Room2Room: Enabling Life-Size Telepresence in a Projected Augmented Reality Environment

Tomislav Pejsa[1,2], **Julian Kantor**[1,3], **Hrvoje Benko**[1], **Eyal Ofek**[1], **Andrew Wilson**[1]

[1]Microsoft Research
Redmond, WA

[2]University of Wisconsin-Madison, Madison, WI

[3]University of Southern California, Los Angeles, CA

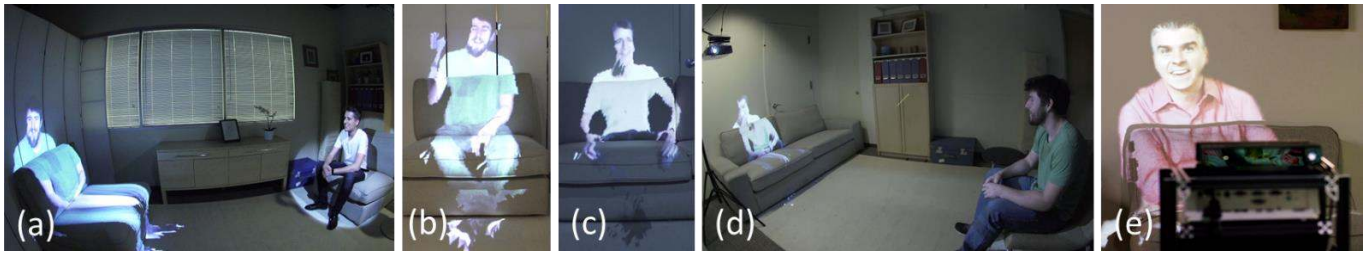

Figure 1. Room2Room uses projected augmented reality to enable co-present interaction between remote participants: (a, d) remote participants are represented as life-size virtual copies projected into the physical space; (b, c) each participant sees their partner's virtual copy with correct perspective and they can communicate naturally using speech and nonverbal cues; (e) following the feedback from our user study, our alternate hardware implementation improves the image quality of projected participants.

ABSTRACT

Room2Room is a telepresence system that leverages projected augmented reality to enable life-size, co-present interaction between two remote participants. Our solution recreates the experience of a face-to-face conversation by performing 3D capture of the local user with color + depth cameras and projecting their life-size *virtual copy* into the remote space. This creates an illusion of the remote person's physical presence in the local space, as well as a shared understanding of verbal and non-verbal cues (e.g., gaze, pointing.) In addition to the technical details of two prototype implementations, we contribute strategies for projecting remote participants onto physically plausible locations, such that they form a natural and consistent conversational formation with the local participant. We also present observations and feedback from an evaluation with 7 pairs of participants on the usability of our solution for solving a collaborative, physical task.

Author Keywords

Telepresence; spatial interfaces; spatial augmented reality; projection-mapping; projector camera system.

ACM 978-1-4503-3592-8/16/02…$15.00
DOI: http://dx.doi.org/10.1145/2818048.2819965

ACM Classification Keywords

H.5.2 [Information Interfaces and Presentation]: User Interfaces - Graphical user interfaces;

INTRODUCTION

Current videoconferencing applications (e.g., Skype, FaceTime) are limited in many ways: they afford only partial views of remote participants, in 2D, on a flat screen, and at a reduced scale. These technical constraints limit the sense of co-presence and the ability to communicate naturally using gaze, gesture, posture, and other nonverbal cues – what Buxton has referred to as shared *person space* [5]. Furthermore, while some applications support the notion of a shared *task space* (e.g., desktop sharing feature in Skype), this task space is typically completely virtual and separate from the person space. Finally, there is limited or no support for the use of nonverbal cues (such as pointing) to refer to objects in the task space–a capability known as *reference space* [5]–which limits many collaborative tasks. Previous research in telepresence systems offered solutions to some of these restrictions: e.g., enabling 3D, view-dependent rendering of participants [2, 15] and supporting gesturing and pointing in the task space [2, 22].

In this paper we propose a novel telepresence system, called *Room2Room*, which uses projected augmented reality (AR) to achieve true integration of person space, task space, and reference space at the life-size scale. We extend an existing spatial AR system, RoomAlive [3, 12], with the ability to capture *virtual copies* of real people and objects in a remote environment and project them into a local physical environment using commodity projectors (Figure 1).

In contrast to traditional videoconferencing approaches, the virtual copy of the remote participant is projected directly

into the physical environment, rendered at life-size scale, and in a view-dependent, perspective-corrected way, such that the local participant can see them from different viewpoints as they move. Furthermore, remote participants are rendered on top of existing real furniture which makes them appear as if they are inhabiting the same space. This facilitates more natural interaction, since people can see each other fully and make better use of nonverbal cues such as gaze, posture, and gestures. Our solution does not require users to wear any display or tracking equipment, nor does it represent them as *avatars* – their appearance and movements are faithfully reproduced on their virtual copies (to within sensor limits).

Room2Room uses a set of three ceiling-mounted projector + camera units at each location capable of projecting on most surfaces of the room. In this room-size setup, virtual copies can be projected onto numerous physical seating affordances or standing in the room, and we explore strategies to their placement such that they form a natural conversational formation with the local participant that is consistent across both spaces. As shown in our user study, the system innately supports collaborative tasks such as physical assembly, since both the participants and task objects are situated in a common space. We also contribute a simplified implementation consisting of a single integrated projector + two camera unit, which allows for easy deployment while limiting flexibility of participants' locations.

Our work makes the following contributions:

- The technical foundation for life-size view-dependent telepresence based on color + depth capture, 3D reconstruction, and view-dependent, projected AR.
- Two hardware implementations: the first optimizes seating flexibility and room-size collaborations; the second focuses on ease of deployment and image quality.
- A discussion of design challenges and solutions for placing the virtual copy of a remote participant into an environment, such that a consistent, integrated person-task space is constructed between two participants.

RELATED WORK

Previous efforts in the development of telepresence systems have focused on improving the sense of shared person space and improving integration of person and task spaces.

Shared Person Space Systems

Pioneering efforts to solve the problem of shared person space are Hydra [19] and MAJIC [9] systems. Hydra is a multiparty videoconferencing system that simulates a round-table meeting by placing Hydra units (combined camera, monitor, and speaker) as proxies for remote participants. Participants are spatially situated and aware of each other's gaze and head turns, but are not life-sized. MAJIC enables similar multi-party interactions where participants are projected life-size on a large, curved screen.

More recently, stereoscopic display technologies and depth cameras have been used to improve the illusion of spatial co-presence between two remote participants. Maimone et al. [15] and Jones et al. [11] respectively use a depth camera array and a 3D face scanner to acquire 3D video of the participant which is displayed to their partners in a view-dependent way. Both employ autostereoscopic displays, so display of remote participants is constrained to a rectangular screen. Another telepresence system by Maimone et al. [16] enables remote participants to be rendered at life-size scale and situated within the physical environment with proper occlusion, but also requires their partner to wear an optical see-through display, greatly limiting field of view and occluding facial expressions.

We propose projecting virtual copies of people directly onto (potentially irregular) surfaces in the physical environment. Raskar et al. [18] demonstrate projection of virtual content and textures onto arbitrary physical surfaces, turning the entire space into an immersive display. Content is rendered with the correct perspective, using a magnetic tracker to obtain the user's viewpoint. More recently, the RoomAlive system [3, 12] features similar capabilities, tracking viewpoint with Kinect sensors rather than wearable devices.

A challenging problem in room-scale telepresence is determining where to place avatars of remote participants without disrupting natural interaction or violating laws of physics. Lehment et al. [14] propose an automated method which aligns remote environments, such that they minimize discrepancies in furniture layout and other features. In later sections we propose solutions to our similar problem of placing remote participants.

Lastly, we acknowledge the large body of work on enabling life-size telepresence in virtual environments (e.g., MASSIVE [7] and blue-c [8]). Furthermore, Benford et al. [1] explores the intersection of the collaborative virtual and mixed reality environments and provide a good taxonomy to understand such hybrid spaces. In contrast to all such collaborative virtual experiences, our work focuses on placing people in their respective physical environments using augmented reality technology and minimal modification of the captured video and depth information.

Person-Task Space Integration

While the works described above situate representations of remote participants within local physical space, they do not consider the integration of task space and participants' shared person space. Buxton identified the space where participants nonverbally refer to task objects (e.g., pointing) as *reference space* [5] – an important capability for remote collaboration. Examples of teleconferencing systems that attempt to include both person and task spaces in teleconferencing are ClearBoard [10], Video Whiteboard [23], DigitalDesk [25] and, more recently, IllumiShare [13]. These systems support joint activities such as drawing by rendering participants' hand movements as they gesture and manipulate objects in task space. However, the variety of tasks is limited by the available 2D surface.

The availability of inexpensive depth cameras has led to systems that support new forms of interaction within shared task space. For example, Sodhi et al. [22] demonstrate the use of mobile devices equipped with depth cameras to capture task objects and particpants' gestures during collaborative 3D assembly tasks. In this approach, reference space is separate from the physical task space and restricted to the small screen of the mobile device. Zillner et al. [27] present a system similar in spirit to ClearBoard [10], representing remote participants as virtual, depth-captured copies; however, these are displayed "behind" the 3D-board display, not within the local participant's space.

Closely related to the present work, MirageTable [2] enables hands-free interactions with captured 3D objects in a reference space and supports 3D capture and display of remote participants on a curved screen at life-size scale. MirageTable is similar to our system in terms of technology, but display and interaction are restricted to the small area above the physical screen.

In summary, while previous work addresses many challenges in enabling life-size telepresence, our solution is, to our knowledge, the most comprehensive effort to develop a range of features needed to support natural, co-present interaction between remote participants in a shared space at the scale of an entire room, without the use of obtrusive equipment or wearable devices.

ROOM2ROOM TELEPRESENCE SYSTEM

The Room2Room telepresence system uses the RoomAlive infrastructure [3, 12] and extends its capabilities to two separate locations. A complete, room-scale, projected AR system is deployed in each room comprised of three ceiling-mounted projector-camera units or *procams* (Figure 2). Each procam includes of a Microsoft Kinect v2 color + depth camera and a commodity wide field-of-view projector (BenQ 770ST). Kinect sensors capture the geometry and appearance of the environment and people in it while projectors display virtual content in the environment, including virtual copies of people and objects. Each Kinect is hosted by a PC which serves Kinect sensor data such as depth, color, body tracking (user's skeleton joint positions), and audio to clients via the network.

In order to project virtual content in precise alignment with the physical environment, the system must first be calibrated. A cloud of 3D points is captured during the calibration process, which is then used to reconstruct the static 3D geometry and appearance of the room (Figure 2c) – this includes stationary features such as the walls, floor, and furniture. Given this information, virtual content may be precisely aligned with physical objects in the room. The calibration is further described in RoomAlive paper [12] and is publicly available as the RoomAlive Toolkit[1].

[1] http://github.com/Kinect/RoomAliveToolkit

Room2Room uses the Unity game engine to author interactive AR experiences that take place in the room (Figures 2c and 3d). The recovered room geometry is loaded into the Unity workspace and virtual content is placed into the scene and projected into the physical environment of the room.

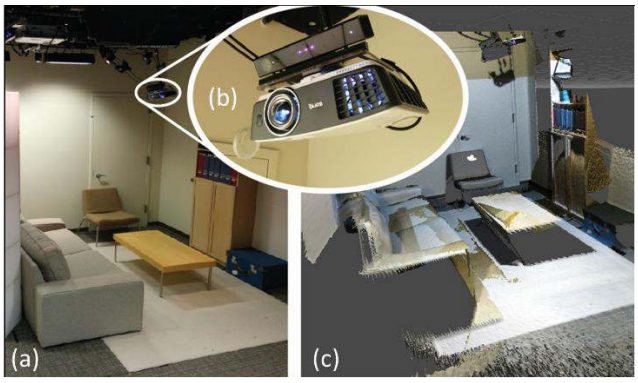

Figure 2. An example Room2Room installation: (a) 3 procam units are deployed in the ceiling of our room; (b) an example procam, consisting of a projector and Kinect; (c) recovered 3D room geometry visualized in Unity editor.

As the user moves about the room, virtual objects are rendered from their viewpoint as found by Kinect tracking [12]. This approach gives a partial illusion of three-dimensionality – virtual objects can be viewed from different angles and have motion parallax, although depth perception is incomplete due to lack of stereopsis. In practice, the approach works quite well as long as virtual objects are positioned close to physical projection surfaces – an important consideration in the design of our system. This is particularly advantageous for seated users as their bodies and their virtual copies tend to be close to the surface of the chair they are sitting in (see Figure 1). We could achieve a better 3D effect by using stereo projectors and shutter glasses [2, 26], but we choose to avoid the use of wearable devices that might restrict the user's field of view and obstruct communication using facial cues.

The remote procams capture the color, depth, and body tracking data of the person in the remote room and are used to reconstruct their 3D geometry and appearance locally, thus obtaining their virtual copy. The virtual copy is then positioned and oriented in the local room and projected into its physical environment. The projection is rendered in a view-dependent manner, based on the viewpoint of the local participant inferred using the local Kinects. In addition, the client also connects to the audio stream from one of the procams in the remote room and thus obtains the speech of the other participant. The analogous procedure is applied on the remote client to obtain and render a virtual copy of the participant in the local room; thus the real person in each room is able to see and hear the virtual copy of the person from the other room in real time.

To facilitate remote, dyadic interaction, Room2Room system addresses two key challenges: (1) capturing people and

objects in the local environment, and (2) positioning and orienting their virtual copies in the remote environment.

Capturing People and Objects

It is relatively straightforward to use the procams to capture people and real-world objects and reconstruct their virtual copies. This is achieved by using the Kinect as a real-time 3D capture device. The procedure involves several steps:

1. *Background acquisition*. The depth texture of the static room (containing no people or non-stationary objects) is captured and averaged over multiple frames, yielding a background depth texture.
2. *Foreground extraction*. At runtime the captured depth is compared to the stored background depth. Objects that are closer to the camera relative to the background are extracted as foreground objects.
3. *3D reconstruction*. Reconstruction of the 3D geometry and appearance of the captured person or object occurs during rendering on the client's end (implemented as a GPU shader). A textured mesh for the user is created from the foreground depth data and the color texture obtained from the RGB camera (similarly to [2, 3]). This textured mesh is projected into the room.

Streaming both depth and color information at runtime requires substantial network bandwidth. Color textures in particular are very large, due to their high resolution. To conserve bandwidth, color textures are JPEG compressed.

The capture process will acquire not just people moving around the room, but also any real-world objects that were not present in the room during calibration–we refer to such objects as *dynamic objects*. The participant can handle dynamic objects and their partner will see their virtual copies in the remote room, which is an important feature for remote collaboration. For example, the cubes in our evaluation study are implemented this way.

Situating People

Having captured and streamed a virtual copy to the local client, we need to project it at the appropriate location in the local room. At the low level, this is a simple matter of translating, rotating, and rendering the virtual copy's depth mesh. However, determining its position and orientation is challenging in several ways. First, the virtual copy should be positioned close to static room geometry (e.g., close to a wall and on top of a couch) that will serve as a projection surface, otherwise projection quality will degrade. While a stool in the center of the room might be a valid seating affordance, it is a poor choice for placing a virtual copy, as it lacks vertical projection surfaces in the vicinity. In our experience, degradation in quality becomes noticeable at distances greater than about 1m from the projection surface. Second, the copy must be situated in a physically plausible way, such that it does not float in the air or intersect other virtual or real-world objects and people in the room. Third, the copy must maintain natural conversational formation with the real person–the virtual and real person should face each other, they should sit (stand) at an appropriate interpersonal distance from one another, and there should be no obstacles blocking the line of sight between them. Lastly, relative geometric relations between the participants should be as similar as possible in *both* rooms, otherwise the shared person-task space will appear inconsistent, i.e., distance between the participants might be different in each room, nonverbal cues such as gaze and pointing might have incorrect direction, making it impossible for the participant to correctly indicate objects in the task space.

In the current implementation, we made several design choices that simplify the problem of situating participants and aligning the shared space. First, rather than automatically determine suitable seating or standing spots for the placement of virtual copies, we require that the designer label them manually using the editor. Furthermore, Room2Room does not situate and show participants' virtual copies until they have settled into a relatively stationary seating or standing position. While participants are walking around their local rooms, they remain invisible in the remote space to avoid appearing as if they are walking through furniture or floating in midair. We use an action inference model to determine when both participants are sitting down or standing still, and only then do we map their virtual copies to suitable seating (standing) affordances. If the situation changes (e.g., one person changes seats), Room2Room automatically remaps the virtual copy of the remote user to the next best position (Figure 3).

Uncertainty in placement of the virtual copy is conveyed to the local user by making the virtual copy invisible while moving. As the virtual copy is introduced to the scene or removed, a flickering and fade-in effect (Figure 4c) simulates the appearance of analog interference; this creates some anticipation and adds to the "hologram aesthetic".

Labeling Affordances

When determining where to situate and project the virtual copy of the remote participant, the system chooses from a

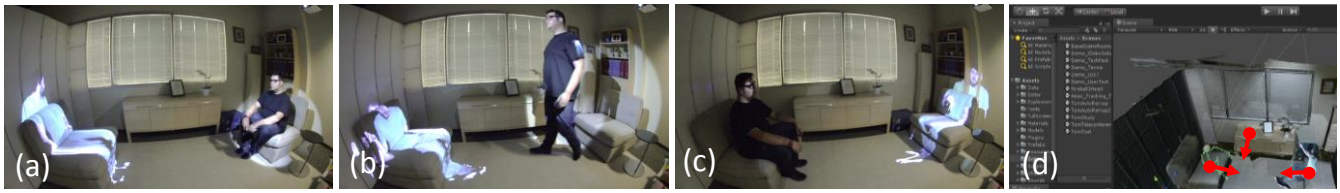

Figure 3. A sequence showing Room2Room placement capabilities: (a-c) when the local user changes their seat, Room2Room automatically remaps the remote user to the other available seat; (d) the view of that same interaction in the Unity editor, which is also used to annotate seating and standing locations (indicated with red arrows).

discrete set of available seating and standing locations in the room. Currently, these locations must be labeled manually by the designer in an authoring step, using the Unity editor. As illustrated in Figure 3d, the designer has a 3D view of the room's geometry in the editor. They label an affordance by creating a special tagged object at its location and adjusting its orientation (indicated by red arrows in Figure 3d). We note that any location can be labeled as a seating affordance, even if it does not correspond to a chair. For example, a seated, remote participant could be projected onto an empty wall as long as that location is labeled as a valid seating affordance. To avoid participants appearing as if they are sitting in mid-air, their physical chair in the remote room can be treated as a dynamic object and captured and projected along with the participant.

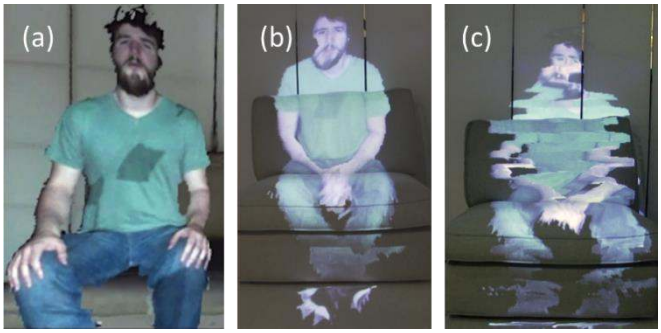

Figure 4. (a) A virtual copy of the remote user shown in our editor; (b) the same virtual copy when projected in the real chair; (c) the effect of flickering and fading into existence.

Participant Action Inference

In order to determine how a person's virtual copy should be situated in the remote room, we must first characterize their movement in their local room. We have implemented a set of simple heuristics for determining the person's current movement action (*walking*, *standing-in-place*, or *sitting*) and their movement target. i.e., the local affordance that they are sitting (standing) on. Despite their simplicity, these heuristics have proven sufficiently robust for our purposes.

We determine the movement action by analyzing the velocity (walking or not) and height (sitting or standing) of the person's root joint, obtained from the Kinect skeleton data stream. Next, we infer the movement target. Even as the person is walking, we try and predict where they will sit down. The advantage of doing so is that we can situate virtual copies sooner: as the person approaches a seating affordance in order to sit down, the remote person can see the virtual copy walk up to and sit down on a chair in their own room. Target inference examines all candidate affordances and chooses the most likely target based on Euclidean distance of the target from the person and the ray projected in the direction of the person's movement.

Mapping to Affordances

Once the movement and intent of each person in the local room is known, a telepresence connection between the two rooms may be established. As part of this connection, virtual copies must be mapped to physical affordances in their remote rooms. Two mapping schemes are supported: *predefined* and *optimal facing* mapping. *Predefined* mapping is useful when we know in advance where people are going to be seated or standing in their respective rooms, so we can also predefine the locations where their virtual copies are going to be placed. For example, we use a predefined mapping in our user study and we designate in each room a suitable target seating area for the virtual copy.

Under *optimal facing* mapping, the system computes all valid mapping configurations, where "valid" means each virtual copy is assigned a seating (standing) affordance not occupied by another participant. These configurations are then evaluated according to participants' positions relative to one another. The evaluation metric takes into account interpersonal distance (*i.e.,* the distance between the real person and the remote person's virtual copy) and relative orientation (*i.e.,* the angle at which the virtual copy and the real person face each other). The metric is configurable – a designer using the Unity editor can set the values of optimal facing angle and interpersonal distance to achieve different conversational formations and levels of intimacy.

Having determined the mappings of virtual copies to affordances, we rigidly translate and rotate their depth meshes to bring them into alignment with the affordances. If the person is already seated (standing), we simply apply the position and orientation of the assigned affordance to the virtual copy. If the person is walking toward an affordance in their local room, we compute the translational and rotational offset between their root and the target local affordance, and we position and orient the virtual copy such that it has the same translational and rotational offset relative to the assigned affordance in the remote room. That way, when the real person sits down on a local chair, its virtual copy will appear to sit down on the remote chair.

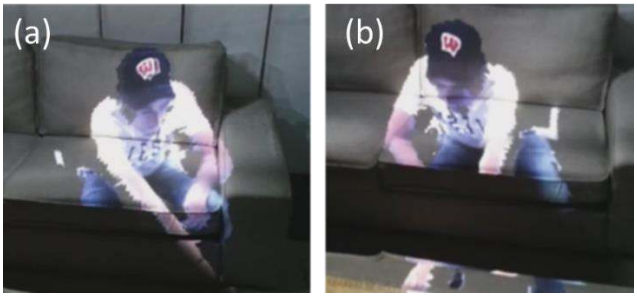

Figure 5. Example of virtual copy mirroring. In the image on the right, the projection of the virtual copy is mirrored and their pointing direction is flipped.

Participants must have similar orientation relative to each other in both rooms. At the minimum, this means they must be situated to the same side of each other in both rooms, otherwise virtual copies' nonverbal cues (eye gaze, pointing gestures) will be oriented incorrectly relative to the real person. This may be impossible to achieve in some room layouts. As depicted in Figure 5, we deal with such situa-

tions by mirroring both virtual copies along the horizontal axis, which inverts the direction of their nonverbal cues. We implement mirroring as a transformation applied to virtual copy's depth mesh and tracked skeleton. When determining optimal facing mapping, we consider mirrored virtual copies among the possible configurations.

Because we want the conversational formation between participants to be consistent across both rooms, it is necessary to determine the mapping configuration in both rooms simultaneously. For this reason, one of the two telepresence clients is designated as *master*. As depicted in Figure 3, the master also has access to affordance labels in the remote room. It determines the mappings of virtual copies in both rooms and supplies the remote client with the position and orientation of its virtual copy. Once a mapping configuration has been computed and applied to the virtual copies, there is no need to re-compute it until one or both participants have moved to a different spot.

Visual Quality Issues

Projections of virtual copies suffer from visual artifacts that are a consequence of the capture process. There are several causes for these artifacts. First, the resolution of the Kinect depth camera is 512x424 pixels, which is too low to capture high-frequency detail on the person's face and body from ceiling mounted cameras. As a result, the geometry of the person's face appears quite amorphous and most of the facial detail comes from the color texture. Second, although our setup incorporates multiple Kinect sensors, we currently use only one sensor at a time to capture the depth and color image of the person. Other researchers have proposed methods of obtaining higher-quality capture using multiple sensors [15], but such extensions are beyond the scope of the current work. Third, our system can be affected by line of sight and occlusion issues, which might result in missing some parts of the virtual copy if they are not visible by the capture camera. Lastly, the contour of the captured virtual copy is the area where the noise in capture and segmentation is most visible. We have improved the visual quality in the alternate implementation discussed below following feedback from our user study.

The visual appearance of the projected virtual copy is also sensitive to lighting conditions in both the remote room (where the person was captured) and the local room (where their copy is being projected). Room illumination needs to be bright enough to acquire the color texture of the local person, but not so bright as to make the projection of the remote person difficult to see. We address this issue by using relatively dim lighting in both rooms and shining *a projected spotlight* onto the person in each room. The spotlight follows the person as they move around and dynamically adjusts its beam width to only illuminate the person and the small area around them. This ensures that each person is captured with relatively bright illumination.

SYSTEM EVALUATION

We evaluated the Room2Room system in a study with 7 pairs of participants (14 total, 6 female), who engaged in a collaborative assembly task. 11 participants reported they were frequent teleconferencing (Skype) users. The goals of our evaluation were to observe how life-size, room-size scale benefited the participants' sense of co-presence when interacting using our system, how integration of person-task space and reference space support would benefit the users' performance in a collaborative, three-dimensional assembly task, and how satisfied the users were with our system.

Design

The participants were asked to construct complex, three-dimensional shapes out of cubic or prismatic blocks. One of the participants was designated *instructor* and handed a schematic of the shape that needed to be constructed (Figure 6a and 6c). The other participant was the *assembler*– they were seated by a green table in the right room (Figure 6b and 6d) with all the available blocks laid out in front of them and their job was to construct the shape by following the instructor's guidance, choosing the correct blocks, and assembling them into the correct shape. The instructor was permitted to provide verbal and nonverbal instructions–e.g., they could describe the correct blocks to the participants, they could gaze at or point to blocks–but they were not allowed to show the schematic to the assembler or to physically move the blocks. There were ten blocks total, but only five blocks were used in each shape. To make the task more challenging, the blocks had different patterns of shapes and colors on each side. Figure 6e-g shows the three patterns of 5 blocks used in the task.

Our study followed a within-participants design with three conditions:

1. *Room2Room* – Participants were seated in separate rooms and used our system to communicate while solving the task (Figure 6c-d). They could see virtual copies of each other in their respective rooms, projected onto seating spots opposite them. The instructor, who was seated in the left room, could also see a virtual copy of the table with blocks.
2. *Skype* – Participants were seated in separate rooms and used a pair of tablets (Microsoft Surface) with Skype to communicate while solving the task (Figure 6a-b). Participants were allowed to hold, move, and set down the tablets however they liked. Both tablets had front-facing and back-facing cameras, and participants were allowed to switch between them at will. By default, the assembler's Skype instance was configured to use the back-facing camera, so the assembler could more easily point it at the task space.
3. *Face-to-face* – Participants were both seated in the right room and communicated face-to-face.

We chose to compare Room2Room to the two possible extremes of the interaction space: the face-to-face condition is what our system is trying to emulate, and the Skype condi-

tion that represents the current "standard" teleconferencing experience. While comparisons to other prototype life-size systems (e.g., [2] or [27]) could also offer novel insights, they remain future work.

Setup

Due to space and complexity constraints, we evaluated the system by splitting our lab space into two separate "rooms" using a physical divider. We refer to the rooms as the "left" and "right" room, respectively. Both rooms had a similar layout consisting of a pair of chairs and a chair and sofa, respectively, placed opposite one another (Figure 6). Room2Room was deployed in each room and consisted of three procams each. Each room also had its own, dedicated computer, which rendered all the graphics.

While our system has full audio streaming capabilities, we disabled audio streaming in all conditions since, in our set-up, the participants could hear each other very well due to proximity of our "rooms" to each other and audio streaming therefore created an undesirable echo effect.

Procedure

We used permuted-blocks assignment to assign dyads to conditions and task shapes. Participants' assigned roles (instructor or assembler) remained fixed throughout the study. Each pair of participants was ushered into the right room, where they were informed about the task and the telepresence tools they would be using to solve it in each condition. We recommended a strategy for solving the task, suggesting that the participants first identify the five blocks needed for the current shape and then figure out how to assemble the shape. Before the start of each condition, the experimenter would arrange the task blocks on the assembler's table in a pseudorandom fashion.

Participants were initially seated at the start of each condition, but we told them they were allowed to move around during the task, as long as they did not leave their designated room. In *Room2Room* condition, they were acquainted with the system's limitations and how movement might impact projection quality on the other end.

We timed the participants during each trial of the task and recorded their completion time. After the third and final trial, participants completed a subjective questionnaire and were interviewed about their experience. Finally, each participant was given their payment (a $10 gift card). The study took about one hour to complete, including the time needed for the questionnaire and interview.

Measures

The study included one objective and two subjective measures. The objective measure was *completion time*–the time it took the participants to complete the task. For the subjective measures we used a modified version of the questionnaire from [13]. While there are several widely-used presence questionnaires developed for virtual environments (e.g., ITQ [21]), we chose not to use them, since they have been shown to be unreliable for comparison of experiences across environments (e.g., comparing virtual to real environments) [24].

Our questionnaire consisted of 13 questions per condition (39 total), asking the participants to rate aspects of their experience with each of the three systems. All questions utilized a 7-point rating scale. We ran a maximum likelihood factor analysis on the data and found that most of the variance in the responses between our conditions could be explained by two sets of highly correlated questions. We labeled these aggregated subjective measures "Presence" and "Efficiency of Communication", respectively:

1. *Presence* – Two-item measure of the participant's feeling of presence (Cronbach's $\alpha = 0.941$). Questions contributing to this measure: "*I felt like my partner was in the room with me*" and "*It felt like I was communicating face-to-face with my partner*".
2. *Efficiency of Communication* – Two-item measure of the participant's feeling that communication with the other participant was fluid and efficient (Cronbach's $\alpha = 0.862$). Questions contributing to this measure: "*Interaction with my partner was fluid and efficient*" and "*I was able to get my partner to understand me*".

We were unable to construct a reliable scale for measuring user satisfaction, so we relied on participants' qualitative comments to assess their satisfaction.

Results

We analyzed the data from our measures using one-way ANOVA. We first analyzed the objective measure, *completion time*. We found that completion time was significantly lower in *Face-to-face* condition than in *Room2Room* condition, $F(1,33) = 15.43, p = 0.0004^{*}$. Furthermore, completion time was significantly lower in *Room2Room* condition than in *Skype* condition, $F(1,33) = 5.47, p = 0.0255^{*}$.

Figure 6: Experimental task setup: (a) Instructor in *Skype* condition. (b) Assembler in *Skype* condition assembling a shape based on her partner's instructions. (c) Instructor in *Room2Room* condition; note the virtual copy of the task space projected in front of her. (d) Assembler performing the task in *Room2Room* condition. (e-g) Target shapes used in the assembly task.

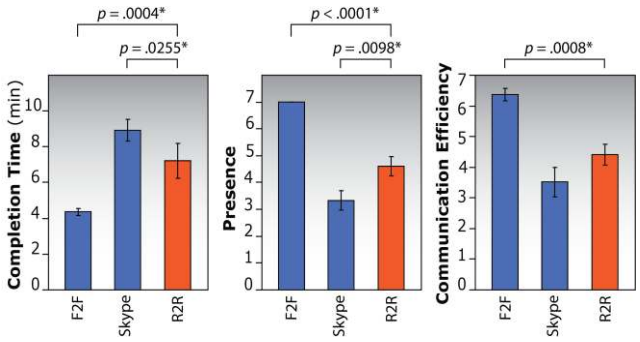

Figure 7. Results from *completion time*, *presence*, and *communication efficiency* measures in each condition.

We also used ANOVA to analyze the subjective measures, in accordance with findings that summative ratings from Likert scales can be treated as interval data and reliably analyzed using parametric tests [6]. Participants rated the sense of *presence* in *Room2Room* condition significantly lower than in *Face-to-face* condition, $F(1,26) = 27.14, p < .0001^*$, but it was also significantly higher in *Room2Room* condition than in *Skype* condition, $F(1,26) = 7.76, p = 0.0098^*$. Moreover, participants found the *efficiency of communication* to be significantly greater in *Face-to-face* than in *Room2Room* condition, $F(1,26) = 14.25, p = 0.0008^*$. There was no significant difference in reported efficiency of communication between *Room2Room* and *Skype*. Figure 7 shows the results from all three measures.

We also analyzed *completion time* with respect to task shape (1, 2, or 3, see Figure 6e-g) and confirmed that neither shape was significantly easier or more difficult to complete than others. We also found no significant effect of task role (instructor vs. assembler) on either *presence* or *efficiency of communication*. Finally, trial number had no significant effect on *completion time*, indicating there were no learning effects in repeated trials of the task.

Discussion

Quantitative results suggest that, predictably, face-to-face communication is superior in terms of task completion time, sense of presence, and efficiency of communication. Furthermore, using our system instead of Skype led to *significant improvements in task completion time,* suggesting that our system makes spatially situated tasks easier to perform. Finally, participants felt their task partners were *significantly more present* when projected into their physical space using Room2Room, than when they were displayed on a 2D screen in Skype. We note that while physical proximity of the rooms might have skewed the presence measure in *Room2Room* and *Skype* conditions, the effect would have been equally present in both conditions and thus unlikely to affect comparison results.

The subjective questionnaire also included four open-ended questions, asking users to describe what they liked and disliked about their experiences with Skype and Room2Room, respectively. Moreover, we interviewed all participants about their experiences, asking them to qualify their use of speech and nonverbal cues, ease of understanding and making themselves understood, use of space and different viewpoints, fidelity of virtual copy representation and its impact on the experience, and their overall satisfaction with the Room2Room system.

Participants' answers suggest a heavy reliance on verbal communication to solve the task, even in *Face-to-face* condition. Instructors made use of deictic gestures (e.g., pointing at a block) and iconic gestures (e.g., describing how a block should be rotated), but assemblers made use of these gestures less frequently in *Room2Room* condition due to low fidelity of the virtual representation. This was confirmed by reviewing video data from the study.

Visual fidelity issues with Room2Room, such as low resolution, incomplete reconstruction, warping, and gaps in projection, were brought up as major problems by almost all our participants. These issues interfered with some usage aspects of the system, e.g., making it harder to visually identify blocks. Participants liked to engage in a joint behavior where the assembler picked up a block from the table and rotated it so the instructor could see it from all sides. However, projection quality would substantially degrade when this happened, whereas no such problems occurred when using Skype. This visual quality feedback was a major motivator for our alternate system implementation discussed below.

Comparisons with Skype have brought to light several advantages and disadvantages of Room2Room. Participants liked the ability to view blocks from different sides by simply getting up and moving. They also remarked on the benefits of reference space, especially the ability to see the partner's hands as they moved the blocks around. A major shortcoming of Room2Room in comparison with Skype was participants' inability to see what their partner was seeing. While Skype interface shows views of both the current user's camera and their partner's camera, our system lacks such a feature, which made it more difficult for participants to gauge projection quality on the other end and whether the partner could see their nonverbal cues. In general, participants had insufficient knowledge of the system's limitations and, as a result, were hesitant to take full advantage of its capabilities–e.g., refraining from using pointing gestures or getting up from their seats in order to get closer to the task objects.

Overall, participants were satisfied with Room2Room and judged it as useful for collaborative assembly tasks. They expressed interest in using a similar system instead of traditional videoconferencing. Some participants saw the potential of the system to turn formerly solitary activities, such as online gaming or watching television, into intimate experiences–e.g., two friends could watch television in the company of each other's virtual copies, while each physically sits in their own living room. These responses underscore

that the current task merely scratches the surface of potential applications of the Room2Room system.

ALTERNATE IMPLEMENTATION

Based on user feedback from the experiment and our own implementation experiences we designed an alternate Room2Room implementation to address two major shortcomings: (1) the complexity of the system, and (2) the low reproduction quality of the virtual copies.

One obstacle to deploying Room2Room in homes and offices is the required installation of ceiling-mounted projectors and cameras. We simplified the hardware requirements by integrating the necessary hardware into a single physical unit per room (Figure 8a) consisting of one forward-facing, short-throw projector (Optoma W305ST), two Kinect v2 cameras (one forward-facing and another rear-facing), a set of speakers, and two small-form factor computers driving the experience. The two computers are only required due to the fact that a PC can only host a single Kinect v2 sensor.

We designed this unit as a standalone device, which can be easily placed in the middle of the room, between the real person and the location where the virtual copy is projected (Figure 8b). The forward Kinect captures the geometry of the environment to enable correct projections of the virtual copy, while the rear Kinect captures the local user and tracks their head position for view-dependent rendering.

We also improved the capture and reproduction quality in several ways (Figure 8c-d). First, the capture was dramatically improved simply by placing the Kinect cameras lower from the ceiling (so their viewpoint is more aligned with viewpoints of people in the room) and bringing them closer to the people being captured. This resulted in higher resolution capture, with fewer artifacts due to off-axis camera placement or occlusion of body parts. Second, visually distracting artifacts due to noise in the depth image were reduced by a series of image processing steps, including upsampling the depth image threefold, bilateral filtering to obtain smoother depth estimates, smoothing the contour of the foreground mask, and finally, feathering the filtered contours in final rendering. These improvements resulted in noticeably better visual quality of virtual copies.

In contrast to the earlier, room-size version, this configuration restricts the projection of the virtual copy to the relatively small area in front of the projector and therefore precludes any virtual copy movement in space. Furthermore, camera arrangement mandates that participants interact face-to-face. While this works best in a symmetrical arrangement (e.g., two opposing chairs on both ends), having similar furniture is not required. Our projection mapping algorithm can accurately project the image of the remote person onto any non-transparent and non-reflective surface (e.g., corner, wall, or bookcase). While a full evaluation remains future work, preliminary user feedback suggests that the simplicity of this setup and improved visual quality make for a better user experience.

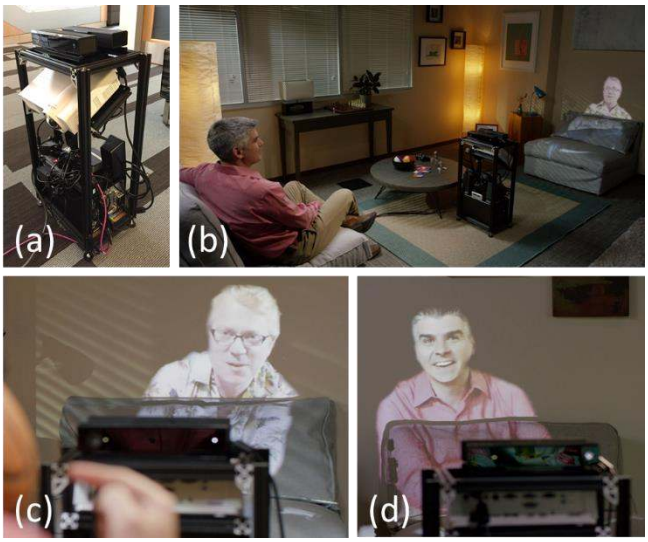

Figure 8. Alternate Room2Room implementation: (a) an integrated projector + 2xKinect + 2xPC unit; (b) the unit is placed between the participants; (c-d) images of the remote participants show improved image quality.

CONCLUSIONS AND FUTURE WORK

We have developed Room2Room to enable remote participants, represented as life-size virtual copies projected into each other's physical environment, to engage in real-time, co-present interaction. Our system does not require participants to wear any specialized equipment, it enables them to move freely and view each other from different angles with correct perspective, and implicitly gives them a common reference space where they can interact naturally using nonverbal cues.

The system is currently limited to one-on-one interactions in spaces with compatible layouts. Our goal is to extend Room2Room to multiparty interactions in larger, more complex, and more diverse spaces, such as large living rooms, boardrooms, and classrooms. This requires the development of automated techniques for identifying seating affordances, extensions to view-dependent rendering to support multiple viewpoints (e.g., [3]), and more advanced mapping of virtual copies to affordances.

While our evaluation has focused on collaborative work, we also hope to investigate how our system might facilitate more intimate and empathic interactions. Nonverbal cues afforded by life-size telepresence, such as posture and proxemics, are important contributors to empathy, and their effects have been demonstrated in recent studies [17].

We believe that Room2Room is an important step toward high-fidelity, life-size 3D telepresence. Rather than relying on technologies which take the user out of their environment (e.g., collaborative VR solutions), we pursue a vision which inserts remote participants into the user's environment, exploiting the environment's affordances, and emulating the experience of face-to-face conversation.

REFERENCES

1. Steve Benford, Chris Greenhalgh, Gail Reynard, Chris Brown, and Boriana Koleva. 1998. Understanding and constructing shared spaces with mixed-reality boundaries. *ACM Trans. Comput.-Hum. Interact.* 5, 3 (Sept. 1998), 185-223.
2. Hrvoje Benko, Ricardo Jota, and Andrew Wilson. 2012. MirageTable: freehand interaction on a projected augmented reality tabletop. In *Proc. ACM CHI '12*.199-208.
3. Hrvoje Benko, Andrew Wilson, and Federico Zannier. 2014. Dyadic projected spatial augmented reality. In *Proc. ACM UIST '14*. 645-655.
4. Judee K. Burgoon, David B. Buller, Jerold L. Hale, and Mark A. de Turck. 1984. Relational messages associated with nonverbal behaviors. *Human Communication Research 10*, 3 (1984), 351-378.
5. Bill Buxton. 2009. Mediaspace – Meaningspace – Meetingspace. Harrison, S. (Ed) *Media Space 20 + Years of Mediated Life*, Springer.
6. James Carifio and Rocco Perla. 2008. Resolving the 50-year debate around using and misusing Likert scales. *Medical Education 42*, 12 (2008), 1150-1152.
7. Chris Greenhalgh and Steven Benford. 1995. MASSIVE: a collaborative virtual environment for teleconferencing. *ACM Trans. Comput.-Hum. Interact.* 2, 3 (Sept. 1995), 239-261.
8. Markus Gross, Stephan Würmlin, Martin Naef, Edouard Lamboray, Christian Spagno, Andreas Kunz, Esther Koller-Meier, Tomas Svoboda, Luc Van Gool, Silke Lang, Kai Strehlke, Andrew Vande Moere, and Oliver Staadt. 2003. blue-c: a spatially immersive display and 3D video portal for telepresence. In *Proc. ACM SIGGRAPH '03*. 819-827.
9. Yusuke Ichikawa, Ken-ichi Okada, Giseok Jeong, Shunsuke Tanaka, and Yutaka Matsushita. 1995. MAJIC videoconferencing system: experiments, evaluation and improvement. In *Proc. ECSCW '95*. 279-292.
10. Hiroshi Ishii and Minoru Kobayashi. 1992. ClearBoard: a seamless medium for shared drawing and conversation with eye contact. In *Proc. ACM CHI '92*. 525-532.
11. Andrew Jones, Magnus Lang, Graham Fyffe, Xueming Yu, Jay Busch, Ian McDowall, Mark Bolas, and Paul Debevec. 2009. Achieving eye contact in a one-to-many 3D video teleconferencing system. *ACM Trans. Graph. 28*, 3 (2009).
12. Brett Jones, Rajinder Sodhi, Michael Murdock, Ravish Mehra, Hrvoje Benko, Andrew Wilson, Eyal Ofek, Blair MacIntryre, Nikunj Raghuvanshi, and Lior Shapira. 2014. RoomAlive: magical experiences enabled by scalable, adaptive projector-camera units. In *Proc. ACM UIST '14*.
13. Sasa Junuzovic, Kori Inkpen, John Tang, Mara Sedlins, and Kristie Fisher. 2012. To see or not to see: a study comparing four-way avatar, video, and audio conferencing for work. In *Proc. ACM GROUP '12*. 31-34.
14. Nicolas Lehment, Daniel Merget, and Gerhard Rigoll. 2014. Creating automatically aligned consensus realities for AR videoconferencing. In *Proc. IEEE ISMAR '14.*
15. Andrew Maimone and Henry Fuchs. 2011. Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras. In *Proc. IEEE ISMAR '11.* 137-146.
16. Andrew Maimone, Xubo Yang, Nate Dierk, Andrei State, Mingsong Dou, and Henry Fuchs. 2013. General-purpose telepresence with head-worn optical see-through displays and projector-based lighting. In *Proc. IEEE VR '13*. 23-26.
17. David T. Nguyen and John Canny. 2009. More than face-to-face: empathy effects of video framing. In *Proc. ACM CHI '09*. 423-432.
18. Ramesh Raskar, Greg Welch, Matt Cutts, Adam Lake, Lev Stesin, and Henry Fuchs. 1998. The office of the future: a unified approach to image-based modeling and spatially immersive displays. In *Proc. ACM SIGGRAPH '98*. 179-188.
19. Abigail J. Sellen. 1992. Speech patterns in video-mediated conversations. In *Proc. ACM CHI '92*. 49-59.
20. John Short, Ederyn Williams, and Bruce Christie. 1976. *The Social Psychology of Telecommunications*. Wiley, New York, NY.
21. Mel Slater. 1999. Measuring Presence: A Response to the Witmer and Singer Presence Questionnaire. *Presence: Teleoper. Virtual Environ. 8, 5* (Oct. 1999), 560-565.
22. Rajinder S. Sodhi, Brett R. Jones, David Forsyth, Brian P. Bailey, and Giuliano Maciocci. 2013. BeThere: 3D mobile collaboration with spatial input. In *Proc. ACM CHI '13*. 179-188.
23. John C. Tang and Scott Minneman. 1991. VideoWhiteboard: video shadows to support remote collaboration. In *Proc. ACM CHI '91*. 315-322.
24. Martin Usoh, Ernest Catena, Sima Arman, and Mel Slater. 2000. Using Presence Questionnaires in Reality. *Presence: Teleoper. Virtual Environ. 9(5)* (Oct. 2000), 497-503.
25. Pierre Wellner. 1993. Interacting with paper on the DigitalDesk. *Commun. ACM 36*, 7 (1993), 87-96.
26. Andrew D. Wilson, Hrvoje Benko, Shahram Izadi, and Otmar Hilliges. 2012. Steerable augmented reality with the Beamatron. In *Proc. ACM UIST '12*. 413-422.
27. Jakob Zillner, Christoph Rhemann, Shahram Izadi, and Michael Haller. 2014. 3D-Board: A shared workspace featuring remote 3D virtual embodiments. In *Proc. ACM UIST '14*. 471-479.