Streams to Support Remote Collaboration on Physical Tasks. *Human-Computer Interaction* 19, 3 (Sep 2004), 273–309.

[9] Lei Gao, Huidong Bai, Weiping He, Mark Billinghurst, and Robert W. Lindeman. 2018. Real-Time Visual Representations for Mobile Mixed Reality Remote Collaboration. In *SIGGRAPH Asia 2018 Virtual & Augmented Reality (SA '18)*. Association for Computing Machinery, New York, NY, USA, Article Article 15, 2 pages. DOI:http://dx.doi.org/10.1145/3275495.3275515

[10] Lei Gao, Huidong Bai, Gun Lee, and Mark Billinghurst. 2016. An Oriented Point-Cloud View for MR Remote Collaboration. In *SIGGRAPH ASIA 2016 Mobile Graphics & Interactive Applications (SA '16)*. Association for Computing Machinery, New York, NY, USA, Article Article 8, 4 pages. DOI: http://dx.doi.org/10.1145/2999508.2999531

[11] Lei Gao, Huidong Bai, Robert W. Lindeman, and Mark Billinghurst. 2017. Static Local Environment Capturing and Sharing for MR Remote Collaboration. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications (SA '17)*. Association for Computing Machinery, New York, NY, USA, Article Article 17, 6 pages. DOI:http://dx.doi.org/10.1145/3132787.3139204

[12] Steffen Gauglitz, Cha Lee, Matthew Turk, and Tobias Höllerer. 2012. Integrating the Physical Environment into Mobile Remote Collaboration. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '12)*. Association for Computing Machinery, New York, NY, USA, 241–250. DOI: http://dx.doi.org/10.1145/2371574.2371610

[13] Steffen Gauglitz, Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. 2014. World-Stabilized Annotations and Virtual Scene Navigation for Remote Collaboration. In *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST '14)*. Association for Computing Machinery, New York, NY, USA, 449–459. DOI: http://dx.doi.org/10.1145/2642918.2647372

[14] William W. Gaver, Abigail Sellen, Christian Heath, and Paul Luff. 1993. One is Not Enough: Multiple Views in a Media Space. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems (CHI '93)*. Association for Computing Machinery, New York, NY, USA, 335–341. DOI:http://dx.doi.org/10.1145/169059.169268

[15] Kunal Gupta, Gun A. Lee, and Mark Billinghurst. 2016. Do You See What I See? The Effect of Gaze Tracking on Task Space Remote Collaboration. *IEEE Transactions on Visualization and Computer Graphics* 22, 11 (Nov 2016), 2413–2422.

[16] Pavel Gurevich, Joel Lanir, Benjamin Cohen, and Ran Stone. 2012. TeleAdvisor: A Versatile Augmented Reality Tool for Remote Assistance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. Association for Computing Machinery, New York, NY, USA, 619–622. DOI:http://dx.doi.org/10.1145/2207676.2207763

[17] Chad Harms and Frank Biocca. 2004. Internal Consistency and Reliability of The Networked Minds Measure of Social Presence. In *M. Alcaniz & B. Rey (Eds.), Seventh Annual International Workshop: Presence 2004*. Valencia: Universidad Politecnica de Valencia.

[18] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183.

[19] Steven Johnson, Madeleine Gibson, and Bilge Mutlu. 2015. Handheld or Handsfree? Remote Collaboration via Lightweight Head-Mounted Displays and Handheld Devices. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1825–1836. DOI: http://dx.doi.org/10.1145/2675133.2675176

[20] Shunichi Kasahara, Shohei Nagai, and Jun Rekimoto. 2014. LiveSphere: Immersive Experience Sharing with 360 Degrees Head-Mounted Cameras. In *Proceedings of the Adjunct Publication of the 27th Annual ACM Symposium on User Interface Software and Technology (UIST'14 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 61–62. DOI: http://dx.doi.org/10.1145/2658779.2659114

[21] Seungwon Kim, Gun A. Lee, Nobuchika Sakata, and Mark Billinghurst. 2014. Improving Co-presence with Augmented Visual Communication Cues for Sharing Experience Through Video Conference. In *Proceedings of the 2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR '14)*. IEEE Computer Society, Washington, D.C., USA, 83–92. DOI: http://dx.doi.org/10.1109/ISMAR.2014.6948412

[22] Seungwon Kim, Gun A. Lee, Nobuchika Sakata, Andreas Dünser, Elina Vartiainen, and Mark Billinghurst. 2013. Study of Augmented Gesture Communication Cues and View Sharing in Remote Collaboration. In *Proceedings of the 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR '13)*. IEEE Computer Society, Washington, D.C., USA, 261–262. DOI: http://dx.doi.org/10.1109/ISMAR.2013.6671795

[23] David Kirk, Tom Rodden, and Danaë Stanton Fraser. 2007. Turn It This Way: Grounding Collaborative Action with Remote Gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 1039–1048. DOI: http://dx.doi.org/10.1145/1240624.1240782

[24] David Kirk and Danae Stanton Fraser. 2006. Comparing Remote Gesture Technologies for Supporting Collaborative Physical Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. Association for Computing Machinery, New York, NY, USA, 1191–1200. DOI: http://dx.doi.org/10.1145/1124772.1124951

[25] Nikolina Koleva, Sabrina Hoppe, Mohammad Mehdi Moniri, Maria Staudte, and Andreas Bulling. 2015. On The Interplay Between Spontaneous Spoken Instructions and Human Visual Behaviour in An Indoor Guidance Task. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society (CogSci '15)*.

[26] Takeshi Kurata, Nobuchika Sakata, Masakatsu Kourogi, Hideaki Kuzuoka, and Mark Billinghurst. 2004. Remote Collaboration Using a Shoulder-Worn Active Camera/Laser. In *Proceedings of the Eighth International Symposium on Wearable Computers (ISWC '04)*. IEEE Computer Society, Washington, D.C., USA, 62–69. DOI: http://dx.doi.org/10.1109/ISWC.2004.37

[27] Gun A. Lee, Theophilus Teo, Seungwon Kim, and Mark Billinghurst. 2017a. Mixed Reality Collaboration through Sharing a Live Panorama. In *SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications (SA '17)*. Association for Computing Machinery, New York, NY, USA, Article 14, 4 pages. DOI: http://dx.doi.org/10.1145/3132787.3139203

[28] Gun A. Lee, Theophilus Teo, Seungwon Kim, and Mark Billinghurst. 2017b. Sharedsphere: MR Collaboration through Shared Live Panorama. In *SIGGRAPH Asia 2017 Emerging Technologies (SA '17)*. Association for Computing Machinery, New York, NY, USA, Article Article 12, 2 pages. DOI: http://dx.doi.org/10.1145/3132818.3132827

[29] Gun A. Lee, Theophilus Teo, Seungwon Kim, and Mark Billinghurst. 2018. A User Study on MR Remote Collaboration Using Live 360 Video. In *Proceedings of the 2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR '18)*. IEEE Computer Society, Washington, D.C., USA, 153–164. DOI: http://dx.doi.org/10.1109/ISMAR.2018.00051

[30] Bo Li, Lionel Heng, Kevin Koser, and Marc Pollefeys. 2013. A Multiple-camera System Calibration Toolbox Using a Feature Descriptor-based Calibration Pattern. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 1301–1307.

[31] Andrew Maimone and Henry Fuchs. 2011. A First Look at A Telepresence System With Room-sized Real-time 3d Capture And Life-sized Tracked Display Wall. *Proceedings of the 21st International Conference on Artificial Reality and Telexistence* (2011), 4–9.

[32] Katsutoshi Masai, Kai Kunze, Maki sugimoto, and Mark Billinghurst. 2016. Empathy Glasses. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1257–1263. DOI: http://dx.doi.org/10.1145/2851581.2892370

[33] Romy Müller, Jens R Helmert, and Sebastian Pannasch. 2014. Limitations of Gaze Transfer: Without visual context, eye movements do not to help to coordinate joint action, whereas mouse movements do. *Acta psychologica* 152 (2014), 19–28.

[34] Romy Müller, Jens R Helmert, Sebastian Pannasch, and Boris M Velichkovsky. 2013. Gaze Transfer in Remote Cooperation: Is it always helpful to see what your partner is attending to? *The Quarterly Journal of Experimental Psychology* 66, 7 (2013), 1302–1316.

[35] Ye Pan and Anthony Steed. 2016. Effects of 3D Perspective on Head Gaze Estimation with a Multiview Autostereoscopic Display. *International Journal of Human-Computer Studies* 86 (Feb 2016).

[36] Thammathip Piumsomboon, Arindam Dey, Barrett Ens, Gun A. Lee, and Mark Billinghurst. 2017. [POSTER] CoVAR: Mixed-Platform Remote Collaborative Augmented and Virtual Realities System with Shared Collaboration Cues. In *Proceedings of the Adjunct Publication of the 2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR' 17 Adjunct)*. IEEE Computer Society, Washington, D.C., USA, 218–219. DOI: http://dx.doi.org/10.1109/ISMAR-Adjunct.2017.72

[37] Thammathip Piumsomboon, Gun A. Lee, Andrew Irlitti, Barrett Ens, Bruce H. Thomas, and Mark Billinghurst. 2019. On the Shoulder of the Giant: A Multi-Scale Mixed Reality Collaboration with 360 Video Sharing and Tangible Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 228, 17 pages. DOI:http://dx.doi.org/10.1145/3290605.3300458

[38] Rodrigo M. A. Silva, Bruno Feijó, Pablo B. Gomes, Thiago Frensh, and Daniel Monteiro. 2016. Real Time 360° Video Stitching and Streaming. In *ACM SIGGRAPH 2016 Posters (SIGGRAPH '16)*. Association for Computing Machinery, New York, NY, USA, Article Article 70, 2 pages. DOI: http://dx.doi.org/10.1145/2945078.2945148

[39] Rajinder S. Sodhi, Brett R. Jones, David Forsyth, Brian P. Bailey, and Giuliano Maciocci. 2013. BeThere: 3D Mobile Collaboration with Spatial Input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 179–188. DOI: http://dx.doi.org/10.1145/2470654.2470679

[40] Anthony Steed, David Roberts, Ralph Schroeder, Ilona Heldal, and others. 2005. Interaction Between Users of Immersion Projection Technology Systems. In *Proceeding of HCI International 2005, the 11th International Conference on Human Computer Interaction (HCI '05)*. 22–27.

[41] William Steptoe, Robin Wolff, Alessio Murgia, Estefania Guimaraes, John Rae, Paul Sharkey, David Roberts, and Anthony Steed. 2008. Eye-Tracking for Avatar Eye-Gaze and Interactional Analysis in Immersive Collaborative Virtual Environments. In *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work (CSCW '08)*. Association for Computing Machinery, New York, NY, USA, 197–200. DOI: http://dx.doi.org/10.1145/1460563.1460593

[42] Patrick Stotko, Stefan Krumpen, Matthias B Hullin, Michael Weinmann, and Reinhard Klein. 2019. SLAMCast: Large-Scale, Real-Time 3D Reconstruction and Streaming for Immersive Multi-Client Live Telepresence. *IEEE Transactions on Visualization and Computer Graphics* 25, 5 (May 2019), 2102–2112.

[43] Matthew Tait and Mark Billinghurst. 2015. The Effect of View Independence in A Collaborative AR System. *Computer Supported Cooperative Work (CSCW)* 24, 6 (Dec 2015), 563–589. DOI: http://dx.doi.org/10.1007/s10606-015-9231-8

[44] Franco Tecchia, Leila Alem, and Weidong Huang. 2012. 3D Helping Hands: A Gesture Based MR System for Remote Collaboration. In *Proceedings of the 11th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry (VRCAI '12)*. Association for Computing Machinery, New York, NY, USA, 323–328. DOI: http://dx.doi.org/10.1145/2407516.2407590

[45] Theophilus Teo, Louise Lawrence, Gun A. Lee, Mark Billinghurst, and Matt Adcock. 2019. Mixed Reality Remote Collaboration Combining 360 Video and 3D Reconstruction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 201, 14 pages. DOI: http://dx.doi.org/10.1145/3290605.3300431

[46] Peter Vorderer, Werner Wirth, Feliz R Gouveia, Frank Biocca, Timo Saari, Futz Jäncke, Saskia Böcking, Holger Schramm, Andre Gysbers, Tilo Hartmann, and others. 2004. MEC Spatial Presence Questionnaire (MEC-SPQ): Short Documentation and Instructions for Application. *Report to the European community, project presence: MEC (IST-2001-37661)* 3 (2004).

[47] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. DOI: http://dx.doi.org/10.1145/1978942.1978963