Interacting with Distant Objects in Augmented Reality

Matt Whitlock[*]
University of Colorado Boulder

Ethan Hanner[†]
University of Colorado Boulder

Jed R. Brubaker[‡]
University of Colorado Boulder

Shaun Kane[§]
University of Colorado Boulder

Danielle Albers Szafir[¶]
University of Colorado Boulder

Abstract

Augmented reality (AR) applications can leverage the full space of an environment to create immersive experiences. However, most empirical studies of interaction in AR focus on interactions with objects close to the user, generally within arms reach. As objects move farther away, the efficacy and usability of different interaction modalities may change. This work explores AR interactions at a distance, measuring how applications may support fluid, efficient, and intuitive interactive experiences in room-scale augmented reality. We conducted an empirical study ($N = 20$) to measure trade-offs between three interaction modalities–multimodal voice, embodied freehand gesture, and handheld devices–for selecting, rotating, and translating objects at distances ranging from 8 to 16 feet (2.4m-4.9m). Though participants performed comparably with embodied freehand gestures and handheld remotes, they perceived embodied gestures as significantly more efficient and usable than device-mediated interactions. Our findings offer considerations for designing efficient and intuitive interactions in room-scale AR applications.

Index Terms: Human-centered computing—Interaction design—Interaction design process and methods—User interface design

1 Introduction

Augmented reality (AR) technologies are being used for a growing range of applications. AR head-mounted displays (AR-HMDs), such as the Microsoft HoloLens, support immersive applications that embed interactive virtual content throughout the visible environment. For example, architects can use AR to monitor progress at construction field sites [14] and prototype large structures [28], facility managers can access building information *in situ* [19], and consumers can analyze the state of connected devices in Internet-of-Things (IoT) networks [21]. For these applications to provide seamless interactive experiences, designers must craft intuitive and efficient interactions with virtual content at extended spatial scales, including virtual objects well beyond a user's reach [33].

While prior research has explored interactions with close virtual content (see Piumsomboon et al. for a survey [30]), research in *proxemic interaction* suggests that optimal interactions with technologies may shift as a function of the distance between the user and the technology [3, 9]. The spatial scale of an application can change users' cognitive and interactive experiences and therefore requires careful consideration during design [4]. In this work, we explore the effect of different input modalities on distal interactions in room-scale AR focusing on three modalities: multimodal voice commands, embodied freehand gestures, and handheld remotes.

AR applications allow designers significant freedom in terms of how interactions may occur. Prior research in near-range interaction suggests that embodied approaches, such as freehand gestures, can lead to more natural manipulations and increase feelings of immersion and intuitiveness [8, 37]. However, pointing and gesture-based selection become more error prone as objects move further away [23]. Integrating voice-based interactions alongside pointing or gestures can help resolve these ambiguities [20], but the efficacy of voice interactions alone may be limited for spatial tasks like object rotation and translation [18].

Distal interaction studies in virtual reality, pervasive computing, and large displays offer additional insight into potential trade-offs between these modalities at extended spatial scales. However, immersive AR-HMDs offer new challenges not considered by prior work. For example, differences in depth perception for physical and virtual objects may effect pointing accuracy. Visible environments may also introduce social constraints that cause discomfort or contexts that complicate displacement methods. To address these challenges, we aim to directly measure the usability and efficiency of distal interaction modalities for use with AR-HMDs.

We conducted a controlled laboratory study focusing on three modalities contextualized in an IoT scenario. For each modality, we had participants perform three direct manipulation tasks: selection, rotation, and translation. Overall, participants were faster and more accurate with freehand gestures and handheld remotes than voice-based interaction. While we found no significant performance differences between gestures and remotes, people felt significantly faster with and expressed a strong preference for embodied interactions. Our findings offer initial insight into how designers might leverage different input technologies for AR applications integrated into large environments.

2 Related Work

While distal interaction may offer unique challenges for AR, fields such as virtual reality and pervasive computing also often require users to interact with content beyond reach. We surveyed literature from close-range AR interactions and distal interactions in related fields to inform our experimental design.

2.1 Interaction in Augmented Reality

An extensive body of research in AR explores modalities for immersive and intuitive interactions with virtual content at close range (see Zhou, Dou, & Billinghurst [44] and Sanz & Andujar [35] for surveys). Most studies find that people are more efficient with and express preference for direct manipulation through freehand gestures [8, 15, 32]. For example, Seo et al. [37] introduces a technique for direct hand manipulation in near-space interactions that increases perceptions of immersion and naturalness compared to device-based approaches. Distal interactions with AR-HMDs add new considerations for these findings. For example, pointing to occluded objects may require nonlinear visual and spatial mapping in noisy environments [10]. Modeling tasks may manipulate geometries that extend beyond a user's reach [17, 28]. Precision with distal pointing in immersive environments, such as those offered by AR-HMDs, degrades quadratically with target size and, as a result, distance to the

[*] e-mail: matthew.whitlock@colorado.edu
[†] e-mail: ethan.hanner@colorado.edu
[‡] e-mail: jed.brubaker@colorado.edu
[§] e-mail: shaun.kane@colorado.edu
[¶] e-mail: danielle.szafir@colorado.edu

target [23]. These challenges complicate the transfer of near-range findings to distal interactions in AR.

Multimodal inputs may help address these challenges. Freehand gestures provide better spatial input, while descriptive speech commands offer better system control [18, 22]. As an example, the Gesture-Speech technique lets users gesture to identify an object and speak to perform an action [31]. Kaiser et al. [20] integrate multiple modalities, including voice and freehand gesture, to offer more accurate target specification than either modality offers individually. They also note that ambiguities introduced by the use of freehand gestures arise predominantly with increased distance causing direct hand manipulation metaphors to break down.

Ambiguities in distal interactions become particularly problematic when AR systems require precise manipulation of visual elements. Badam et. al [2] measure how well different interaction modalities support different visualization tasks, and map modalities to tasks based on performance. For example, touch and freehand gestures excel at selecting individual items, whereas voice commands are better suited for creating new visualizations. Wither & Hollerer [42] investigated the relationship between modality and efficacy in 3D cursor placement. Unlike in near-range AR studies, they found participants preferred handheld remotes and performed comparably with remotes and embodied headtracked cursors. Our research builds on these studies to understand preference and performance in manipulation tasks for distant objects.

2.2 Distal Interaction in Other Contexts

Other display modalities can also require interactions with distal objects. For example, pervasive computing often involves placing interactive technologies throughout an environment. Designers can consider a user's proximity to each device to optimize interactions with these technologies [3, 24]. In this context, distal pointer and device-mediated interactions allow users to quickly switch between multiple devices [39]. Wearable devices also allow centralized control over distributed physical devices [13]. Similarly, AR can visually extend the users' body to mediate interactions with physical devices [11, 33]. However, these techniques for distal interaction in pervasive computing environments focus on predefined interactions specific to physical objects.

Large-scale 2D displays suffer similar ambiguity errors to distal interaction with virtual content in AR. Many large display techniques leverage multimodal voice and freehand gesture to resolve these ambiguities [6, 7]. Alternative approaches augment specific objects with visual or auditory cues to support more accurate selection [40], or use techniques such as snapping to or enlarging target regions [25, 27]. However, these techniques primarily facilitate selection and provide limited support for other object manipulation tasks. Other techniques bring distal content within reach [5, 26] or onto an intermediate device [36] to allow more complex interactions within the personal space.

Work in virtual reality has explored similar object displacement and scaling techniques, including virtual arm extension [33, 34], VoodooDolls [29], control-display gain manipulation [1, 12], and world-in-miniature approaches [38]. However, removing these techniques from a fully virtual environment may have negative consequences. Approaches like surrogate objects [31] mimic these ideas, but also decontextualize virtual content from the physical world. AR applications often use virtual content to augment objects in the physical world, so this decontextualization may create cognitive challenges. We instead focus on selecting and manipulating content that remains at a fixed distance, and build on prior findings in these spaces to measure precision and ease of use in interacting with distant objects in AR.

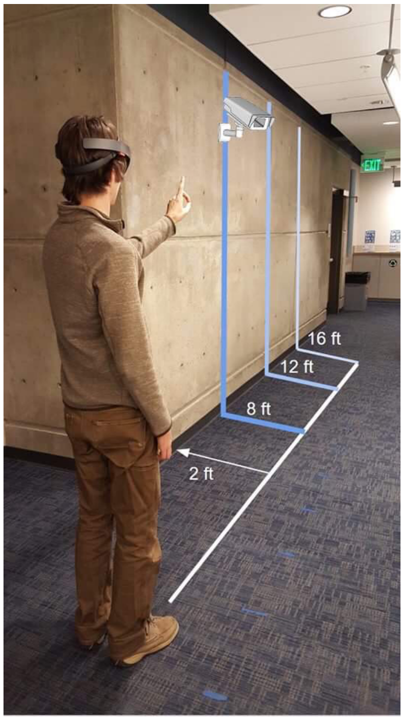


Figure 1: We used an Internet of Things scenario to measure the speed, accuracy, and perceived usability of three modalities for interacting with virtual content at 8, 12, and 16 feet.

3 Experimental Overview

Work in VR, AR, and HCI has noted clear challenges for interacting with objects at a distance, including reduced visual angle, ambiguous pointing, and paradigm shifts for embodied interactions. Work on the effects of proxemics on interactions with large displays and pervasive computing environments suggests not only that increased distance between users and objects affects efficiency but also that being within different proxemic zones changes users mental models for these interactions [3, 9]. While significant research explores AR interactions within the intimate (0.5ft - 1.5ft) and personal (1.5ft - 4ft) zones, relatively little investigates the social (4ft - 12ft) and public (larger than 12ft) proxemic zones where visual, cognitive, and proxemic design factors may shift the effectiveness of different interaction modalities. We tested interactions with virtual content placed 8 feet, 12 feet and 16 feet away (2.4m, 3.7m, and 4.9m; Fig. 1). These distances provide coverage of the social (8 ft) and public (16 ft) proxemic zones as well as the distance between them.

Hands-free AR-HMDs allow users to leverage a variety of possible input mechanisms for interaction, including embodied interactions (e.g., hand gestures and headtracking), external devices (e.g., handheld devices) and multimodal inputs (e.g., combined voice and gesture). In this paper, we examine three interaction modalities commonly available in AR-HMDs: multimodal headtracking and voice, embodied headtracking and free-hand gesture, and a pointer-based handheld remote. While prior studies offer initial insight into these modalities at near distances, we have comparably little understanding of their efficacy for distal interactions. For example, the factors that make embodied interaction modalities effective for close object may change as objects move across different proxemic zones [42]. Prior work in multimodal AR interfaces shows that voice interactions could make distal interactions more intuitive and efficient for some tasks [2, 31]. Embodied freehand gestures support effective near-range interactions as users understand the metaphor of interacting with virtual content in the same space as physical objects [8, 15, 37]. However, the scalability of freehand gestures to distant AR interactions where freehand manipulation is not directly on the object is not well understood. Pointing devices commonly used in distal interactions in pervasive computing environments may also support effective distal object manipulation in AR [39, 42]; however, angular errors in pointing are amplified at a distance [23]. Prior studies in virtual reality offer preliminary hypotheses for the efficacy of these modalities in distal interactions, but perceptual and contextual factors discussed in Section 2 may limit the extensibility of these results in practice.

We evaluated distal interactions for these modalities using three interaction primitives: selection, rotation, and translation. These primitives form the basis of most interactions with virtual objects and menu systems, and have been studied in other contexts (e.g., [20,31]). Focusing on these tasks allows us to evaluate the effects of different modalities on meaningful interactions, including those where the decreased size of distant objects may play a role.

People's limited capabilities for estimating object depth may further complicate distal interaction in AR [28]; to interact with distant content, participants must *perceive* the content as existing beyond their reach. To overcome this challenge, we situate our exploration in the context of virtual smart-home devices [21]. We used tasks involving a thermostat and security camera as these objects have a familiar shape and size, which allowed us to use angular size as a depth cue. They are also often mounted at eye level or higher, avoiding field of view limitations in many AR-HMDs. Finally, the basic functionality of these devices is generally familiar to users, which allowed us to design natural user interfaces to ground our evaluative tasks. Contextualizing selection, rotation, and translation in two different task scenarios also provides insight into these tasks across multiple contexts and operational framings.

We collected objective and subjective metrics associated with efficacy and usability for each modality. Our objective measures focused on completion time and accuracy, while subjective metrics explored perceived usability, physical and cognitive demands, and preference. We hypothesized that:

H1: Handheld remotes will provide qualitatively less frustrating interactions than embodied freehand gestures.

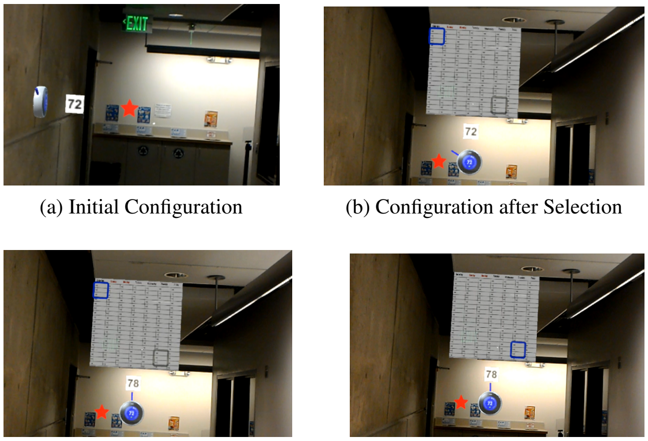
Gesture-based interactions require participants to actively engage their bodies to complete a task by virtually "grasping" and manipulating objects, often occluding objects as people bring their hands between their eyes and the object. We hypothesize that this occlusion combined with fatigue associated with the user holding their arms up [16] will result in more frustration than with a pointer-based remote that can be held at a user's side and provides an analog to conventional technologies, such as television remotes.

H2: Multimodal voice interactions will be robust to distance, while embodied freehand gestures and handheld remotes will be slower and less accurate as distance increases.

Controlling a raycasted cursor, as with headtracked or device-mediated interactions, becomes more challenging as distance from an object increases [41]. While our multimodal voice implementation leverages some positional interaction through a headtracked cursor, we anticipate distance will be more problematic for freehand gesture and remote-based interactions that require additional spatial manipulation to complete. These manipulations may compound inaccuracy in aiming due to shifts in the user's head orientation during movement, whereas no motion is required for the voice commands.

H3: Perceived & objective efficiency will mirror perceived usability.

Ideal interactive experiences require that input modalities not only be quick and accurate, but also natural and intuitive. Given the differences in metaphor provided by the tested methodologies, we predict that efficiency (measured by time to completion and target accuracy) will correlate directly with perceived usability.



(a) Initial Configuration

(b) Configuration after Selection

(c) Configuration after Rotation

(d) Configuration after Translation

Figure 2: The smart thermostat scenario asked participants to select a thermostat to initialize a menu, rotate the thermostat to adjust the room temperature, and translate a box to program a schedule.

4 Methods

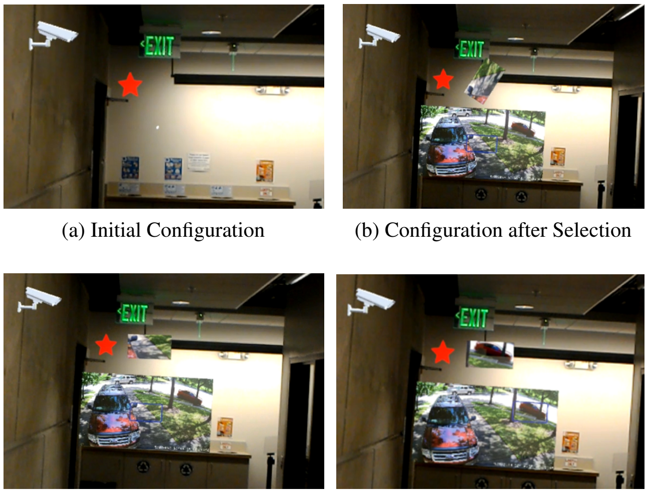
We conducted a 3 (modality) × 3 (distance) × 2 (device scenario) mixed-factors study to evaluate the utility of three different input modalities for distal interactions. Modality was treated as a within-subjects factor, while distance and scenario were between-subjects factors counterbalanced between participants. We used a full-factorial design with interaction modality and distance as independent variables. Dependent variables included completion time, accuracy, number of errors, and subjective usability (c.f., §4.4).

4.1 Experimental Tasks

Drawing on the set of basic interactions identified in prior literature (e.g., [20, 31]), we tested three different interactions comprising common operations in graphical user interfaces and AR applications: selection, rotation, and translation. Participants completed tasks involving each interaction for two virtual smart-home devices: a thermostat and a security camera. These two scenarios allow us to explore the utility of our tested modalities across different contexts, magnitudes, and question framings to generate more robust insights into the performance of each modality.

Scenario 1: Smart Thermostat The thermostat scenario asked participants to adjust the temperature and schedule settings on a virtual thermostat (Figure 2). The thermostat first appears on a wall next to a red star indicating task status. Participants make two selections: one to move the thermostat off the wall to face the user and a second to open a calendar menu. Participants then rotate the thermostat face to adjust the temperature from 72° to 78°, corresponding to a 60° rotation of the model. Finally, participants updated the thermostat schedule by translating the blue viewport around the hours of 6am to 9am on Saturday to surround 8pm to 11pm on Thursday. We constrained the blue box to the bounds of the calendar menu and placed the viewport's target off any edge to avoid potential effects from this constraint.

Scenario 2: Security Camera The security camera scenario asked participants to open and adjust a security camera feed (Figure 3). The camera first appears on the wall next to a red star. The participant then makes two selections: one displays the camera's "feed" (an image of a driveway), and the second reveals the camera's focus, indicated by a thumbnail above the image and blue box on the feed. The focus is initialized at $-60°$ from horizontal. Participants rotate the feed thumbnail to align with the full image. They then translate



(a) Initial Configuration

(b) Configuration after Selection

(c) Configuration after Rotation

(d) Configuration after Translation

Figure 3: The security camera scenario asked participants to initialize a feed using selection, rotate the camera to correct the feed, and translate the camera to adjust the camera focus.

the viewport's focus by moving a blue box in the primary feed from the driveway to a red car in the upper right of the image.

4.2 Apparatus

We conducted the experiment using a Microsoft HoloLens, a popular commercial see-through stereographic AR-HMD, with a 30° × 17° field of view. We leveraged the head-tracking, voice, and gesture tracking capabilities native to the HoloLens to assist with our interactions whenever possible to ensure reproducibility. Because the remote device included with the HoloLens used in this study had limited degrees of freedom, we opted to use a Nintendo Wiimote as our handheld input device. Wiimote inputs were collected using a series of wireless IR sensors.

We conducted our study using a custom experimental framework built in Unity (available at https://github.com/CU-VisuaLab/FullARInteractionsStudy). This application allowed for integrated data collection, control over presentation order, and interchangeable scenarios. We monitored framerates for each modality to ensure consistent performance (minimum of 14fps). In piloting, we found no noticeable perceived performance degredation due to the minimum framerate.

4.2.1 Interaction Modalities

We used a combination of the sensors integrated into the HoloLens and the Wiimote to implement interaction functionality for each of the three modalities. When possible, we used default behaviors from existing systems (e.g., the HoloLens' Air Tap selection) to mimic interactions participants may already be familiar with. We conducted a pilot study with six lay users to determine interaction designs when default behaviors were not available. Because we aimed to evaluate specific modalities rather than to craft optimal interactions, we designed each interaction's functionality to mimic functionality in the other modalities as closely as possible (e.g., translation always involved a click-and-drag paradigm). While inferring optimal designs is important future work, this choice mitigated potential confounds from design choices.

Voice-based Interactions: Our multimodal voice interface paired a headtracked cursor for object specification with voice commands to initiate specific actions. A white circular cursor appeared in the middle of the participant's field of view and moved with changes in head position. Verbal commands were processed using the Microsoft HoloToolkit Keyword Manager.[1] To select objects, the user adjusted the cursor position to hover over the object and said "Select." Participants translated objects by hovering over the object and saying "Move up/down/left/right," initiating object movement along a 2D plane perpendicular to the wall ($z = 8ft$, $z = 12ft$, or $z = 16ft$). at 44mm per second until the participant said "Stop moving" with the cursor positioned over the object. Participants rotated objects by hovering over the object and saying "Rotate left/right." Objects rotated at 11° per second until participants said "Stop rotating."

Freehand Gesture-Based Interactions: As with voice, embodied freehand gestures used the default HoloLens headtracked cursor for object identification. We used the HoloLens' Air Tap gesture for object selection: participants held their hand in front of the device with index finger pointed directly up, then tapped their index finger against their thumb to select that object.[2] Translation initiated by again depressing the index finger, moving the closed hand to reposition the target object, and raising the index finger to stop. As orientation tracking proved unreliable with closed hands, we implemented rotation as a two-handed gesture beginning with the same air-tap gesture with both hands. The object's rotation adjusted as the angle between participants' hands changed, corresponding to changes in slope between hands. The interaction completed when either or both index fingers pointed upwards.

Handheld Remote Interactions: We connected the Wiimote to the experimental framework by relaying sensor and input data through the HoloToolkit. The Wiimote cursor corresponded to where participants pointed the Wiimote. Participants selected objects by hovering the cursor over the object and pressing the "A" button. Translation occurred by holding the "B" button or trigger on the back of the device and moving the cursor by pointing the Wiimote to a new location. Translation ceased when the button was released. Rotation required participants to hover over an object and again hold the button. They rotated the object by rotating the Wiimote and releasing the button to stop.

4.3 Procedure

Participants first provided informed consent and were screened for stereoblindness to eliminate potential confounds from stereoscopic viewing. Each participant then completed a training phase, testing phase, and questionnaire for each input modality, resulting in three blocked phases presented in a random order. After completing all three blocks, participants completed a questionnaire comparing the three modalities and were compensated with a $10 gift card. On average, the experiment took 63 minutes to complete.

The experiment began with a training phase intended to orient each participant to the set of interactions they would be using. The training scene consisted of a cube and a sphere placed 3m in front of the participant. Participants were verbally instructed on performing selection, translation, and rotation for the current modality. They could practice those interactions for up to 5 minutes until they felt comfortable. Participants saw the same training scene for each modality, three times total.

During the testing phase, participants completed one selection, rotation, and translation task using the active modality. Each task occurred at either 8, 12, or 16ft, with participants seeing each distance once per modality. As a quality check on the perceived proxemic zone of our interactions, participants reported their perceived distance from the objects at the end of each phase. On average, participants perceived targets to be between 6.9ft and 14.1ft away. These estimates are consistent with prior research, and span our intended proxemic zones. Scenario and distance order were counterbalanced between participants using a Latin square design.

[1] https://github.com/Microsoft/MixedRealityToolkit-Unity

[2] https://developer.microsoft.com/en-us/windows/mixed-reality/gestures

Prior to each testing phase, participants were positioned at a fixed point in the experimental space and instructed to look at a specific calibration target to align the virtual objects and physical environment (Figure 1). For each task, participants were first told how to complete the task and to complete each task as quickly and accurately as possible. They indicated that they were ready to begin the task by using the voice command "Start", which turned the visual indicator green (Figures 2 & 3). When complete, the participant said "Done," turning the visual indicator red. We chose a voice-based transition to provide a consistent framing across modalities and allow participants time to refine each task. After completing each testing phase, participants completed a subjective questionnaire about the active modality. After all three blocks, participants completed a final questionnaire comparing the different input modalities and self-reporting demographic information.

4.4 Measures

We used three objective measures to capture user performance: completion time, accuracy, and the number of times participants attempted each interaction. We measured completion time as the time elapsed between the "Start" and "Done" commands for each task. We cross-referenced video logs to correct for any errors in keyword identification impacting completion time. In rotation and translation, we measured accuracy as distance from the defined target on the x-y plane. We counted the number of attempts as the number of times the participant performed a complete action between "Start" and "Done" commands (e.g., number of Air Taps, "Start/Stop" commands, or button presses). We analyzed objective metrics using a four-factor (modality, scenario, distance, and participant's self-reported familiarity with AR) ANCOVA for each task. We included modality order and distance order as random covariates to mitigate the influence of learning effects. Tukey's Honest Significant Difference Test (HSD) was used for post-hoc analysis.

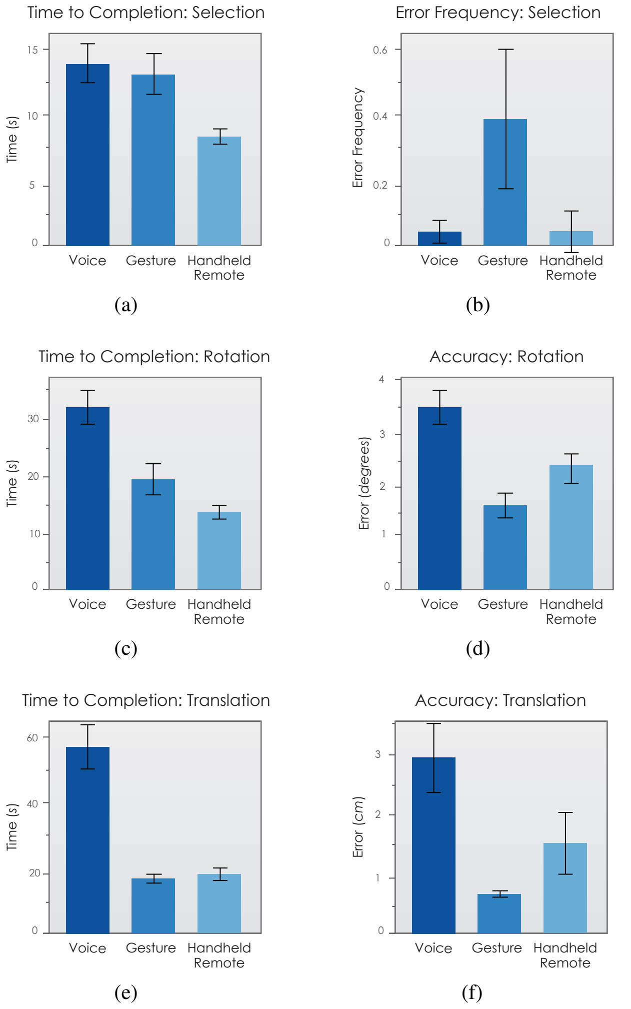
Our subjective metrics described factors of perceived effectiveness and usability collected through four questionnaires: one for each modality and one comparing all three modalities. The modality questionnaire included a version of the System Usability Scale (SUS) with language modified to match our system implementation, 19 supplemental questions using 5-point Likert scales, and six free-response questions to gather data on frustration, intuitiveness of selection, rotation and translation. We constructed four scales from the subjective responses pertaining to naturalness, concentration, frustration, and social acceptability (Cronbach's $\alpha > .75$ for all factors). Our closing questionnaire collected perceptions of relative efficacy and preference across modalities as well as demographic information and data about participants' previous experience with relevant technologies. Scales were analyzed using a two-factor ANCOVA (modality and prior familiarity with AR) with modality ordering and distance ordering as random covariates and Tukey's HSD for post-hoc analysis. We analyzed rankings from the comparative questionnaire using a Wilcoxon Signed Rank Test. Questionnaires are available at cmci.colorado.edu/visualab/ARDistanceInteractions.

4.5 Participants

We collected data from 20 participants recruited from the local campus community (11 male, 9 female). Participants ranged in age from 18-65 years ($\mu_{age} = 21.3, \sigma_{age} = 6.0$). Participants reported low prior familiarity with AR ($\mu = 2.5$), the HoloLens ($\mu = 2.1$), and gesture input ($\mu = 2.4$); medium familiarity with voice-based inputs ($\mu = 3.2$); and high familiarity with Wiimote input ($\mu = 4.4$). All participants were native English speakers to avoid potential ambiguity around voice commands.

5 Results

We analyzed effects of modality, distance, scenario, and prior AR experience on objective performance and subjective usability and preference. We summarize main findings here and have made the data available at cmci.colorado.edu/visualab/ARDistanceInteractions. Figure 4 summarizes our objective results.



(a) (b) (c) (d) (e) (f)

Figure 4: Our objective results indicated that embodied gestures and device-mediated interactions supported faster and more accurate interactions at a distance overall. Error bars represent standard error.

5.1 Objective Results

Selection: We found a significant main effect of modality on completion time for selection tasks ($F(2,18) = 6.36, p = .002$), with handheld remotes (8.45s ± 1.19) being significantly faster than voice (14.00s ± 2.74) or embodied gestures (12.97s ± 2.82) (Fig. 4a). Participants reporting moderate to high prior familiarity with AR performed selections faster on average than those with low familiarity ($F(1,18) = 10.81, p = .001$). We found no significant effect of distance on time to completion for selection.

We also found a significant main effect of modality on the number of times participants missed the selection target ($F(2,18) = 5.58, p = .005$). Participants using embodied gestures(0.38 ± 0.28) registered significantly more off-target selections than with voice (0.05 ± 0.05) or handheld remotes (0.067 ± 0.105). The thermostat task had more off-target selections overall than the camera task ($F(1,18) = 4.45, p = .046$). We did not evaluate accuracy data for selection, as selections were either successful or not.

Rotation: We found a significant main effect of modality on rotation completion time ($F(2,18) = 14.96, p < .0001$): participants performed rotation tasks significantly more quickly with handheld remotes (13.32s ± 2.36) and embodied gestures (19.16s ± 5.4) than with voice interactions (31.73s ± 5.86). Participants with higher self-reported familiarity with AR performed faster overall ($F(1,18) = 10.48, p = .001$).

We found a significant main effect of modality on rotation accuracy ($F(2,18) = 14.76, p < .0001$), with voice interactions (3.69° ± .65) leading to greater error than embodied gesture (1.61° ± .51) or handheld remotes (2.51° ± .56). A significant interaction effect between modality and scenario indicated that voice-based rotations were especially error-prone in the security camera scenario ($F(1,18) = 17.45, p < .0001$). Participants with prior AR experience were marginally more accurate overall ($F(1,18) = 3.05, p = .083$). We also found a significant effect of modality on the number of attempts needed to successfully complete the task ($F(2,18) = 9.32, p < .0001$), with voice interactions taking a significantly higher number of attempts. We found a marginally significant improvement with prior AR familiarity ($F(1,18) = 3.46, p = .065$), but no significant effect of distance on performance.

Translation: We found a significant main effect of modality ($F(2,18) = 41.53, p < .0001$) and scenario ($F(1,18) = 9.21, p = .003$) on translation completion time with handheld remotes (17.66s ± 3.61) and embodied gestures (15.96 ± 2.03) outperforming voice (55.92s ± 12.62), and participants completing thermostat tasks more quickly overall (Camera: 23.79 ± 3.88, Thermostat: 35.77s ± 9.34). We again found evidence that prior AR experience leads to faster completion times ($F(1,18) = 9.21, p = .003$). We found a significant interaction of modality and scenario ($F(2,18) = 9.42, p = .0001$), with voice-based interactions being significantly slower in the thermostat condition (75.05s ± 22.74). We found a three-way interaction between modality, scenario, and AR familiarity, with less experienced participants taking longer to complete voice-based translations in the thermostat scenario ($F(2,18) = 3.80, p = .024$).

We found a significant main effect of modality on translation accuracy ($F(2,18) = 6.44, p = .002$), with embodied gestures (6.23mm ± 1.04) leading to more accurate translations than voice (14.69mm ± 9.80). Voice led to significantly larger errors for the thermostat scenario (38.38mm ± 22.20) than the camera scenario (18.85mm ± 4.24, $F(2,18) = 3.59, p = .03$). We also found a significant main effect of modality on the number of attempts made ($F(2,18) = 33.34, p < .0001$), with voice again requiring significantly more attempts on average. The thermostat scenario required more attempts to complete ($F(1,18) = 13.29, p = .004$), especially with voice input ($F(2,18) = 8.67, p = .0003$). We found a significant interaction of modality, scenario, and distance ($F(2,18) = 3.37, p = .037$), with handheld remote errors increasing and voice errors decreasing as distance increased.

5.2 Subjective Results

System Usability Scale: SUS results indicated that embodied gestures had significantly higher perceived usability than voice interactions ($F(2,18) = 6.33, p = .003$). Embodied gestures had the highest perceived usability (79.0 ± 8.8) overall, followed by handheld remote interactions (72.25 ± 7.75) and voice (59.5 ± 8.25).

Per-Task Ease of Interaction: We created scales from our Likert responses for highly correlated elements. We normalized these scales such that 1 corresponded to a negative experience (e.g., high fatigue, low enjoyment) and 5 corresponded to a positive experience (e.g., low fatigue, high enjoyment). We analyzed these scales using a two-factor ANCOVA (modality and AR familiarity) and Tukey's HSD for post-hoc comparisons.

We created one scale for selection from perceptions of frustration and annoyance for both selection and rotation. We found a significant effect of prior familiarity with AR on the scale for selection ($F(1,18) = 2.19, p = .006$), with less experienced participants reporting more positive experiences overall. We also found a significant effect of modality on this scale for rotation ($F(2,18) = 1.12, p = .037$) with participants scoring embodied hand gestures the least frustrating (3.65 ± .56), followed by handheld remotes (3.00 ± .53) and then voice (2.58 ± .71).

We created two scales from subjective perceptions of translation: one combing responses about whether translation was tiring, frustrating, and annoying (fatigue/frustration) and a second combining responses to questions on difficulty in translation and whether translation required heavy concentration into a second scale (ease of interaction, 1 = easy, 0 = difficult). We found a significant effect of modality ($F(2,18) = 16.8, p < .0001$) on fatigue/frustration, with participants scoring embodied gestures the highest (4.58 ± .23), followed by handheld remotes (3.85 ± .64), and then voice interactions (2.57 ± .51). We also found a significant main effect of input modality on ease of interaction ($F(2,18) = 25.38, p < .0001$). Translation with hand gestures was perceived as easiest (.98 ± .052), with handheld remotes as moderately difficult (.68 ± .22), and with voice input as most difficult (.25 ± .16). We found a significant interaction of modality with AR experience ($F(2,18) = 3.65, p = .024$), with more experienced participants scoring the handheld remote interactions lower than less experienced participants.

Enjoyment Responses to whether interactions felt natural and fun formed an enjoyment scale. We found a significant effect of modality ($F(2,18) = 4.34, p = .018$) with participants scoring embodied gestures the highest (4.28 ± .31), followed by handheld remotes (3.98 ± .46), and then voice (3.4 ± .53).

Comfort in Social Setting Responses to how comfortable participants would feel using the system in shared spaces (e.g., workplace) or public spaces (e.g., a mall or store) formed a social comfort scale. We found a significant effect of input modality ($F(2,18) = 3.17, p = .050$) with participants scoring embodied gestures the highest (3.6 ± .43), followed by handheld remotes (3.2 ± .54), and then voice (2.72 ± .60). We also found a significant effect of experience with AR ($F(1,18) = 6.11, p = .017$), with more experienced participants giving lower scores overall.

5.3 Direct Comparisons

We asked participants to directly compare their perceptions of how easily and how quickly the three modalities allowed for selection, rotation and translation. Across all three tasks, participants consistently reported embodied freehand gestures as the easiest to use ($Z_{sel} = -58.5, p = .035$, $Z_{rot} = -67.5, p = .013$, $Z_{trn} = -4.33, p = .0004$) and voice-based interactions as most difficult ($Z_{sel} = 75.00, p = .004$, $Z_{rot} = -66.00, p = .01$, $Z_{trn} = 84.50, p = .001$). We found the same ranking for perceived speed across rotation and translation, with gestures perceived as fastest ($Z_{rot} = -84.50, p = .001$, $Z_{trn} = -4.33, p = .0004$) and voice commands as slowest ($Z_{rot} = 71.50, p = .0074$, $Z_{trn} = 92.00, p = .0001$).

6 Discussion

We examined how three different modalities–multimodal voice, embodied gesture, and handheld remote–support effective and usable interactions with distal virtual content. We conducted a controlled user study revealing

- Voice interaction was least efficient and least preferred,
- Embodied gestures were perceived as the most usable and strongly preferred, and
- Both handheld remotes and embodied gestures enabled fast and accurate interactions.

In free-form responses, participants found both gesture and remote-based interactions to support a more intuitive "pick up, put down" metaphor for the spatial interactions tested in our study. While our results and prior work suggest voice is more robust to distance, participants found the predictive needs associated with completing tasks using voice cognitively demanding. However, open-ended responses identified significant potential for voice commands as "better oriented towards actions that lacked accuracy," (P7), complimenting recent findings [2]. Others stated that these interactions were "similar to talking to smart phone so it felt both natural and intuitive to me" (P9). However, we found no indication that prior familiarity with these technologies led to systematic preference for any given modality, providing preliminary evidence that relative performance is inherent to each modality.

Contrary to H1, participants consistently preferred embodied gestures to handheld remotes, perceiving them as more usable and efficient despite a lack of objective efficiency differences. While handheld remotes commonly support real-world distal interactions, our subjective feedback indicates that locating and maintaining a cursor was significantly easier with embodied gestures, where the cursor appears in the center of the field of view. Cursors felt most "stable" when following head motion (as in our voice and embodied gesture conditions). Participants found it "difficult to find the mouse using the wii remote because it would go outside the [field of view boundary]" (P3) and noted "using a remote control for the cursor did not feel natural" (P9). These findings run contrary to preferences for cursor positioning in Wither & Hollerer [42]. We also found that participants missed the target when selecting objects significantly more often with embodied gestures than with either voice or handheld remotes despite their expressed preference. We hypothesize this is due to relative inexperience with hand gestures, but further exploring these discrepancies is important future work.

We found little support for H2: increased distance only improved performance for voice interactions and degraded performance for handheld remotes and freehand gestures when considered in combination with other factors. The lack of degradation is likely explained by Kopper et. al.'s observation that precision in distal interactions may degrade quadratically with distance [23]: our targets may have been sufficiently far away that the 4 foot gap between distances did not significantly increase expected error. Participants perceived distance to have the strongest effect during the handheld remote interactions. This perceived degradation could relate to the same challenges that led to lower perceived usability overall (e.g., finding and stabilizing the cursor).

H3 was partially supported: voice-based interaction performed worst both objectively and subjectively. Our results replicate prior studies of voice interactions for spatial tasks [18]. Despite comparable objective performance, embodied freehand gestures had higher perceived usability than the handheld remote in all subjective metrics. This finding mirrored participants' relative comparisons. It is supported by prior literature in close-range AR interactions, but contrasts with Wither & Hollerer's findings for cursor positioning [42] that also found no significant performance differences between device-mediated and embodied interaction, but a preference for device-mediated interaction. We anticipate that this difference arises from our use of two spatial manipulation tasks in addition to selection: while participants performed selections more effectively with handheld remotes, subjective feedback reflects user preferences across all three tasks.

6.1 Limitations & Future Work

In measuring the usability and efficacy of different modalities for distal interactions, we made several decisions for experimental control that raise potential directions for future research. We designed the interactions to be consistent across each modality. For example, each translation followed a continuous "click and drag" metaphor. We used this approach to focus on the affordances of each modality rather than try to identify an optimal design for each task. Feedback from participants suggests that those decisions were sensible and intuitive; however, both considering the space of possible interactions afforded by each modality and extending our study to future input devices provide promising directions for future work. Results from our subjective metrics and free-response questions can provide preliminary insight into new interaction designs across these modalities as the basis for future participatory design studies [32, 43]. Future extensions of this study may also include additional interaction tasks such as zoom, pan and resize and exploring objects placed outside of the immediate field of view.

Using a third-party handheld remote required significant integration into the experimental apparatus compared to utilizing native HoloLens headtracking, gesture and voice interactions. This introduced possible latency issues. While framerate was comparable across all modalities, some participants reported a minor lag when performing Wiimote rotations. Future iterations on our infrastructure could use the Wiimote's internal gyroscopes rather than IR to monitor rotation. We also did not provide supplemental virtual depth cues such as cast shadows. Instead, we used virtual object sizes corresponding to the modeled physical artifacts to allow participants to infer depth based on their understanding of the object's size which may introduce some variation in perceived depth.

Finally, we focused on a small set of tasks sampled across room-scale distances. However, future work should explore an expanded library of spatial scales, contexts, and configurations. For example, we focused on interaction along a 2D plane to measure interaction capabilities at fixed distances; however, AR affords interactions that manipulate objects in three dimensions and at varying heights. Exploring interactions along a continuous range of distances crossing multiple proxemic zones and spatial trajectories would further explore how proxemic ranges change the efficacy of different interaction modalities. Additionally, these explorations may further clarify the role of experience on distal interaction design. We anticipate our findings and infrastructure will enable new research in these directions.

7 Conclusion

We explore how interaction modalities influence AR interactions with distal content, focusing on content in the social and public proxemic zones. We conducted a controlled user study to evaluate performance and preference for selecting, rotating, and translating virtual objects. Both embodied freehand gestures and mediated handheld remote interactions led to more efficient interactions than voice control. However, people strongly preferred embodied interactions, perceiving them as faster and easier to use. These findings offer initial insight into designing systems for distal interaction in AR, where immersive environments offer environment-scale interactions with digital content embedded in the physical world.

Acknowledgements

We thank Emily Howell for her assistance. This research was funded by NSF IIS-1652907 and a grant from CU Innovative Seed Program.

References

[1] C. Andújar and F. Argelaguet. Friction surfaces: scaled ray-casting manipulation for interacting with 2d guis. In *Proc. EGVE*, pp. 101–108. The Eurographics Association, 2006. doi: 10.2312/EGVE/EGVE06/101-108

[2] S. K. Badam, A. Srinivasan, N. Elmqvist, and J. Stasko. Affordances of input modalities for visual data exploration in immersive environments. In *Proc. Workshop on Immersive Analytics at IEEE VIS*, 2017.

[3] T. Ballendat, N. Marquardt, and S. Greenberg. Proxemic interaction: designing for a proximity and orientation-aware environment. In *Proc. ITS*, pp. 121–130. ACM, New York, 2010. doi: 10.1145/1936652.1936676

[4] E. Barba and R. Z. Marroquin. A primer on spatial scale and its application to mixed reality. In *Proc. ISMAR*, pp. 100–110. IEEE Computer Society, Washington, DC, 2017. doi: 10.1109/ISMAR.2017.27

[5] A. Bezerianos and R. Balakrishnan. The vacuum: facilitating the manipulation of distant objects. In *Proc. CHI*, pp. 361–370. ACM, New York, 2005. doi: 10.1145/1054972.1055023

[6] M. Billinghurst. Put that where? voice and gesture at the graphics interface. *SIGGRAPH Comput. Graph.*, 32(4):60–63, Nov. 1998. doi: 10.1145/307710.307730

[7] R. A. Bolt. "put-that-there": Voice and gesture at the graphics interface. In *Proc. SIGGRAPH*, pp. 262–270. ACM, New York, 1980. doi: 10.1145/800250.807503

[8] V. Buchmann, S. Violich, M. Billinghurst, and A. Cockburn. Fingartips: gesture based direct manipulation in augmented reality. In *Proc. GRAPHITE*, pp. 212–221. ACM, New York, 2004. doi: 10.1145/988834.988871

[9] T. Dingler, M. Funk, and F. Alt. Interaction proxemics: Combining physical spaces for seamless gesture interaction. In *Proc. PerDis*, pp. 107–114. ACM, New York, 2015. doi: 10.1145/2757710.2757722

[10] A. O. S. Feiner. The flexible pointer: An interaction technique for selection in augmented and virtual reality. In *Proc. UIST*, pp. 81–82. ACM, New York, 2003.

[11] T. Feuchtner and J. Müeller. Extending the body for interaction with reality. In *Proc. CHI*, pp. 5145–5157. ACM, New York, 2017. doi: 10.1145/3025453.3025689

[12] S. Frees, G. D. Kessler, and E. Kay. Prism interaction for enhancing control in immersive virtual environments. *TOCHI*, 14(1):2, may 2007. doi: 10.1145/1229855.1229857

[13] M. Gandy, T. Starner, J. Auxier, and D. Ashbrook. The gesture pendant: A self-illuminating, wearable, infrared computer vision system for home automation control and medical monitoring. In *Proc. ISWC*, p. 87. IEEE Computer Society, Washington, DC, 2000. doi: 10.1109/ISWC.2000.888469

[14] M. Golparvar-Fard, F. Peña-Mora, and S. Savarese. D4ar - a 4-dimensional augmented reality model for automating construction progress monitoring data collection, processing and communication. *Electronic Journal of Information Technology in Construction*, 14:129–153, june 2009.

[15] T. Ha, S. Feiner, and W. Woontack. Wearhand: Head-worn, rgb-d camera-based, bare-hand user interface with visually enhanced depth perception. In *Proc. ISMAR*, pp. 219–228. IEEE Computer Society, Washington, DC, 2014. doi: 10.1109/ISMAR.2014.6948431

[16] J. D. Hincapié-Ramos, X. Guo, P. Moghadasian, and P. Irani. Consumed endurance: a metric to quantify arm fatigue of mid-air interactions. In *Proc. CHI*, pp. 1063–1072. ACM, New York, 2014. doi: 10.1145/2556288.2557130

[17] T. N. Hoang, R. T. Smith, and B. H. Thomas. Ultrasonic glove input device for distance-based interactions. In *Proc. ICAT*, pp. 46–53. IEEE Computer Society, New York, 2013. doi: 10.1109/ICAT.2013.6728905

[18] S. Irawati, S. Green, M. Billinghurst, A. Duenser, and H. Ko. Move the couch where?: developing an augmented reality multimodal interface. In *Proc. ISMAR*. IEEE Computer Society, Washington, DC, 2006. doi: 10.1109/ISMAR.2006.297812

[19] J. Irizarry, M. Gheisari, G. Williams, and B. N. Walker. Infospot: A mobile augmented reality method for accessing building information through a situation awareness approach. *Automation in Construction*, 33:11–23, aug 2013. doi: 10.1016/j.autcon.2012.09.002

[20] E. Kaiser, A. Olwal, D. McGee, H. Benko, A. Corradini, X. Li, P. Cohen, and S. Feiner. Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality. In *Proc. ICMI*, pp. 12–19. ACM, New York, 2003. doi: 10.1145/958432.958438

[21] S. Lanka, S. Ehsan, and A. Ehsan. A review of research on emerging technologies of the internet of things and augmented reality. In *Proc. I-SMAC*, pp. 770–774. IEEE Computer Society, Washington, DC, 2017. doi: 10.1109/I-SMAC.2017.8058283

[22] M. Lee, M. Billinghurst, W. Baek, R. Green, and W. Woo. A usability study of multimodal input in an augmented reality environment. *Virtual Reality*, 17(4):293–305, nov 2013. doi: 10.1007/s10055-013-0230-0

[23] R. Lopper, D. A. Bowman, M. G. Silva, and R. P. McMahan. A human motor behavior model for distal pointing tasks. *International Journal of Human-Computer Studies*, 68(10):603–615, 2010. doi: 10.1016/j.ijhcs.2010.05.001

[24] N. Marquardt and S. Greenberg. Informing the design of proxemic interactions. *IEEE Pervasive Computing*, 11(2):14–23, 2012. doi: 10.1109/MPRV.2012.15

[25] M. J. McGuffin and R. Balakrishnan. Fitts' law and expanding targets: Experimental studies and designs for user interfaces. *TOCHI*, 12(4):388–422, dec 2005. doi: 10.1145/1121112.1121115

[26] B. A. Myers, R. Bhatnagar, J. Nichols, C. H. Peck, D. Kong, R. Miller, and A. C. Long. Interacting at a distance: measuring the performance of laser pointers and other devices. In *Proc. CHI*, pp. 33–40. ACM, New York, 2002. doi: 10.1145/503376.503383

[27] J. K. Parker, R. L. Mandryk, M. N. Nunes, and K. M. Inkpen. Tractorbeam selection aids: improving target acquisition for pointing input on tabletop displays. In *Proc. INTERACT*, pp. 80–93. Springer-Verlag, Berlin, 2005. doi: 10.1007/11555261_10

[28] W. Piekarski and B. H. Thomas. Interactive augmented reality techniques for construction at a distance of 3d geometry. In *Proc. EVGE*, pp. 19–28. ACM, New York, 2003. doi: 10.1145/769953.769956

[29] J. S. Pierce, B. C. Stearns, and R. Pausch. Voodoo dolls: seamless interaction at multiple scales in virtual environments. In *Proc. I3D*, pp. 141–145. ACM, New York, 1999. doi: 10.1145/300523.300540

[30] T. Piumsomboon. *Natural hand interaction for augmented reality*. PhD thesis, University of Canterbury, New Zealand, 2015.

[31] T. Piumsomboon, D. Altimira, H. Kim, A. Clark, G. Lee, and M. Billinghurst. Grasp-shell vs gesture-speech: A comparison of direct and indirect natural interaction techniques in augmented reality. In *Proc. ISMAR*, pp. 73–82. IEEE Computer Society, Washington, DC, 2014. doi: 10.1109/ISMAR.2014.6948411

[32] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn. User-defined gestures for augmented reality. In *Proc. CHI EA*, pp. 955–960. ACM, New York, 2013. doi: 10.1145/2468356.2468527

[33] I. Poupyrev, M. Billinghurst, and S. W. T. Ichikawa. The go-go interaction technique: non-linear mapping for direct manipulation in vr. In *Proc. UIST*, pp. 79–80. ACM, New York, 1996. doi: 10.1145/237091.237102

[34] I. Poupyrev, T. Ichikawa, S. Weghorst, and M. Billinghurst. Egocentric object manipulation in virtual environments: Empirical evaluation of interaction techniques. *Computer Graphics Forum*, 17(3):41–52, 1998. doi: 10.1111/1467-8659.00252

[35] F. A. Sanz and C. Andujar. A survey of 3d object selection techniques for virtual environments. *Computers and Graphics*, 37(3):121–136, 2013. doi: 10.1016/j.cag.2012.12.003

[36] J. Seifert, A. Bayer, and E. Rukzio. Pointerphone: Using mobile phones for direct pointing interactions with remote displays. In *Proc. INTERACT*, pp. 18–35. Springer, Berlin, 2013. doi: 10.1007/978-3-642-40477-1_2

[37] D. W. Seo and J. Y. Lee. Direct hand touchable interactions in augmented reality environments for natural and intuitive user experiences. *Expert Systems with Applications*, 40(9):3784–3793, july 2013. doi: 10.1016/j.eswa.2012.12.091

[38] R. Stoakley, M. J. Conway, and R. Pausch. Virtual reality on a wim: interactive worlds in miniature. In *Proc. CHI*, pp. 265–272. ACM Press/Addison-Wesley Publishing Co., New York, 1995. doi: 10.1145/223904.223939

[39] C. Swindells, K. M. Inkpen, J. C. Dill, and M. Tory. That one there! pointing to establish device identity. In *Proc. UIST*, pp. 151–160. ACM, New York, 2002. doi: 10.1145/571985.572007

[40] D. Vogel and R. Balakrishnan. Distant freehand pointing and clicking on very large, high resolution displays. In *Proc. UIST*, pp. 33–42. ACM, New York, 2005. doi: 10.1145/1095034.1095041

[41] C. A. Wingrave and D. A. Bowman. Baseline factors for raycasting selection. In *Proc. HCI International*, 2005. doi: 10.1.1.97.1738

[42] J. Wither and T. Hollerer. Evaluating techniques for interaction at a distance. In *Proc. ISWC*, pp. 124–127. IEE Computer Society, Washington, DC, 2004. doi: 10.1109/ISWC.2004.18

[43] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. In *Proc. CHI*, pp. 1083–1092. ACM, New York, 2009. doi: 10.1145/1518701.1518866

[44] F. Zhou, H. B.-L. Duh, and M. Billinghurst. Trends in augmented reality tracking, interaction and display: A review of ten years of ismar. In *Proc. ISMAR*, pp. 193–202. IEEE Computer Society, Washington, DC, 2008. doi: 10.1109/ISMAR.2008.4637362