Journal of Universal Computer Science, vol. 26, no. 8 (2020), 972-995
submitted: 12/1/20, accepted: 8/7/20, appeared: 28/8/20 *CC BY-ND 4.0*

Comparing Collaboration Fidelity between VR, MR and Video Conferencing Systems: The Effects of Visual Communication Media Fidelity on Collaboration

Sangsun Han
(Hanyang University, Ansan, South Korea
hanss1@hanyang.ac.kr)

Kibum Kim
(Hanyang University, Ansan, South Korea
kibum@hanyang.ac.kr)

Seonghwan Choi
(Keimyung University, Daegu, South Korea
shjy1669@naver.com)

Mankyu Sung
(Keimyung University, Daegu, South Korea
mksung@kmu.ac.kr)

Abstract: Video/audio conferencing systems have been used extensively for remote collaboration over many years. Recently, virtual and mixed reality (VR/MR) systems have started to show great potential as communication media for remote collaboration. Prior studies revealed that the creation of common ground between discourse participants is crucial for collaboration and that grounding techniques change with the communication medium. However, it is difficult to find previous research that compares VR and MR communication system performances with video conferencing systems regarding the creation of common ground for collaborative problem solving. On the other hand, prior studies have found that display fidelity and interaction fidelity had significant effects on performance-intensive individual tasks in virtual reality. Fidelity in VR can be defined as the degree of objective accuracy with which the real-world is represented by the virtual world. However, to date, fidelity for *collaborative tasks* in VR/MR has not been defined or studied much. In this paper, we compare five different communication media for the establishment of common ground in collaborative problem-solving tasks: Webcam, headband camera, VR, MR, and audio-only conferencing systems. We analyzed these communication media with respect to collaborative fidelity components which we defined. For the experiments, we utilized two different types of collaborative tasks: a 2D Tangram puzzle and a 3D Soma cube puzzle. The experimental results show that the traditional Webcam performed better than the other media in the 2D task, while the headband camera performed better in the 3D task. In terms of collaboration fidelity, these results were somehow predictable, although there was a little difference between our expectations and the results.

Keywords: Virtual Realty, Mixed Reality, Fidelity, Collaboration, Common Ground, Human-Computer Interaction
Categories: H.5.0, H.5.2, H.5.3, L.3.0, L.3.1, L.5.0, L.6.2

1 Introduction

Nowadays, virtual reality and augmented reality (VR/AR) technologies are expanding the limits of space and of time. Although most current VR/AR platforms have been developed for single player use, an increasing number of collaborative virtual environments in which local or remote participants can share collaborative experiences are being marketed. By using VR/AR, users can remotely share a workspace and utilize augmented information or virtual objects to collaborate. For example, HTC Vive published a multiplayer mode VR game titled Front Defense: Heroes™ [Vive Studios, 2019], and Facebook announced a social VR system, Facebook Horizon™ [Facebook, 2020] which supports interactions between remote partners as if they were in the same virtual space.

The construction of common ground and the associated convergent conceptual changes among discourse participants are critical to efficient cooperative work and collaborative learning processes. The literature indicates that shared visual information promotes efficient communication among discourse participants as well as "grounding" – the joint process of establishing mutual belief [Clark, 2004][Fussell, 2000][Fussell, 2003b][Gale, 2002][Kraut, 2002]. Communication media influence the quantity, quality and context of the visual information exchanged. In this paper, we study five different communication media for creating common ground in conferencing systems: virtual reality (i.e., HTC Vive), mixed reality (i.e., Microsoft HoloLens), fixed camera (i.e., Webcam), movable camera (i.e., headband camera), and audio-only (as a baseline). These communication media are different in terms of fidelity. Fidelity refers to how accurately a technology creates a virtual world experience that mimics a real-world experience. Through a case study of collaborative learning tasks, we studied how different components of fidelity in visual communication media play their roles in creating common ground between discourse participants. During our experiment, two participants were asked jointly to complete visual block puzzles. In their collaboration, one person played a teacher's role and directed a partner while the other person, a novice in relation to the puzzles, took a student role and arranged the 2D and 3D puzzle pieces. We identified and evaluated the holistic effects of a range of fidelity factors for the diverse communication media in collaborative learning tasks and then discussed issues for user performance that we discovered from the experiment.

1.1 Fidelity of Individual's Experience in VR

Many researchers have used the term "fidelity" to characterize how similar the users' virtual world experiences are to their real world experiences [Ahmed, 2017][Kaya, 2018][Mania, 2006][McMahan, 2011][McMahan, 2012][Sziebig, 2009]. For example, a VR HMD (head mounted display) supports higher fidelity than a desktop PC's mouse because changing perspective on a scene by rotating the HMD is more akin to the way we change our view in the real world than by dragging a mouse.

Previous research on fidelity has focused exclusively on the individual's experience in VR [Laha, 2014][McMahan, 2012]. According to a previous study, fidelity can be divided into two types – display fidelity and interaction fidelity [McMahan, 2011]. Display fidelity presents how well a display system reproduces real-world visual stimuli. The components of display fidelity include stereoscopy, field of view (FOV), field of regard (FOR), display resolution, display size, refresh rate, and

frame rate [Bowman, 2012][McMahan, 2011]. FOV refers to the angular extent that can be viewed instantaneously by the user, i.e., without shifting, panning, or tilting the eyes. On the other hand, FOR refers to the total size of the visual field in degrees of visual angle surrounding the user [Bowman, 2007]. For example, a three wall CAVE provides roughly a 120° horizontal FOV and a 180° horizontal FOR. In a previous study, increasing the FOV afforded faster search times because the narrow FOV required users to move their heads more often [Ren, 2016]. In our experiments, HTC Vive HMD in the VR condition has a wide FOV (110°), while Microsoft HoloLens in the MR condition has a narrow FOV (35°). We expect that the narrow FOV of HoloLens will limit the user's performance in manipulating virtual objects.

Alternatively, interaction fidelity addresses how precisely a VR system reproduces real world interaction experiences. McMahan defined interaction fidelity as *the objective degree of exactness with which real-world interactions can be reproduced in an interactive system* [McMahan, 2011]. The components of interaction fidelity include the positional accuracy of an input device (i.e., how accurately an input device represents a position), rotational accuracy, force accuracy, form factor (i.e., the shape of a hand held input device), device location (i.e., the location of the input device relative to the user), update rate of interaction result, number of control dimensions, and integration of dimensions (i.e., how the degree of freedom (DOF) is interpreted together) [McMahan, 2011]. In our experiments, HoloLens in the MR condition delivers higher interaction fidelity than Vive in the VR condition. This is because participants assemble *real* puzzle pieces inside a *virtual* silhouette hint (e.g., see white lines in Fig. 3(e)) in the MR condition, while participants assemble *virtual* puzzle pieces inside a *virtual* silhouette hint (e.g., see black lines in Fig. 3(d)) in the VR condition. In the Webcam, headband camera, and audio only conditions, participants assemble *real* puzzle pieces inside a *real* outline hint (e.g., see black lines in Figs. 3(b) and 3(c)). Therefore, the interaction fidelity for the Webcam, headband camera, and audio conditions are high; for MR, it is medium; and for VR, it is low, as indicated in Fig. 1 and 2.

1.2 Sharing Visual Contexts to Facilitate Collaborative Learning

Classroom learning has long been oriented toward face-to-face interaction. However, extended reality (XR) (i.e., augmented reality, virtual reality and mixed reality) technologies offer students new opportunities to increase their potential for interacting with remote peers and to facilitate distant collaborative learning. For example, CIVE (Collaborative Immersive Virtual Environment) was developed and studied for the cognitive and social aspects of collaboration in shared, immersive virtual environments for geography education [Šašinka, 2018]. Bowman et al. explained the differences among VR, AR, and MR [Bowman, 2004]. VR fully immerses the user in virtual environments. AR enhances (augments, complements) the user's view of a real-world environment with virtual objects or information. MR mingles real and virtual information, where physical and digital objects co-exist and the user interacts with objects in real time.

Vygotsky emphasized that learning is fundamentally a social process [Cole, 1978]. If students can solve a problem after a teacher offers leading questions or initiates the solution, or after they have collaborated with other peers, then students are in the zone of proximal development. This joint construction of knowledge requires creating a

process of cognitive, social, and emotional interchange. Learning in the zone of proximal development involves the shared contexts in which participants are interpersonally engaged. Therefore, socio-cultural contexts are central to the creation of the zone of proximal development.

The particular set of problems we consider in this study is designed to promote the rapid acquisition of visual contexts (i.e., workspaces) and the sharing of meaning among ad hoc social groups using XR technology. Sharing meaning is the fundamental issue for collaborative learning. In the classroom, sharing meaning has two practical forms: "formative assessment", in which a teacher attempts to assess the progress of an ongoing activity and to intervene if necessary, and "peer-monitoring", in which a student attempts to learn what other students are doing by observing peers. In our experiments, we explored how different communication media could provide the benefits of sharing visual contexts during collaborative learning tasks, and we showed the impact of diverse collaboration fidelity factors on *common ground*.

1.3 Contribution

Our study makes the following five contributions. First, we evaluated commercial off-the-shelf VR/MR products as communication media for collaboration. HTC Vive and Microsoft HoloLens are among the most popular extended reality (XR) products nowadays. However, not much formal evaluation has been done on how these devices function as collaboration media. By comparing these state-of-the-art products with traditional communication media, such as video and audio conferencing systems, we have identified and defined the pros and cons of VR and MR products as next generation communication media.

Second, we extended the existing fidelity of individual VR settings and suggested new components for the fidelity of collaborative VR. These components are : consistent view, immediacy, perspective alignment, view control, interaction fidelity and FOV for hints (clues). Through experiments, we verified the effects of the suggested fidelity components on collaboration. These components are evaluated so that they may be ranked in order to produce the highest degree of fidelity and functionality in a collaborative VR system.

Third, we also applied collaborative fidelity components to analyze traditional communication media such as video and audio-conferencing systems. This would provide an evaluation framework for comparing VR and MR not only with their real-world counterparts (video/audio conferencing systems) but also with the new media of the future.

Fourth, we investigated the development of common ground by repeating the puzzle trials in our experiment. The Tangram puzzle has been widely used to explore the creation and maintenance of common ground in psycholinguistic literature [Clark, 1996]. According to Clark's conversation grounding theory, participants accumulate common ground and complete tasks more quickly by repeating trials. These phenomena agreed with our experiment results, so we determined that Clark's grounding theory could be applied to VR and MR communication media as well.

Fifth, we found some drawbacks with VR/MR systems in collaborative tasks. The Vive (VR) had a problem in directly manipulating the virtual object and the HoloLens (MR) had a problem in sharing the same view of a workspace between the student and

the teacher due to the student's shifting gaze. We discuss these problems based on prior study and our experiment.

In the remaining part of this paper, we review related work, explain the details of the experiment, show experiment results, discuss findings, and present our summary conclusion.

2 Related Work

We studied prior works in respect of various remote visual collaboration systems and diverse fidelity components.

2.1 Remote Visual Collaboration

Various media have been used for remote conferencing systems. Isokoski and Akkil created a video conferencing environment in which the helper could point using a mouse or via eye gazing [Isokoski, 2019]. Mouse pointing brightened the area upon which the helper's mouse pointer dwelled; eye gaze pointing brightened the area that was followed by the helper's gaze. The user study using a complicated 3D puzzle task showed that both methods of pointing were better than no pointer. Other researchers enabled an effective collaboration by adding virtual objects to a remote workspace. Huang et al. made a collaborative system called HandsIn3D in which a helper could see a workspace with VR and a worker could see an instruction indicated by the helper's virtual hand [Huang, 2013]. The 3D camera, which was located at the worker's space, captured and transmitted the workspace to the remote helper's HMD. The helper was then able to send back virtual hand instruction feedback to the worker's LCD monitor. The system showed the 'video mix' of the worker's workspace and the helper's virtual hand motion in real time to both participants. In Nguyen et al.'s system, the worker used a VR device to find directions inside a building, while the helper had blueprint information about the structure of the building on a desktop computer [Nguyen, 2012]. The helper was able to lead the worker to a target by showing the virtual direction marker in the worker's view. Furthermore, the helper was able to help the worker by highlighting a particular area or by teleporting the worker to the target position.

An MR-based collaborative system helps users to get additional information intuitively by overlapping a real workspace with virtual objects. Holstein et al. constructed a collaborative system called Lumilo, in which a teacher could see students' questions via HoloLens [Holstein, 2018]. In a classroom, students solved math quizzes with a desktop computer while a teacher was monitoring their progress by looking at virtual question marks on their heads using HoloLens. If the teacher clicked the question marks, s/he could easily access students' questions.

Feick et al. developed a collaborative system in which two users used VR and MR respectively [Feick, 2018b]. Using an MR camera, one user showed an object that was related to a task and the other user could see the object on the smartphone VR. In addition, as the VR user moved the virtual object, the MR user could see the movement as a virtual proxy. In this study, the researcher tried to achieve improved collaborative performance either by showing the helper's motion via a hologram or by providing virtual hints making the procedure easy to understand.

2.2 Fidelity of Video Conferencing Systems

There are studies about the fidelity of shared workspace views in traditional video conferencing systems. Gergle et al. analyzed different display configurations when showing the worker's workspace in the helper's monitor [Gergle, 2013]. The experimental task was to assemble a 2D puzzle using square pieces. They evaluated immediacy, perspective alignment, field of view, and view control options. Immediacy refers to the system's ability to update the worker's changes immediately. Perspective alignment was about updating the worker's changes by randomly flipping puzzle pieces in a vertical or horizontal direction and then by rotating them 45, 90, and 135 degrees. The field of view referred to the area that was visible to the helper. In the experiment, the field of view conditions had the options of full view, large square (which showed some puzzle pieces), and small square (which essentially showed one single piece). View control referred to the ability to change the area that was visible in the display. The results showed that collaboration performance was better with regard to three specific conditions: immediate update rather than with delay, non-rotated rather than rotated, and updated full screen rather than updated partial view of the workspace. This work is closely related to our study, but it researched video conferencing systems only. By contrast, we evaluated both VR and MR conferencing systems for collaboration fidelity.

Fussell et al. suggested that showing the workspace to the helper was an essential element for the efficient performance of remote collaboration [Fussell, 2000]. They compared an audio-conferencing system, a video conferencing system showing the whole workspace, and a face-to-face collaboration. Their video conferencing system was set up with the helper using a desktop computer to view the shared workspace and the worker using a camera attached to the worker's head showing the workspace. The experimental task was to repair a bike. The researchers found that the video condition delivered worse task completion times than the face-to-face condition, but better task completion times than the audio-only condition. In the video condition, the helper asked the worker to move his/her head to reset the workspace view because the worker's head camera sometimes did not show the workspace. Fussell et al. also compared different workspace capturing settings between a helper and a worker in a collaboration [Fussell, 2003a]. The three workspace capture settings were a camera fixed to a table that showed the workspace and the worker; a worker's head-mounted camera with eye tracker, which showed the workspace jointly with the worker's focus of attention; a video mix of these two views. The task for collaboration was to make a robot model. The results showed that the fixed camera condition was better than either the worker's head camera view or the mixed view setting. The reason for this outcome was that the eye tracker often slipped off the head, and the head mounted camera had a narrow field of view.

Johnson et al. compared two collaborative media: a head-mounted camera and a tablet camera[Johnson, 2015a]. The task used for comparison was to build a toy car. The results showed that the tablet was better than the head mounted camera. They argued that the reason why the head-mounted camera performed poorly was that it did not maintain a consistent view. The helper needed to keep asking the worker to move the head to adjust the visible area of the workspace. Feick et al. developed a robot mediated communication system called Remote Manipulator [Feick, 2018a]. They compared a traditional video conferencing system with Remote Manipulator, in which the local user managed an object and then the remote user could see that a collocated

robot arm manipulated the object in the same way. Their experiment results showed that the same perspective of the shared object had better performance than the mirrored perspective of the shared object.

Unlike previous research on remote visual collaboration systems, our study conducted experiments, which directly compared VR, MR, and traditional video and audio conferencing systems to determine each system's fidelity for cooperative 2D and 3D puzzle-solving tasks.

3 Collaboration Fidelity in Our Experiment

In our experiment, we identified and evaluated the following six components as fidelity components in collaborative VR/MR systems - *consistent view*, *immediacy*, *view control*, *perspective alignment*, *field of view,* and *interaction fidelity*.

- *Consistent view*: If the camera was mounted on the student's head, the shared workspace view would likely be intermittently out of sight on the teacher's monitor as the student looked around during the collaboration. With HoloLens, the teacher's shared workspace view was disrupted as the student's head moved to refocus on a virtual hint in the workspace. We defined the fidelity of consistent view as high if the system had a continuous and stable view of the pertinent shared workspace. The Webcam produced high fidelity for consistent view because the camera was fixed and offered a stable view of the workspace. Audio had low fidelity because there was no shared workspace view. The rest of the conditions delivered medium fidelity because the camera moved with the student's head and the shared view was sometimes not pertinent to the critical workspace.
- *Immediacy*: The fidelity of immediacy is high when the system allows a student to share his/her workspace view with a teacher immediately (i.e., with no lag). The immediacy fidelity of Webcam, headband camera and Vive was high because there is no latency in transmission of views from the student's workspace to the teacher's monitor. HoloLens had medium fidelity because there is about a two-second latency between the student's workspace and the teacher's monitor when setting up with the current off-the-shelf network utility (i.e., Windows Device Portal). Lastly, audio produced low fidelity because there is no shared workspace view.
- *View control*: Ideal view control allows the collaborators to share the workspace from any view they prefer. In the headband camera, Vive and HoloLens conditions, the teacher could ask the student to move his/her head to look at a side view of the 3D Soma cube puzzle. However, in the Webcam condition, the camera was fixed on the monitor and the teacher could see only the front view of the 3D Soma cube puzzle. For the 2D Tangram puzzle task in which the puzzle pieces were placed on top of the desk, the view control took a less important role with regard to camera angle because the Webcam was set up to capture the whole desk space. We considered that the fidelity of view control was high if the system delivered a comprehensive range of views of the workspace. For view control in 2D, all the systems except audio were high because these systems can give a comprehensive view of the workspace.

In 3D, headband camera, HoloLens and Vive had high fidelity of view control because the teacher could see all the sides of Soma cube puzzles by asking the student to move his/her head. Webcam produced medium fidelity because the teacher could only see the front side of the cube puzzles. Audio had low fidelity because there was no shared workspace view.

- *Perspective alignment*: If the camera is mounted on the student's head, it is possible for both student and teacher to share the same perspective on the workspace. However, with Webcam, left and right in the student's shared workspace view was mirrored in the teacher's view because the mounted camera viewed the workspace from the reverse side to the student's view. Therefore, for perspective alignment fidelity, Webcam had medium fidelity. Audio produced low fidelity because there was no workspace view. The headband camera, HoloLens and Vive had high fidelity perspective alignment because they showed the identical view as seen by the student.

- *Field of view (FOV)*: In a previous study, narrow FOV afforded slower search times because the narrow FOV required users to move their heads more often [Ren, 2016]. In our 2D task, the student's FOV in audio, Webcam and headband camera were high because students could see the whole hint (a silhouette of the final shape of the Tangram); for Vive, the FOV (110°) of a virtual hint was medium and for HoloLens, the FOV (35°) was low. In our 3D task, the FOV for audio, Webcam and headband camera was low because there was no hint; the fidelity for HoloLens was medium because a virtual hint was provided with a narrow FOV; the fidelity for Vive was high because a virtual hint was provided with a wide FOV.
- *Interaction fidelity*: In our experiment, participants in the MR condition assembled real puzzle pieces inside a virtual silhouette hint, while participants in the VR condition assembled virtual puzzle pieces inside a virtual silhouette hint. In the Webcam, headband camera, and audio conditions, participants assembled real puzzle pieces on a real silhouette hint. Therefore, the interaction fidelity for the Webcam, headband camera, and audio conditions are high (i.e., interacting with real pieces and a real hint); fidelity is medium for MR (i.e., interacting with real pieces and a virtual hint) and it is low for VR (i.e., interacting with virtual pieces and a virtual hint).

Fig. 1 and Fig. 2 summarize the various degrees of fidelity for each communication media on the 2D task and the 3D task used in our experiment.

4 Hypotheses

Prior works pointed out that media with higher fidelity would produce better performance [Mania, 2006][McMahan, 2011][Ragan, 2015]. Therefore, we expect that media with more "High" fidelity ratings (in Fig. 1 and 2) would produce better performance.

As seen in Fig. 1, the number of "High" rankings in the 2D task are ordered as follows:

- Audio (2) < HoloLens (3) < Vive (4) < Webcam, Headband Camera (5)

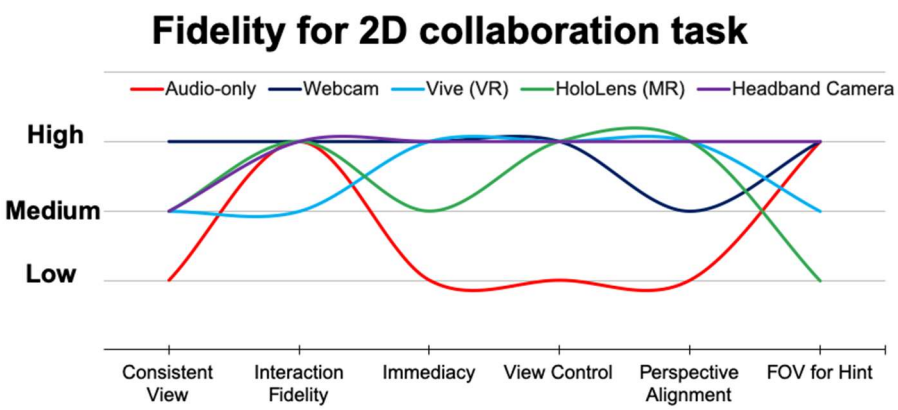

Figure 1: The parallel coordinates showing degrees of fidelity for communication media in the 2D experimental task. On the y-axis, High indicates higher fidelity and Low indicates lower fidelity.

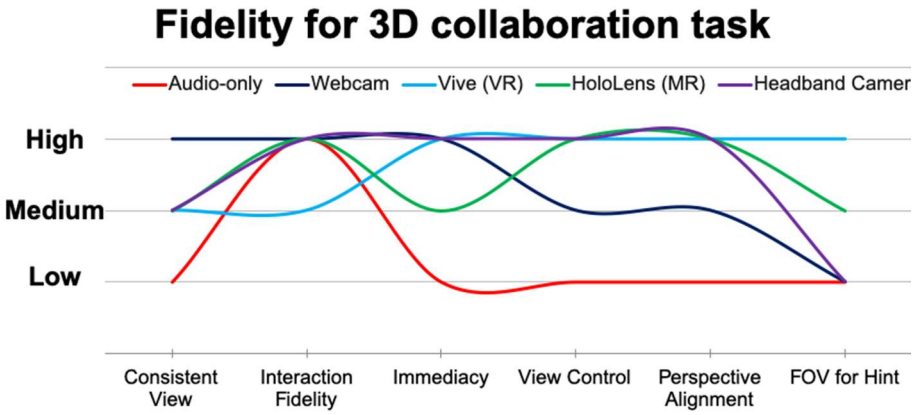

Figure 2: The parallel coordinates showing degrees of fidelity for communication media in the 3D experimental task. On the y-axis, High indicates higher fidelity and Low indicates lower fidelity.

In Fig. 2, the number of "High" rankings in the 3D task are ordered as follows:

- Audio (2) < HoloLens (3) < Webcam (4) < Headband Camera, Vive (5)

Therefore, our hypotheses for overall performance are as follows.

H1. Task completion time will increase with the descending order of the numbers of "High" fidelity rankings:

- 2D: Webcam, Headband Camera<Vive<HoloLens<Audio
- 3D: Headband Camera, Vive<Webcam<HoloLens<Audio

H2. Error rate will increase with the descending order of the numbers of "High" fidelity rankings:

- 2D: Webcam, Headband Camera<Vive<HoloLens<Audio
- 3D: Headband Camera, Vive<Webcam<HoloLens<Audio

H3. User satisfaction will increase with the ascending order of the numbers of "High" fidelity rankings:

- 2D: Audio<HoloLens<Vive<Webcam, Headband Camera
- 3D: Audio<HoloLens<Webcam< Headband Camera, Vive

5 Experiment

In this section, we describe the experiment participants, apparatus and settings, design, tasks, procedure, and measurement.

5.1 Participants

Prior to the beginning of recruitment, all procedures were approved by the university institutional review board (IRB). Sixty participants (32 Males, 28 Females; aged between 18 and 35 years, with a mean age of 24) were recruited through advertisements. All participants were paid for their participation. Thirty-one participants reported that they wore glasses, and they kept wearing glasses during the experiment if they so preferred. Forty-seven of the participants were college students. One participant had previous experience with HoloLens and thirty participants had experience with VR HMD. Fifty-two participants were right-handed. All sixty participants took the role of the "student" in the experiment task. In addition, three "teachers" were trained in advance, so that they knew the solutions for the puzzles and so they could manage the trials consistently. All teachers participated equally by turns in the experiment.

5.2 Apparatus and Settings

The teacher and the student sat at opposite ends of the large table arrangement which was made by connecting two small tables. A divider in the middle of the table arrangement prevented them from looking at each other's workspaces directly. At the student's table there were seven pieces of a Tangram puzzle for the 2D task and seven pieces of a Soma cube puzzle for the 3D task. However, in the VR condition in which the student uses the Vive, virtual puzzle pieces were provided to the student. In addition, a hint for the solution of the Tangram puzzle was given to the student in the form of a large silhouette of the final shape on a piece of paper (the strawboard in Fig. 3). Vive and HoloLens presented the hint as a virtual silhouette (fourth and fifth pictures in Fig. 3). For the Soma cube puzzle in the 3D task, Vive and HoloLens gave participants a hint in the form of a virtual 3D silhouette (fourth and fifth pictures in Fig. 4). However, a 3D silhouette hint was not feasible for Webcam and headband camera conditions.

Fig. 5 shows the setting for the experiment in the MR condition. The teacher used a PC monitor for all the conditions except the audio condition, and the student used a different medium depending on the given experiment condition. In the audio condition,

the student did not visually share the workspace with the teacher, but s/he collaborated with the teacher using verbal communication. In the Webcam condition, a camera was set up to face the student's workspace so that throughout the trials the teacher could see the workspace from the opposite perspective (i.e., the second picture in Figs. 3 and 4). In the headband camera condition, the student attached a camera to the forehead with a headband. The camera was adjusted so that the teacher was able to see the workspace from the same perspective as the student (i.e., the third picture in Figs. 3 and 4). In the VR condition, the student used the Vive HMD headset and two controllers. The student manipulated the virtual puzzle pieces with the controllers (i.e., the fourth picture in Figs. 3 and 4). The teacher looked at the virtual workspace on the large monitor with the same view appearing on the student's HMD using the 'Game pane' in the Unity editor. In the MR condition, the student was able to see the real puzzle pieces overlaid with the virtual silhouette hint via the Microsoft HoloLens (i.e., the fifth picture in Figs. 3 and 4). On the large monitor, the teacher was able to see the workspace and the virtual hint which appeared on the student's HoloLens using the 'Windows Device Portal' network utility.

5.3 Experiment Design

We used the mixed experiment design. The between-subjects' components were five different communication media and the within-subject components were repeated trials during which a participant solved five different puzzles for each 2D and 3D task. Sixty participants were divided into 5 groups of 12 participants for each communication media, and they solved five different puzzles both for 2D and for 3D tasks. Fig. 3 shows five different Tangram puzzles in the 2D task. Fig. 4 shows five different Soma cube puzzles in the 3D task. The five communication media used in the experiment were as follow:

- Audio: This is the baseline condition. The teacher could not see student's workspace. Instead, the teacher and the student collaborated using only verbal communication. Students moved and rotated puzzle pieces by hand (see Figs. 3(a) and 4(a)).
- Webcam: The camera was fixed on the monitor to capture the student's workspace area in the 2D task so that the teacher could see the entire workspace. For the 3D task, the camera was fixed on the table, so that the teacher could see only the front view of the assembled puzzle pieces. The student moved and rotated puzzle pieces by hand (see Figs. 3(b) and 4(b)).
- Headband camera: The teacher used the view from student's forehead camera. The various view perspectives on the workspace could be shared by the teacher asking the student to move his/her head. Students moved and rotated puzzle pieces by hand (see Figs. 3(c) and 4(c)).
- Vive HMD in VR: The virtual puzzle pieces and a virtual silhouette hint were given to students through Vive HMD. The teacher shared the same workspace view as the student's Vive which was displayed on the teacher's monitor. Students used Vive controllers to move and rotate virtual puzzle pieces. In this environment, students used the mid-air manipulation technique for manipulating virtual puzzle pieces (see Figs. 3(d) and 4(d)).

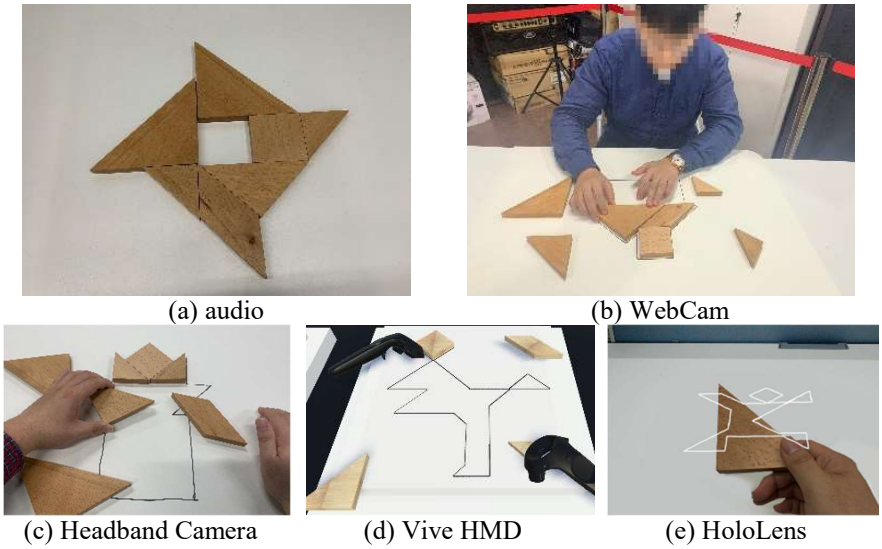

Figure 3: Five Tangram puzzles used in the 2D task (from left to right, it shows the audio, Webcam, headband camera, VR, and MR conditions)

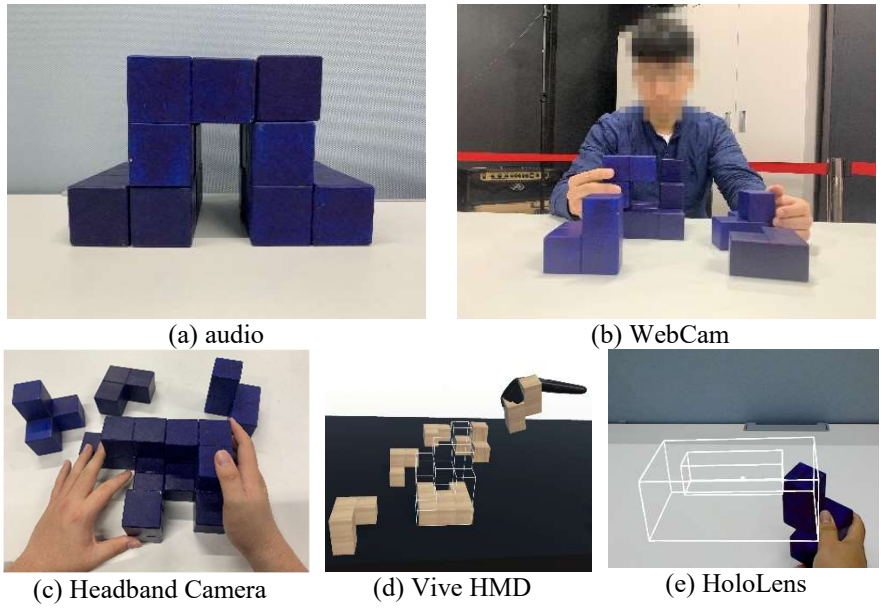

Figure 4: Five Soma cube puzzles used in the 3D task (from left to right, it shows the audio, Webcam, headband camera, VR, and MR conditions)

- HoloLens in MR: The virtual silhouette hint was given to students through HoloLens HMD. Students moved and rotated real puzzle pieces with their hands. The student's HoloLens view was displayed on the teacher's monitor (see Figs. 3(e) and 4(e)).

The treatment orders for five different communication media, five different puzzles, and 2D or 3D tasks were all counterbalanced using a Latin square. For the puzzle trial factor, we used a within-subjects design and each participant solved a set of five different puzzles using one of the communication media. For the communication media factor, we used a between-subjects design to avoid a learning effect due to repeated trials because participants needed to solve the same set of five different puzzles for each communication media.

5.4 Tasks

The experiment task was to solve the 2D and 3D spatial puzzles within as short a time as possible. Students needed to find out the expected shape and position of each puzzle piece by following the teacher's instructions. For the 2D task, the Tangram puzzle was provided. The puzzle pieces consisted of two quadrangle shapes (a square and a parallelogram) and five different sizes of triangles. By positioning and rotating the pieces, students could make various 2D shapes. Our task was to make five specific 2D shapes using all seven pieces. The five Tangram puzzle shapes used in the experiment are shown in Fig. 3. For the 3D task, the Soma cube puzzle was provided. The Soma cube puzzle consisted of seven basic blocks, which were used to make various 3D shapes. Our task was to make five specific 3D shapes using all seven basic blocks. The five Soma cube puzzle shapes used in the experiment are shown in Fig. 4.

5.5 Procedure

At the beginning of the experiment, each participant signed a consent form. Then, the overview of the experiment was explained. Participants answered questionnaires regarding their demographic background and their experience with VR and MR media. After that, each participant took the student's role and had a practice session to become familiar with the apparatus before starting the main task. For the main session, students were asked to construct five different puzzle shapes for each of the 2D and 3D puzzles by using one form of communication media (i.e., audio, Webcam, headband camera, VR, and MR). A teacher assisted students in solving the puzzle by monitoring the student's workspace on the large display (except for the audio condition) and by giving instructions to the students. Students were free to ask the teacher questions while performing the tasks. Each trial had a time limit of seven minutes. If students could not complete a task on time, the trial was counted as unfinished and students moved on to the next trial. The order of five puzzle shapes was counterbalanced. The order of 2D and 3D tasks was also counterbalanced. Students answered the NASA Task Load Index (TLX) questionnaires after completing either the 2D or 3D task ([NASA, 2020], see Appendix I). They then took a five-minute break. Students did five more trials with the other 3D or 2D task. Students then answered another NASA TLX questionnaire. When

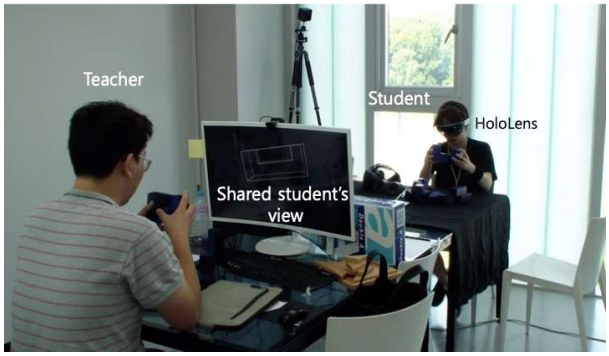

Figure 5: The teacher and the student are collaborating in the MR condition.

both the 2D and the 3D tasks were finished, students completed the final survey about their overall experiences.

5.6 Measurements

We evaluated user performance with quantitative and qualitative measurements. The quantitative measurements evaluated task completion time and error rate. Task completion time was the time taken to solve the puzzle. Error rate referred to the number of puzzle pieces that were positioned incorrectly with regard to position and orientation, or which remained unplaced at the end of each round. When each trial was finished, we recorded the task completion time and the error rate. The other quantitative measurements were made with NASA TLX questionnaires and post-experiment surveys. NASA TLX is a much-used instrument for the evaluation of the perceived workload [Mcgill, 2015][Wang, 2011]. It consists of six subjective subscales: mental demand, physical demand, temporal demand, effort, performance and frustration (see Appendix I). The NASA TLX questionnaires were given once after all 2D tasks were completed and again after all 3D tasks were completed. When all 2D and 3D tasks were completed, students completed the final 7-point Likert scale survey. This survey included an "Inclusion of Others in the Self" (IOS) scale to see how close students felt to the teacher in each communication media [Culbertson, 2016][Qiu, 2019][Woosnam, 2010].

For qualitative measurements, we interviewed students and got detailed explanations of their experiences with the experiment. When students completed the questionnaire and survey, they answered a few questions relating to difficulty using devices, communicating, and about the task itself. We used open-ended questions to ensure variety in the information regarding the participants' experimental experiences.

6 Results

Task completion time was analyzed using mixed ANOVA. A power analysis showed that the study had sufficient power (i.e., greater than 0.80). For the 2D task, results indicated significant main effects both for different communication media (F(4,55)

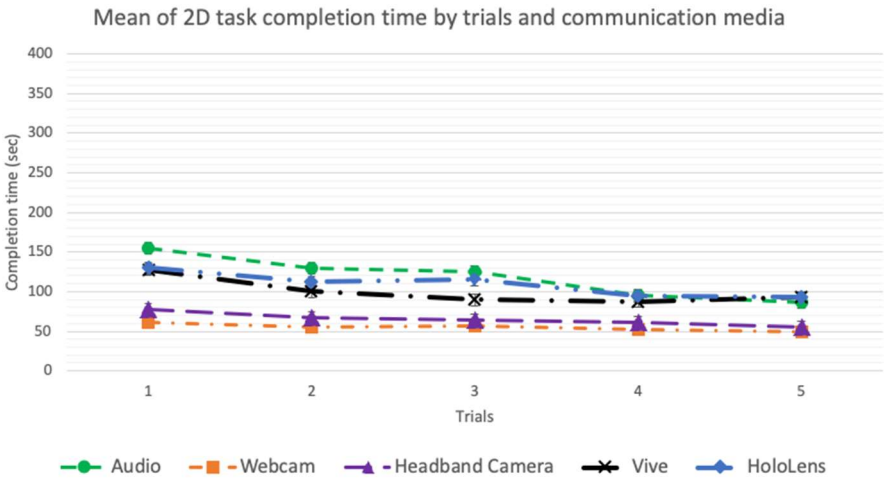

Figure 6: Average task completion time for each communication media on the 2D Tangram puzzle by repeated trials.

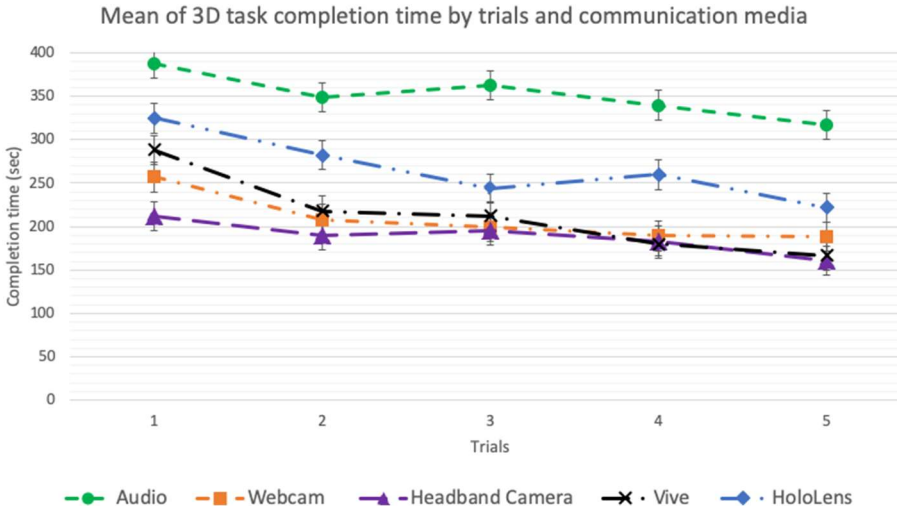

Figure 7: Average task completion times for each communication media on the 3D Soma cube puzzle by repeated trials.

=20.767, p<0.001) and for repeated trials (F(4,220) =10.632, p<0.001). The post-hoc pairwise comparisons of communication media with Bonferroni correction indicated that Webcam and headband camera were significantly faster than Vive, HoloLens and audio (p<0.05). Table 1 shows mean and standard errors for the 2D Tangram puzzle task completion time, revealing significant differences between media. The double split line in Table 1 indicates a significant difference between two groups. The average task completion time of the 2D task for each communication media by each trial is graphed in Fig. 6.

For the 3D task, the results also indicated significant main effects both for different communication media (F(4,55)= 14.208, p<0.001) and for repeated trials (F(4,220) =22.342, p<0.001). The post-hoc pairwise comparisons of communication media with the Bonferroni correction indicated that audio was significantly slower than Vive, HoloLens, headband camera and Webcam (p<0.01). In addition, there was a significant difference between the headband camera and the HoloLens (p<0.05). Table 2 shows mean and standard errors for the 3D Soma cube puzzle task completion time, revealing significant differences between media. The double split line in Table 2 indicates a significant difference between two groups. The average task completion time for the 3D task for each communication media by each trial is graphed in Fig. 7.

In summary, the order of task completion time for the 2D task was:

- Webcam < Headband camera < Vive < HoloLens < Audio

It is the same order as we expected in H1:

- Webcam, Headband Camera < Vive < HoloLens < Audio

Also, the order of task completion time for the 3D task was:

- Headband camera < Webcam < Vive < HoloLens < Audio.

The order is similar to what we expected in H1, except for Vive and Webcam:

- Headband Camera, Vive < Webcam < HoloLens < Audio

Regarding errors, the audio condition resulted in a total of nine errors in the 2D task. No error was found in the other communication media for the 2D task. The 3D task showed that audio (129) had more errors than HoloLens (10), headband camera (9), Webcam (0), and Vive (0). Three participants who used the HoloLens could not complete the task on time. Notably, 11 participants out of 12, who were in the audio condition and worked on the 3D Soma cube puzzles, could not complete the puzzle at least once out of five trials. Audio had the highest error rate among five communication media.

The NASA TLX questionnaire results were analyzed using one-way ANOVA for the different communication media. In the 2D task, the mental demand (F (4,55) = 2.618, p < 0.05) and the physical demand (F (4,55) = 4.597, p < 0.01) showed significant differences. A Tukey post-hoc test revealed that the mental demand for audio was significantly higher than for Webcam (p=0.037); the physical demand for Vive was significantly higher than audio (p=0.007), Webcam (p=0.007), and headband camera (p=0.019). In the 3D task, the mental demand (F (4,55) = 3.029, p<0.05), the temporal demand (F(4,55) = 3.516, p<0.05), and the performance (F(4,55) = 3.056, p<0.05) showed significant differences. A Tukey post-hoc test revealed that the mental demand for audio was significantly higher than for Webcam (p=0.048), Vive (p=0.048),

Media	Webcam	Headband Camera	Vive (VR)	HoloLens MR	Audio
completion time in seconds	55.18 (7.589)	65.0 (7.589)	99.65 (7.589)	108.75 (7.589)	118.3 (7.589)

Table 1: Mean and standard error of the 2D task completion time for each communication media. The middle double line spilt shows the significant difference between two groups ($p < 0.05$).

Media	Headband camera	Webcam	Vive (VR)	HoloLens (MR)	Audio
completion time in seconds	188.37* (16.931)	208.53 (16.831)	212.93 (16.831)	266.37* (16.831)	351.4 (16.831)

Table 2: Mean and standard errors for the 3D task completion time of each communication media. The middle double line spilt shows the significant difference between audio and all the other media ($p < 0.01$). There was a significant difference between headband camera and HoloLens ($p < 0.05$).*

and HoloLens (p=0.048); the temporal demand for audio was significantly higher than for

HoloLens (p=0.009); the performance for audio was significantly more oriented toward failure than for headband camera (p=0.019). Overall, audio demanded the most for each NASA TLX question items. Table 3 shows mean and standard errors indicated by the NASA TLX questionnaires. In the IOS (Inclusion of Others in the Self) survey, the ANOVA analysis showed no statistically significant difference among communication media for any of the questions. In addition, we statistically analyzed the number of teacher's comments which called for adjustments in the students' puzzle pieces. There were significant differences among communication media both in 2D and 3D puzzles (2D: $F(1,4)=5.394$, $p<0.01$, 3D: $F(1,4)=3.854$, $p<0.01$). In the 2D puzzle, audio required a higher number of teacher's comments than Webcam ($p<0.01$), HoloLens ($p<0.05$) and headband camera ($p<0.01$). Vive also had a higher number than Webcam ($p<0.01$) and headband camera ($p<0.01$). For the 3D puzzle, audio required a higher number of teacher's comments than Webcam ($p<0.05$), HoloLens ($p<0.05$) and headband camera ($p<0.01$). Overall, in the audio condition, teachers persistently asked students about the puzzle pieces because there was no shared view of the workspace. In addition, for Vive, the teacher needed to repeatedly give feedback about placing and rotating puzzle pieces because students had difficulty in manipulating virtual objects.

7 Discussion

In this section, we discuss the following four issues for user performance that we discovered from the experiment: FOV of hint, interaction fidelity, view control and common ground.

	NASA TLX Statement	HoloLens (MR)	Vive (VR)	Headband camera	Webcam	Audio
2D Task	How mentally demanding was the task?	1.92 (0.358) **	2.42 (0.434)	2.08 (0.288)	1.25 (0.131)	2.58 (0.313) **
	How physically demanding was the task?	1.75 (0.411)	2.67 (0.527) *, **	1.25 (0.179) **	1.08 (0.083) *	1.08 (0.083) *
3D Task	How mentally demanding was the task?	3.17 (0.49) **	3.17 (0.49) **	3.5 (0.359)	3.17 (0.458) **	5 (0.477) **
	How hurried or rushed was the task?	2 (0.389) *	2.5 (0.469)	2.75 (0.463)	2.58 (0.499)	4.5 (0.68) *
	How successful were you in accomplishing what you were asked to do?	1.92 (0.417)	2.25 (0.411)	1.5 (0.195) **	2.75 (0.494)	3.5 (0.597) **

Table 3: Mean and standard errors of NASA TLX questionnaires (1–Very Low, 7–Very High, $^{*}p < 0.01$, $^{**}p < 0.05$)

7.1 Field of View (FOV) for Hint: A narrow FOV in HoloLens leads to bad performance

Vive has a FOV around 110° and HoloLens has a FOV around 35° [Piumsomboon, 2018]. During the experiment with HoloLens, students could see only a part of the hint at a glance because of the narrow FOV. Some students who wore HoloLens asked a teacher to make the whole hint visible in order to overcome the effects of the narrow FOV. This inconvenience can be explained by a theory relating to narrow FOVs in collaboration from Johnson et al.'s work [Johnson, 2015b]. "Cognitive tunnel" is a phenomenon in which observers cannot immediately comprehend the presented information and must focus on the information in order to understand it [Thomas, 2000]. Johnson et al. showed that a narrow FOV increased the effect of cognitive tunnel and led to increasing errors in judgment [Johnson, 2015b]. Another phenomenon is the "keyhole effect" which is caused when viewing large areas with limited angular view [Woods, 2004]. The keyhole effect increases the difficulty in navigating and understanding spatial environments. Johnson et al. also showed that a narrow FOV increased the keyhole effect and led to longer completion times for spatial collaborative tasks [Johnson, 2015b]. In our experiment, these phenomena happened in HoloLens vs. Vive conditions. The HoloLens condition, which has a narrower FOV, showed longer completion times than the Vive condition which has a wider FOV. It seems that HoloLens's narrow FOV caused a keyhole effect by showing only part of the silhouette hint in a glance and this created a cognitive tunnel effect which made students take more time and effort to understand the teacher's instructions.

7.2 Interaction Fidelity: Direct manipulation has difficulty of precise manipulation in the Vive condition

Our experiment tasks demanded the precise manipulation of puzzle pieces; otherwise, the puzzles could not be completed. Using Vive, students took too much time controlling the pieces precisely. This problem is an already known weakness of Vive's direct manipulation method [Mendes, 2017][Tuachob, 2018]. The direct manipulation

method allows users to manipulate objects naturally, but it lacks precision [Mendes, 2018]. To compensate for the effects of this weakness, we deactivated collision and gravity. If collision and gravity are activated in the virtual environment, there is a possibility of students accidentally pushing previously positioned pieces away while maneuvering another piece into place. Should this happen, the original pieces would have to be repositioned, and time would be wasted. Therefore, for this experiment, we deactivated collision and gravity in order to maintain the position and rotation of pieces wherever they had been properly placed by students. However, as a side effect, students needed to consider the height of each piece. Without gravity and collision, students needed to consider the z-axis to put the puzzle pieces either on the virtual table without overlapping them or on the top of other virtual pieces without floating them in the air. According to Chan et al.'s study, the values for mid-air touching surfaces of virtual objects were difficult to determine [Chan, 2010]. Touching accuracy and values on the z-axis were harder to determine than touching values on either x or y-axes. From their experiment on touching surfaces for virtual objects in mid-air, they found that people have around 13mm error for virtual objects on the z-axis. In our Vive condition, students commented that they had trouble placing puzzle pieces vertically and that they therefore consumed more time. For example, in one case, even though the teacher told the student, "The piece is in the right position and direction. You're okay to move to other pieces," the student continued to work on positioning the piece. Overall, the Vive condition, which uses virtual puzzle pieces, induced time wasting in placing puzzle pieces too precisely.

7.3 View Control: View control results in good performance in 3D puzzles

View control is the feature that allows a camera to move around the object and show various views of the object from different angles. This makes it easier to understand the whole shape of a 3D puzzle piece. View control produced different results between the 2D and 3D tasks in the Webcam condition. For the 2D condition, view control was not crucial because the entire workspace could be captured easily by the Webcam, but for the 3D condition, view control was necessary because the teacher couldn't see all facets of the assembled puzzle pieces with the Webcam. Within the fixed Webcam condition in the 3D task, the teacher could see only the front side of the puzzle, and the teacher could understand different sides only by asking students questions about the puzzle pieces. As a result, while using the webcam, the teacher asked more questions than when using the other communication media, such as the headband camera, Vive, and HoloLens, all of which supported view control in the 3D task. From this result, we concluded that the task could be completed more quickly when the various views on the 3D object were supported by view control. Previous studies also confirmed that view control significantly affected performance, especially when the object was 3D [Kim, 2012].

7.4 Accumulation of Common Ground: Accumulated common ground leads to better performance

According to Clark's conversation grounding theory, discourse participants accumulate their common ground by repeating trials, so that they can communicate more efficiently and complete the task quickly as they do more trials [Clark, 1991][Clark, 1996][Clark,

2004]. Through our experiment, we also found that communication times became shorter as the trials were repeated. Notably, the teacher's statements became even more concise as they repeated 3D task instructions than as they repeated the 2D task instructions. Soma cube puzzles in the 3D task demanded more complex sentences to indicate their shape, size, and orientation than the case of Tangram puzzles in the 2D task. Therefore, the effect of accumulating common ground with trials was more recognizable in the 3D task than in the 2D task. After common ground was formed by repeating the trials, the teacher was able to describe the puzzle pieces in much more concise terms, and students became familiar with the terminology used, and so more easily chose the correct pieces. In our experiment, common ground reduced task completion time progressively as trials were repeated, as represented in Figs. 6 and 7.

8 Conclusion

We compared five communication media, including not only traditional collaborative media like audio, Webcam, and headband camera, but also high-end collaborative media such as VR and MR. We compared these media using two different puzzle tasks: a 2D task using the Tangram puzzle and a 3D task using the Soma cube puzzle. Students solved five Tangram puzzles and five Soma cube puzzles.

Our experiment results show that traditional camera-based communication media performed better than VR or MR media. The main reason for degradation of MR media was that the narrow FOV of the HoloLens reduced collaboration efficiency. On the other hand, the main reason for VR's degradation was that VR controllers offered poor interaction fidelity compared to natural hand manipulation. By repeating trials, accumulated common ground progressively and significantly reduced the task completion times. The complex 3D tasks showed more improvement than the relatively easy 2D tasks by accumulating common ground.

Finally, we identified various fidelity components for collaborative VR and found that these fidelity components significantly affected cooperative problem-solving both in the 2D and in the 3D tasks. By identifying these fidelity components through user studies for the first time we might also have established evaluation criteria for other collaborative VR experiments. We look forward to applying our approach regarding collaborative fidelity to other new communication media that will emerge in the future.

Acknowledgements

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (No. 2018R1D1A1A02085645). Corresponding author: Kibum Kim (kibum@hanyang.ac.kr).

References

[Ahmed, 2017] Ahmed, I., Harjunen, V., Jacucci, G., Ravaja, N., and Spapé, M.M.: "Total Immersion: Designing for Affective Symbiosis in a Virtual Reality Game with Haptics, Biosensors, and Emotive Agents". Proc., Springer International Publishing (2017), 23-37.

[Bowman, 2007] Bowman, D.A. and McMahan, R.P.: "Virtual Reality: How Much Immersion Is Enough?"; Computer, 40, 7, 36-43 (2007)

[Bowman, 2012] Bowman, D.A., McMahan, R.P., and Ragan, E.D.: "Questioning naturalism in 3D user interfaces"; Communications of the ACM, 55, 9, 78-88 (2012)

[Bowman, 2014] Bowman, D.A., Kruijff, E., LaViola J.J., and Poupyrev I.: "3D User Interfaces: Theory and Practice" Addison Wesley Longman Publishing Co., Inc. (2014), USA.

[Chan, 2010] Chan, L.-W., Kao, H.-S., Chen, M.Y., Lee, M.-S., Hsu, J., and Hung, Y.-P.: "Touching the void: direct-touch interaction for intangible displays"; Association for Computing Machinery, Atlanta, Georgia, USA (2010).

[Clark, 1996] Clark, H.H.: "Using language", Cambridge (1996).

[Clark, 1991] Clark, H.H. and Brennan, S.E.: "Grounding in communication"; American Psychological Association, (1991).

[Clark, 2004] Clark, H.H. and Krych, M.A.: "Speaking while monitoring addressees for understanding"; Journal of Memory and Language, 50, 1 (2004), 62-81.

[Cole, 1978] Cole, M., John-Steiner, V., Scribner, S., and Souberman, E.: "Mind in society"; Mind in society the development of higher psychological processes. Cambridge, MA: Harvard University Press, (1978).

[Culbertson, 2016] Culbertson, G.R., Andersen, E.L., White, W.M., Zhang, D., and Jung, M.F.: "Crystallize: An Immersive, Collaborative Game for Second Language Learning". Proc. Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW '16, (2016), 635-646.

[Facebook, 2020] Facebook: "Facebook horizon"; Retrieved 2020.01.31, Access from https://www.oculus.com/facebookhorizon/

[Feick, 2018a] Feick, M., Mok, T., Tang, A., Oehlberg, L., and Sharlin, E.: "Perspective on and Re-orientation of Physical Proxies in Object-Focused Remote Collaboration". Proc. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, ACM Publishing, 3173855 (2018), 1-13.

[Feick, 2018b] Feick, M., Tang, A., and Bateman, S.: "Mixed-Reality for Object-Focused Remote Collaboration". Proc. The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings, ACM Publishing, 3266102 (2018), 63-65.

[Fussell, 2000] Fussell, S.R., Kraut, R.E., and Siegel, J.: "Coordination of communication: Effects of shared visual context on collaborative work". Proc. 2000 ACM conference on Computer supported cooperative work, ACM (2000), 21-30.

[Fussell, 2003a] Fussell, S.R., Setlock, L.D., and Kraut, R.E.: "Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks". Proc. SIGCHI Conference on Human Factors in Computing Systems, ACM Publishing, 642701 (2003), 513-520.

[Fussell, 2003b] Fussell, S.R. and Siegel, J.: "Visual Information as a Conversational Resource in Collaborative Physical Tasks"; Human–computer interaction, 18, 1-2 (2003), 13-49.

[Gale, 2002] Gale, C.: "A Look Is Worth a Thousand Words: Full Gaze Awareness in Video-Mediated Conversation"; Discourse Processes, 33, 3 (2002), 257-278.

[Gergle, 2013] Gergle, D., Kraut, R.E., and Fussell, S.R.: "Using Visual Information for Grounding and Awareness in Collaborative Tasks"; Human–computer interaction, 28, 1 (2013), 1-39.

[Holstein, 2018] Holstein, K., Mclaren, B.M., and Aleven, V.: "Student Learning Benefits of a Mixed-Reality Teacher Awareness Tool in AI-Enhanced Classrooms". Proc., Springer International Publishing (2018), 154-168.

[Huang, 2013] Huang, W., Alem, L., and Tecchia, F.: "HandsIn3D: Supporting Remote Guidance with Immersive Virtual Environments". Proc. Human-Computer Interaction – INTERACT 2013, Springer Berlin Heidelberg (2013), 70-77.

[Isokoski, 2019] Isokoski, P. and Akkil, D.: "Comparison of Gaze and Mouse Pointers for Video-based Collaborative Physical Task"; Interacting with Computers, 30, 6 (2019), 524-542.

[Johnson, 2015a] Johnson, S., Gibson, M., and Mutlu, B.: "Handheld or Handsfree?: Remote Collaboration via Lightweight Head-Mounted Displays and Handheld Devices". Proc. 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, ACM Publishing, 2675176 (2015), 1825-1836.

[Johnson, 2015b] Johnson, S., Rae, I., Mutlu, B., and Takayama, L.: "Can You See Me Now? How Field of View Affects Collaboration in Robotic Telepresence". Proc. 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15, (2015), 2397-2406.

[Kaya, 2018] Kaya, E., Alacam, S., Findik, Y., and Balcisoy, S.: "Low-fidelity prototyping with simple collaborative tabletop computer-aided design systems"; Computers & Graphics, 70 (2018), 307-315.

[Kim, 2012] Kim, K., Bolton, J., Girouard, A., Cooperstock, J., and Vertegaal, R.: "TeleHuman: effects of 3d perspective on gaze and pose estimation with a life-size cylindrical telepresence pod". Proc. SIGCHI Conference on Human Factors in Computing Systems, ACM Publishing, 2208640 (2012), 2531-2540.

[Kraut, 2002] Kraut, R.E., Fussell, S.R., Brennan, S.E., and Siegel, J.: "Understanding effects of proximity on collaboration: Implications for technologies to support remote collaborative work"; MIT Press, Cambridge, MA, US (2002).

[Laha, 2014] Laha, B., Bowman, D.A., and Socha, J.J.: "Effects of VR System Fidelity on Analyzing Isosurface Visualization of Volume Datasets"; IEEE Transactions on Visualization and Computer Graphics, 20, 4 (2014), 513-522.

[Mania, 2006] Mania, K., Wooldridge, D., Coxon, M., Robinson, A.J.I.T.O.V., and Graphics, C.: "The effect of visual and interaction fidelity on spatial cognition in immersive virtual environments"; 12, 3 (2006), 396-404.

[Mcgill, 2015] Mcgill, M., Boland, D., Murray-Smith, R., and Brewster, S.: "A Dose of Reality". Proc. 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15, (2015), 2143-2152.

[McMahan, 2011] McMahan, R.P.: "Exploring the effects of higher-fidelity display and interaction for virtual reality games"; Virginia Tech, Doctor's Thesis, (2011)

[McMahan, 2012] McMahan, R.P., Bowman, D.A., Zielinski, D.J., and Brady, R.B.: "Evaluating Display Fidelity and Interaction Fidelity in a Virtual Reality Game"; IEEE Transactions on Visualization and Computer Graphics, 18, 4 (2012), 626-633.

[Mendes, 2018] Mendes, D., Caputo, F.M., Giachetti, A., Ferreira, A., and Jorge, J.: "A Survey on 3D Virtual Object Manipulation: From the Desktop to Immersive Virtual Environments"; Computer Graphics Forum, 38, 1 (2018), 21-45.

[Mendes, 2017] Mendes, D., Sousa, M.I., Lorena, R., Ferreira, A., and Jorge, J.: "Using custom transformation axes for mid-air manipulation of 3D virtual objects". Proc. 23rd ACM Symposium on Virtual Reality Software and Technology, ACM Publishing, 3139157 (2017), 1-8.

[Nasa, 2020] NASA: Nasa Task Load Index, Retrieved 2020.05.12. Access from https://humansystems.arc.nasa.gov/groups/TLX/downloads/TLXScale.pdf

[Nguyen, 2012] Nguyen, T.T.H., Fleury, C., and Duval, T.: "Collaborative exploration in a multi-scale shared virtual environment". Proc. 2012 IEEE Symposium on 3D User Interfaces (3DUI), (2012), 181-182.

[Piumsomboon, 2018] Piumsomboon, T., Lee, G.A., Hart, J.D., Ens, B., Lindeman, R.W., Thomas, B.H., and Billinghurst, M.: "Mini-Me: An Adaptive Avatar for Mixed Reality Remote Collaboration". Proc. 2018 CHI Conference on Human Factors in Computing Systems, ACM Publishing, 3173620 (2018), 1-13.

[Qiu, 2019] Qiu, S., Hu, J., Han, T., Osawa, H., and Rauterberg, M.: "Social Glasses: Simulating Interactive Gaze for Visually Impaired People in Face-to-Face Communication"; International Journal of Human–Computer Interaction, (2019), 1-17.

[Ragan, 2015] Ragan, E.D., Bowman, D.A., Kopper, R., Stinson, C., Scerbo, S., and McMahan, R.P.: "Effects of Field of View and Visual Complexity on Virtual Reality Training Effectiveness for a Visual Scanning Task"; IEEE Transactions on Visualization and Computer Graphics, 21, 7 (2015), 794-807.

[Ren, 2016] Ren, D., Goldschwendt, T., Chang, Y., and Höllerer, T.: "Evaluating wide-field-of-view augmented reality with mixed reality simulation". Proc. 2016 IEEE Virtual Reality (VR), IEEE (2016), 93-102.

[Šašinka, 2018] Šašinka, Č., Stachoň, Z., Sedlák, M., Chmelík, J., Herman, L., Kubíček, P., Šašinková, A., Doležal, M., Tejkl, H., Urbánek, T., Svatoňová, H., Ugwitz, P., and Juřík, V.: "Collaborative Immersive Virtual Environments for Education in Geography"; ISPRS International Journal of Geo-Information, 8, 1 (2018), 3.

[Vive Studios, 2019] Vive Studios: "Front Defense: Heroes"; Retrieved 2019.03.29, Access from https://store.steampowered.com/app/763430/Front_Defense_Heroes/

[Sziebig, 2009] Sziebig, G.: "Achieving Total Immersion: Technology Trends behind Augmented Reality- A Survey". Proc. WSEAS International Conference on SIMULATION, MODELLING AND OPTIMIZATION, World Scientific and Engineering Academy and Society (WSEAS) (2009).

[Thomas, 2000] Thomas, L.C. and Wickens, C.D.: "Effects of display frames of reference on spatial judgments and change detection"; D. DOCUMENT, (2000).

[Tuachob, 2018] Tuachob, M. and Nilkhamhang, I.: "Command-Based Object Manipulation in Virtual Reality for Visualization and Design Tasks". Proc. 2nd International Conference on Digital Signal Processing, ACM Publishing, 3193054 (2018), 93-97.

[Wang, 2011] Wang, X. and Dunston, P.S.: "Comparative Effectiveness of Mixed Reality-Based Virtual Environments in Collaborative Design"; IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 41, 3 (2011), 284-296.

[Woods, 2004] Woods, D.D., Tittle, J., Feil, M., and Roesler, A.: "Envisioning human-robot coordination in future operations"; IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 34, 2 (2004), 210-218.

[Woosnam, 2010] Woosnam, K.M.: "The inclusion of other in the self (ios) scale"; Annals of Tourism Research, 37, 3 (2010), 857-860

APPENDIX I. NASA TLX questionary

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High