Dyadic Projected Spatial Augmented Reality

Hrvoje Benko
Microsoft Research
Redmond, WA, USA
benko@microsoft.com

Andrew D. Wilson
Microsoft Research
Redmond, WA, USA
awilson@microsoft.com

Federico Zannier
Microsoft
Redmond, WA, USA
fzannier@microsoft.com

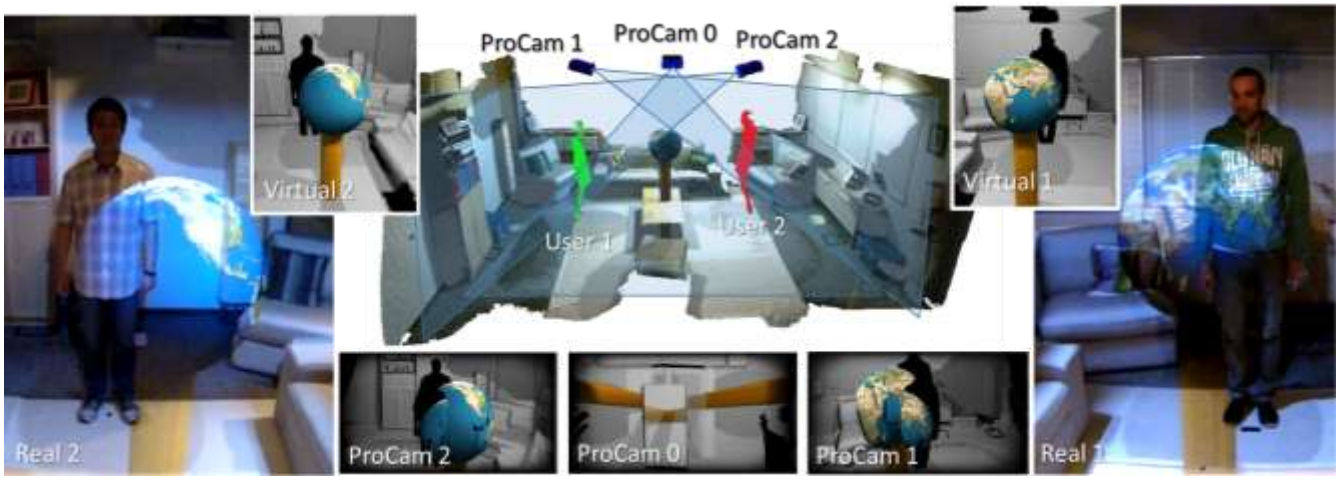

Figure 1. Dyadic projected spatial augmented reality enables two users to interact with a shared virtual scene and each other in a face to face arrangement. Center: Room geometry, user geometry and projector camera pairs are illustrated. Virtual 1 & 2: The desired view for each user is rendered offscreen. ProCam 0, 1 & 2: Projected graphics are warped to account for surface geometry, including the other user. Real 1 & 2: Each user's resulting view compares well with the desired view (Virtual 1 & 2).

ABSTRACT

Mano-a-Mano is a unique spatial augmented reality system that combines dynamic projection mapping, multiple perspective views and device-less interaction to support face to face, or dyadic, interaction with 3D virtual objects. Its main advantage over more traditional AR approaches, such as handheld devices with composited graphics or see-through head worn displays, is that users are able to interact with 3D virtual objects and each other without cumbersome devices that obstruct face to face interaction. We detail our prototype system and a number of interactive experiences. We present an initial user experiment that shows that participants are able to deduce the size and distance of a virtual projected object. A second experiment shows that participants are able to infer which of a number of targets the other user indicates by pointing.

Author Keywords

Augmented reality; projector camera system; depth cameras

ACM Classification Keywords

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

UIST '14, October 05 - 08 2014, Honolulu, HI, USA

http://dx.doi.org/10.1145/2642918.2647402

INTRODUCTION

"Spatial Augmented Reality" (SAR) [26] techniques create augmented reality experiences by changing the look of the physical environment with projected light [4, 25, 35]. Whereas many augmented reality approaches involve rendering graphics over a live video feed on hand held or head worn devices, SAR places the augmenting graphics over the physical object itself and so does not divert the users' attention from the real world. "See-through" head worn displays can achieve similar effects, but presently support a limited field of view and require wearing bulky equipment that can hinder face to face interaction among users.

Under the right circumstances SAR systems can change the surface appearance of an object to make it appear as if it is made of a different material. For example, a carpeted floor can be made to look like a mossy bog. With the knowledge of the user's head position SAR systems can project light over the physical environment so that a virtual 3D object appears correctly over the real world. In both cases it is necessary to have a precise geometric model of the physical environment. Intuitively, this model is used to alter the projected graphics to account for the distortion of the projected image caused by variation in the real world surface geometry. When the physical environment includes moving

objects (e.g., people), this model must be updated in real time. Commodity depth cameras such as the Microsoft Kinect now make it relatively easy to compute geometric models of dynamic indoor environments. Because it can be used to render virtual 3D objects and scenes, SAR may seem like a practical, available alternative to holographic video. However, in comparison to true holographic techniques, SAR appears very limiting, since by its use of view-dependent rendering it seems to support a single view and therefore only a single user. However, when using projection techniques in larger environments (e.g., throughout a room) there may be configurations where what is visible to one user is not visible to another, thus creating the possibility for multiple simultaneous views.

In this paper we explore one such configuration that is particularly useful: face to face, or dyadic, interaction. We demonstrate *Mano-a-Mano*, a projected SAR system which supports separate perspective views for two users when they are arranged face to face in a room with several feet of space between them. *Mano-a-Mano* renders virtual 3D objects as if they are hovering in the space between the two users. Moreover, various interactions with these objects are supported. For example, in a combat-style game, a player can summon a fireball to their raised hand and throw it at the other player. Both players see the appropriate view of the action.

The dyad configuration of *Mano-a-Mano* allows for multiple simultaneous views by assuming that each user is largely unaware of graphics projected on the wall behind them, or on their own bodies. Instead, those graphics are intended for the other user. For example, an object held in front of the body by one user is rendered twice: first on the far wall to give that user the view of the object in their hand, and second on their body so that the other user also sees the object in the first user's hand (Figure 5a).

We envision *Mano-a-Mano* to be useful in a variety of dyadic interactions. We anticipate collaborative scenarios that rely on the shared understanding of the 3D layout and motion of virtual objects, such that if one user points at an object, the other user sees them pointing at that object. Of further interest is to what extent *Mano-a-Mano* gives users the impression of *object presence* [11]*:* that is, whether virtual objects appear to be the correct size and distance, even when a number of familiar depth cues are absent or in conflict (e.g., stereoscopy). In short, we would like to know if users "buy" the effect. We make the following contributions:

- A room-scale SAR system supporting dyadic interactions, including a number of early demonstrations.
- A study which tests users' ability to correctly perceive the distance and size of 3D virtual objects rendered in projected SAR.
- A study which tests users' ability to determine which of several objects is indicated by the other user's pointing gesture.

After reviewing related work, we detail the *Mano-a-Mano* system, describe a number of demonstration interactions, present the two user studies and lastly discuss a variety of issues regarding dyadic projected SAR.

RELATED WORK

Providing large-scale immersive experiences is a major thread of virtual reality research. A larger display enables a larger field of view to the user, which induces a more powerful sense of presence and immersion in the experience (which can also induce simulator sickness) [15, 30]. CAVE [8] has been a popular method of creating large virtual reality experiences by projecting content on three to six walls of a cube shaped empty room. In contrast to CAVE which completely surrounds and isolates the user, our system builds on the concept of Spatial Augmented Reality (SAR) [4] to envelop a pair of users in a shared projector-augmented experience, enabling them to interact with virtual objects in a normal room without the use of special glasses and equipment.

Our approach also differs from previous efforts to support computer-aided face-to-face interactions employing transparent screens [14, 15, 17] or "immaterial" fog screens [22] between users. For example, ClearBoard [17] uses a window metaphor for face-to-face collaboration between two remote users.

Spatial Augmented Reality

Raskar et al. [25, 26, 27] first demonstrated a working implementation of SAR by registering a virtual 3D model with the underlying 3D physical object in order to overlay virtual content. This was followed by Underkoffler et al.'s concept of an I/O Bulb [35], a projector-camera pair capable of both sensing and providing computational illumination (i.e., augmentation). This I/O Bulb vision has been further developed with room [18, 28, 37], tabletop [3, 36], mobile [12, 24] and steerable form factors [23, 38].

With the emergence of widely available Microsoft Kinect depth cameras, research into larger scale SAR interactions has accelerated [3, 19, 36, 37, 38]. Our work draws inspiration from our earlier project, called LightSpace [37], which first demonstrated a system which combines three projector and depth camera pairs to augment the surfaces in the environment. The augmentations and interactions in LightSpace were constrained to the available surfaces and no perspective views were available.

Availability of real-time depth capture enables a new generation of SAR systems which can compensate for moving projection surfaces as well as provide correct perspective views [3, 34, 38]. For example, MirageTable [3] offers a perspective stereoscopic view to a single user working on a projected SAR tabletop by compensating the projection for the user's hands and any other physical objects on the table. Our work extends the state of the art by both addressing the entire room and by providing multiple simultaneous perspective views to two users.

Immersive Displays with Multiple Perspective Views

Providing multiple people with personal perspective views in augmented or virtual environments has traditionally been achieved through head-worn displays [7], where each user has their own view on their own personal display. In practice, the obstruction of the face with such displays and their narrow field of view can lead to unnatural face-to-face interactions. Our solution avoids this difficulty. Alternatively, projection-bases systems rely on either *time-* or *space-multiplexing* the projected images to show multiple independent views. Agrawala et al. [2] demonstrates a time-multiplexed approach with synchronized shutter glasses to enable two stereoscopic user perspectives on an interactive tabletop. While conceptually simple, this approach tends to suffer from low image brightness and sometimes perceivable flicker since each eye only gets a small slice of the available light in each frame.

In contrast, Bimber et al.'s Virtual Showcase [5] uses multi-plane beam combiners to enable up to four independent perspectives on a spatially-multiplexed tabletop: each view appears at a different location of the display and is optically combined with four mirrors. Similarly, IllusionHole [21] uses spatial-multiplexing to provide multiple participants correct perspective views around the tabletop. Their display greatly reduces the visible display area for each user to ensure minimal image overlap. Our work also employs spatial multiplexing approach to support multiple views. In contrast, we exploit the fact that in the face-to-face arrangement, most of the surfaces that each person sees are on their partner's body or in their partner's background. These surfaces are good for projections of individual views, as they are not easily perceived by their partner.

Depth Perception and Object Presence

The ability of our system to provide two users with a sense of spatial presence in a room-scale augmented reality hinges on the human ability to perceive depth from a perspective view without binocular cues. Conventional measures of person- and object-presence have mostly been defined for virtual environment displays that surround and isolate the user [31]. Stevens and colleagues found that the users can experience a measurable sense of object-presence with projection-augmented models [33]. However, their questionnaire-based study examined only planar projections and not perspective views.

The relative importance of various depth cues in perception of virtual objects [9] is an important consideration for our system since we do not offer stereoscopic vision. Sollenberger and Milgram [32] showed a large improvement of head-coupled stereo over static non-head-coupled non-stereo displays while Arthur et al.'s experiments [1] showed that users greatly preferred head coupling without stereo to stereo head-coupled displays in fish-tank VR. How such results apply to perspective SAR configurations remains unclear. The most closely related work in this space is the depth perception study of MirageTable [3] which showed substantial accuracy in users' estimates of depth on a SAR tabletop with head-coupled stereo view. Also related is a pilot experiment reported by Broecker et al. [6] who investigated a variety of cues affecting depth perception in a view-dependent near-field SAR, but found no statistically significant results.

While focusing on projected tabletops, Hancock et al. [11] evaluated people's ability to judge object orientation under different projection conditions. Their work highlights the importance of correct perspective on judgment of objects' spatial presence especially in multi-user scenarios. There is also a long line of related research in cognitive psychology on understanding the relative importance of different cues for depth perception (e.g., [9, 10, 30]). The complete review of this work is beyond the scope of this paper, but the three volume book by Howard [16] offers the definitive summary of knowledge on the topic.

SYSTEM DESCRIPTION

We describe our prototype dyadic SAR system, including hardware configuration, scene modeling, dynamic projection mapping and support for multiple simultaneous views. Technical contributions of our work include: a particular configuration of pro-cam units to support dyadic interaction, a graphics pipeline that blends views from two simultaneous perspectives while supporting projection onto dynamic depth maps (e.g., people), and the design of interactions and experiences that showcase these capabilities.

Hardware Configuration

Our prototype dyadic SAR system employs three HD video projectors (BenQ W1080ST), each paired with a Kinect for Windows v2 sensor. Their mounting was chosen to both display and sense around two users that are approximately facing each other in a large room. Two of the projector and camera pairs are mounted on the ceiling, about two feet above the heads of each of the two users. These are oriented so that they approximately face one another, covering the opposite walls and part of the floor (Figure 1). Roughly speaking, each user's view is rendered by the projector above them. The precise surface geometry necessary for dynamic projection mapping for a user's view is provided by the Kinect paired with the projector above them. Meanwhile, body tracking of that user is supported by the opposite facing Kinect camera. While mounted significantly higher than in most applications, Kinect body tracking works well in this configuration. This symmetric arrangement of projectors and cameras follows the symmetric nature of dyadic interaction. The third projector and camera pair is mounted on the ceiling, facing downwards, to cover the area between the areas covered by the first two projectors.

Our current implementation is primarily hosted on a single PC which drives all three projectors. As the current Kinect for Windows v2 SDK can support only one camera per PC, we have three additional PCs which send the Kinect depth images and other image processing results (e.g., body tracking, optical flow) to the main PC via network. All depth

data is merged into a single scene in Unity 3D (http://unity3d.com). Unity provides a convenient overview of all sensed geometry and virtual objects in the room, and can be used to quickly script new interactive applications. Furthermore, a variety of Unity surface shaders such as lighting, shadows and procedural texturing methods are available to make virtual objects appear more realistic.

Calibration and Scene Model

Precise dynamic projection mapping requires the pose, focal length and optical center of each projector and Kinect camera. Our prototype system recovers this information in an offline automatic procedure whereby each projector in turn displays a series of Gray code patterns. These patterns are observed by the color camera of the paired Kinect sensor. Given the coordinate mapping functions of the Kinect SDK, this Gray code pattern is used to establish the precise mapping from a 3D point in the Kinect camera's coordinate frame to the corresponding point in the projector's image.

The relative pose of each projector camera pair is established by having all Kinect color cameras additionally observe the Gray code patterns of all other projectors, noting regions where the other projectors overlap with the camera's own paired projector, and computing the transform that brings corresponding 3D points of overlapping regions into alignment. This process results in a world coordinate system for all projectors and cameras. The calibration procedure is further described in [19].

Dynamic projection mapping also requires the precise surface geometry of the physical environment. The Kinect for Window v2 sensor includes a time-of-flight depth camera which generally produces more precise depth maps than the original Kinect sensor. In particular, the precision of depth data for the v2 camera is approximately constant over the range of depth (0.5m-4.5m), whereas the depth precision of the original Kinect sensor degrades with distance [20].

As with related projects [37, 3] moving objects such as user's bodies are handled separately from the static geometry of the room. This is primarily for two reasons: first, the system may project onto regions of the room that are otherwise occluded by moving objects. Second, static geometries support various offline analysis and optimization procedures that are difficult to perform on dynamic geometry. Furthermore, this model of the environment is pre-processed to provide detailed collision geometry used in rigid body physics simulations.

Dynamic Projection Mapping

Given the parameters of all projectors and depth cameras, as well as the geometry of the room, it is relatively straightforward to render graphics that change the surface appearance of physical objects in the room [27, 37]. This can be implemented as a single vertex shader which employs the mapping from 3D world coordinate point to the 2D point in a projector's image, as computed from calibration.

Rendering a virtual 3D object so that it appears correct from a given user's viewpoint is more complex and can be implemented as a multi-pass rendering process, whereby the virtual objects and real physical geometry are rendered offscreen from the desired viewpoint (see Figure 1 Virtual images). This offscreen rendering is then combined with the surface geometry and projector parameters in a standard projective texturing procedure, where only the physical geometry is rendered. We use a process similar to that described in [3]. The user's viewpoint is set to follow the head position determined by Kinect body tracking.

Supporting Multiple Views

The placement of the two opposite facing projectors was chosen so that each projector primarily displays the view corresponding to the user standing under it. A virtual object placed several feet above the ground and between the users, for example, is rendered twice, once for each user. Because the users are looking in opposite directions, one user is unlikely to see the graphics intended for the other user, because it will appear on the wall behind them, or on their own bodies. In this work we do not consider configurations with more than two users, or where users are not arranged face to face. In these more complex configurations, it is likely that one user will see the projection intended for another user. We speculate that such visual "cross-talk" may be disruptive to users, depending greatly on the task and configuration.

Figure 2. Examining virtual objects: a) airplane model in hand, b) race car on the table, c) hand spinning the globe. Note that while the objects are projected over diverse physical geometry, they appear correct to the user.

When the virtual object is placed nearer the ground between the users, the view of the object intended for one user may be seen by the other. Our rendering procedure models the view of each user as a standard perspective graphics camera. Where the physical surfaces addressed by each view overlap, the renderings are blended so that both are ultimately visible. For example, two views of the same object may appear on the floor. In the case when the object is placed on the ground, the two renderings of the object will overlap and meet at the ground. The impact of this "double rendering" may depend greatly on the application and is a matter of future research.

In some cases it may be more appropriate to render the object only once from a neutral viewpoint above and between the users as suggested by Hancock et al. [11].

Interactions

In addition to seeing virtual 3D objects rendered correctly, users may also interact with them in various ways. We have explored a number of simple interactions supported by Kinect body tracking and lower-level features derived from depth and infrared images. For example, "touching" a virtual object can be implemented by intersecting either the tracked hand position or points taken from the user's depth map, with the geometry of the virtual object.

An important and compelling interaction is the ability for the user to hold a virtual object in their hand (Figure 2a). The virtual object may be scripted to follow a point just above the hand as it is tracked in 3D by Kinect. The multi-view rendering described above renders the object once for each user's view, as described above. As the user holds their hand up in front of their body, their view will generate a large projection at the far surface of the room, possibly spanning the other user's body. Meanwhile, that second user's view will generate a small projection of the object, possibly over the first user's torso (see example in Figure 1).

Held objects can be dropped or thrown by meeting some conditions for release. For example, a held ball may be thrown if the velocity or acceleration of the hand exceeds some threshold. At the moment of release, the ball might take the velocity of throwing hand. Catching or "picking up" might be implemented by simply testing when an object is sufficiently close to the tracked hand.

Some applications may benefit from detecting the collision of virtual objects with the room or the user, leading to a realistic collision response. Our prototype uses pre-computed collision geometries for the static parts of the room, and approximates the shape of moving objects such as the user with a number of sphere colliders [3] (Figure 3).

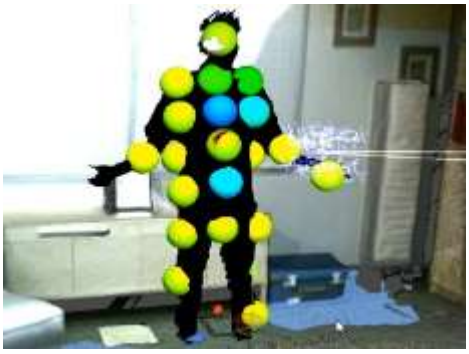

Figure 3. A Unity view showing sphere collider proxies used to determine run-time collisions between game objects (e.g., the incoming fireball) and dynamic surface geometry (e.g., users).

While with our rendering pipeline it is straightforward to apply a texture to a static part of the room, applying a texture to a moving object such as the user requires real time tracking of the projection surface. We have experimented with using low level motion features such as optical flow with encouraging results. Real time optical flow is computed from the Kinect infrared video. A texture can be applied to moving objects by determining an initial placement of the texture and then following the surface over many frames using the motion estimated by optical flow (Figure 4c).

EXPERIENCES

In this section we describe a number of initial interactive *Mano-a-Mano* experiences. The accompanying video includes shots taken from the user's point of view.

Rendering Virtual Objects

As an initial demonstration of dynamic projection mapping, we placed a number of static models above the coffee table in the middle of the room. For example, a model of an airplane, a racecar or a globe is shown in Figure 2. While both users can get a good sense of the airplane, the globe is more challenging because each user views different sides. We incorporated a rudimentary ability to spin the globe by "touching" or intersecting with the globe.

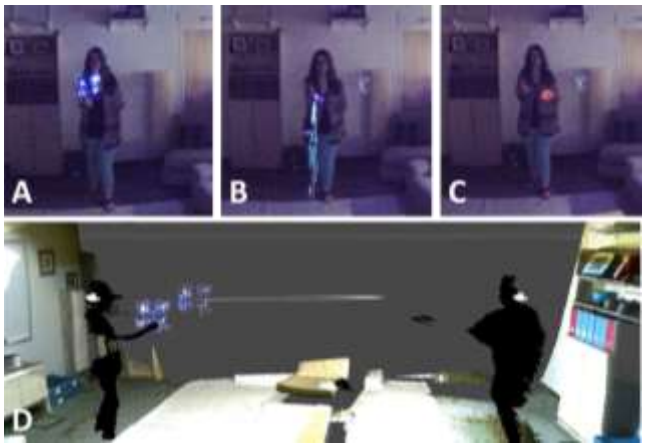

Figure 4. Fireball game: a) the user sees their partner holding a fireball; b) partner throws the fireball, the trail is visible; c) when the partner is hit, blood damage sticks to their body and its position gets updated with optical flow; d) Unity view showing flying and held fireballs.

Fireball

To test the ability to hold a virtual object and throw it accurately, we created a fun, fast-paced combat-style game where players summon a fireball by raising their hand. The fireball appears to hover over the hand a few inches. The player can quickly throw the fireball at the opposing player. If it hits the wall, damage is temporarily rendered at the point of collision. If it strikes the user, a blood texture is applied to the user. This visual effect tracks the projection surface on the player using optical flow as described above, and the attacking player scores a point (Figure 4). Release of the fireball is triggered by exceeding a threshold on the velocity of the hand holding the fireball. The direction is determined by computing the ray from the tracked head position through the hand. With a little practice, players can accurately direct their shot.

Catch

Two users can play catch with a virtual tennis ball (Figure 5). This experience extends the previous example by adding a rudimentary means to catch the ball: if the ball is sufficiently close to the hand, it is considered "caught" and is placed in the hand. The player may then throw it back to the other player.

If the player is unable to catch the ball, it may collide with the user or other objects in the room. In this case the ball will bounce and roll in a physically plausible way. The ability of the player to catch the ball hinges directly on their ability to perceive the virtual object in space.

Figure 5. Playing catch with a virtual ball in *Mano-a-Mano*: a) ball in player's hand, b) ball in partner's hand. Note: from the partner's viewpoint the ball is rendered on the table in front.

USER EXPERIMENTS

To evaluate the effectiveness of our dyadic projected SAR system, we conducted two user experiments that focus on the most relevant aspects of a dyadic SAR system: first, can users correctly perceive virtual objects, and second, can users understand their collaborators' interactions with virtual objects?

Our studies offer initial support to the claim that dyadic SAR can support effective interaction with virtual objects between two face-to-face participants.

The first experiment focuses on examining the effectiveness of a single-user monoscopic perspective view to convey the sense of the virtual object's spatial presence. In particular, we examine whether the participants can perceive projected virtual objects as *spatial* rather than appearing only at the projection surface (*projected)*, and what factors affect their perception. The second study quantifies how well two collaborators understand each other's pointing references when discussing virtual objects between them in a face-to-face scenario.

We recruited 11 participants from our organization (5 female, mean age 36 years, std. dev. = 10 years). The participants performed both experiments in order and then completed a questionnaire to gather subjective feedback. The participants were compensated with a $10 coupon and the total session took approximately 30 minutes to complete.

Experiment 1: Object Presence

While previous SAR research [3] demonstrates that users can perceive perspective-projected virtual objects as *spatial*, rather than *projected*, those experiments were done with stereoscopic projections and virtual objects were always close to the projection surface. In our room-scale environment, it is unclear whether the users will perceive the spatial characteristics of virtual objects when they are projected without stereo and far away from the projection surface. This ability to perceive virtual object's presence in mid-air is crucial for any view-dependent larger-scale SAR scenario.

Experiment Design

As a measure of the virtual object presence, we asked the participants to rate the distance to and size of projected virtual objects (similar to [3]). The projected test object was a green cube of three different sizes (small=10cm, medium=15cm and large=20cm edge) and virtually positioned at three different distances from the participant (near=1.5m, middle=2.5m, and far=3.5m). From the participant's point of view, the image of the virtual object was always projected on the back wall of the room (approximately 3.75m away from the user) (Figure 6).

It is important to note that the object's sizes were highly confusable across our tested distances (this was intentional). In fact, when projected at the nearest location the smallest object subtended roughly the same visual angle as the largest object at the farthest distance.

In addition to varying size and distance, we introduced two different conditions for performing this task: with and without physical markers. The *with markers* condition included three black poles placed exactly at the location where the virtual object could appear, while the *no markers* condition had those poles removed (Figure 6). If observed correctly, the virtual object would appear to sit on the physical marker. Our goal was not to tell the participant where the object is with the physical marker, but rather to aid them by giving them a real-world physical anchor that they can compare to the virtual object. The participants were still required to determine which of the three markers the virtual object is attached to. In addition to our main hypothesis that the participants are able to correctly understand the spatial placement of projected virtual objects (*H1*), we hypothesized that having a physical object marking the possible location of the virtual object makes the rating task simpler (*H2*). This was motivated by our observations that virtual objects projected in the collaborator's hand always seemed a bit more spatial and real, than the objects purely placed in mid-air. The physical markers in our experiment served as a controlled proxy for the user's hand.

Lastly, we hypothesized that participants are more accurate in rating objects closer to the projection surface (*H3*), i.e., further away from the participant, since the real world and the virtual object's location are in closer agreement in such cases. This is contrary to the real world behavior where humans are better in rating closer objects as objects further away form a smaller visual angle and are therefore harder to see [10].

In summary, our experiment design was: Size (3) x Distance (3) x Condition (2). The participants performed 4 ratings for each of these combinations resulting in a total of 72 ratings per participant. We recorded the participant's rating of both size and distance of the object as well as their response time.

Figure 6. Experiment 1: a) participant performing the "with markers" condition, where black poles serve as physical markers for the location of the virtual cubes; b) Unity view showing the virtual object at the far distance; c) three real-word cubes sit on the table in front of the participant for comparison.

Procedure

Each participant was first given a brief introduction to our system, and then asked to stand facing the side of the room where their perspective projected view would be shown. In front of them was a short coffee table on top of which we place three physical models of the three cubes. Those physical cubes precisely matched the scale of projected virtual cubes. The three possible locations where the virtual object would appear were marked on the floor with a number (1-near, 2-mid, 3-far).

We asked the participants to verbally rate each object's size (indicating "small", "medium", or "large") and distance (indicating "1", "2", or "3", corresponding to the marked floor location). The coordinator recorded their ratings and advanced to the next trial. To ensure the same amount of stimulus across all participants, the object was projected for exactly 5 seconds, after which it disappeared. The participants were instructed to give their ratings as soon as they felt confident, and their response time was recorded as the time from the object's appearance to the time of coordinator's entry. For evaluation simplicity, the trials were grouped by condition since setup of physical markers required time and would render collection of large dataset difficult. The presentation of size and distance trials was randomized within each condition block, and the order of conditions was counterbalanced across users to reduce the effects of ordering. Before each condition, the participants were given a practice session where they were shown each object at each size and distance combination (without the 5 second limit) and the study coordinators gave them feedback on their ratings.

The participants were initially positioned such that the objects appeared at the same physical location on the back wall. They were explicitly told that they could move around if that helped them make a decision, but were asked to remain roughly within a step away from the initial position marked on the floor. While the projection was monoscopic, participants could use other depth cues including shading, perspective, shadows, size, and motion parallax.

Results

Overall, participants were more accurate in rating Distance (88.8%) than Size of virtual objects (70.7%). When considering both Distance and Size, they provided a correct rating in 66.5% of trials, which is significantly better than chance (1 out of 9 combinations). These findings support our *H1* hypothesis that users can and do perceive virtual objects as having spatial presence in a perspective SAR scenario.

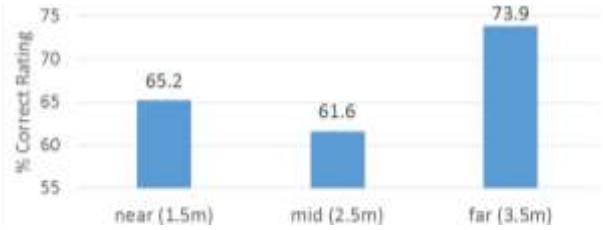

Figure 7. Effect of Distance on the participant's overall ratings of objects depth placement and size.

Out of 792 total trials, only 9 ratings were more than one option away from the ground truth in either size or distance (e.g., mistaking a "small" size for "large") and those could be considered outliers. Given only three possible options in each category, this is not surprising, but this effectively means that the participants were either right on target ("correct") or one target off ("incorrect").

We therefore coded the participant's responses into a binary variable ("correct" or "incorrect") for each of the size, distance, as well as size and distance combined rating. Given such binary responses, standard linear regression models or analysis of variance (ANOVA) are not appropriate since they assume scalar responses drawn from a normal distribution. The appropriate method of statistical analysis for correctness of user's ratings is *repeated measures logistic regression* [40]. We used IBM SPSS v.21 to calculate our statistics. In our case, logistic regression computed the correlation between the varied factors (e.g., different sizes, distances, or conditions) and a binary outcome ("correct" or "incorrect"). The significance metric for logistic regression is Wald Chi-Square (χ^2).

We ran our analysis on the model containing the following factors: Condition, Size, and Distance. When analyzing the overall correctness (i.e., correct for both size and distance), we found significant effects for Distance ($\chi^2 = 11.746$, df = 2, p = .003; a p-value <=0.05 shows a statistically significant effect), but not for Size or Condition. Distance had a strong effect on the ratings with *far* distance (3.5m away) being significantly more accurate than *near* and *mid* (Figure 7). This result runs contrary to the real-world behavior where closer objects tend to be easier to rate, but it confirms our *H3* hypothesis, that being closer to the projection surface makes it easier for the user to correctly perceive the spatial characteristics of the projected 3D virtual object. In our

experiment, the *far* distance was only 25cm away from the back wall where the object was being projected.

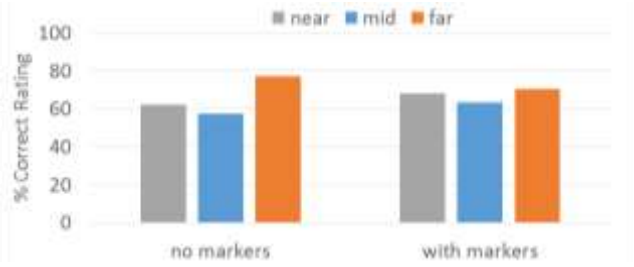

Figure 8. Participants' performance across different Conditions at different Distances.

While participants were more correct in the condition with physical markers (67.4%) versus no markers (65.6%), Condition was not found to have a statistically significant effect. This result failed to confirm our hypothesis *H2*. We further analyzed our results to understand why. Presence of markers improved participants' ratings on average 6% when objects were far away from the projection surface (at *near* and *mid* distances), but actually hurt their performance in the *far* distance (Figure 8). This might be explained by the fact that at the *far* location, participants already had a large physical reference, the projection wall itself, to help them judge distance and were potentially just distracted by the presence of multiple poles in front of the projected object. We speculate that when the virtual object was further away from the projection surface, the presence of physical markers may have been beneficial.

Furthermore, interaction of Size*Distance had a highly significant effect ($\chi^2 = 47.482$, df = 4, $p < .001$). This was not surprising, since some of the combinations were easily identifiable (e.g., the small cube at the far distance was the smallest projected object) while some were highly confusable (e.g., the medium cube at the near position is easily confused with the large cube at mid position).

We also identified an effect of gender on our results ($\chi^2 = 6.524$, df = 1, p = .011) with males outperforming female participants by 12%. While this agrees with results of other experiments that evaluate spatial ability (e.g., Mental Rotation Test [29]), note that our sample size is rather small (5 female and 6 males) and therefore it is difficult to draw gender conclusions from our experiment.

Response time includes the time for the coordinator to log the result and advance the trial (approximately 1 second). The average response time was 6.47s (std. dev. = 2.27s). We performed repeated measures ANOVA on the response time and found that it closely correlates with our rating analysis: *participants were significantly faster in responding to conditions where they were also found to be more accurate*. This indicates that our results do not fall under a speed-accuracy tradeoff common to many targeting experiments. Since these results are closely correlated and exhibit similar statistically significant effects, we omit detailed analysis of the response time.

Subjective Feedback

While our quantitative data does not show statistically significant benefits to the presence of physical markers in the first experiment, 9 out of 11 participants stated that having physical markers helped them complete the task.

Our questionnaire captured the participants' feedback on the statement "I could tell where the objects were in space" as the average rating of 3.81 (with markers) vs. 3.0 (without markers) (t-test $p<0.005$) on a 5 point Likert scale (5 indicating agreement). The responses for "I am confident that my guesses are correct", yielded similar responses on the same scale: average rating of 3.09 (with markers) vs. 2.54 (without markers) (t-test $p<0.007$). Participants were asked to comment on any strategies they used to complete the trials. While most of the participants reported using movement and size to rate the objects, 5 out of 11 participants self-reported relying heavily on the position of virtual shadows as well.

Experiment 2: Understanding Collaborator's Spatial References in a Dyad

The main purpose of our SAR system is to allow two people to interact in an unobstructed face-to-face manner (e.g., see our *Fireball* or *Catch* experiences above). The effectiveness of that experience hinges on the user's ability to understand what their partners are doing, so that they may respond appropriately. We designed the second experiment to explore how well the user understands the actions and intentions of their collaborator when discussing virtual objects in a face-to-face scenario in Dyadic SAR.

Experiment Design

We designed a task in which the participant observed their *partner* (one of the study coordinators) raise a short pole to their eye level and point at one of 16 spheres that appeared between the participant and the coordinator. Participants verbally indicated which sphere they believed their partner was pointing at. The spheres were arranged in a 4x4 grid and were clearly numbered (Figure 9). In this configuration, the spheres in the grid were projected partially on the human bodies in space and partially on the walls behind them.

Figure 9. Experiment 2: a) side view of the participant and their partner pointing at a target; b) participant's perspective (recorded with a helmet camera).

Each sphere in the grid was 10cm in diameter and the spheres were 11 cm apart (center to center). The grid floated 1.5 meters above the floor. This particular arrangement of targets was found in our pilot experiment to be dense enough to be

potentially confusable while offering enough spatial resolution to mimic the requirements in many real world face-to-face tasks (e.g., two people discussing an architectural model between them). While targeting with a rifle-like aiming gesture is not a natural pointing style, we chose this pose to avoid the ambiguity of individual person's pointing style. We are not interested in understanding how well can the participants understand pointing gestures in general, but rather how well can they understand a particular spatial reference when performed with respect to a virtual object in front of them. In our pilot experiments, this targeting style was selected as the least ambiguous.

Rather than comparing targeting performance among multiple conditions, this experiment was designed to quantify the overall accuracy of participant's understanding of their partner's references. We therefore measured the error (in meters) in their estimate as the distance from the sphere they indicated to the actual targeted sphere.

Procedure

The same group of participants from the first experiment participated in this experiment. Each participant stood at the same location as in the first experiment. Their collaborator stood on the other side of the room (~2.5m away). At the start of the trial, the collaborator was silently prompted by the system to point at a specific numbered ball in the grid. The participant then verbally indicated which sphere they believe their collaborator was pointing at. The trial was not time limited, but participants were instructed to respond as soon as they felt confident in their rating. One of the study coordinators entered their response to conclude the trial. The collaborator returned to a neutral pose (non-targeting) between each trial. Before running the experiment, each participant was given a set of 15 practice trials during which they were given feedback on their performance.

We randomized the order of presentation of target spheres and each participant gave two ratings for each sphere condition for a total of 32 ratings.

Results

Participants identified the correct target in 54.5% of 352 total trials. While this might appear low, it is significantly higher than chance (1 out of 16) and is more impressive when one considers the distance between the selected target and the true target. Averaged over all trials this spatial error was 0.056m (std. dev. = 0.063m). This low value indicates that when participants selected the incorrect target, they overwhelmingly indicated one of the nearest neighbor targets (the targets themselves were 0.11m away). Errors were not uniformly distributed. For example, the lower right target exhibited more than twice the error compared to targets in the upper right corner (Figure 10); however, our data does not offer any explanation for this effect.

The angular difference in targeting two adjacent target spheres was approximately 4.5°. That participants could tell the target reference to within a 12cm radius is impressive and provides evidence that our system is capable of presenting virtual information between the two unencumbered users in a way that enables mutual spatial understanding.

4 0.0638	3 0.0492	2 0.0409	1 0.065
8 0.0483	7 0.0516	6 0.0425	5 0.0454
12 0.052	11 0.0736	10 0.0635	9 0.065
16 0.0824	15 0.0582	14 0.0482	13 0.0524

Figure 10. The 4x4 target grid (left) is shown from the participant's perspective (Unity3D view). The table (right) with average amount of error (in meters) for the participant's rating of each of the spheres in a grid. The numbers are color-coded to indicate performance (green=best, red=worst).

DISCUSSION

The figures in this paper and the accompanying video demonstrate that the dynamic projective texturing process is accurate enough to generate good monocular views of virtual objects for each user.

Less clear is to what extent, and under what circumstances, users of our system have a good sense of the virtual object's presence at the correct distance and size. While the first user experiment demonstrates that users are able to perform much better than chance on determining object size and distance in a controlled setting, subjects clearly performed worse at this task than if real objects were presented. Stereo projection would likely enhance the correct perception of size and distance (as in [3]). However, today's stereo projection technologies generally require wearing glasses, obstructing the face and impinging on the fluidity of face to fact interaction.

In practice, the sense of object presence may depend on a number of factors which are not considered in the user experiments [9]. For example, an object of familiar size and shape such as a tennis ball may be easier to correctly perceive than an abstract, featureless cube. Our own experience with the *Fireball* and *Catch* application examples suggest that interacting directly with a virtual object lends a stronger sense of the position and size of the object. For example, when holding a virtual object, it seems as if the higher level, cognitive realization that the object is being held in the hand helps in seeing the virtual object in the hand. This effect may be similar to that of "visual capture" studied in psychology, whereby, for example, a ventriloquist's voice appears to be coming from a dummy's mouth. Note that this cue may be stronger than the markers provided in the first experiment: in that task, participants determined which of three physical references indicated the distance of the object. When holding an object, the hand is the one possible physical reference.

While holding and moving a virtual object gives a stronger sense of object presence, it is also our experience that

watching the other user hold the object similarly helps. In fact, the effect is often stronger, since the projection of the held object may lie on the other user's body, placing its projection closer to the object's simulated position. As shown in the first experiment participants found the task easier when the object was nearer the projection surface.

Participant's success in the second experiment indicates a rudimentary shared understanding of the layout of virtual objects between them, and thereby a basis for more advanced collaborative interactions. As with the first experiment, our experience pointing and gesturing towards virtual objects may be rather different than that of the experiment, as the context of the task and arrangement of objects can have a great impact, as well as the various ways in which people gesture in more natural tasks.

CONCLUSION

Mano-a-Mano is a unique spatial augmented reality system that combines dynamic projection mapping, multiple perspective views and device-less interaction to support dyadic interaction with virtual 3D objects. We show a few initial interactive experiences with the system that demonstrate some fundamental interactions, but clearly there is the potential to investigate applications that go beyond gaming. Of particular interest are those that leverage the system's unique capabilities to support collaboration and co-reference of 3D virtual objects, especially where uninterrupted face to face interaction is valuable.

ACKNOWLEDGMENTS

We would like to thank Brett Jones, Rajinder Sodhi, Michael Murdock, Eyal Ofek, Ravish Mehra, Blair MacIntyre and Nikunj Raghuvanshi for their foundational work on the RoomAlive system which enabled this project.

REFERENCES

1. Arthur, K., Booth, K.S., and Ware, C. 1993. Evaluating 3D Task Performance for Fish Tank Virtual Worlds. In *ACM Trans. on Information Systems*, 11(3). 239–265.
2. Agrawala, M., Beers, A.C., McDowall, I., Fröhlich, B., Bolas, M., and Hanrahan. P. 1997. The two-user Responsive Workbench: support for collaboration through individual views of a shared space. In *Proc. of ACM SIGGRAPH '97*. 327–332.
3. Benko, H., Jota, R. and Wilson, A. D. 2012. MirageTable: Freehand Interaction on a Projected Augmented Reality Tabletop. In *Proc. of ACM CHI '12*. 199–208.
4. Bimber, O. and Raskar, R. 2005. Spatial Augmented Reality: Merging Real and Virtual Worlds. A. K. Peters, Ltd., Natick, MA, USA.
5. Bimber, O., Fröhlich, B., Schmalstieg, D., and Encarnacao, L. M. 2002. The Virtual Showcase. *IEEE Comput. Graph. Appl.*, 21(6) (November 2001). 48-55.
6. Broecker, M., Smith, R.T., and Thomas, B. 2014. Depth Perception in View-Dependent Near-Field Spatial AR. *Australasian User Interface Conf. (AUIC '14).* (poster).
7. Cakmakci, O., Member, S., Rolland, J., and Member, A. 2006. Head-Worn Displays: A Review. *IEEE Journal of Display Technology.*
8. Cruz-Neira, C., Sandin, D.J., and DeFanti, T.A. 1993. Surround-screen projection-based virtual reality: The design and implementation of the CAVE. In *Proc. of ACM SIGGRAPH '93.* 135–142.
9. Cutting, J.E., and Vishton, P.M. 1995. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth. In W. Epstein and S. Rogers (eds). *Handbook of perception of space and motion,* 69–117.
10. Gilinsky, A.S. 1951. Perceived Size and Distance in Visual Space. *Psychological Review*, 58, 460–480.
11. Hancock, M., Nacenta, M., Gutwin, C., and Carpendale, S. 2009. The effects of changing projection geometry on the interpretation of 3D orientation on tabletops. In *Proc. of ACM ITS '09*. 175–182.
12. Harrison, C., Benko, H., and Wilson, A.D. 2011. OmniTouch: Wearable Multitouch Interaction Everywhere. In *Proc. of ACM UIST '11*. 441–450.
13. Heo, H., Park, H.K., Kim, S., Chung, J., Lee, G. and Lee, W. 2014. Transwall: a transparent double-sided touch display facilitating co-located face-to-face interactions. In *CHI '14 Extended Abstracts*. 435–438.
14. Hirakawa, M. and Koike, S. 2004. A Collaborative Augmented Reality System Using Transparent Display. In *Proc. of the IEEE Int. Symposium on Multimedia Software Engineering* (ISMSE '04). 410–416.
15. Hou, J., Nam, Y., Peng, W., and Lee, K.M. 2012. Effects of screen size, viewing angle, and players' immersion tendencies on game experience. *Comp. in Human Behavior*, 28(2). 617–623.
16. Howard, I. 2012. Perceiving in Depth. New York: Oxford University Press. ISBN 978-0-199-76414-3.
17. Ishii, H. and Kobayashi, M. 1992. ClearBoard: a seamless medium for shared drawing and conversation with eye contact. In *Proc. of ACM CHI '92*. 525–532.
18. Jones, B., Benko, H., Ofek, E., and Wilson, A. D. 2013. IllumiRoom: Peripheral Projected Illusions for Interactive Experiences. In *Proc. of ACM CHI '13.*
19. Jones, B., Sodhi, R., Murdock, M., Mehra, R., Benko, H., Wilson, A., Ofek, E., MacIntyre, B., Raghuvanshi, N., Shapira, L. 2014. Roomalive: Magical Experiences Enabled by Scalable Adaptive Projector Camera Units. In Proc. of *ACM UIST '14.*
20. Khoshelham, K. 2012. Accuracy Analysis of Kinect Depth Data. In *Proc. Of Int. Achieves of Photogrammetry, Remote Sensing and SI Sciences*. Aug.
21. Kitamura, Y., Konishi, T., Masaki, T. and Kishino, F. 2001. IllusionHole: A Stereoscopic Display for Multiple Observers. In *Proc. SPIE vol. 4297, Stereoscopic Displays and Virtual Reality Systems VIII*.

22. Olwal, A., DiVerdi, S., Candussi, N., Rakkolainen, I., and Hollerer, T. 2006. An Immaterial, Dual-sided Display System with 3D Interaction. In Proc. of IEEE Conference on Virtual Reality (VR '06). 279–280.
23. Pinhanez, C. S. 2001. The Everywhere Displays Projector: A Device to Create Ubiquitous Graphical Interfaces. In *Proc. of ACM UBICOMP '01*. 315–331.
24. Raskar, R., van Baar, J., Beardsley, P., Willwacher, T., Rao, S., and Forlines, C. 2003. iLamps: Geometrically aware and self-configuring projectors. *ACM TOG 223*, 809–818.
25. Raskar, R., Welch, G., Cutts, M., Lake, A., Stesin, L. and Fuchs, H. 1998. The office of the future: a unified approach to image-based modeling and spatially immersive displays. In *Proc. of ACM SIGGRAPH '98.* 179-188.
26. Raskar, R., Welch, G. and Fuchs, H. 1998. Spatially Augmented Reality. In *Proc. of IEEE Workshop on Augmented Reality (IWAR'98).*
27. Raskar, R., Welch, G., Low, K.-L., and Bandyopadhyay, D. 2001. Shader Lamps: Animating Real Objects with Image-Based Illumination. In *Proc. of the Eurographics Workshop on Rendering Techniques*. 89-102.
28. Rekimoto, J. and Saitoh, M. 1999. Augmented Surfaces: A Spatially Continuous Work Space for Hybrid Computing Environments. In *Proc. of ACM SIGCHI '99.* 378–385.
29. Richardson, J.T.E. 1994. Gender differences in mental rotation. *Perceptual and Motor Skills*, 78. 435-448.
30. Sekular, R.B. Perception. 2nd Edition. (1990).
31. Sheridan, T.B. 1992. Musings on telepresence and virtual presence. *Presence: Teleoperators and Virtual Environments*, 1(1). 120-125.
32. Sollenberger, R.L. and Milgram, P. 1991. A comparative study of rotational and stereoscopic computer graphics depth cues. In *Proc. of Human Factors Society 35th Annual Meeting.* 1452–1456.
33. Stevens, B. Jerrams-Smith, J., Heathcote, D. and Callear, D. 2002. Putting the Virtual into Reality: Assessing Object-Presence with Projection-Augmented Models. *Presence*, 11(1). 79–92.
34. Tang, Y., Lam, B., Stavness, I., and Fels, S. 2011. Kinect-based augmented reality projection with perspective correction. In *ACM SIGGRAPH '11*. (poster).
35. Underkoffler, J., Ullmer, B., and Ishii, H. 1999. Emancipated pixels: Real-world graphics in the luminous room. In *Proc. of ACM SIGGRAPH '99.* 385–392.
36. Wilson, A. 2007. Depth-Sensing Video Cameras for 3D Tangible Tabletop Interaction. In *Proc. of IEEE International Workshop on Horizontal Interactive Human-Computer Systems (TABLETOP '07).* 201–204.
37. Wilson A. and Benko H. 2010. Combining multiple depth cameras and projectors for interactions on, above and between surfaces. In *Proc. of ACM UIST'10.* 273-282.
38. Wilson, A. D., Benko, H., Izadi, S., and Hilliges, O. 2012. Steerable Augmented Reality with the Beamatron. In *Proc. of ACM UIST '12.* 413-422.
39. Wilson, A. D., Izadi, S., Hilliges, O., Garcia-Mendoza, A., and Kirk, D. 2008. Bringing physics to the surface. In *Proc. of ACM UIST '08.* 67-76.
40. Winer, B.J., Brown, D. and Michels, K. 1991. Statistical Principles in Experimental Design, 3rd Ed. New York, NY. McGraw-Hill.