VISUALLY ANNOTATED RESPONSIVE DIGITAL TWINS FOR REMOTE COLLABORATION IN MIXED REALITY ENVIRONMENTS

Louis Baumgartner[1], Luca Brägger[1], Kathrin Koebel[1], Joe Scheidegger[2], Arzu Çöltekin[1*]

[1]Institute of Interactive Technologies, University of Applied Sciences and Arts Northwestern Switzerland, Switzerland
[2]SBB Center of Competence Extended Reality, Bern, Switzerland

Commission IV WG IV/9

KEY WORDS: Digital twins, mixed reality, remote collaboration, visual annotations

ABSTRACT:

Various forms of extended reality might empower remote collaboration in ways that the current de facto standards cannot facilitate. Especially when combined by a digital twin of the remote physical object, mixed reality (MR) opens up interesting new ways to support spatial communication. In this study, we explore the use of a digital twin to facilitate visuospatial communication in an expert-guided repair and maintenance operation scenario, supported by visual annotations. We developed two MR prototypes, one with a digital twin of the object of interest, and another where a first-person camera view was shown additionally. We tested these prototypes in a study with 19 participants (9 pairs) against a state-of-the art solution as a baseline and measured their usability, and obtained qualitative user feedback. Our findings suggest that digital twin supported mixed reality enriched with real time visual annotations can potentially improve remote collaboration tasks.

1. INDTRODUCTION AND BACKGROUND

Different forms of extended reality (XR), such as virtual (VR) or mixed reality (MR), are emerging as promising collaboration tools for tasks that require visuospatial and tactile tasks that are difficult with verbal or more traditional digital tools. For example, MR as a collaborative tool has been explored object manipulation in industrial design (Ong and Shen, 2009), show-and-tell for teaching three-dimensional concepts in classrooms or in space missions (Fairchild et al., 2016; Giraudeau et al., 2019), exploring historical artefacts in archaeology or gaming (Benko et al., 2004; Pulver et al., 2020; Zhou et al., 2019) due to its potential to benefit --accelerate and improve-- productivity and memorability depending on the context. Nonetheless, collaboration in XR still presents a number of yet-to-be-solved technology, and specifically, human-computer interaction (HCI) challenges. In this paper, we present a remote collaboration concept using digital twins of physical objects to facilitate visuospatial communication in tasks such as guided maintenance and repair, as well as other contexts that require remote instructions by an expert for someone to carry out a task in the field. Specifically, we explore a responsive digital twin implementation which could be visually annotated in real time to improve communication between two parties working on the same task in different locations as a team. Our contributions in this paper include a concise review of the interdisciplinary literature on collaboration, mixed reality and digital twins followed by an implementation of prototypes which we developed based on user-centered design principles in multiple design iterations; and usability tests along with some exploratory trials, including qualitative feedback from the participants (n=18, 9 pairs).

1.1. Collaboration

Contemporary textbook definitions of collaboration refer to processes in which multiple individuals or organizations work together and share responsibility of the outcomes (Appley and Winder, 1977; Camarihna-Matos and Afsarmanesh, 2008; Martinez-Moyano, 2006). Collaboration is studied by various disciplines, such as sociology or social, work and organizational psychology (Landy and Conte, 2016). Even though there is not a single formula for successful collaboration (Bennett and Gadlin, 2012), there are basic models of collaboration that can help characterize collaboration and make team effectiveness more comprehensible. One of the most important models for this is McGrath's (1964) "input-process-output model" (IPO model) (McGrath, 1964) (Figure 1).

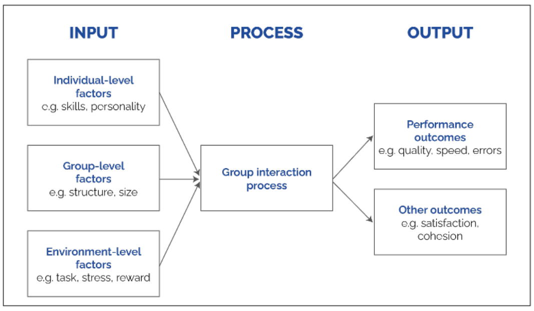

Figure 1. McGrath's (1964) input-process-output model (redrawn).

As Figure 1 shows, McGrath's (1964) model divides collaboration into three primary categories: *input*, *process*, and *output*. *Input* in this collaboration model is the composition of a group or team (*e.g.,* size of the team or the abilities of individual team members). *Process* refers to necessary activities for fulfilling a goal-oriented team task (*e.g.,* establishing standards, communication, coordination, cohesion, and decision-making). These processes depend on input factors, and they influence the output. *Output* describes to what extent a particular goal has been achieved through the cooperation, *e.g.,* performance,

[*] corresponding author arzu.coltekin@fhnw.ch

innovation and the well-being of the individual team members (Brodbeck, 1996). The processes of the IPO model are based on the fact that goal-oriented tasks are developed in teams. McGrath (1984) later provided another model that specifies the processes and tasks with his "group task circumplex" (McGrath, 1984) (Figure 2). McGrath's models are extensively used in *computer supported collaborative work* (CSCW) research to examine processes and to define tasks. Identifying the characteristics of a task is crucial to understand and predict group effectiveness (Straus, 1999). According to McGrath (1984), collaborative tasks can be linked to four processes: *generate*, *select*, *negotiate*, and *execute*. Each of these processes is divided into two additional task categories, thus a total of eight categories are arranged in a circumplex (Figure 2) where horizontal axis indicates the extent to which the task includes cognitive/behavioral performance requirements, and the vertical the extent to which the task is cooperative or conflict prone.

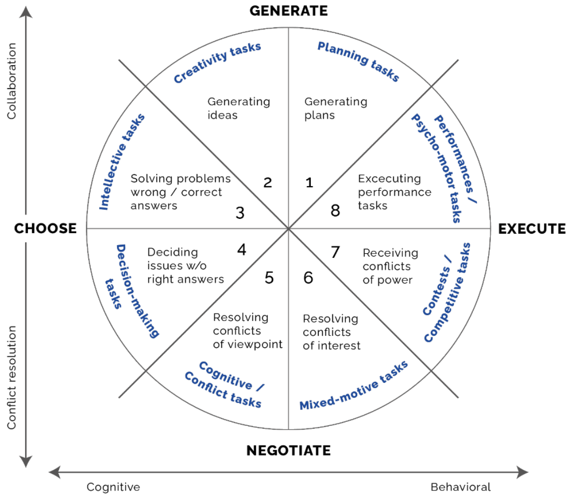
GENERATE
Collaboration
Creativity tasks
Planning tasks
Generating ideas
Generating plans
Intellective tasks
Performances / Psycho-motor tasks
Solving problems wrong / correct answers
2
1
Excecuting performance tasks
3
8
CHOOSE
EXECUTE
Deciding issues w/o right answers
4
7
Receiving conflicts of power
Decision-making tasks
5
6
Contests / Competitive tasks
Resolving conflicts of viewpoint
Resolving conflicts of interest
Conflict resolution
Cognitive / Conflict tasks
Mixed-motive tasks
NEGOTIATE
Cognitive
Behavioral

Figure 2. McGarth's (1984) group task circumplex (figure redrawn).

1.2. Collaboration in virtual teams

Remote collaboration has become more feasible in the last two decades than ever before due to the recent technological developments. Soon after Greif & Cashman (1988) coined the phrase CSCW decades ago (Johansen, 1988), the concept *groupware* was born. Johansen (1988) places groupware in a space-time matrix which provides a useful framework in studying collaborative MR (Johansen, 1988) (Figure 3).

	Synchronous (Same time)	Asynchronous (Different time)
Colocated (Same place)	Face to face interactions	Continuous task
Remote (Different place)	Remote interactions	Communication & Coordination

Figure 3. Johansen's (1988) space-time matrix that classifies the key contexts for tech-supported collaboration (figure redrawn).

As Figure 3 demonstrates, Johansen's (1988) model treats four spatiotemporal contexts: whether the collaboration takes place in one place (colocated) or in different places (remote), and whether the collaboration takes place simultaneously (synchronous) or at different times (asynchronous) (Johansen, 1988). Today, even when geographically separated by long distances, a considerable number of people can reliably hold location distributed meetings in video and audio conferences, and process online documents simultaneously using online drawing boards or text processing software. Such services enable building virtual teams and such virtual teams can mean time savings, reduced transportation costs, and in a larger scale, opens the global labor market, and consequently working in virtual teams can increase productivity, provide access to globally distributed experts, and create environmental benefits (Cascio, 2000). However, the lack of physical presence impairs sharing of large portions of essential in-situ verbal, non-verbal, visuospatial, tactile and other sensory information in current tools used by virtual teams. These elements of situational awareness are prerequisites for the quality of the communication, and for the coordination of cooperation. Such shortcomings introduced by *a lack of presence* can reduce trust within a team, which is an essential factor for cooperation (Landy and Conte, 2016). Assuming the technical and HCI challenges are overcome, XR holds enormous potential for solving many such problems experienced by virtual teams today as they can mimic reality and create a better sense of presence (Billinghurst and Kato, 1999; Chenechal et al., 2016).

1.3. Digital twins and extended reality

The digital twin concept involves a virtual representation of a physical phenomenon, which is often (but not always) visuospatial, conveying information about the phenomenon's appearance and status digitally. Ideally, changes to the phenomenon are synchronized with the digital twin, and *vice versa*. Digital twins are being adopted in many contexts, e.g., industry, financial management, e-learning, virtual tours, shopping and social interactions on the Internet (El Saddik, 2018). Remote support with MR has the potential to make work more efficient, which in turn should have a positive impact on costs (Henderson and Feiner, 2009). However, little has been reported on the benefits of digital twins with respect to MR (Ens et al., 2019), thus fundamental questions on the advantages and disadvantages of the use of a digital twin in MR versus traditional technologies such as videoconferencing remain unanswered. Furthermore, there is a lack of research on the way instructions are visualized on a physical counterpart of a digital twin used in collaborative XR. XR promises the benefits of a shared visuospatial context, and can contain cues that allow inferring the remote partner's actions, which are known as *awareness cues* (Piumsomboon et al., 2017). These cues can be especially useful especially if a digital twin is employed. Given the realistic representations of 3D objects and possibilities to interact with them, a collaborative XR environment should deliver such awareness cues much more naturally than current state-of-art digital collaboration tools, and unlock exciting new possibilities, especially for remote teams (Bai et al., 2020; Ladwig and Geiger, 2018; Orts-Escolano et al., 2016). However, despite its promise, there are still many challenges attached to collaboration using XR. As mentioned earlier, some of them are technology (specifically, HCI) challenges, while others are human and social psychology challenges, e.g., individual and group differences in visuospatial cognition such as age, expertise, anxiety, spatial abilities (Çöltekin et al., 2017; Lokka and Çöltekin, 2020; Thoresen et al., 2016). Among the different types of XR, MR stands out as a collaboration environment because the virtual objects can be spatially registered to their (hypothetical or real) physical locations, enabling a range of field applications that are otherwise costly and complex (Çöltekin et al., 2020b, 2020a). For example, medical interventions in the field in the case of disasters, remote operation of complex vehicles with a non-expert in the field, infrastructure maintenance and similar tasks would benefit from real-time MR applications. Given the background summarized above, we explore a digital twin supported collaborative MR solution collaboration in a remote setting. A specific goal in this study is to explore the advantages and disadvantages of a responsive mixed reality digital twin compared to a 2D desktop scenario for remote assistance.

2. OUR EXPERIMENT

We developed a scenario based on interviews with experts at the Swiss Federal Railways (SBB), in which a remote expert assists a technician in the field for repair and maintenance on a task that requires visuospatial communication. Even though this publication features a qualitative exploratory study, we will use the terminology for controlled studies in this section, because we consider the study described here for coherence and readability, as the first phase of a larger user experiment. With that in mind, our independent variables were Microsoft's *Remote Assist (RA)* app to represent the baseline video-communication in collaboration, and two digital twin implementations supported by mixed reality: *digital twin (DT)*, and a *digital twin with first-person perspective (DTF)*. As dependent variables, we collected response time, number of words for verbal communication, number of deictic word use i.e., terms like *here, there* or *that*, which can be successfully used if the mutual situational awareness is high (Gutwin and Greenberg, 2004), and number of recognized errors. We also conducted semi-structured qualitative interviews including questions on subjective experience, recommendations for improvements and preference, and administered the *system usability scale* (SUS) (Brooke, 1996) in for the three prototypes.

Despite the exploratory (i.e., *not* confirmatory) nature of the study, based on the literature review and iterative interim usability tests with a small group of participants, we expected that: 1) The DT and DTF should require less time to complete the test task than the RA; 2) With the DT and DTF, the participants should need a smaller number of words for communication than the RA; 3) DTF should allow the most frequent use of deictic utterances; 4) Confirmation of the technician by visuospatial means ('virtual confirmation') leads to increased error detection by the expert with DT and DTF in comparison to RA; 5) Expert users may express higher satisfaction with the DT and DTF than the RA in SUS; 6) DTF should be overall superior to both RA and DT since it provides additional relevant information in comparison to both.

2.1. Participants

We recruited 18 participants (6 women, 12 men, 20-29 years old), who were students or graduates of business or computer science programs. This educational profile was mostly due to convenience sampling as pandemic (covid19) made it more complicated to conduct user experiments. However, the tasks also require certain level of expertise, so we believe the participants' educational profile might match reasonably well with the target population. For each run, two participants formed a virtual team (*i.e.,* 9 pairs worked in collaboration). None of the participants had previous experience with MR.

2.2. Materials: Prototype implementations

The prototype development from the first draft to the final version were based on the concept *bodystorming* (Schleicher et al., 2010) and agile user-centered design methods. We used the Unity 3D engine platform with LTS-Version 2019.2.21f1 to develop virtual visualizations and interactions for MR. In addition, the Mixed Reality Toolkit (MRTK) version 2.3.0 from Microsoft was integrated. The MRTK provides basic building blocks for the development of MR applications. These include pre-built interaction systems or user interface elements. For specific object recognition, we selected the Vuforia Engine, which allows the HoloLens (1st gen) to recognize the physical control box and extend it in MR with virtual information. For the communication between two HoloLens headsets, Photon Unity Network (PUN) was used for multiplayer application development. The context of the study is a remote support for the operation of a control box (Figure 4), and it involves two participants working at different workplaces: the physical control box on the side of *the technician,* and a remote workplace for *the expert* with a digital twin of the control box.

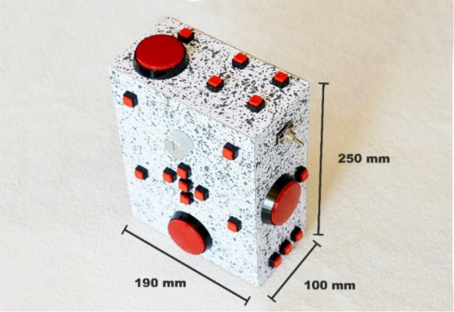

Figure 4. Physical control box with the dimensions.

To represent the complexity of real-world engineering operations, we first constructed a physical control box with a total of 30 unlabeled switch buttons for the purposes of experimentation. We installed pushbuttons, toggle switches and rotary switches. More specifically, the control box was fitted with 23 small, rounded pushbutton switches, three large round switches, three toggle switches and one rotary switch. The buttons are installed on all sides, except on the bottom and the back. The surface of the box has a high-contrast irregular texture, which allows for optimal image tracking of the box. After completing the physical prototype, we built three virtual prototypes (*remote assist*, *digital twin*, and *digital twin with first-person perspective*), and conducted an exploratory user study to get first insights on user performance, preference and usability of DT, RA and DTF. We detail these prototypes and the exploratory user study in the following sections.

Remote Assist (RA). The RA is a video conference system developed by Microsoft for MR applications that can facilitate communication between a HoloLens and a computer. The communicating person (i.e., "expert" in our scenario) at the computer can freeze and edit the current view of the HoloLens as an image. In doing so, arrows can be placed in the frozen image. The visualizations are shown to the receiving person (i.e., "technician" in our scenario) on the HoloLens. In addition, the face of the expert can be displayed on the HoloLens via video transmission. Figure 5 illustrates the interaction between the expert and the technician in the RA condition.

Digital Twin (DT). In the case of the digital twin, the expert and the technician both wear a HoloLens (1st gen). They cannot see each other but can communicate with each other via speech (Figure 6). The expert sees the digital twin of the physical control box in front of her or him. The technician stands in front of the physical control box. Through the digital twin, the expert transmits *visual instructions* to the technician, and these visual instructions are displayed on the physical object as annotations via MR. The technician performs what she or he is instructed on the physical control box.

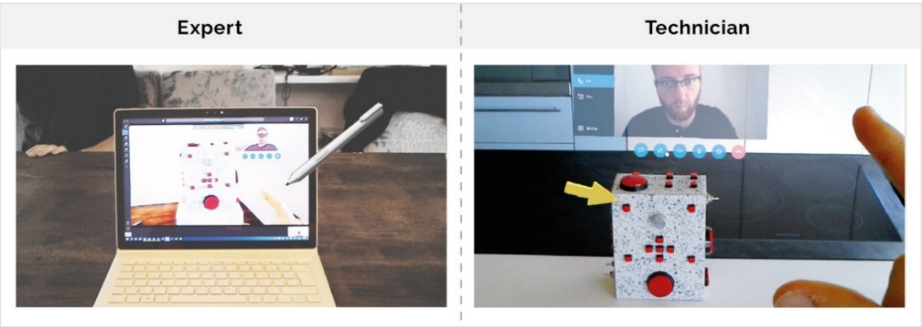

Figure 5. Role-specific perspectives of the RA (photo use permitted by Louis Baumgartner).

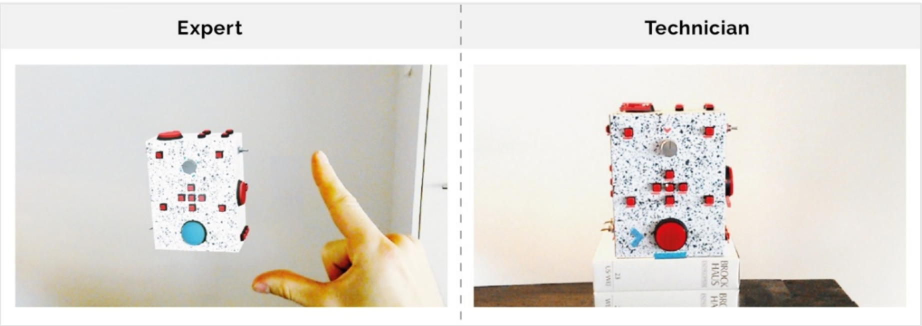

Figure 6. Role-specific perspectives of the DT

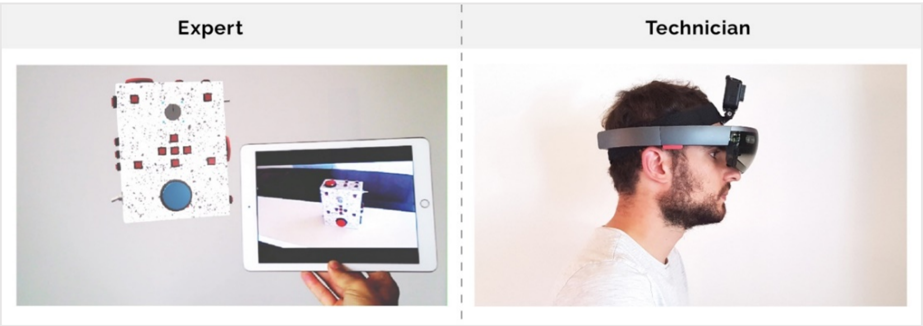

Figure 7. Role-specific perspectives of the DTF (photo use permitted by Luca Brägger).

Digital Twin with first-person perspective (DTF). For the DTF, a GoPro Hero5 was mounted on the technician's head in addition to the HoloLens (1st gen) so that the expert could see the technician’s perspective via a video transmission on an Apple iPad (5th generation) (Figure 7). Like the DT condition, in the DTF, the expert gives visual instructions to the technician using the digital twin and the technician executes them on the physical control box. With the DTF, the expert can additionally check the technician's actions at any time through the iPad.

Visual annotations. Available in all prototypes, we enabled remotely controlled visual annotations for the expert to better guide the technician. These involve a) highlighting the task relevant areas, b) pointing at the specific button with an arrow that also shows the direction of movement. Virtual red bars appear as under the interactive objects, indicating that they can be selected or manipulated, as soon as Vuforia’s image tracking correctly detects the box. At the concept development and design phase, we considered various visualization options for visualizing interactions (Figure 8, left). Based on the qualitative feedback in the interim user-tests, we implemented animated arrows, a transparent box to outline interactive areas, and several virtual markings to indicate that users need to air-tap (Fig 8, right). The buttons and the visual annotations change from red to blue (or blue to red) depending on their state, where blue means selected (Figure 7). We animated button-press and implemented a clicking sound as feedback mechanisms.

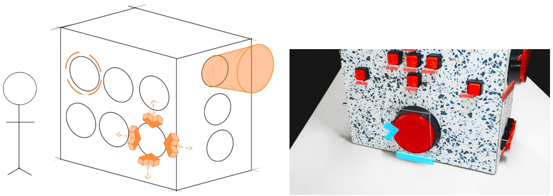

Fig 8. Left: Design options considered for visual annotations and highlighting. Right: Animated arrow shows the button of interest and direction of movement.

2.3. Procedure

Upon arriving the laboratory, participants signed a consent form, and received a brief training on how to operate the Microsoft HoloLens. For the collaborative exercises, we paired them in teams of two. The overall scenario is that in a remote support team, the remote expert instructs the technician in which order to press which button physically on a control box.

Training session. To ensure that participants' operating ability was sufficient for the experiment, an interaction test was performed. The pair was given the task of alternately instructing and confirming as many buttons as possible without comments within 30 seconds, which fits to McGarth's (1984) classification as a psychomotor / performance task with some micro decision-making steps. The expert indicated one button at a time on the front side of the control box and waited till the technician confirmed it. The experts have been told to only select each button once. In the case of DT and DTF, the technician confirmed the highlighted button in MR, and in the case of the RA, she or he physically pressed the button. Participants received the same amount of time for training.

Main experiment. After the training session, participants were randomly assigned to the roles of the expert and the technician. The expert was told to teach the technician to press a specific sequence of buttons on the control box. The technician, on the other hand, had the task of interpreting the instruction correctly and to physically press the buttons on the control box. For each of the three prototypes (DT, RA, DTF) we designed comparable but slightly different tasks (modified order in the sequence but similar in level of difficulty, sequence length, changes in perspective *etc.*) to counter against the learning effect. At the beginning of the task, instructions were shown to the expert, placed as images with instructions in front of her or him on the table. Whenever a task was completed, the prototype changed (in rotated order), and the pair received new instructions. The technician was instructed to produce two errors *on purpose* (by pressing a pre-defined 'wrong button') to check whether the expert recognizes this error and how she or he reacts to it. Because of the hardware limitation with Microsoft HoloLens (1st gen), the participants could not directly manipulate the virtual objects with hand interaction. Therefore, the participants needed to perform two steps: First, the haptic pressing of the button, and then the virtual confirmation with the HoloLens air-tap gesture. To emulate the voice communication supported in remote collaboration environments, we allowed participants to communicate verbally during the experiments to clarify the instructions. Each session (RA, DT, DTF) was concluded with a post questionnaire including interview questions and the SUS.

3. RESULTS

Task duration. The average task durations are summarized in Figure9. Participants (n=18, 9 pairs), on average, took about half a minute longer with the DT than with the other two solutions While participants were fastest with the DTF, there was only a slight difference (two seconds) between the RA and the DTF.

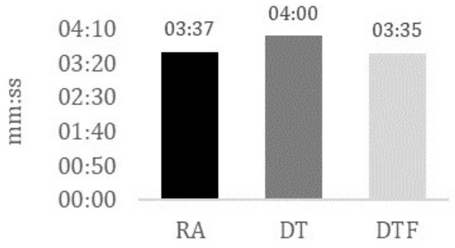

Figure 9. Participants' average task durations with the RA, DT and DTF.

Number of deictic words used by the participants. Frequent use of deictic words, *i.e.,* words that take context-dependent meaning such as 'here', 'there', 'that', can indicate greater situational awareness (Bermes, 1999; Gutwin and Greenberg, 2004), thus we also measured number of deictic words used by the participants as they worked with each solution. Figure 10 shows the average number of deictic uses for each prototype. The RA has the highest average value.

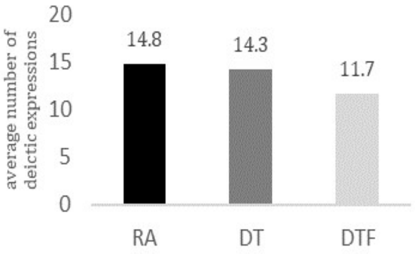

Figure10. Average number of deictic uses.

Number of words in verbal communication. Abundant verbal communication may indicate poor visuospatial support, thus we measured the number of words used by the participants in the role of the expert. Figure 11 summarizes the number of words used by *the expert* for verbal communication.

Number of errors noticed by the expert. Since the 'technician' was instructed to commit two international errors, we observed how often the 'expert' was able to catch them. We assumed that the more errors the expert can detect, the better the system facilitates communication. The average number of errors noticed by the participant who took the role of the expert is displayed in Figure12. The average value of the noticed errors is highest with the DT, closely followed by DTF, where the RA appears to facilitate error detection the least.

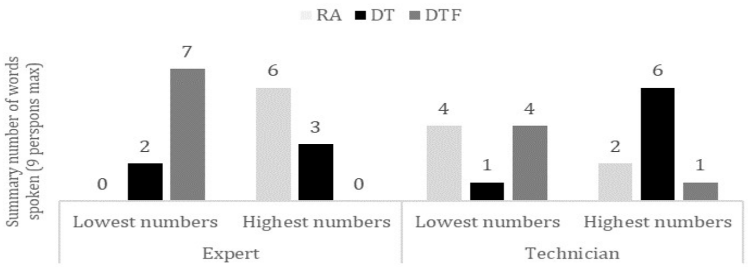

Figure 11. Number of spoken words per participant, and per system. With the DTF, seven out of nine experts used the *lowest* number of words, closely followed by the DT (six out of nine), whereas the RA led to the *highest* number of words spoken (six out of nine experts).

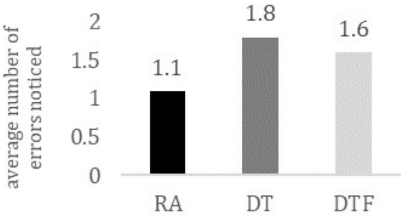

Figure12. Average number of errors noticed by the expert.

SUS questionnaire. The outcomes from the SUS questionnaire is summarized in Figure13. The experts were most satisfied with the DT, and least with the RA. For the technician, however, the RA had the highest subjective value, closely followed by the DTF. The highest possible SUS score is 100, and if the SUS score is below 68, the tested system needs improvements (Brooke, 1996). Our results suggest that the DT and DTF systems are usable, whereas the RA is not satisfying at least one user group.

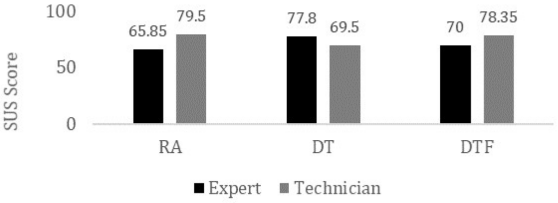

Figure13. Summary of SUS questionnaire.

Interview / post-experiment questionnaires. Following the exploratory questions quantitatively summarized above, we conducted qualitative, semi-structured interviews. Key observations from these interviews are as follows:

- Most participants share the opinion that the DTF supported mutual understanding best. Specifically, the expert was able to intervene preventively in case of doubt due to the additional first-person perspective from the technician, and the technician felt more secure as a result.
- All participants found the visual markings (annotations) of the buttons on the physical and virtual control box helpful in the collaboration irrespective of their role.
- More than half of the participants mentioned that there was difficulty distinguishing whether the expert or the technician selected the highlighted buttons when multiple buttons were selected.
- All participants who took the role of the expert mentioned that with DTF there was too much information at one time which involved controlling the iPad, reading instructions, and carrying out the instructions on the digital twin.
- Most technicians mentioned that the expert's face in video transmission was more personal, but not necessarily helpful for the task.
- Many technicians suggested that the confirmation of a physically pressed button should be done without an additional air-tap.
- Another suggestion of improvement is to unite all processes such as video transmission and the tasks to only one medium as multiple devices were distracting.

In the preference question, 5/9 experts preferred the DTF, and an equal number of, 5/9, technicians preferred the RA. Only 3/9 technicians preferred the DTF, and only 1/9 expert preferred RA. DT was preferred by 3 experts and 1 technician.

4. DISCSSION

About the digital twin prototypes. Overall, given the arguments in the relevant work about how being able to share perspective and manipulate objects directly improves feeling or presence (Billinghurst and Kato, 1999; Chenechal et al., 2016; Landy and Conte, 2016), we expected that both digital twin implementations (DT and DTF) would be superior to Remote Assist (RA), and DTF would be superior to DT. The latter is because with the DTF, the expert can see what the technician sees (thus, a shared perspective is enabled); not only her or his face, or not only the object of interest. Furthermore, participants should need fewer verbal instructions to coordinate the tasks with our digital twin prototypes, therefore we expected that participants would use overall less time finish the task. The preliminary results show that participants took longest with the DT, whereas DTF and RA have similar task times (Figure 9). However, it is important to note that a) we have a small sample of participants, b) speed should not be interpreted on its own: If the expert did not notice the errors committed by the technician, they may be overall faster in finishing the task, which does not imply "better performance". The more the expert had to correct the technician, the time needed, and the number of words both increased. Given that need for frequent verbal communication might suggest poor visuospatial support, number of words matter. However, we see that the expert notices the least number of errors with RA (Figure 12), RA forces the participants to use the highest number of words (Figure 11), and also generates the highest number of deictic word use (Gutwin and Greenberg, 2004) (Figure 10). Taken together, participants' speed with the RA is a consequence of RA's inability to facilitate the detection of errors, and digital twin solutions both might to be superior to RA as we speculated. As a next step, controlled laboratory studies should be conducted to confirm this observation.

In general, a retrospective analysis of the videos suggests that the operation with the HoloLens (1st gen) would benefit from a bit more practice by the participants. In addition to physically pressing the buttons, the technician had to confirm them virtually, which interrupted the workflow. When participants used the system a second time, they needed less time for this task type, (though it should be noted that we randomized the order of the comparted porotypes, and this training effects should not bias our comparative statements). We believe that better hand tracking solutions will help with this problem, as it has been shown previously that proper hand tracking allows for more intuitive interaction in MR (Yeo et al., 2015). In our scenario, hand tracking would have a particular impact on the technician, who could perform the virtual confirmation at the same time as the physical pressing of the button. For the expert, in turn, the interaction with knobs could be significantly simplified, since with the current implementation a selection via air-tap and subsequently a gesture is required for turning. Precise rotations seem to require further improvement too, where hand tracking and gesture recognition could also be useful (Huesser et al., 2021). When it comes to subjective evaluations, we expected that DTF would be the most popular solution. However, the SUS scores strongly suggest *the role* that the participant played (thus their context and specific tasks) led to opposing preferences (Figure 13). The experts rated the DT with the highest score, followed by the DTF while the technicians rated the RA best, closely followed by the DTF. Technicians' rating of the DT is bordering the 'usable' threshold (69.5), and the experts' rating of the RA is even worse

(65.85). Reasons for the high SUS scores for DT by experts may be due to its simpler content (contains less information), which they may have deemed easier for an inexperienced user to process. The high rating of the RA by the technicians can be explained by the simpler operability. Both groups might also be more familiar with the RA as it essentially a video-based system which is commonplace. The almost equally high rating of the DTF, on the other hand, is possibly due to an assurance in communication. As some of the technicians stated in the interviews, with the RA, they were not really sure whether the expert could view their actions and give the right feedback. In sum, the video transmission of the first-person perspective, combined with an additional confirmation for the expert, can increase the possibilities of control. This was also evident in subjective opinions of the participants: In the interviews, the experts strongly preferred the DTF, while the technicians often preferred the RA. The experts justified their opinions that the DTF gave them the most comprehensive access to information about the work environment. From the technicians' perspective the RA seemed easier to operate, because it required fewer interactions, because it did not require virtual confirmation. In addition, some of the technicians mentioned that the visual representation in RA was tidier but somewhat less precise, than with the DT and DTF.

Verbal communication in remote collaboration. We expected the RA would require higher number of words than both digital twin systems, which our results suggest might indeed be the case (Figures 10 and 11). Since participants worked with two kinds of digital twin implementations (DT and DTF), despite the systematic rotation, the second time they used a digital twin solution, they may have used needed fewer words (which would help with both DT and DTF, to a small degree). However, the primary reason—as deduced from the interview answers and the transcripts—seems to be about coordination-related expressions. In the case of the RA, the expert was asked to instruct the technician to look at a certain side of the control box, or to adjust the viewing angle slightly before she or he was able to edit the still image. With DT and DTF, the expert's field of view is independent of that of the technician. Consequently, the advantages mentioned by Billinghurst already in 1999 regarding the independent field of view enabled by MR appear to be still relevant in this scenario (Billinghurst and Kato, 1999). Against our expectations, we see a higher number of deictic expressions in the RA (Figure 10). We assume that the number of deictic expressions is higher with the RA, because the RA has an overall larger number of spoken words (Figure 11). For both roles, in the interviews, participants stated a higher dependence on verbal communication was evident for the RA. To better establish the cause-effect relationships between the deictic word use vs. overall word use, more data is needed, thus a future controlled study could confirm our assumptions.

Number of errors noticed by the expert. Since the goal in the scenario was to enable collaborative work where an expert could guide a technician and give feedback, as well as check their work, it is very important that the errors that are committed by the technician are noticed by the expert. We expected that the expert would notice a higher number of errors on the DT and DTF than in the RA, and our results suggest that might be the case (Figure 12). Interviewees pointed out that one of the main reasons for this is that the physical actions of the technician were partially missed in the video transmissions in the RA condition. Just one or two seconds of inattention were enough to miss the action. In the case of the digital twins, the experts could always manage to find out which button was pressed at least partly because of the "awareness cues" (Piumsomboon, et al., 2017) we implemented i.e., in this case, the visual annotations. One reason participants performed worse with the DTF than the DF is, according to interview statements, an information overload consisting of simultaneous display of the instructions on the iPad and on the digital twin. This probably affected the concentration of the experts at the DTF more than at the DT. In general, there was praise for the fact that the video transmission made it possible to have a preventive influence on errors made by the technician. A solution to counter against this would be to display the first-person perspective and the guidance both virtually in the MR.

Another design issue that emerged was about the use of color: For both roles, it became difficult as soon as several buttons were selected at the same time, as the user cannot easily determine which buttons have been pressed by the expert or by the technician. A future design should ideally have a role-specific color coding and consider alternative solutions to annotate / label the actions by people who are in different roles.

5. CONCLUSIONS AND OUTLOOK

We examined the use of XR in collaborative work in remote setups supported by responsive, visually annotated digital twins. Our qualitative observations and usability examinations of the prototypes we suggest that XR and digital twin combination supported by visual annotations is in general a promising approach with potential to improve collaboration processes, however, design decisions remain important *e.g.,* it makes a difference if a button is highlighted using an individual color based on role. Our insights allow us designing well-informed to controlled experiments in near future to quantitatively assess XR implementations supported with real time visual annotations on digital twins of object of interest. Another clear next step is implementing more advanced digital twin prototypes that interact with the state-of-the-art MR systems based on the lessons learned in this preliminary set of experiments; and possibly combine the digital twin solution with gaze support which might be especially useful when people's hands are busy. We believe our efforts in better understanding how to design collaborative MR systems may be relevant and useful in many domains and for many interdisciplinary projects.

ACKNOWLEDGEMENTS

We are grateful to our participants and reviewers. This project was partially supported by the SBB Center of Competence for Extended Reality and Innosuisse project 50321.1.

REFERENCES

Appley, D.G., Winder, A.E., 1977. An evolving definition of collaboration and some implications for the world of work. The Journal of Applied Behavioral Science 13, 279–291.

Bai, H., Sasikumar, P., Yang, J., Billinghurst, M., 2020. A User Study on Mixed Reality Remote Collaboration with Eye Gaze and Hand Gesture Sharing, in: Proc. of the CHI '20: ACM, Honolulu HI USA, pp. 1–13.

Benko, H., Ishak, E.W., Feiner, S., 2004. Collaborative mixed reality visualization of an archaeological excavation, in: Third IEEE and ACM International Symposium on Mixed and Augmented Reality. IEEE, pp. 132–140.

Bennett, L.M., Gadlin, H., 2012. Collaboration and team science: from theory to practice. Journal of Investigative Medicine 60, 768–775.

Bermes, C., 1999. Arnim Regenbogen/Uwe Meyer: Wörterbuch der philosophischen Begriffe.

Billinghurst, M., Kato, H., 1999. Collaborative mixed reality, in: Proc. of the First International Symposium on Mixed Reality. pp. 261–284.

Brodbeck, F., 1996. Work group performance and effectiveness: Conceptual and measurement issues. Handbook of work group psychology 285–315.

Brooke, J., 1996. SUS-A quick and dirty usability scale. Usability evaluation in industry 189, 4–7.

Camarihna-Matos, L.M., Afsarmanesh, H., 2008. Concept of collaboration, in: Encyclopedia of Networked and Virtual Organizations. IGI Global, pp. 311–315.

Cascio, W.F., 2000. Managing a virtual workplace. Academy of Management Perspectives 14, 81–90.

Chenechal, M.L., Duval, T., Gouranton, V., Royan, J., Arnaldi, B., 2016. Vishnu: virtual immersive support for HelpiNg users an interaction paradigm for collaborative remote guiding in mixed reality, in: the 2016 3DCVE, IEEE, Greenville, SC, USA, pp. 9–12.

Çöltekin, A., Brychtová, A., Griffin, A.L., Robinson, A.C., Imhof, M., Pettit, C., 2017. Perceptual complexity of soil-landscape maps: a user evaluation of color organization in legend designs using eye tracking. International Journal of Digital Earth 10, 560–581.

Çöltekin, A., Griffin, A.L., Slingsby, A., Robinson, A.C., et al., 2020a. Geospatial information visualization and extended reality displays, in: Manual of Digital Earth. Springer, Singapore, pp. 229–277.

Çöltekin, A., Lochhead, I., Madden, M., et al., 2020b. Extended reality in spatial sciences: a review of research challenges and future directions. ISPRS International Journal of Geo-Information 9, 439.

El Saddik, A., 2018. Digital twins: The convergence of multimedia technologies. IEEE MultiMedia 25, 87–92.

Ens, B., Lanir, J., Tang, A., Bateman, S., Lee, G., Piumsomboon, T., Billinghurst, M., 2019. Revisiting collaboration through mixed reality: The evolution of groupware. International Journal of Human-Computer Studies 131, 81–98.

Fairchild, A.J., Campion, S.P., García, A.S., Wolff, R., Fernando, T., Roberts, D.J., 2016. A mixed reality telepresence system for collaborative space operation. IEEE Transactions on Circuits and Systems for Video Technology 27, 814–827.

Giraudeau, P., Olry, A., Roo, J.S., Fleck, S., Bertolo, D., Vivian, R., Hachet, M., 2019. CARDS: a mixed-reality system for collaborative learning at school, in: Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces. pp. 55–64.

Gutwin, C., Greenberg, S., 2004. The importance of awareness for team cognition in distributed collaboration., in: S Team Cognition: Understanding the Factors That Drive Process and Performance. American Psychological Association, Washington, pp. 177–201.

Henderson, S.J., Feiner, S., 2009. Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret, in: 2009 8th IEEE International Symposium on Mixed and Augmented Reality. IEEE, pp. 135–144.

Huesser, C., Schubiger, S., Çöltekin, A., 2021. Gesture Interaction in Virtual Reality: A Low-Cost Machine Learning System and a Qualitative Assessment of Effectiveness of Selected Gestures vs. Gaze and Controller Interaction, INTERACT 2021, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 151–160.

Johansen, R., 1988. Groupware: Computer support for business teams. The Free Press.

Ladwig, P., Geiger, C., 2018. A literature review on collaboration in mixed reality, in: International Conference on Remote Engineering and Virtual Instrumentation. Springer, pp. 591–600.

Landy, F.J., Conte, J.M., 2016. Work in the 21st century: An introduction to industrial and organizational psychology. John Wiley & Sons.

Lokka, I.E., Çöltekin, A., 2020. Perspective switch and spatial knowledge acquisition: effects of age, mental rotation ability and visuospatial memory capacity on route learning in virtual environments with different levels of realism. Cartography and Geographic Information Science 47, 14–27.

Martinez-Moyano, I., 2006. Exploring the dynamics of collaboration in interorganizational settings. Creating a culture of collaboration: The International Association of Facilitators handbook 4, 69.

McGrath, J.E., 1984. Groups: Interaction and performance. Prentice-Hall Englewood Cliffs, NJ.

McGrath, J.E., 1964. Social psychology: A brief introduction. Holt, Rinehart and Winston.

Ong, S., Shen, Y., 2009. A mixed reality environment for collaborative product design and development. CIRP annals 58, 139–142.

Orts-Escolano, S., Rhemann, C., Fanello, S., Chang, W., et al., 2016. Holoportation: Virtual 3D Teleportation in Real-time, in: Proc. UIST '16: The 29th Annual ACM Symposium on User Interface Software and Technology, ACM, Tokyo Japan, pp. 741–754.

Piumsomboon, T., Day, A., Ens, B., Lee, Y., Lee, G., Billinghurst, M., 2017. Exploring enhancements for remote mixed reality collaboration, in: SIGGRAPH Asia 2017 Mobile Graphics & Interactive Applications. pp. 1–5.

Pulver, Y., Merz, C., Koebel, K., Scheidegger, J., Cöltekin, A., 2020. Telling engaging interactive stories with extended reality (XR): Back to 1930s in Zurich's main train station. ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences 5.

Schleicher, D., Jones, P., Kachur, O., 2010. Bodystorming as embodied designing. Interactions 17, 47–51.

Straus, S.G., 1999. Testing a typology of tasks: An empirical validation of McGrath's (1984) group task circumplex. Small Group Research 30, 166–187.

Thoresen, J.C., Francelet, R., Coltekin, A., Richter, K.-F., Fabrikant, S.I., Sandi, C., 2016. Not all anxious individuals get lost: Trait anxiety and mental rotation ability interact to explain performance in map-based route learning in men. Neurobiology of Learning and Memory 132, 1–8.

Yeo, H.-S., Lee, B.-G., Lim, H., 2015. Hand tracking and gesture recognition system for human-computer interaction using low-cost hardware. Multimedia Tools and Applications 74, 2687–2715.

Zhou, Z., Márquez Segura, E., Duval, J., John, M., Isbister, K., 2019. Astaire: A Collaborative Mixed Reality Dance Game for Collocated Players, in: Proc. of the CHI PLAY '19: The Annual Symposium on Computer-Human Interaction in Play, ACM, Barcelona Spain, pp. 5–18.