RemoteCoDe: Robotic Embodiment for Enhancing Peripheral Awareness in Remote Collaboration Tasks

MOSE SAKASHITA, E. ANDY RICCI, JATIN ARORA, FRANÇOIS GUIMBRETIÈRE, Cornell University, USA

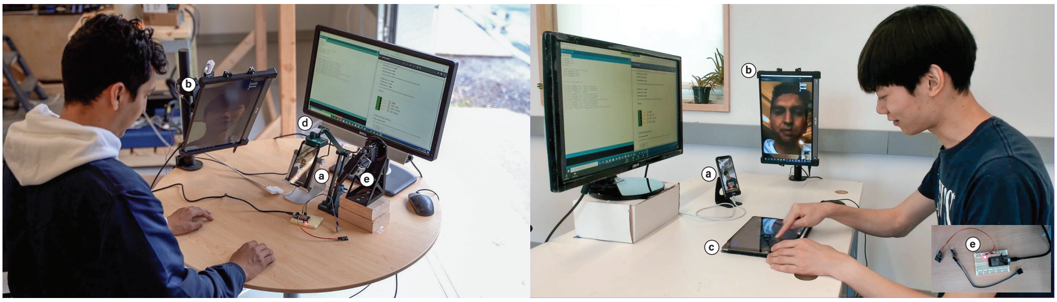

Fig. 1. A typical *RemoteCoDe* debugging session. At both sites, a TrueDepth camera (a) tracks the local user's head, and an embodied proxy (b) renders the remote user's locus of attention. The user can observe and point to the circuit using a tablet (c) connected to a document camera (d) and an actuated laser pointer (e).

Many collaborative design activities are centered around a shared artifact such as a low fidelity model of a building, or a circuit implemented on a breadboard. In such settings, collocation is extremely valuable as collaborators can easily infer each other's focus of attention through peripheral awareness. These cues are often lost in traditional video conferencing systems such as Zoom. In this paper, we present an embodied remote presence system designed to support design activities that involve a physical artifact by enhancing peripheral awareness. In our system, a remote user's focus of attention is tracked by an iPhone's TrueDepth camera and rendered locally using an articulated display. Our implementation can track the wide range of head movements that occur when one switches attention between a physical artifact, a laptop, and their collaborator. To support discrepancies between local and remote workspace layouts, head movements are remapped as robotic movements to correspond to these key elements in the local user's space. We report on the results of an evaluation characterizing the system's remapping accuracy and its ability to support peripheral awareness of a remote participant's locus of attention.

CCS Concepts: • **Human-centered computing → Collaborative and social computing devices**.

Additional Key Words and Phrases: Robotic Embodiment; Remote Collaboration; Design;

1 INTRODUCTION

Collaborative design activities are often centered around physical artifacts. Depending on the design activity, this can be the model of a building, paper crafts, carving artwork, or a new circuit to be debugged and evaluated. In a typical setting, collaborators are seated around a table and divide their attention between the design artifact under review, at least one laptop supporting measurements and information foraging, and of course their collaborators. Although these activities involve complex sets of tools and configurations, people can easily work together when they are present in the same space. This is because the physical presence of a partner affords peripheral awareness to inform where the partner's attention is and what they are doing. This peripheral awareness allows collaborators to coordinate actions and manage coupling to achieve a shared

Author's address: Mose Sakashita, E. Andy Ricci, Jatin Arora, François Guimbretière, Cornell University, USA, Ithaca, NY, USA, ms3522@cornell.edu,ericci@cs.cornell.edu,ja787@cornell.edu,fvg3@cornell.edu.

task [23]. For example, it is quite easy to know when your partner switches their focus from a breadboard to you as a request to start a face to face discussion.

While current video-conferencing services (such as Zoom [68] and Skype [57]) have made great strides in reliability and overall quality, they often fall short of supporting peripheral awareness of what other people are doing or where they are looking. For example, when working on a hands-on task, it is challenging to know what the other person is doing without looking directly at the video stream. It often requires a user to pause their task and switch their focus from the physical artifact to the computer monitor that displays the other person [7].

Previous research in telepresence systems have presented solutions in an attempt to address some of these constraints. Video conferencing work such as VideoWhiteboard [60] and Clearboard [27] have proposed the use of a large display to project both a remote collaborator and a shared drawing to provide awareness. However, this solution may not work well for more complex scenarios in which a physical artifact is involved as users have to look away from the display to work on a physical prototype, which makes it difficult to perceive their collaborator's presence. Telepresence robots [29, 43, 59] or kinetic proxies [63, 66, 67] have been explored as promising approaches to address the lack of physical presence, but controlling these telepresence proxies add a distracting workload to the task at hand [49, 62]. Thus, this type of robotic proxy is not suited for design activities in which collaborators are required to divide their attention among multiple areas to accomplish a design task. To reduce the cognitive load for control, several systems incorporate an automatic control mechanism into a telepresence robot [1, 38, 56], but these systems are optimized for face-to-face conversations and not for offering ability to share multiple task areas or flexibility of having a different configuration in each location.

In this paper, we present *RemoteCoDe*, a robotic proxy designed to support scenarios where remote collaborators share a wide-spread task area for physical artifacts (e.g., breadboard). Our system allows users to be aware of their remote partner's locus of attention through an articulated display without the need for explicit control. To accommodate a wide range of head movements, we use a mobile motion camera to capture the 3D mesh and texture of the remote user's face as well as their head rotations. Based on this information, our robotic proxy is able to display a frontal shot of a remote user's face on a rotating and tilting display synchronized to the remote user's head movement. To manage the fact that collaborators often have different workspace configurations, *RemoteCoDe* uses a framework that remaps remote users' head movements to the local configuration. We illustrate how our remapping framework can further enhance access to the remote space with an overview feature allowing a remote user to visually explore the local space. To use this feature, the remote user only needs to focus their attention on the overview window. While they are looking at this window, their head movement causes the robot to turn to the corresponding direction in the room. This provides a simple way to look around and search for a tool or part, for example, on a shelf.

We also report findings and design implications based on the evaluation of the system we conducted. Given the previous work supporting the benefits of automated control to lower cognitive load [38, 49, 62], our primary focus in our evaluation is to confirm the validity of the key features designed and implemented in the *RemoteCoDe* system. More specifically, we ran a user study to measure how accurately *RemoteCoDe* is able to express the locus of attention of a remote user and to investigate its ability to support peripheral awareness of this locus of attention. The first study found that participants could perceive three general attention areas (workbench, laptop, and partner) with a 99.7% accuracy rate, and nine detailed areas with a 93.5% accuracy rate. A second study show that participants were able to take advantage of peripheral awareness to infer their partner's locus of attention while focusing on an other task.

2 RELATED WORK

We use the framework presented by Buxton et. al. [7] to conceptualize previous work in this area. The framework introduces two shared spaces: *person space* and *task space*. The *person space* is a "collective sense of copresence between/among group participants", this includes member's facial expressions, bodily gestures, and gaze. The *task space* is defined as "copresence in the domain of the task being undertaken", in the case of hardware debugging this could be a shared view of the breadboard.

2.1 Video-based Systems

Besides standard videoconferencing systems [57, 68], a number of systems have investigated ways to provide a more accurate visual rendering of a remote user. Solutions based on video overlay such as the VideoWhiteboard [60] and Clearboard [27] are well adapted to whiteboard collaboration rendering both person and task spaces together on a single display but this may not support a design task in which a wide area of task space is required, such as bench work. Hydra [54] represents a remote user using a rigid camera/monitor pair. Other solutions proposed projecting life sized images on a wall (Office Portal [65]), on a curved display (MAJIC [36]) or on a multi-viewpoint directional display (MultiView [35, 41]). These systems render a person space very well supporting eye-contact communication but are not as effective with respect to a task space. Because of their reliance on visual cues they might not support strong peripheral awareness. Our work explores how embodiment can provide a similar sense of presence using a simpler configuration. Our system is probably closest to SharedARK [58]. SharedARK presents a person and task space on separate monitors, but this setup makes it challenging to distinguish whether the remote user is focused on the computer display to work on a task or if they are focused on the local user to establish eye contact [7]. Thus our system renders a person space by an articulated display to afford peripheral awareness as to the state of the remote collaborator.

2.1.1 Eye and Head Movements. Previous work has demonstrated the value of gaze tracking to identify the locus of attention [11, 12, 17, 21, 26] and provide peripheral awareness within a screen [3, 11, 31]. This is particularly important in settings where the users can access most information in a task space with minimal head movements as in collaborative tasks such as solving puzzles [8, 12, 40], pair programming [11], and image search [6]. Our work attempts to cover tasks in which the task space is larger, such as a desk or bench. In such a setting, head tracking is a good indicator of the current locus of attention and it can provide greater physical movement, which contributes to enhance the peripheral awareness. Additionally, gestural head movements contribute to interactions by signaling agreement/disagreement, enhancing communicative attention, communicating backchannel information [30], and indicating interest and empathy [25].

2.2 Physical Embodiment

Paulos and Canny's Personal Roving Presence [43] is one of the earliest contributions in this category and was followed by many others ranging from small robots placed on a meeting table (e.g. [55, 66, 67]), to near-human-size devices that move around in a physical space [43, 59]. These robots have been used to physically render a person space in various contexts such as conferences [50], workplaces[29, 63], home care [32], and education [18, 42]. One of the main drawbacks of these systems in support of design activities is the relatively high cognitive load they require, because the remote users have to control them manually [1, 38, 56, 62]. In our experience with Beam [43, 59] and Kubi [66, 67] this makes their integration into a design discussion very distracting , this is consistent with the findings of Rae et. al. [49].

The focus of our work is closely related to systems providing automatic user tracking such as MMSpace [38], MeBot [1] and implicit controlled kinetic robot [56]. MeBot [1] can be controlled with head rotations but requires a remote user to keep looking at a single screen during control, which limits the area of their activity. MMSpace [38] employs several fixed positions for looking at each user and a robotic proxy is controlled based on which user is looked at, but this system focuses on face-to-face conversations, and therefore, only renders the person space. Although MeBot [1] presents a task space, it is shared only on a single monitor through which a remote user controls a robot. In contrast, our work pursues a solution to share a widespread task space while rendering the person space through a robotic embodiment.

MMSpace [38] also requires both users to have more or less the same configuration. This setup does not function well in the context of a collaborative design. In reality, remote collaborators often have different workspace configurations. For example, the sizes and positions of monitors or the layout of a breadboard in a workspace may differ between locations. Our work seeks to demonstrate how the previous approaches can be extended to collaborative design by expanding their flexibility [38], and the range of head motions they support [1, 63]. Our work also accommodates different workspace configurations through automatic remapping.

2.2.1 Face Rendering on Robotic Proxy. Projecting a face on a flat monitor can be misleading in conveying detailed gaze direction, which is known as Mona Lisa effect [15]. This effect is also reported in a kinectic robot similar to our proxy and is even intensified by the complex combination of the rotations of the proxy and of the user's face and eyes displayed on the monitor [28]. Our face re-projection technique is designed to mitigate this ”double rotation” issue. ThirdEye [39] introduces an interesting approach to compensate the effect by affording eye-gaze awareness through a “third“ eye, but does not afford much peripheral awareness. Other work explored a curved face avatar to alleviate the effect [13, 34], but these systems require a user to fix their face in front of a camera. Our face rendering approach allows users to freely move their head.

2.3 Virtual and Augmented Reality

Here we focus on systems that are designed to support remote interactions requiring users to wear a VR or AR headset such as the *WearCom* [5]. Rendering a person space as a virtual avatar over physical objects in a see-through headset helps users be aware of their collaborators during remote collaboration [45–47, 61]. More recently fully VR based solutions have also been proposed [37, 44]. Room2Room [44] projects a life sized person on furniture in real life, while Holoportation [37] renders an entire environment captured by depth cameras in an AR headset. These VR/AR systems can render both task and person spaces in a fully digital manner and thus can be useful for some physical tasks allowing collaborators to shift focus or stay aware of each others' area of attention. While these solutions have great potential, they are often very resource intensive. We also note that current systems have a somewhat restricted field of view, which limits peripheral awareness. Systems requiring headsets mask users' faces preventing collaborators, both remote and co-located, from seeing each other's faces (but face tracking while using VR masks is an active area of research [4]). We propose possible alternatives to these solutions to foster better understanding of the pros and cons of each approach among the HCI community.

3 *RemoteCoDe* DESIGN

Our design goal for a remote collaborative system is to deliver a smooth design session for geographically distributed collaborators. To identify essential features for such systems, we use the special case of breadboard design review as this is a good representative example of a complex setting in which multiple focus areas must be handled seamlessly.

In a typical breadboard debugging session, collaborators shift their focus between (Fig. 1): the breadboard itself; one or more computer displays used for firmware programming, documentation or instrumentation and, of course, their collaborators. In a co-located situation, peripheral awareness of each other makes it quite easy for each party to understand what the other party is focusing on and to bring the other's attention to some specific area with the help of pointing. Each party is also free to look around in the room, for example to check if a piece of equipment is available.

Looking at this setting from a requirement perspective we note that:

- Users should be able to quickly perceive where the other collaborator's attention is;
- People are focusing on a difficult task which requires significant focus and attention. Thus, the system should be as simple as possible, particularly with respect to its control interface, to avoid cognitive overload;
- As shown Fig. 1, the collaborators' workspace configurations can be substantially different. For example, one person might be in a lab using a bench, while their collaborator is limited to a home office. The system must take these different setups into consideration to correctly communicate each collaborator's area of attention;
- To be able to look at these different areas of interest, a system must accommodate large head movements, for example a user may change focus from the breadboard on the table to their collaborator on the side of the table;

To fulfill the requirements highlighted above, we designed **RemoteCoDe**, a remote embodied collaborative system for supporting design activities. **RemoteCoDe** uses an articulated display as a robotic proxy to represent the remote collaborator. The display reflects the orientation of the remote collaborator's head while always displaying a rendering of the front of the user's face. To accommodate a wide working area, the system includes a tracking system based on an iPhone's TrueDepth camera. The tracking system captures the user's head orientation as well as a 3D model of their face which allows their face to be rendered from different points of view. To support different physical settings in each location, the system also has a remapping feature which fluidly maps one physical configuration to another, while avoiding unwanted movements during the transition when one switches their attention between areas. This remapping feature also enables easy integration of fisheye camera overview for the scenario where one has to look around a remote space to explore or search for a specific object, as well as to look at multiple people in the space. We describe each part of the system in detail in the following section.

3.1 Kinetic Embodiment Proxy

We design a 2-DOF kinetic robot with a monitor as an embodiment proxy (see Fig. 1.b). Our robotic proxy is designed to cover a wide working area with wide horizontal and vertical movements. Observing that many design activities such as breadboard debugging take place on a table, we attach the robot to the table using a sturdy desk mount. This allows the robotic platform to be rigid and reduces shaking movements when simulating rapid human head motion.

While typical robotic proxies are manually controlled, our proxy is automatically controlled by user's normal head movements using the tracking system described below. This automatic control minimizes user input allowing them to focus on the design task [62].

3.2 Head Tracking & Front Shot Generation

To allow the wide range of head movements required to look around multiple areas in the environment, we use an iPhone TrueDepth camera for tracking head orientation and capturing the user's face. This mobile sensor device is more accessible and compact for use on a table than other depth

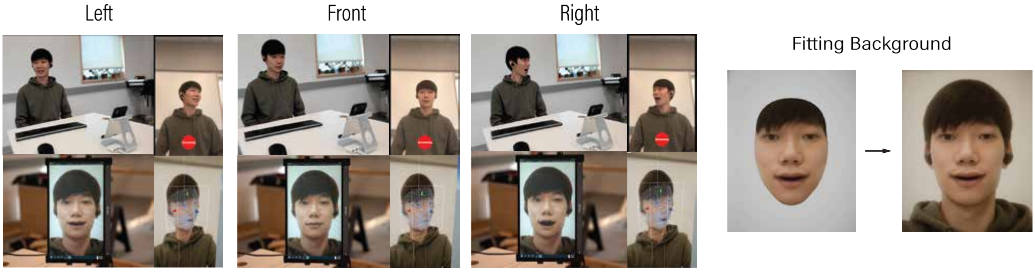

Fig. 2. Creating a front shot. *Left*: we show our pipeline for 3 examples. For each example, starting bottom left and going counterclockwise we show: the user's pose; the image captured by the phone; the 3D model created from the TrueDepth sensor; the same 3D model rendered as a front shot on the display. *Right*: we add a fixed background to complete the rendering.

sensors (e.g., MOCAP, Kinect). The iPhone's depth sensor can support a relatively wide range of head rotations and has been explored for multi-monitor interactions [64].

The system also needs to generate a front shot of the user's face to be rendered on the robotic proxy. The front shot prevents users from misinterpreting or overestimating their partner's head rotation as the robot moves [28]. The TrueDepth system is capable of inferring head movements and facial expressions even from partially occluded faces. Thus we can reconstruct the geometry of the face even if it is off axis from the camera. We use this reconstruction model, combined with a texture provided by the camera, to create a front shot of a face. We show typical results of our system in Fig. 2.

Note that the TrueDepth sensor can only capture the full face texture when the head is looking straight at the camera. To avoid missing data when the user looks sideways, our system records a face texture in the beginning of an interaction to have the full texture of the face for the duration of the interaction. Currently this process is triggered manually but the system could also automatically refresh the texture information when the head is looking directly at the camera. As shown in Fig. 2 right, we use the image from the face texture recording as a background for the face shot to include external elements (e.g., hair, neck, clothing, and environment).

3.3 Remapping Robot Movements

With the elements described so far, the system can only accept symmetrical configuration setups, similar to the setup explored in MMSpace [38]. However, it is often the case that each setting has a unique configuration and simple mapping of head movements to a robot does not work.

To address this problem, *RemoteCoDe* has a remapping feature that maps the head movements of a remote user to match the workspace configuration of the local user. For example, if Alice has her breadboard on the desk in front of her, but the remote user Bob has the document camera view of the breadboard (see Section 3.4) to his right, when Alice looks down at her breadboard, the proxy for Alice in the Bob's space will actually turn left to match the intent of this movement in the local context (See Fig. 3).

When conducting design collaboration, there is a suitable formation such as side-by-side or face-to-face formations depending on the task [27]. This remapping method makes our system flexible with regard to the formation. Although we picked a side by side setting as it is very common for hardware debugging, our remapping feature allows users to customize their configurations to create preferable formation.

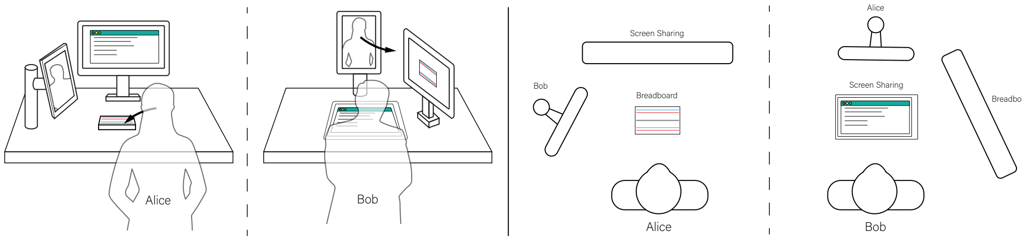

Fig. 3. Example of using the remapping feature: Alice has a breadboard on a desk and the Bob's proxy to her left, while Bob has the shared board to his right and the Alice's proxy to the front. When Alice looks down to focus on a breadboard on the desk, the Alice's proxy on the Bob's desk rotates to left to be consistent with the intention of her head movement.

3.3.1 Calibration Process. The calibration process is used to compute a homography matrix for each area of interest in the users' workspaces. These matrices are used to map face rotations to robot movements. We designed the calibration process to only take a few minutes to register calibration data for each new user and workspace configuration.

The calibration procedure has two phases. In the first phase, a user calibrates the position of each attention area from the articulated display's point of view. To do so, they simply move the articulated display by hand to face toward each calibration point, which is a similar process to Learning from Demonstrations (LfD) in HRI [53]. A user collects data points at four corners for each area (e.g., monitor, desk) and one directed towards themself. During this phase, we found it useful to attach a laser pointer on the front of the display, so that users can visually see the direction of the robot for a precise calibration as shown in Fig. 3. In the second phase, a user calibrates the position of the same areas from their perspective. At the start of this phase, they set up their phone so that it can track head rotations for all areas. Then the user looks al all four corners of each area to collect the required head rotation data.

3.3.2 Smoothing. Simply using homography conversion matrices for robot motion can result in rapid and somewhat unpredictable robot motions during the transition between calibrated areas. This can be distracting to a user. Thus the robot does not follow instantaneous head movements, but instead switches between two different states. In the normal state where a robot moves within a same area using the same remapping matrix, we use a window size of 15 frames to filter subtle noise from the sensor. For the transition state when a robot changes its direction from one area to the other, we use a window size of 30 frames to eliminate any abrupt movements when reaching a far goal position.

Besides the transition issue, we also need to handle a situation where the head direction does not correspond to any recognized areas. In these cases, the system first identifies the closest area based on the current face rotation and constrains the rotation within the identified area. This process can prevent the robot from applying movements out of the expected range and creating abrupt or illegible movements.

3.4 Remote Task View & Pointing

Referring to a specific object in a physical space is essential for remote collaborative work involving a physical task [19]. Especially for collaborative design work like circuit debugging, it is hard to explain what is being referred to with only verbal communication.

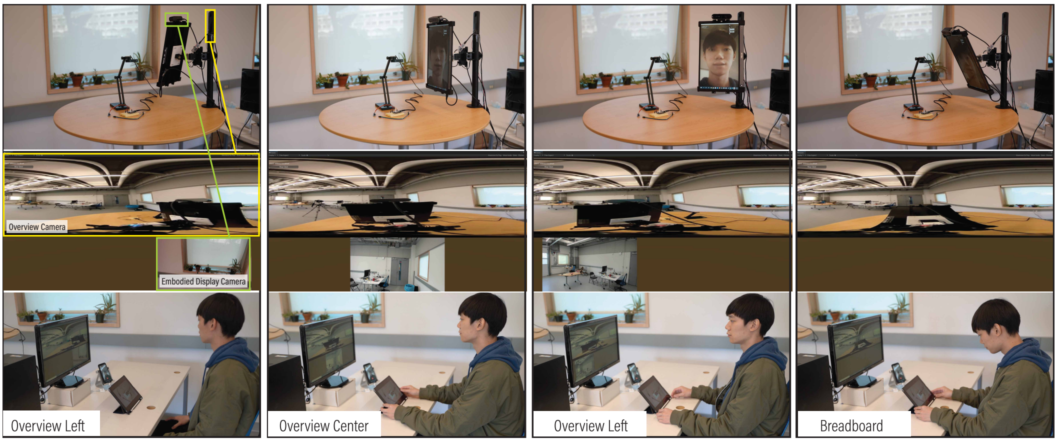

Fig. 4. Overview feature: For this mode we add a fixed fisheye view camera and attach a webcam on top of the articulated display. The remote user sees the overview on one of his displays and the webcam view just below it. We use our remapping mechanism to translate head movement, in the direction of the overview display, to embodied display movement. This provides an easy exploration mechanism for the remote viewer while providing clear feedback to the local user. Note that the webcam view follows the remote user's head movement and is only displayed when his focus of attention is on the overview.

Thus, our system includes a document camera along with a 2-DOF laser pointer on a desk where a physical artifact is, providing the remote user with a close-up view of the artifact and the ability to point to areas on the artifact, similar to TeleAdvisor [22]. The system streams the document camera showing the task area to a tablet on the remote user's desk. The remote user can simply point anywhere on the image to have the actuated laser pointer point at the same location on the desk of the local user (see Fig. 1). Note that this feature can also be used to share the screen of a desktop, but this capability is available through commercial applications.

3.5 Overview Capability

When testing our system we found that situational awareness is important for design collaboration and can be fostered by observing the physical space of a remote collaborator[16]. In collocated design collaborations, one can look around and search for a tool or component to help build their design artifact. Our system provides a similar capability for remote collaboration by allowing a user to control the direction of the embodied display (and camera) by simply looking at a specific area of the overview stream.

Fig. 4 shows an example of using the overview feature. This feature is activated as soon as the user looks at the overview display. As the remote user changes their focus on the overview stream shown on the monitor, the embodied display looks in the same direction using the remapping function. The embodied display camera's stream is displayed just below the overview which helps the remote user to see a close up view of the space. This configuration presents the advantage that the local user is aware that the remote user is looking around the room, while also providing better awareness of the local environment for the remote user. When the remote user's focus is on a different physical task area away from the overview, the embodied display camera view disappears to give the user the notion that the robot is also looking toward the task area. Using this feature,

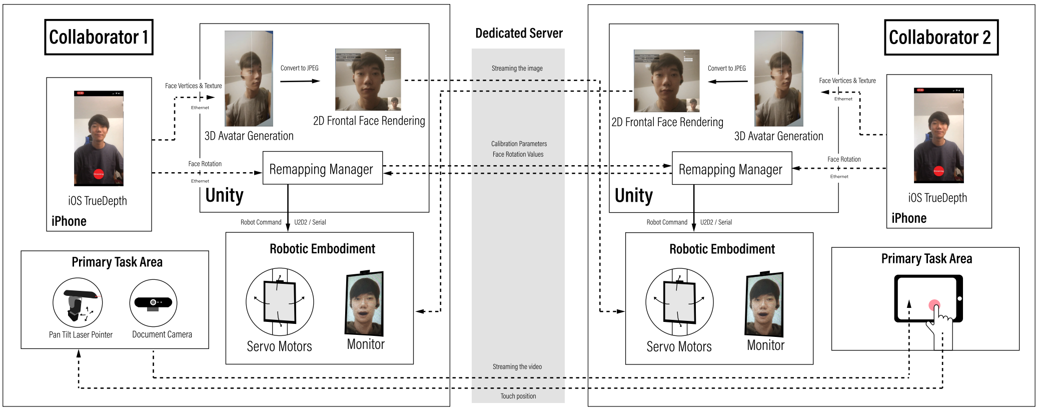

Fig. 5. System architecture. Local Unity clients are connected through a Mirror server. Each client processes the information received from a iOS TrueDepth sensor to create a front shot for the local user and detect the orientation of their head. These pieces of information are sent to the other client which sends commands to the servo motors. The document camera is processed independently. The overview processing is not represented, but is similar to the document camera subsystem.

the system can support one-to-many remote interactions where a remote user sees multiple people on a screen while the local people can see who the remote user is looking at.

4 IMPLEMENTATION

In this section, we highlight the implementation of the *RemoteCoDe* system. As show in Fig. 5, the hardware of *RemoteCode* includes an articulated display, iPhone depth sensor, and a pan-tilt laser pointer, and a overview camera. The software of the system processes the sensor information for face rendering and sends the encoded video, head rotations, and a pointing position through a dedicated server to the other client. The received head rotation is remapped by the remapping manager and sent to a micro-controller to control the local robot.

4.1 Software

Our system is implemented using the Unity gaming framework. Unity supports a wide variety of platforms including PC (Windows, Mac, Linux), smartphone (Android and iOS) and even VR headsets. This flexibility proved very useful to quickly add new features to our system such as a VR interface for the robotic embodiment. The different Unity clients are connected using Mirror Networking [33], an open-source networking library designed for Unity Engine. In our system the Mirror server is used to exchange face images, voice, calibration data, and facial rotations across different private networks. Once two parties get connected, each client sends calibration data to the other client through the server and calculates homography matrices for remapping. Then, the clients start sending facial rotations and remapping the other client's movements to the robot.

We use iPhone X's depth sensor to capture a face rotation and facial geometry. We construct the face mesh from 2120 vertices and a camera image, instead of using blendshapes [1], to get detailed face animations. In every frame, we send this information to the local Unity client to generate a real-time front shot rendering of the local user.

[1] https://developer.apple.com/documentation/arkit/arfaceanchor/2928251-blendshapes

4.2 Hardware

Our embodied proxy uses an animated display (15.6" ZenScreen MB16AC) to represent the remote user's head movement. The proxy is controlled using a U2D2 controller connected to two servo motors (Dynamixel XH430-W210-T, XL430-W250-T) used for pan and tilt directions, respectively. The proxy is attached to a desk mount which is fixed onto a desk or table where multi monitors or design artifacts are located.

We use two Dynamixel AX-12A motors also interfaced with a U2D2 controller for the laser pointer in the remote pointing feature. The system sends the pointing position from the iPad to the remote partner and manipulates the local 2DOF laser pointer using a conversion matrix to point at the corresponding position.

5 EVALUATION

A primary goal of *RemoteCoDe* is to improve remote collaboration by enabling seamless peripheral awareness of a partner's locus of attention. Given that our system supports automatic control and that the literature indicates automation reduces cognitive load [38, 49, 62], our evaluation focuses on accuracy and affordance of peripheral awareness to validate the primary contribution of the system. The two main objectives of our study are to 1) measure the legibility of attention awareness using the remapping method, and 2) observe whether our system affords peripheral perception for identifying locus of attention. To meet these goals, our evaluation is composed of two tasks, completed by participants in a single, in-person laboratory study. In picking our tasks, we aimed to emulate as much as possible a typical debugging session for which we optimized our system design in Section 3 while limiting possible confounding factors.

5.1 Task 1: Attention Awareness Task

Task 1 is designed to measure how accurately users can infer locus of attention by observing the orientation of the *RemoteCoDe* display. We measure two levels of accuracy: 1) the general area of attention and 2) quadrants within each working area. We measure more detailed accuracy in 2) as it is often useful to know where on an artifact a collaborator is focusing, in addition to the fact that their attention in on the artifact.

The three general areas used in this task are: a monitor, a workspace on a desk, and the remote participant; the layout of these elements are shown in Fig. 6 and are picked as to reflect that of a

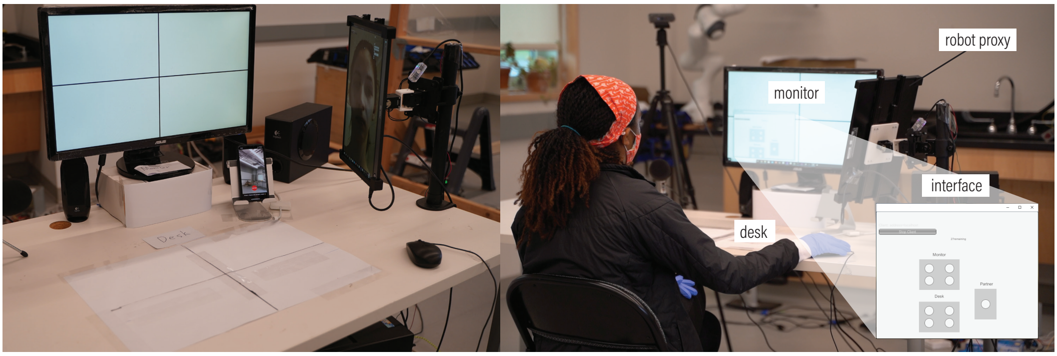

Fig. 6. The configuration for Task 1. The quadrants of the monitor and desk are shown in the right image, and the GUI participants use to submit guesses is shown in the left image.

typical debugging session. The monitor and desk workspace are each divided into four quadrants to measure the more detailed accuracy of the system, resulting in 9 areas of attention.

In the task, participants are told to guess which of the 9 areas *RemoteCoDe* is focusing on whenever they hear a beep sound. Participants submit their responses using an interface on the main monitor, shown in Fig. 6. The movements of the robot, as well as the accompanying face on the robot are prerecorded so that the movements are consistent across participants. All areas are looked at three times, resulting in 27 total iterations. The order of the areas are randomized for each participant.

5.2 Task 2: Peripheral Awareness Task

The goal of the second task is to evaluate to what extent *RemoteCoDe* affords peripheral awareness of a remote collaborator's locus of attention. A video baseline is used to confirm that our system provides more peripheral awareness than a video that might be seen during video conferencing. In the video condition, a pre-recorded video of the remote user working at a desk is played on the *RemoteCoDe* display.

While the system was originally designed for circuit debugging, we found it difficult to find participants with an uniform level of understanding of circuit basics. As a result, we use a timed jigsaw puzzle game as a distractor activity to capture peripheral awareness. The jigsaw puzzle game is a simple activity that simulates a design situation where a user is focusing on a hands-on task locally while also working with a remote collaborator. Additionally, the puzzle game does not require participants to have technical or design skills (e.g., firmware programming or circuit building) and can be presented consistently across participants and conditions to preserve internal validity. In this task, participants worked on puzzles on the iPad from the iOS app "Jigsaw Puzzle on iPad & iPhone"[2]. The puzzle task was equivalent between conditions as the same jigsaw puzzle app and puzzle difficulty level were used in all conditions, but different puzzles were given for each condition, the order of conditions and puzzles were counterbalanced to account for possible learning effects.

In this task, participants worked individually on the game and were periodically prompted to verbally guess where their remote partner is looking. We used verbal input instead of a user interface to minimize distractions to the participant's puzzle game task. The three areas the robot could be looking are: the monitor, the workspace, and the participant. The participants are interrupted

[2] https://apps.apple.com/us/app/jigsaw-puzzle-on-ipad-iphone/id642831690

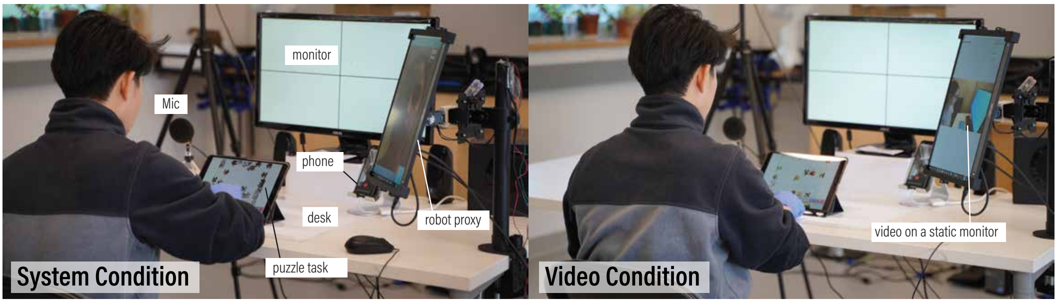

Fig. 7. The configuration for Task 2. In both conditions, we have a microphone to record the participant's guesses. The robotic proxy rotating its monitor in the system condition is shown in the left image, and a static monitor showing a video of a remote partner in the video condition is shown in the right image.

by a beep sound 9 times per condition, with each of the three areas being referenced three times. There are 30 seconds between each beep interruption. Two random patterns of the ordering of areas were prepared in advance of the study, and each participant was given each pattern one time. The ordering of the patterns and conditions are counterbalanced. The setup for task 2 is shown in Fig. 7.

The video, face rendering, and robot motions used during this task were pre-recorded to enhance repeatability and consistency across participants. We consider the use of pre-recordings to be acceptable since our focus in this task is to investigate on peripheral awareness, not conversational interactions. In the video condition the video of the remote user was recorded with a 90 degree field of view to capture the remote user as well as their environment (e.g., task area, monitor). In the video condition the remote user's workspace configuration is similar to the participant's due to the inability to remap standard video, but the system was placed on the opposite side of the remote user when collecting data for the embodiment condition so that we could test the remapping feature on distinct configurations.

5.3 Procedure

The study began with participants completing an informed consent form and a demographic survey, after which they were seated at a table with the embodiment system, a monitor, and a workspace (shown in Figs. 6 and 7) and given a brief introduction to *RemoteCoDe* and the study tasks.

5.3.1 Task 1 Procedure. After the introduction, instructions were given for the first task and the experimenter ran a brief demonstration showing all of the places the *RemoteCoDe* proxy could focus on in a given order (partner, monitor areas, and task areas). During the demo participants were also introduced to the interface they would use to make guesses. Then the participants completed a practice round where they observed the display and guessed attention areas using the interface. During this practice round the areas were presented in a random order. After the participant stated that they felt comfortable with the interface and robotic movements, the participants completed the task as described in Section 5.1.

5.3.2 Task 2 Procedure. After completing task 1, the experimenter prepared the workspace for task 2 and gave the participants instructions for the task. Then the participants practiced the task for 1-2 minutes to get familiar with the interface, after which they completed the first condition. After the first condition, the experimenter prepared the workspace for the second condition and participants were given another 1-2 minutes to practice the task. Then they complete the second condition. The details for the task are detailed in Section 5.2 and the order of the conditions (*RemoteCoDe* and videoconferencing) was counter-balanced. After each condition, participants answered a NASA TLX questionnaire [24] that takes about a minute. Finally each participant was debriefed for about two minutes.

5.3.3 Safety Procedures. Due to COVID-19, participants were required to wear personal protective equipment (e.g., mask, gloves) except in Task 2 during which we needed to track their head orientation. For these portions of the study the researcher left the participant alone in the experiment room.

5.4 Participants

We recruited 13 participants (7 female and 6 male) ages between 19 and 41 (mean 25.38). All the participants had videoconferencing experience and one of them had previously used a telepresence robot. We discarded one of the participant's data due to technical difficulties. Participants received course credit or $15 for their participation in the hour long study.

5.5 Data Collection & Analysis

Here we explain how we analyzed the collected data. To compute a statistical difference between the two conditions in each measurement for Peripheral Awareness (Section 5.5.2) and Task Load (Section 5.5.3), we applied a paired t-test as each participant ran both conditions, and used the Bonferroni correction.

5.5.1 Attention Awareness Accuracy. We compared the guesses submitted by the participants to the correct target areas in each of the 9 areas. For example, if a participant guess is the top right but the actual target is the top left we count this as an error. The same process was repeated for the 3 general areas (monitor, workspace and desk). We computed the system accuracy using the equation:

$$\frac{\#total\ trials - \#errors}{\#total\ trials} \times 100$$

5.5.2 Peripheral Awareness. We recorded audio during the Task 2 to derive guess accuracy and time from these recordings. In the coding process, the coder was unaware of the condition and correct answers to minimize potential biases. Listening to each audio clip, the coder checked the time of the beginning of the beep sound and that of when a participant started saying their answer. We timed when they started answering rather than finishing their answer because speech speed differs among individuals. The timings excluded filler language (such as “um”) and were measured to the nearest hundredth of a second. The timings were calculated using a visualization of the audio in the program Audacity [2]. The label sounds function in Audacity was used to find the exact moment when the a new sound was made (ie a beep or a speech utterance).

If a participant uses peripheral vision to perceive where their partner is looking, we expect that their head rotation will change very minimally. If they do not use their peripheral vision, participants will need to raise their head from their task to look at their partner, resulting in a large change in head rotation. Thus, head rotation is a good indicator to confirm whether a participant uses peripheral vision to perceive the partner. The participants’ head rotation was captured for the duration of Task 2 using the iPhone TrueDepth camera at a rate of 15 fps. The raw data contains a time stamp and two head rotation values in degrees, one for the vertical (up/down) direction and one for the horizontal (right/left) direction. Video recordings were used to confirm the phenomena observed in quantitative measurements, that is, to check the consistency between the recorded head rotations and how the participant actually moved their head. Using the recorded face rotation, we measured the maximum angle difference (in degrees) of the head rotation from the rotation at the beep sound for the 5 seconds following the beep. For example, when the initial face angle in the up/down direction was 40.5 degree at the beep sound and the furthest angle from that initial angle within the 5 seconds was 103.7 degree, the maximum angle difference is calculated as follows: $|103.7 - 40.5| = 63.2$. We chose to use the period of 5 seconds after the beep as 97.03% of time participants guessed the attention area within 5 seconds (participants guessed after 7 seconds in 3 cases). We discarded 3 samples that were collected when the head tracking was turned off.

5.5.3 Task Load. We also conducted a NASA TLX questionnaire [24] to compare the two conditions in Task 2 in terms of cognitive load required for guessing a partner’s attention.

5.5.4 Qualitative Feedback Through Debriefing Session. After completing all the tasks, the experimenter conducted a two minute in-person debriefing session with the participant to gather qualitative feedback on the embodiment system focusing on the task 2. The primary question was: “How did you feel about the differences between the first condition and the second condition?”

Follow-up questions where asked to get a better understanding of how the participants understood the differences in guessing locus of attention between the conditions, such as: “Was it harder or easier to guess where the person was looking in that condition?” The session was audio recorded and transcribed. The transcribed participants’ comments were repeatedly read looking for similarities and differences between responses and categorized according to recurring answers and topics by a single researcher. The responses all pertained to which condition the participants preferred and why. Therefore the responses could be grouped by preferred condition, as well as the reasons given for the preference. After noting all of the reasons given, we found multiple participants mentioned the same things. These similar responses were grouped into topics. Because each session was only a few sentences long, we only identified three topics raised by participants: peripheral awareness of the robot, appearance of the robot, and noise of the robot. The results of this were used to observe the general perception of the robotic proxy and the consistency with the quantitative analysis.

6 RESULTS

6.1 Attention Awareness Accuracy

Participants guessed the correct general area with great accuracy: 99.30% for the desk and 100% for monitor and partner. They were also able to guess each subarea (2×2 subdivision of the desk and the monitor) with an overall accuracy of 93.51%, as shown in Fig. 8. The overall results of this part of the evaluation confirm that the remapping feature of our system can accurately render the locus of attention of the remote participant. We also note that the top left area of the monitor and the top left and bottom left areas of the desk have relatively low accuracy. This may be related to the fact that these areas are on the far side from the robot. We address this in Section 7.1.

6.2 Peripheral Awareness

6.2.1 Guess Accuracy & Time. In terms of the accuracy of the attention guesses, there were three errors for the video condition (97.22%) and one error for the system condition (99.07%) out of 108

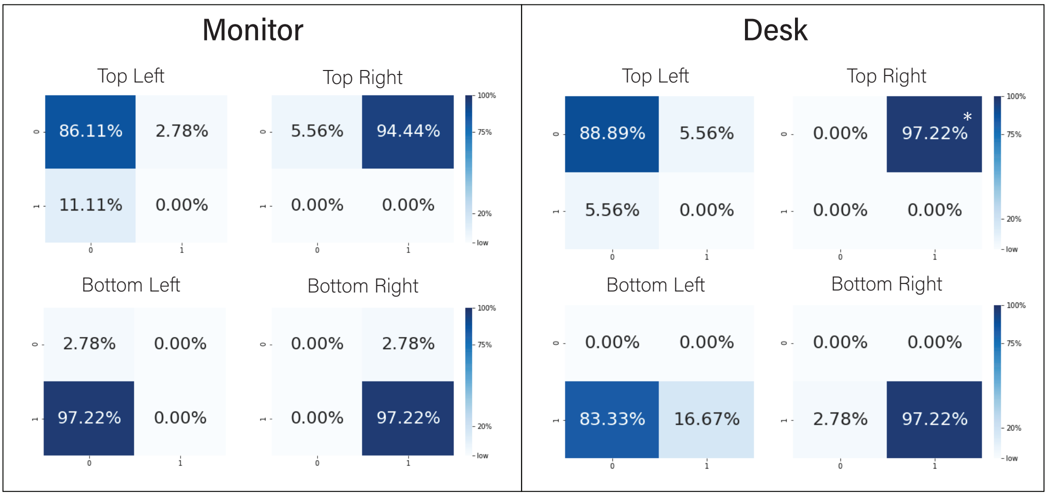

Fig. 8. Attention Awareness Accuracy: We show for the monitor (left) and the desk area (right) the distribution of participants answers for each sub-area. Note that the top right area of the desk does not add up to 100% as 2.78% guessed it was the bottom right corner of the monitor.

guesses for each. A paired samples t-test showed no statistical differences (t(11)=1.48, p = 0.17, two-tailed).

The analysis of completion time for guessing attention area shows that participants took more time to guess in the video condition (MD = 2.41s, SD=0.97) than in our system condition (MD = 1.44s, SD=0.46). A paired t-test showed a significant difference between the two conditions (t(11)=4.44, p = 0.001, two-tailed. The observed power (1- β err prob) is 0.98 (α = .05)). Note that we excluded nine data points when a participant guessed before the beep sound. These results support our hypothesis that by using our system users can spend more time and attention on their primary task, as compared to standard video conferencing systems.

6.2.2 Head Rotation. In the video condition, on average, participants rotated their head 35.00 degrees to the right toward the display and 25.79 degrees up after the beep. In the *RemoteCoDe* condition, however, participants rotated their head only 11.54 degrees to the right toward the display and 11.39 degrees up. Both metrics were found significantly different while running a paired t-test (t(11) = 5.46, p < 0.001, observed power (1- β err prob) = 0.999 (α = .05), and t(11) = 4.20, p = 0.0015, observed power (1- β err prob) = 0.995 (α = .05) respectively).

This supports our hypothesis that users rotate their head less to look at the monitor to determine the attention area, and can observe it through their peripheral vision. Fig. 9 shows example pictures and graphs of pan (to the right toward the display) and tilt (up) face rotations when the participant guesses an attention area after a beep sound. The graphs imply that the participant moved her head a lot to look at the monitor in the video condition whereas she only needed to use her peripheral vision, without moving her head, in the system condition.

The results here indicate the system requires less time and physical effort to stay aware of a remote partner's attention. This enables less distracting interactions and allows collaborators to focus on their primary tasks.

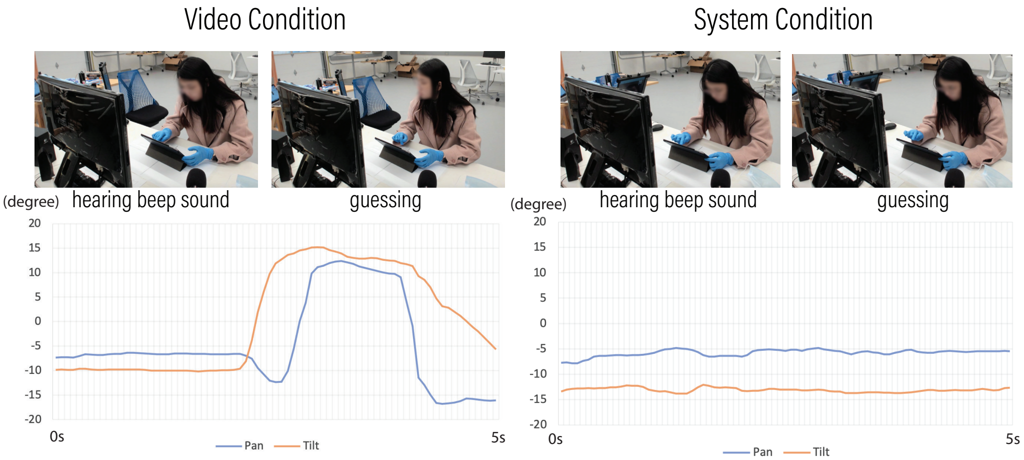

Fig. 9. Head rotation for the peripheral awareness task. The plots show the head rotation values from the beep sound, to 5 seconds after the beep. We can see that the participant looks up from the task in the Video condition, but not while using our system.

Category	Mean (System)	Mean (Video)	t(11)	p-value
Frustration	2.00	3.00	1.56	0.15
Effort	2.50	3.25	1.52	0.16
Performance	5.00	5.33	0.56	0.59
Temporal Demand	2.33	2.75	0.76	0.46
Physical Demand	1.50	2.00	1.25	0.23
Mental Demand	2.58	2.75	0.41	0.69

Table 1. Self-reported task load ratings (NASA TLX) for guessing a partner's locus of attention in both conditions (System, Video).

6.2.3 NASA TLX Questionnaire. To analyze the the self-reported task load ratings for guessing a partner's attention, we performed a two-tailed t-test between the two conditions for each measures of the questionnaire, the values are shown in Table 1. We did not find any significant results.

6.3 Qualitative Feedback

The qualitative inputs from participants during the debriefing sessions support the validity of the quantitative analysis above. Ten out of twelve participants (all except P2 and P9) mentioned that the ability to sense the movement and/or position of the robot through their peripheral vision helped them easily guess the attention areas. As an example P10 commented: *"in the [system condition] even without looking at the robot I could see with my peripheral which direction, . . . sometimes I could do the puzzle and not even look at the robot directly."*

Some participants also commented on the appearance of the robotic proxy. P1 noted the impact of having a face rendered on the screen on his perception of the collaborator: *"For me it's like a real human being somehow. I know its a computer, but having the face of a human being is what makes it different...it's more real."* P6 expressed the sense of interacting with the robot due to its movements: *"The monitor didn't distract me much because I can feel like the robot is actually interacting with me in a way, makes me feel like okay where he or she is looking at through the movement."* P6 also mentioned the positive effect of the noise from the robot saying *"For the rotation of the monitor that one is easier to guess without even looking at it . . . just by hearing the sound somehow its just like . . . we don't even have to look at it and we can guess it correctly."* On the other hand, P2 preferred the video condition saying *"because of the shape of the monitor I still need to have some time to get used to [the system]. (If the monitor was) a real head shape that would probably help me guess"*.

The participants' feedback affirms that our system can support peripheral awareness and offers some implications for the design of future robotic proxies for remote collaboration.

6.4 Limitations

Our study validated a set of key features of the system including the accuracy of remapped head motions and ability to afford peripheral awareness. We believe that the results on the accuracy of attention estimation and the peripheral awareness can be applicable to other types of tasks as long as the robot is visible within their peripheral vision. On the other hand this study has low ecological validity and a follow-up field study involving real-world collaborative design scenarios is required to further investigate the efficacy of our system. For example, how does the peripheral awareness provided by the robot affect user performance in the context of hardware debugging? How does it enhance engagement and sense of co-presence in their collaboration? Can the movements of the proxy help users build trust in their teammate? To answer these questions, we plan to deploy our system in a more realistic setting such as in a hardware design class.

The study used pre-recorded movements of a remote partner where the partner looks at various areas in a randomized order but these movements did not contain spontaneous head movements and other non verbal cues (e.g., nodding). While the purpose of using pre-recorded movements was to have consistency across participants and high internal validity, we believe that further study with pairs of users will uncover benefits and drawbacks of remapping head movements to a robot in actual conversational settings.

We did not focus on the cognitive load of our system because of previous work supporting the benefits of automation [38, 49], but there might be an unexpected induced cognitive load imposed by even a fully automated system. For example, expressive head movements of a proxy might require more efforts on users to interpret the movements leading to a sense of distraction in exchange for reduced workload for control. Thus, a further survey on our automatic control method is required to investigate to what extent our system can reduce cognitive burden in comparison to a typical embodiment system with a control interface.

7 DISCUSSION AND FUTURE WORK

7.1 Embodiment Design

We decided on a fixed articulated display based on our experience teaching hardware design. It indicated that user's full body movements were not as important as head movements for these types of tasks. However, this will be different in large scale activities such as Critique in which people need to move around in a wider space. For these applications, Choi et al.'s work [10] will be of great use to implement robot behavior. The study conducted by Rae et. al. [49] also gives design implications for a mobile robotic platform that can be extended from our system. As we discuss in Section 2.2, current mobile telepresence robots are controlled through a GUI or joystick. A key challenge for future work will be to design an embodied robot that can automatically reproduce rapid human movements in an open-space environment without a control interface.

The results from our evaluation validated the basic design of the system. They also provided us with some design implications to improve our robotic proxy. As pointed out by a participant, the shape of the articulated display can affect the perception of their partner and accuracy of gaze awareness [15]. The results also indicate that the left areas of the monitor and desk tend to have relatively low accuracy. The general direction of the robot could be ambiguous either because it is looking down at the table or far away to the left of the screen. This implies that our articulated display might still cause the Mona Lisa effect to some extent although projecting a front face is implemented to mitigate it. One solution to this can be to use a curved display or face-shaped mask [13, 34]. Using our face re-projection approach, it is easy to project a generated 3D face to these types of displays without the need to fix a face on an equipment [34]. Alternative redesign of the system could consider removing or re-positioning the rendering of the user's face, as Gaver et al.'s findings indicate that viewing a remote collaborator's face is not as important as getting a good view of their working environment and shared task space [20]

7.2 Tracking

While our system is very flexible and easy to set up, it does require re-calibration each time the TrueDepth camera or any of the displays is moved (Section 3.3.1). This was not too much of a problem during development, but it could clearly become problematic during actual use. This could be easily addressed by augmenting the different devices with ArUco tags [48].

Another limitation we observed was the limited field of view of the TrueDepth camera. This is a somewhat more complicated problem that could be addressed by positioning the depth camera carefully, but there may be still a problem when users need to increase a number of monitors or physical artifacts. Ultimately we believe that best solution will be to use a wearable face capture sensor such as NeckFace [9].

Our user evaluation indicates that the head movement was effective in conveying locus of attention, but certainly adding gaze tracking will add to the accuracy of the system and could be done using the eye-tracking feature of TrueDepth API.

7.3 Privacy

During our tests, we also discovered how intrusive the overview feature could be. In that sense local user's should seriously consider that the remote user will be invited to their lab or office when using the system. A related problem is the best way to "enter" one's office when using this system. We believe that an actual system should avoid an instantaneous connection (like that of the discord system [14]) but instead leverage the physical embodiment to indicate that communication might be imminent. For example, before streaming any camera feed or face avatar, a robot can move in a certain way to signal that someone is trying to initiate collaboration.

7.4 Alternative Access Interfaces

A VR headset can be an interesting alternative interface for our embodied robot. Some previous work demonstrates an example of how a VR headset can be used for remote robot control [51, 52]. With the help of our remapping approach, the head movements of a VR user can be remapped to a physical robot in a remote location and vice versa for rendering a virtual avatar. The VR interface can be useful when users want to have a full immersion experience, but could hinder access to local resources (physical artifacts, tools, etc.) although this can be easily solved by blending remote and local views. Facial tracking is difficult to support in AR as a headset covers most of the face, but this capability may be accessible for consumer-level HMDs soon. We are in the process of developing such an interface as it will be useful to conduct a side-by-side study to investigate the pros and cons of VR interface and physical embodiment.

8 CONCLUSION

We present the design of *RemoteCoDe*, which allows remote collaborators to engage in design activities by leveraging their peripheral awareness through a robotic embodiment. Using the system, collaborators can focus on a design task, switching their attention between multiples areas, while peripherally perceiving their collaborator's attention. The remapping feature allows the system to accommodate different configurations in each location. We would like to further explore the potential features of the system and extend its capabilities to support more complex remote design collaborations.

ACKNOWLEDGMENTS

We thank Mike Russo, Bryson Graves, Vivian Li who helped us in the early stage of ideation and prototyping. We also thank Ryo Suzuki for providing insightful feedback on the project idea. This research was supported by NSF Award IIS-1563705 and IIS-1925100, and the Nakajima Foundation.

REFERENCES

[1] S. O. Adalgeirsson and C. Breazeal. 2010. MeBot: A robotic platform for socially embodied telepresence. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 15–22. https://doi.org/10.1109/HRI.2010.5453272

[2] Audacity Team. 2021. Audacity software. Version 3.0.5. https://audacityteam.org

[3] Reynold Bailey, Ann McNamara, Nisha Sudarsanam, and Cindy Grimm. 2009. Subtle Gaze Direction. *ACM Trans. Graph.* 28, 4, Article 100 (Sept. 2009), 14 pages. https://doi.org/10.1145/1559755.1559757

[4] Guillermo Bernal, Tao Yang, Abhinandan Jain, and Pattie Maes. 2018. PhysioHMD: A Conformable, Modular Toolkit for Collecting Physiological Data from Head-Mounted Displays. In *Proceedings of the 2018 ACM International Symposium*

on Wearable Computers (Singapore, Singapore) *(ISWC '18)*. Association for Computing Machinery, New York, NY, USA, 160–167. https://doi.org/10.1145/3267242.3267268

[5] Mark Billinghurst and Hirokazu Kato. 1999. Collaborative mixed reality. In *Proceedings of the First International Symposium on Mixed Reality*. 261–284.

[6] Susan E. Brennan, Xin Chen, Christopher A. Dickinson, Mark B. Neider, and Gregory J. Zelinsky. 2008. Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition* 106, 3 (2008), 1465–1477. https://doi.org/10.1016/j.cognition.2007.05.012

[7] William A. S. Buxton. 1992. Telepresence: Integrating Shared Task and Person Spaces. In *Proceedings of the Conference on Graphics Interface '92* (Vancouver, British Columbia, Canada). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 123–129.

[8] Jean Carletta, Robin L. Hill, Craig Nicol, Tim Taylor, Jan Peter Ruiter, and Ellen Gurman Bard. 2010. Eyetracking for two-person tasks with manipulation of a virtual world. *Behavior Research Methods* 42, 1 (1 Feb. 2010), 254–265. https://doi.org/10.3758/BRM.42.1.254

[9] Tuochao Chen, Yaxuan Li, Songyun Tao, Hyunchul Lim, Mose Sakashita, Ruidong Zhang, Francois Guimbretiere, and Cheng Zhang. 2021. NeckFace: Continuously Tracking Full Facial Expressions on Neck-mounted Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5 (06 2021), 1–31. https://doi.org/10.1145/3463511

[10] Mina Choi, Rachel Kornfield, Leila Takayama, and Bilge Mutlu. 2017. *Movement Matters: Effects of Motion and Mimicry on Perception of Similarity and Closeness in Robot-Mediated Communication*. Association for Computing Machinery, New York, NY, USA, 325–335. https://doi.org/10.1145/3025453.3025734

[11] Sarah D'Angelo and Andrew Begel. 2017. *Improving Communication Between Pair Programmers Using Shared Gaze Awareness*. Association for Computing Machinery, New York, NY, USA, 6245–6290. https://doi.org/10.1145/3025453.3025573

[12] Sarah D'Angelo and Darren Gergle. 2016. *Gazed and Confused: Understanding and Designing Shared Gaze for Remote Collaboration*. Association for Computing Machinery, New York, NY, USA, 2492–2496. https://doi.org/10.1145/2858036.2858499

[13] F. Delaunay, J. de Greeff, and T. Belpaeme. 2010. A study of a retro-projected robotic face and its effectiveness for gaze reading by humans. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 39–44. https://doi.org/10.1109/HRI.2010.5453271

[14] Discord Inc. 2021. Discord. https://discord.com

[15] Jens Edlund, Samer Al Moubayed, and Jonas Beskow. 2011. The Mona Lisa Gaze Effect as an Objective Metric for Perceived Cospatiality. In *Intelligent Virtual Agents*, Hannes Högni Vilhjálmsson, Stefan Kopp, Stacy Marsella, and Kristinn R. Thórisson (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 439–440.

[16] Mica R. Endsley. 1988. Design and Evaluation for Situation Awareness Enhancement. *Proceedings of the Human Factors Society Annual Meeting* 32, 2 (1988), 97–101. https://doi.org/10.1177/154193128803200221 arXiv:https://doi.org/10.1177/154193128803200221

[17] Anna Maria Feit, Shane Williams, Arturo Toledo, Ann Paradiso, Harish Kulkarni, Shaun Kane, and Meredith Ringel Morris. 2017. *Toward Everyday Gaze Input: Accuracy and Precision of Eye Tracking and Implications for Design*. Association for Computing Machinery, New York, NY, USA, 1118–1130. https://doi.org/10.1145/3025453.3025599

[18] Deborah I. Fels, Judith K. Waalen, Shumin Zhai, and Patrice L. Weiss. 2001. Telepresence under Exceptional Circumstances: Enriching the Connection to School for Sick Children. In *INTERACT*.

[19] Susan R. Fussell, Leslie D. Setlock, Jie Yang, Jiazhi Ou, Elizabeth Mauer, and Adam D. I. Kramer. 2004. Gestures Over Video Streams to Support Remote Collaboration on Physical Tasks. *Human–Computer Interaction* 19, 3 (2004), 273–309. https://doi.org/10.1207/s15327051hci1903_3 arXiv:https://doi.org/10.1207/s15327051hci1903$_3$

[20] William W Gaver, Abigail Sellen, Christian Heath, and Paul Luff. 1993. One is not enough: Multiple views in a media space. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 335–341.

[21] Kunal Gupta, Gun A. Lee, and Mark Billinghurst. 2016. Do You See What I See? The Effect of Gaze Tracking on Task Space Remote Collaboration. *IEEE Transactions on Visualization and Computer Graphics* 22, 11 (2016), 2413–2422. https://doi.org/10.1109/TVCG.2016.2593778

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

[23] C. Gutwin and S. Greenberg. 2004. A Descriptive Framework of Workspace Awareness for Real-Time Groupware. *Computer Supported Cooperative Work (CSCW)* 11 (2004), 411–446.

[24] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[25] Dirk Heylen. 2005. Challenges Ahead: Head movements and other social acts in conversations. In *Proceedings of the Joint Symposium on Virtual Social Agents*, Lynn Halle, Peter Wallis, Sarah Woods, Stacy Marsella, Catherine Pelachaud, and Dirk K.J. Heylen (Eds.). The Society for the Study of AI and the Simulation of Behav., 45–52. Joint Symposium on Virtual Social Agents, AISB 2005 Convention : Social Intelligence and Interaction in Animals, Robots and Agents, AISB; Conference date: 12-04-2005 Through 15-04-2005.

[26] Keita Higuch, Ryo Yonetani, and Yoichi Sato. 2016. Can Eye Help You?: Effects of Visualizing Eye Fixations on Remote Collaboration Scenarios for Physical Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5180–5190. https://doi.org/10.1145/2858036.2858438

[27] Hiroshi Ishii and Minoru Kobayashi. 1992. ClearBoard: A Seamless Medium for Shared Drawing and Conversation with Eye Contact. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, USA) *(CHI '92)*. Association for Computing Machinery, New York, NY, USA, 525–532. https://doi.org/10.1145/142750.142977

[28] Ikkaku Kawaguchi, Hideaki Kuzuoka, and Yusuke Suzuki. 2015. Study on Gaze Direction Perception of Face Image Displayed on Rotatable Flat Display. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. ACM, New York, NY, USA, 1729–1737. https://doi.org/10.1145/2702123.2702369

[29] Min Kyung Lee and Leila Takayama. 2011. "Now, I Have a Body": Uses and Social Norms for Mobile Remote Presence in the Workplace. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. ACM, New York, NY, USA, 33–42. https://doi.org/10.1145/1978942.1978950

[30] Evelyn Z. McClave. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics* 32, 7 (2000), 855–878. https://doi.org/10.1016/S0378-2166(99)00079-X

[31] Ann McNamara, Reynold Bailey, and Cindy Grimm. 2009. Search Task Performance Using Subtle Gaze Direction with the Presence of Distractions. *ACM Trans. Appl. Percept.* 6, 3, Article 17 (Sept. 2009), 19 pages. https://doi.org/10.1145/1577755.1577760

[32] François Michaud, Patrick Boissy, Daniel Labonte, Hélène Corriveau, Andrew Grant, Michel Lauria, Richard Cloutier, Marc-André Roux, Daniel Iannuzzi, and M.-P. Royer. 2007. Telepresence Robot for Home Care Assistance. In *AAAI Spring Symposium: Multidisciplinary Collaboration for Socially Assistive Robotics*.

[33] Mirror Networking. 2021. Mirror Networking – Open Source Networking for Unity. https://mirror-networking.com/

[34] Kana Misawa, Yoshio Ishiguro, and Jun Rekimoto. 2012. LiveMask: A Telepresence Surrogate System with a Face-Shaped Screen for Supporting Nonverbal Communication. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (Capri Island, Italy) *(AVI '12)*. Association for Computing Machinery, New York, NY, USA, 394–397. https://doi.org/10.1145/2254556.2254632

[35] David Nguyen, John Canny, and John Canny. 2005. MultiView: Spatially Faithful Group Video Conferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Portland, Oregon, USA) *(CHI '05)*. ACM, New York, NY, USA, 799–808. https://doi.org/10.1145/1054972.1055084

[36] Ken-Ichi Okada, Fumihiko Maeda, Yusuke Ichikawaa, and Yutaka Matsushita. 1994. Multiparty Videoconferencing at Virtual Social Distance: MAJIC Design. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work* (Chapel Hill, North Carolina, USA) *(CSCW '94)*. ACM, New York, NY, USA, 385–393. https://doi.org/10.1145/192844.193054

[37] Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L. Davidson, Sameh Khamis, Mingsong Dou, Vladimir Tankovich, Charles Loop, Qin Cai, Philip A. Chou, Sarah Mennicken, Julien Valentin, Vivek Pradeep, Shenlong Wang, Sing Bing Kang, Pushmeet Kohli, Yuliya Lutchyn, Cem Keskin, and Shahram Izadi. 2016. Holoportation: Virtual 3D Teleportation in Real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (Tokyo, Japan) *(UIST '16)*. ACM, New York, NY, USA, 741–754. https://doi.org/10.1145/2984511.2984517

[38] K. Otsuka. 2016. MMSpace: Kinetically-augmented telepresence for small group-to-group conversations. In *2016 IEEE Virtual Reality (VR)*. 19–28. https://doi.org/10.1109/VR.2016.7504684

[39] Mai Otsuki, Taiki Kawano, Keita Maruyama, Hideaki Kuzuoka, and Yusuke Suzuki. 2017. *ThirdEye: Simple Add-on Display to Represent Remote Participant's Gaze Direction in Video Communication*. Association for Computing Machinery, New York, NY, USA, 5307–5312. https://doi.org/10.1145/3025453.3025681

[40] Jiazhi Ou, Lui Min Oh, Jie Yang, and Susan R. Fussell. 2005. *Effects of Task Properties, Partner Actions, and Message Content on Eye Gaze Patterns in a Collaborative Task*. Association for Computing Machinery, New York, NY, USA, 231–240. https://doi.org/10.1145/1054972.1055005

[41] Ye Pan and Anthony Steed. 2014. A Gaze-preserving Situated Multiview Telepresence System. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. ACM, New York, NY, USA, 2173–2176. https://doi.org/10.1145/2556288.2557320

[42] S. J. Park, J. H. Han, B. H. Kang, and K. C. Shin. 2011. Teaching assistant robot, ROBOSEM, in English class and practical issues for its diffusion. In *Advanced Robotics and its Social Impacts*. 8–11. https://doi.org/10.1109/ARSO.2011.6301971

[43] Eric Paulos and John Canny. 1998. PRoP: Personal Roving Presence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Los Angeles, California, USA) *(CHI '98)*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 296–303. https://doi.org/10.1145/274644.274686

[44] Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew Wilson. 2016. Room2Room: Enabling Life-Size Telepresence in a Projected Augmented Reality Environment. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (San Francisco, California, USA) *(CSCW '16)*. ACM, New York, NY, USA, 1716–1725. https://doi.org/10.1145/2818048.2819965

[45] Thammathip Piumsomboon, Gun A. Lee, Jonathon D. Hart, Barrett Ens, Robert W. Lindeman, Bruce H. Thomas, and Mark Billinghurst. 2018. Mini-Me: An Adaptive Avatar for Mixed Reality Remote Collaboration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 46, 13 pages. https://doi.org/10.1145/3173574.3173620

[46] Thammathip Piumsomboon, Gun A. Lee, Andrew Irlitti, Barrett Ens, Bruce H. Thomas, and Mark Billinghurst. 2019. *On the Shoulder of the Giant: A Multi-Scale Mixed Reality Collaboration with 360 Video Sharing and Tangible Interaction.* Association for Computing Machinery, New York, NY, USA, 1–17. https://doi-org.proxy.library.cornell.edu/10.1145/3290605.3300458

[47] Thammathip Piumsomboon, Youngho Lee, Gun Lee, and Mark Billinghurst. 2017. CoVAR: A Collaborative Virtual and Augmented Reality System for Remote Collaboration. In *SIGGRAPH Asia 2017 Emerging Technologies* (Bangkok, Thailand) *(SA '17)*. Association for Computing Machinery, New York, NY, USA, Article 3, 2 pages. https://doi.org/10.1145/3132818.3132822

[48] S. Garrido-Jurado R. Muñoz-Salinas. 2013. ArUco Library. http://sourceforge.net/projects/aruco/

[49] Irene Rae, Bilge Mutlu, and Leila Takayama. 2014. Bodies in Motion: Mobility, Presence, and Task Awareness in Telepresence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. ACM, New York, NY, USA, 2153–2162. https://doi.org/10.1145/2556288.2557047

[50] Irene Rae and Carman Neustaedter. 2017. Robotic Telepresence at Scale. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 313–324. https://doi.org/10.1145/3025453.3025855

[51] Mose Sakashita, Tatsuya Minagawa, Amy Koike, Ippei Suzuki, Keisuke Kawahara, and Yoichi Ochiai. 2017. You as a Puppet: Evaluation of Telepresence User Interface for Puppetry. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 217–228. https://doi.org/10.1145/3126594.3126608

[52] MHD Yamen Saraiji, Tomoya Sasaki, Reo Matsumura, Kouta Minamizawa, and Masahiko Inami. 2018. Fusion: Full Body Surrogacy for Collaborative Communication. In *ACM SIGGRAPH 2018 Emerging Technologies* (Vancouver, British Columbia, Canada) *(SIGGRAPH '18)*. Association for Computing Machinery, New York, NY, USA, Article 7, 2 pages. https://doi.org/10.1145/3214907.3214912

[53] Stefan Schaal. 1996. Learning from Demonstration. In *NIPS*. 1040–1046. http://papers.nips.cc/paper/1224-learning-from-demonstration

[54] Abigail J. Sellen. 1992. Speech Patterns in Video-mediated Conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, USA) *(CHI '92)*. ACM, New York, NY, USA, 49–59. https://doi.org/10.1145/142750.142756

[55] David Sirkin and Wendy Ju. 2012. Consistency in Physical and On-screen Action Improves Perceptions of Telepresence Robots. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (Boston, Massachusetts, USA) *(HRI '12)*. ACM, New York, NY, USA, 57–64. https://doi.org/10.1145/2157689.2157699

[56] David Sirkin, Gina Venolia, John Tang, George Robertson, Taemie Kim, Kori Inkpen, Mara Sedlins, Bongshin Lee, and Mike Sinclair. 2011. Motion and Attention in a Kinetic Videoconferencing Proxy. In *Proceedings of the 13th IFIP TC 13 International Conference on Human-computer Interaction - Volume Part I* (Lisbon, Portugal) *(INTERACT'11)*. Springer-Verlag, Berlin, Heidelberg, 162–180. http://dl.acm.org/citation.cfm?id=2042053.2042074

[57] Skype Technologies. 2021. Skype. https://www.skype.com

[58] Randall B. Smith, Tim O'Shea, Claire O'Malley, Eileen Scanlon, and Josie Taylor. 1990. *Preliminary Experiments with a Distributed, Multi-Media, Problem Solving Environment.* North-Holland Publishing Co., NLD, 31–48.

[59] Suitable Technologies Inc. 2020. Beam. https://suitabletech.com

[60] John C. Tang and Scott Minneman. 1991. VideoWhiteboard: Video Shadows to Support Remote Collaboration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New Orleans, Louisiana, USA) *(CHI '91)*. ACM, New York, NY, USA, 315–322. https://doi.org/10.1145/108844.108932

[61] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence. In *Proceedings*

of the 32nd Annual ACM Symposium on User Interface Software and Technology (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 161–174. https://doi.org/10.1145/3332165.3347872

[62] Katherine M. Tsui and Holly A. Yanco. 2013. Design Challenges and Guidelines for Social Interaction Using Mobile Telepresence Robots. *Reviews of Human Factors and Ergonomics* 9, 1 (2013), 227–301. https://doi.org/10.1177/1557234X13502462 arXiv:https://doi.org/10.1177/1557234X13502462

[63] Gina Venolia, John Tang, Ruy Cervantes, Sara Bly, George Robertson, Bongshin Lee, and Kori Inkpen. 2010. Embodied Social Proxy: Mediating Interpersonal Connection in Hub-and-satellite Teams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI '10)*. ACM, New York, NY, USA, 1049–1058. https://doi.org/10.1145/1753326.1753482

[64] Simon Voelker, Sebastian Hueber, Christian Holz, Christian Remy, and Nicolai Marquardt. 2020. GazeConduits: Calibration-Free Cross-Device Collaboration through Gaze and Touch. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3313831.3376578

[65] Wei-Chao Wen, H. Towles, L. Nyland, G. Welch, and H. Fuchs. 2000. Toward a compelling sensation of telepresence: demonstrating a portal to a distant (static) office. In *Proceedings Visualization 2000. VIS 2000 (Cat. No.00CH37145)*. 327–333. https://doi.org/10.1109/VISUAL.2000.885712

[66] Xandex Inc. 2021. KUBI Telepresence. https://kubiconnect.com/

[67] Nicole Yankelovich, Nigel Simpson, Jonathan Kaplan, and Joe Provino. 2007. Porta-person: Telepresence for the Connected Conference Room. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems* (San Jose, CA, USA) *(CHI EA '07)*. ACM, New York, NY, USA, 2789–2794. https://doi.org/10.1145/1240866.1241080

[68] Zoom Video Communications Inc. 2021. Zoom. https://zoom.us

Received July 2021; revised November 2021; accepted November 2021