Eliciting Multimodal Gesture+Speech Interactions in a Multi-Object Augmented Reality Environment

Xiaoyan Zhou* Colorado State University
Adam S. Williams† Colorado State University
Francisco R. Ortega‡ Colorado State University

ABSTRACT

As augmented reality technology and hardware become more mature and affordable, researchers have been exploring more intuitive and discoverable interaction techniques for immersive environments. In this paper, we investigate multimodal interaction for 3D object manipulation in a multi-object virtual environment. To identify the user-defined gestures, we conducted an elicitation study involving 24 participants for 22 referents with an augmented reality headset. It yielded 528 proposals and generated a winning gesture set with 25 gestures after binning and ranking all gesture proposals. We found that for the same task, the same gesture was preferred for both one and two object manipulation, although both hands were used in the two object scenario. We presented the gestures and speech results, and the differences compared to similar studies in a single object virtual environment. The study also explored the association between speech expressions and gesture stroke during object manipulation, which could improve the recognizer efficiency in augmented reality headsets.

Keywords: Human computer interaction (HCI), User studies, Mixed / augmented reality, Gestural input, Elicitation, Multimodal

Index Terms: Human-centered computing—Human computer interaction (HCI)—HCI design and evaluation methods—User studies; Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Mixed / augmented reality; Human-centered computing—Human computer interaction (HCI)—Interaction techniques; Human-centered computing—Human computer interaction (HCI)—Empirical studies in HCI Human-centered computing—Human computer interaction (HCI)—Interaction techniques—Gestural input;

1 INTRODUCTION

Easy-to-remember gestures produce high usability interfaces [15]. A gesture set that does not align with users' expectations or mental models often leads to frustrating user experiences [6]. Wobbrock et al. introduced an elicitation methodology to collect proposed gestures from users [33], facilitating intuitive gestures design without implementing perfect recognizers in advance. Prior findings proved that users prefer to choose input modalities based on their needs during the interaction [1, 8, 11]. Previous studies have explored gesture set design in different devices and interfaces [4, 11, 18, 25, 33], and several have been done with multimodal interactions in AR or VR [16, 26]. However, few to no researches have involved multimodal interactions in a multi-object AR or VR environment. These prior works raised multiple questions that we were explored during this study: Does multimodal interaction look different when having multi-object virtual environments? Does a multi-object environment impact the gesture and speech proposals? What gestures do users prefer with multiple object manipulation, and are there any differences from single object manipulation? The raised questions drive our motivation to understand if previous single object studies may transfer to more realistic environments. For this work, an elecitation study was conducted for multimodal interaction in AR with a Wizard of Oz (WoZ) experiment design (i.e., a researcher emulating a live system) [23, 33]. It involved 24 participants, 22 referents (i.e., command) in augmented reality (AR), and a head-mounted display (HMD). It yielded 528 proposals, and we generated a winning gesture set with 25 gestures after utilizing binning and ranking. We compared our single virtual object manipulation proposals to the findings from prior studies in a single object virtual environment [16, 26, 29]. For multiple object manipulation proposals, we compared them with the proposed gestures of single virtual object manipulation in our study. To the best of our knowledge, this is the first study to conduct multi-object mid-air interaction using optical-see through augmented reality headsets.

2 RELATED WORK

Elicitation methodology has been widely used in the HCI field to collect user-defined gestures. Wobbrock et al. popularized an elicitation methodology to collect proposed gestures from users [33], which aims to assist in designing more intuitive [33], guessable [32], learnable, and memorable [14] interaction techniques. Morris et al. found that people prefer gestures proposed by end-users, which were less complex than ones designed by human-computer interaction (HCI) experts [13]. Based on recent literature review results [24], over two hundred studies have adopted the use of an elicitation methodology in their work. Prior findings proved that users prefer to choose input modalities based on their needs during the interaction, such as choosing gestures over speech in an quiet environment [1, 8, 11]. Wobbrock et al. [33] discovered that having synonyms in a user-defined gesture set can increase the guessability of proposed gestures. A multimodal elicitation study provides the opportunity to create multimodal synonyms [11], which can offer users different modalities to achieve the same effect.

Nevertheless, most elicitation studies involving mid-air gestures in augmented reality (AR) only considered single object manipulation in a single object virtual environment [16, 17, 26, 29]. Pham et al. conducted an elicitation study with an AR headset that included a scenario of single building manipulation among multiple buildings [16]. However, the whole model was attached to a physical surface so that the elicited gestures in the study were not mid-air gestures. Moreover, as far as we know, no research has been done in multimodal interactions with multiple object manipulations in AR. Piumsomboon et al. implemented an elicitation study in AR (video-see through) that asked participants to select multiple objects and the elicited gestures were surface gestures [17]. Wittorf et al. adopted an elicitation methodology for exploring mid-air gestures with a wall display [31]. Danielescu and Piorkowski conducted an elicitation study to explore free-space gestures with a projector display that included multiple target selection among a set of photos [5]. However, the referent showed that photos were selected one by one, which could bias the participants' gesture proposal. Wobbrock et al. found that users preferred one hand over two hands for tabletop interaction [33]. We were interested in whether users preferred two

*e-mail: Xiaoyan.Zhou@colostate.edu
†e-mail: AdamWil@colostate.edu
‡e-mail:fortega@colostate.edu

hands for more than one object manipulation. This study aimed to understand the multimodal interaction in a multi-object virtual environment compared to a single object virtual environment. Furthermore, we were interested in the difference between single object manipulation and multiple object manipulation.

3 Study Design

This study conducted the elicitation experiment using a similar process as previous work [21, 23, 33]. 22 tasks (i.e., referents) were used for each modality during this work. Of those, 17 basic referents were selected based on their inclusion in prior works [26, 29], while the other 5 were developed to be multi-object versions of basic referent. Referents included six translations (along x, y, and z axes), six rotations (around x, y, and z axes), three abstracts (create, destroy, and select) and two scales (enlarge and shrink). For multiple object manipulation, only abstract and scale referents were included. There were three experiment blocks in this study, which included modality gesture only (G), speech only (S), and gesture plus speech (GS). Each block took approximately 10 minutes, plus two questionnaires and three surveys. The experiment lasted approximately 45 minutes.

3.1 Participants

The study involved 24 participants (12 female, 12 male). Due to the pandemic, it was difficult to recruit outside of the Computer Science (CS) department, therefore 17 out of 24 participants came from CS. Their ages ranged from 18-34 years (Mean = 23.42, SD = 4.20). All participants had previously used multi-touch devices, nineteen had used motion sense devices (e.g. Xbox Kinect or Nintendo Wii Motion), sixteen had used virtual reality headsets, and three had used augmented reality headsets.

3.2 Setup

The experiment was conducted using Microsoft HoloLens 2 optical see-through AR head-mounted display (HMD). The system used for the experiment was developed in Unity Engine 2019.4.4f1. A GoPro Hero 7 Black was mounted on top of HoloLens 2 to record an ego-centric view of the interactions, as shown in Figure 2. A 4k camera was placed on the front left corner facing participants to record an exo-centric view of the interactions. Two hand-shape icons on the screen were used to indicate if the hand or hands were in the view of the headset [27], as shown in Figure 1. If either hand is out of view, the corresponding hand icon would disappear from the screen. Before starting the experiment, participants were requested to complete the informed consent and demographics questionnaire. Then participants were informed that there would be three experiment blocks with different modalities as input and they can use any interaction they feel is appropriate to execute the command based on presented text referent and input modality. Participants were told to perform gestures inside of the headset view, which they can tell by the hand icons display. The interaction modalities were presented to participants in a counter-balanced order. In each block, referents were presented in random order. The post-study questionnaire was filled out by each participant at the end of the experiment.

3.3 Hypotheses

Our hypotheses were grounded in previous observations in our lab and from previous work [26]: H_1) for the same single object manipulation referent, winning gestures in a multi-object virtual environment will be different from ones in a single object virtual environment; H_2) participants would prefer to use both hands for two object manipulation referents.

4 Results

With the experiment, 528 proposals were collected from each modality. To eliminate the effect of the referent text biasing the speech proposal [26], prior to analysis, speech proposals that were identical

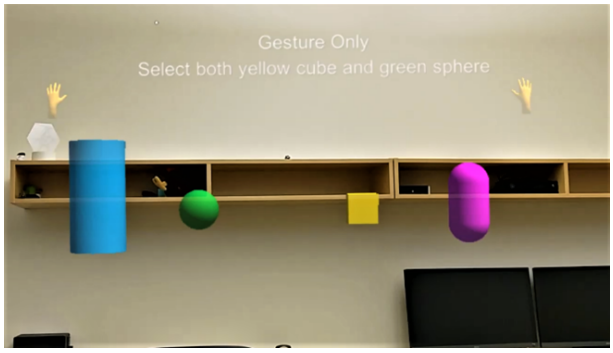


Figure 1: Participant view

Figure 2: Experiment setup

to the text displayed as part of the referent were removed. Resulting in 277 proposals from GS block and 261 proposals from S block.

The agreement rate ($\mathscr{AR}$), co-agreement rate ($\mathscr{CR}$) and (V_{rd}) significance test were used to determine consensus among gesture proposals [22]. $\mathscr{AR}$ is used to quantify consensus of the binned proposals for interaction by referent [30], as shown in Eq. 1. $\mathscr{CR}$ is used to measure the amount of agreement shared between referents [22]. This study adopted Fliess's Kappa coefficient (k_F) and the related chance agreement term (p_e) [21] when presenting the overall agreement rate of gesture proposals. The bootstrapped 95% confidence intervals were calculated to provide an interval estimate of each agreement score [21]. We used the AGATe 2.0 tool (AGreement Analysis Toolkit)[1] to assist our statistical analysis. The consensus-distinct ratio (*CDR*) was adopted to quantify the speech proposals [11]. For a complete treatment on elicitation studies and methods, see Williams et al. [30].

$$\mathscr{AR}_r = \frac{\sum_{P_i \subset P} \frac{1}{2}|P_i|(|P_i|-1)}{\frac{1}{2}|P|(|P|-1)} \quad (1)$$

The agreement rate $\mathscr{AR}$ for each referent r was calculated with Eq. 1. In Eq. 1, P is the set of all proposed gestures for referent r, and P_i are the subsets of identical proposed gestures from P.

The overall agreement rate for gestures from G and GS blocks was .190. Based on the interpretations proposed by Vatavu and Wobbrock [22], our study achieved a medium agreement with 12 referents and a high agreement with 4 referents. The individual agreement rate of gestures from G block and GS block alone were also calculated. The G block has .189 in agreement rate with k_F coefficient of .165. The chance agreement term p_e was .029, which indicates that the probability of agreement occurring by chance was minimal [21]. The GS block obtained .193 agreement rate with

[1] Available at http://depts.washington.edu/acelab/proj/dollar/agate.html

k_F coefficient of .151, and the chance agreement term p_e was .050, which shows evidence of agreement beyond chance. Compared to the previous elicitation study results in the single 3D object environment [26], we have lower agreement rates in general.

4.1 Unimodal Gesture and Unimodal Speech

4.1.1 Gesture Only

We observed a significant effect of referent type on agreement rate in G block ($V_{rd(21,N=528)}$ = 639.363, p <.001). The study found there were 13 referents who obtained a medium to high agreement ($\mathscr{AR} > .10$), which showed significant difference between agreement rates ($V_{rd(12,N=312)}$ = 191.492, p <.001). Accordingly, nine referents have agreement rates below .100, which means they are in low agreement, and no significant difference in agreement rates was found ($V_{rd(8,N=216)}$ = 7.550, p <1.000). The highest agreement rates came from referents Select and Select Both, which are .457. The pointing gesture won the highest agreement rate for Select referent, mostly based on the natural interaction for specifying an object in the real world. The referents Shrink Both and Roll Counter Clockwise (RCC) are also achieved high agreements ($\mathscr{AR}_{ShrinkBoth} = .341$, $\mathscr{AR}_{RCC} = .308$). Among abstract referents, Destroy and Destroy Both got the two lowest agreement rates ($\mathscr{AR}_{Destroy} = .072$, $\mathscr{AR}_{DestroyBoth} = .047$). In rotation referents, Pitch up and Pitch down exhibited the two lowest agreement rates ($\mathscr{AR}_{PitchUp} = .058$, $\mathscr{AR}_{PitchDown} = .062$). For the translation referent, referent Move Up has the lowest agreement rate ($\mathscr{AR}_{MoveUp} = .072$), although Move Down has a much higher agreement rate ($\mathscr{AR}_{MoveDown} = .228$).

A co-agreement analysis for dichotomous referents and one object versus two object referents is shown in Figure 4. The co-agreement rates of one object versus two object referents were in general higher than in dichotomous referents. The referents Select and Select Both achieved a high co-agreement ($\mathscr{AR}_{Select} = .457$, $\mathscr{AR}_{SelectBoth} = .457$, $\mathscr{CR} = .355$), which indicates 78% of all pairs of participants have consistent gesture preference with both referents. Another high co-agreement rate came from Shrink and Shrink Both which showed 76% of all pairs of participants that were in agreement with referent Shrink were also in agreement with gestures for referent Shrink Both ($\mathscr{AR}_{Shrink} = .286$, $\mathscr{AR}_{ShrinkBoth} = .341$, $\mathscr{CR} = .217$).

4.1.2 Speech Only

For speech data, we adopted the binning criterion wherein "enlarge yellow and green" and "enlarge cube and sphere" were equal to "enlarge yellow cube and green sphere". However, "yellow cube pitch down" and "yellow cube rotate down" were counted as different proposals.

Table 1 shows the consensus-distinct ratio (*CDR*) of different categories of referents in the S block. The *CDR* is used to calculate the percent of distinct speech proposals by referent that achieved a consensus threshold of two [11]. The results demonstrated that scale referents have the highest *CDR*, in addition to abstract referents which present a *CDR* that are almost twice high when compared to 24.52% from the previous elicitation study with a single 3D object [26]. The rotation referents hold the lowest *CDR*. Based on the data, a low *CDR* could be caused by different expressions of rotation. For example, "spin" or "rotate" plus gesturing direction was proposed to achieve "roll", "yaw", or "pitch". A similar finding was presented in the previous elicitation study with a single 3D object [26]. There are few alternative phrases for "move up / down / left / right, " which could explain that translation referents have a higher *CDR* than rotation referents. Similarly, less options of replacement for action or status phrases such as "shrink" and "smaller" in scale proposals. Figure 5 presents the syntax formats covered more than 80% of proposals in the S block. It is obvious that $\langle action \rangle$ $\langle object \rangle$ and $\langle action \rangle$ $\langle object \rangle$ $\langle direction \rangle$ are the most common formats for speech proposals. Moreover, rotation and translation referents elicited more variants of syntax, which means that various syntax should be considered while designing unimodal speech commands.

Despite bias from text referents, participants often preferred interaction from left to right with multiple 3D objects. For example, with the referents of "create two objects at the same time", two participants proposed "create green sphere and yellow cube", even though all text referents involving two objects started as "yellow cube and green sphere" in the experiment. It shows that participants favored creating objects starting from the left since the green sphere was placed to the left side of the yellow cube in the scene.

4.2 Multimodal interaction: Speech and Gesture

4.2.1 Gesture in GS

The results additionally demonstrate that the referent type has significant effect on gesture agreement rates in GS block ($V_{rd(21,N=528)}$ = 361.624, p <.001). There were 19 referents who achieved medium to high agreement ($\mathscr{AR} > 0.10$), and presented significant difference between agreement rates ($V_{rd(18,N=456)}$ = 262.325, p <.001). Only 3 referents have low agreement rates ($\mathscr{AR}_{DestroyBoth} = 0.094$, $\mathscr{AR}_{PitchUp} = 0.069$, $\mathscr{AR}_{PitchDown} = 0.087$), and further significant differences among those agreement rates were not found ($V_{rd(2,N=72)}$ = 1.368, p <1.000). The highest agreement rate in the GS block was from referent Select ($\mathscr{AR}_{Select} = 0.42$), and referent Select Both, who was not far behind in rank. ($\mathscr{AR}_{SelectBoth} = 0.395$). As in the G block, referent Shrink also obtained a high agreement rate while combining with speech ($\mathscr{AR}_{Shrink} = 0.308$). Moreover, referents Destroy and Destroy Both showed a similar low agreement as in the G block, compared to other referents ($\mathscr{AR}_{Destroy} = 0.101$, $\mathscr{AR}_{DestroyBoth} = 0.094$). As shown in Figure 3, the two lowest agreement rates in the GS block came from referents Pitch Up and Pitch Down.

In terms of co-agreement, for one object versus two object referents, the average co-agreement rate was 68% without including referent Create Both. This finding indicates that that a high number of participants kept the same preferences for both one object and two object manipulation. The cause of a low co-agreement rate between referent Create and Create Both could be the low agreement rate for Create Both in the GS block ($\mathscr{AR}_{Create} = .192$, $\mathscr{AR}_{CreateBoth} = .120$, $\mathscr{CR} = .036$). Higher co-agreement rates were found for dichotomous referents compared to the G block. As shown in Figure 4, the co-agreement rates of translation referents were increased the most compared to the values in the G block, which indicates multimodal interaction assisted participants achieving more agreement for dichotomous translation referents.

4.2.2 Speech in GS

Figure 6 shows syntax formats covered more than 80% of proposals in the GS block. Compared to the S block, participants have proposed a fair amount of single-word commands with compensation from gestures. These single-word proposals included $\langle action \rangle$ only, $\langle object \rangle$ only, $\langle direction \rangle$ only, and $\langle status \rangle$ only. All proposals with single-word commands account for 39.71% of the total proposals in the GS block. In contrast, the proportion of single-word proposals in the S block were merely 6.48%. The prior study mentioned that part of the $\langle action \rangle$ $\langle object \rangle$ syntax proposed in the S block turned into $\langle action \rangle$ plus gesture proposals in the GS block [26]. In the GS block, the most used syntax format was $\langle action \rangle$ only, shown in all four categories of referents. If the speech does not indicate the target, it can then be assumed that gestures were used for identifying the target object in a multi-object virtual environment. As anticipated, based on the results, 87.5% of proposals with $\langle action \rangle$ only syntax format have involved gestures of"pointing", "tapping", or "grabbing". Furthermore, with $\langle direction \rangle$ only syntax proposals, 84.21% of gestures showed "pointing" or "tapping" to indicate the target object. In contrast, proposals consisting of the $\langle object \rangle$ only syntax format had merely 28.95% of the proposals involving the

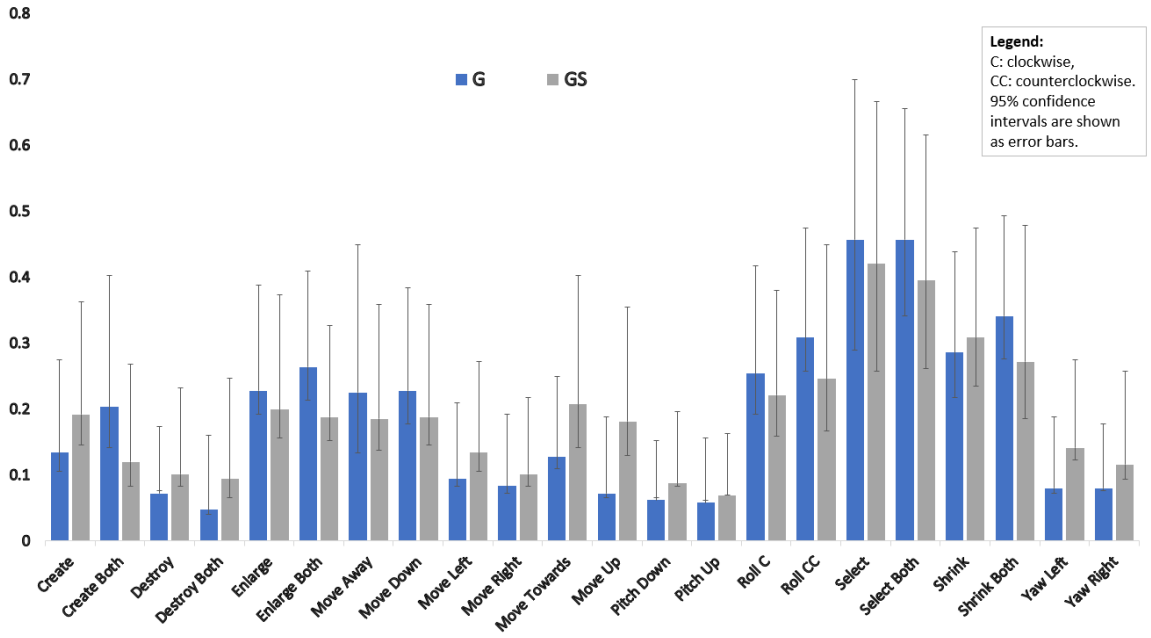


Figure 3: Gestures agreement rates in gesture only (G) block, gesture with speech (GS) block

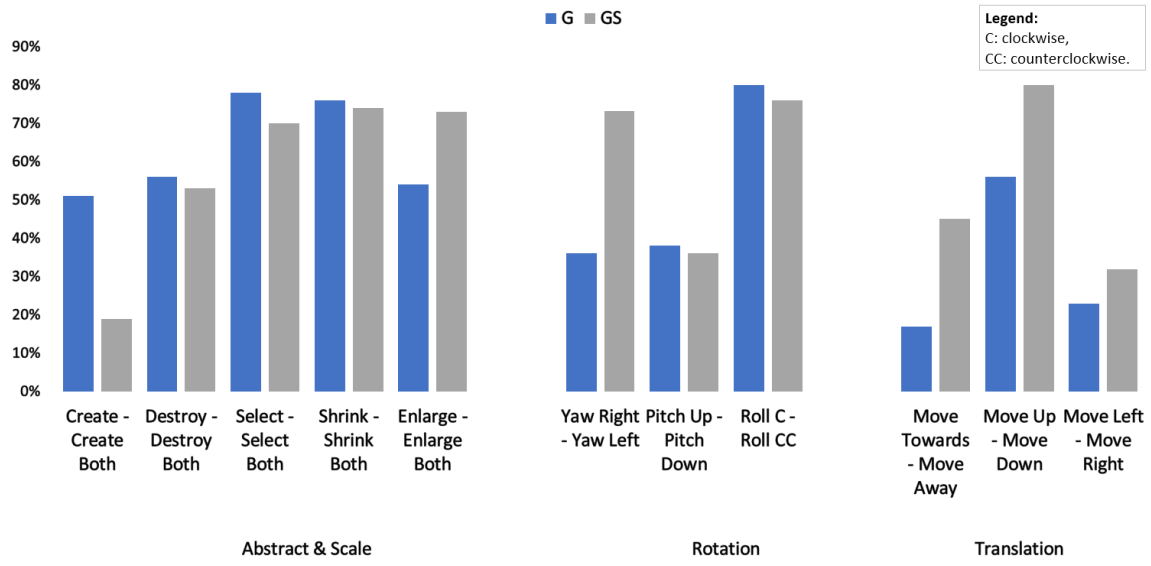


Figure 4: Gestures co-agreement between referents in gesture only (G) block and gesture with speech (GS) block

Table 1: Consensus-distinct ratio (*CDR*) of speech only (S) and Gesture and Speech (GS) block by referent category

Referent Category	Speech Only	Gesture and Speech
Abstract	43.75%	35.21%
Rotation	24.56%	16.44%
Scale	57.14%	32.0%
Translation	42.65%	22.89%

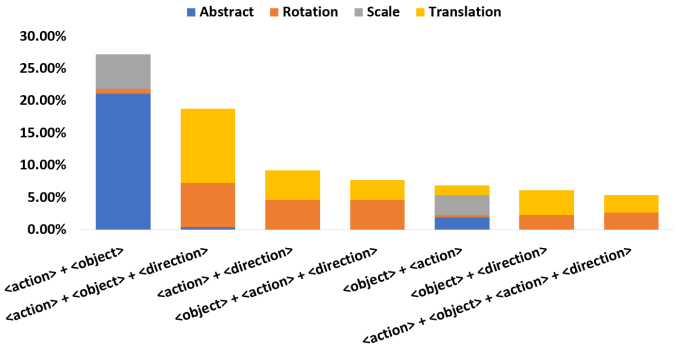


Figure 5: Usage of syntax format by referent type in the speech only (S) block

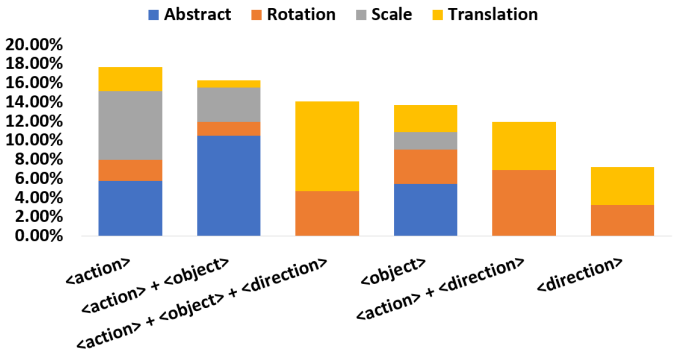


Figure 6: Usage of syntax format by referent type in the gesture and speech (GS) block

gestures "pointing" or "grabbing". This result proved the complementary feature of multimodal interaction. Due to the necessity of identifying the target object in a multi-object environment for manipulation and the flexibility of using speech that multimodal interaction gave participants, less agreement was shown with speech proposals in the GS block compared to in the S block (Table 1).

4.2.3 Gesture and Speech Association

The study looked into the association between the stroke of a gesture proposal and the corresponding speech proposal in the GS block. A stroke is considered the peak of effort for a specific gesture [10], which holds the meaningful content of the gesture. We classified the main speech content into three types of expressions (nominal, deictic, verb) based on prior work from Bourguet and Ando [3]. During the video annotation, recordings were made of the expressions in relation to speech content while the main stroke of the gesture occurred. The study found that strokes for abstract referents were mainly associated with nominal expressions, such as "the yellow cube" or "objects". The referents Destroy and Destroy Both were exceptions, which could be related to the low agreement on gesture proposals. All scale strokes were more synchronized with verb expressions, mostly "enlarge" and "shrink". It should be noted that there were much fewer deictic expressions used in the scale speech proposals, which indicates the limitation of associated expressions. In terms of the translation and rotation referents, 9 out of 12 showed a strong association between strokes and deictic expressions. For example, participants would execute the stoke of pitch up while saying "up". The Move Away and Yaw Right referents were slightly more synchronized with verb expressions, and the stroke of roll counter-clockwise showed more association with nominal expressions.

5 Discussion and Design Guidelines

The results support our hypothesis H_2 that the consensus set of gestures indicates that participants preferred to use both hands for two virtual object manipulation. Chi-square analysis showed that the difference between one hand and two hands adoption in the two virtual environments was statistically significant ($X^2 = 255.33$, $p < .001$). This means multi-object environment increases the usage of both hands. The results support H_1 for some tasks, because there were 11 out of 17 single object manipulation referents resulting with different winning gestures, compared to ones from a single object virtual environment. In general, there are similarity and dissimilarity in multimodal interaction with an multi-object virtual environment and single object virtual environment. The multi-object environment has inspired more physical interaction which came from experience with real world.

Based on our resulting top three proposed variants of each referent, among 17 single object manipulation tasks, six referents have the same winning gestures as prior findings in a single object virtual environment [16, 26]. Another five referents' second place proposals were identical to previous results in a single object virtual environment [16, 26]. Within the six referents that have the same winning gestures in both virtual environments, two of them are Shrink and Enlarge from scale referents, three translation referents are Move Toward, Move Away, and Move Left, and one rotation referent is Yaw Left. Legacy bias is an issue in elicitation study that uses' gesture proposals are biased due to the previous experience with exist interfaces [12]. The legacy bias from interaction with a multi-touch screen could contribute to the identical scaling proposals. The "screwing in a light bulb" gesture for rotating around the Z axis was also found in Williams et al., and Pham et al.'s works [16, 26]. Unsurprisingly, abstract referents have less similarity on their proposals. The winning gesture for creating an object in this work used gathered and then spread fingertips as the original blooming gesture from HoloLens 1 [20], but with the palm facing forward instead of facing up. Due to the difference, we do not consider that our blooming gesture came from legacy bias and more likely was a spontaneous proposal that could inspire future gesture designing for AR interaction.

In terms of speech proposals, our results showed more variety in syntax formats. We have two more single-word syntax formats in GS block compared to ones found in Williams et al. [29], and they were $\langle object \rangle$ only and $\langle status \rangle$ only. For speech only interaction, our study presented $\langle action \rangle$ + $\langle object \rangle$ as the top rank syntax format, compared to the $\langle action \rangle$ only syntax format which has a similar proportion in a single object environment [29]. We believe this result was due to the multiple object environment in our study, and participants tended to specify the target object for interaction.

The results of the study found that participants preferred symmetric bimanual versions of the single-handed gesture for two object manipulation. For example, the winning proposal for shrinking a single object was the zoom in gesture, and the winning gesture for shrinking two objects side by side was to perform zoom in with both hands simultaneously. This result of symmetric bimanual interaction is reasonable since both targets were inside the participant's field of view, which made symmetric action easy to perform [2]. The exception of destroying proposals could be related to the low agreement rate for both destroy referents, which indicates people have less common sense for destroying from reality-based interaction [7]. According to the answers in the post-study questionnaire, 13 out of 24 participants expressed that it was fairly natural to think of using both hands for two object manipulation. Five participants indicated it was harder to develop the proposal for two object interaction compared to the single object manipulation. One participant said that the single hand gesture could be used to replace two hand interaction as needed. Our findings could be used to develop gesture recognizers for a multi-object virtual environment by sensing the user's intent

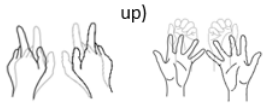
Referent	Category	Winning Gesture	Second Place	Third Place	Coverage
Create	Abstract	bloom gesture (palm face forward) / point finger 	tap finger 	zoom out 	77.09%
Create Both	Abstract	bloom gesture (palm face forward) 	point finger 	tap finger/bloom gesture (palm face up) 	72.34%
Destroy Both	Abstract	squish with fist 	point finger 	zoom in 	43.75%
Destroy	Abstract	point finger 	squish with fist / toss hand forward 	swipe finger to corner 	56.25%
Select	Abstract	point finger 	tap finger 	open hand (palm face forward) 	93.75%
Select Both	Abstract	point finger 	tap finger 	open hand (palm face forward) 	85.41%

Figure 7: Top three proposed gesture variants by abstract referent

Referent	Category	Winning Gesture	Second Place	Third Place	Coverage
Enlarge Both	Scale	zoom out 	bloom gesture (palm face forward) 	bloom gesture (palm face forward) while move hand backward 	79.16%
Enlarge	Scale	zoom out 	bloom gesture (palm face forward) 	extend hands distance diagonally (open hand or pinch gesture) 	81.25%
Shrink Both	Scale	zoom in 	gather all fingertips 	reduce hands distance diagonally (open hand or pinch gesture) 	89.59%
Shrink	Scale	zoom in 	gather all fingertips 	reduce hands distance diagonally (open hand or pinch gesture) 	87.50%

Figure 8: Top three proposed gesture variants by scale referent

Referent	Category	Winning Gesture	Second Place	Third Place	Coverage
Move Towards	Translation	pull open hand	pull pinch gesture	pull finger(s)	58.33%
Move Away	Translation	push open hand	push finger(s)	point finger	64.59%
Move Down	Translation	swipe finger(s) down	push open hand down	grasp and move hand down	74.99%
Move Up	Translation	swipe finger(s) up	push open hand up	grasp and move hand up	58.33%
Move Left	Translation	pinch and move to top then left	point and move to top then left	grasp and move top then left	58.34%
Move Right	Translation	point and move to top then right	grasp and move top then right	pinch and move top then right	52.09%

Figure 9: Top three proposed gesture variants by translation referent

Referent	Category	Winning Gesture	Second Place	Third Place	Coverage
Yaw Left	Rotation	swipe finger or open hand to left / grasp and rotate hand around Z	flick finger to left	flick hand to left	70.83%
Yaw Right	Rotation	swipe open hand to right	grasp and rotate hand around Z / flick finger to right	swipe finger to right	58.33%
Pitch Down	Rotation	flick hand down	pinch and move hand down / flick finger down	swipe finger(s) down/ grasp and move down / pinch and rotate fingers around Y	72.93%
Pitch Up	Rotation	flick finger up	pinch and move hand up	grasping hand move up	41.67%
Roll Clockwise	Rotation	grasp and rotate hand around X	pinch and rotate fingers around X	rotating finger(s) around X	83.33%
Roll Counterclockwise	Rotation	grasp and rotate hand around X	point / pinch and rotate around X	rotating open hand around X	91.66%

Figure 10: Top three proposed gesture variants by rotation referent

based on the hands involved.

Speech recognition with an AR headset is difficult due to the environment noise, unintended commands, and sometimes the accent of the user. With the knowledge of the association between speech expressions and gesture stroke, a more specific hypothesis can be implemented in the recognition system to improve speech detection efficiency and accuracy in AR. While the previous study only focused on pointing gestures [3], our study discovered the association between common manipulation gestures and speech commands for interaction in a multi-object virtual environment.

Design Guidelines – Based on the user-defined gesture sets from our study and literature, while some gestures and speech syntax formats remain similar, there were differences in multimodal interaction between a single object and a multi-object virtual environments. Participants' proposals in our study showed more physical interactions such as pinching or grasping the target object and "turning a doorknob" for rotation tasks. Similar to prior findings suggested to include aliasing for gestures and speech [11, 26, 33], we propose that including aliasing could significantly improve the performance of the recognizer. For example, using the commands "spin" or "rotate" plus gesture indicates direction should be equal to use commands "roll/yaw/pitch". With gestures, performing pinching or grasping then moving the hand for virtual object translation should be equivalent to pointing at the target then moving the finger. Our results indicate that implementing the top three proposed variants (Figure 7, Figure 8, Figure 9, Figure 10) of a gesture could increase the coverage of proposed gestures to 70% on average. The variety of syntax formats in the GS block indicates that various combinations of speech and gesture could be designed for interaction in an augmented reality environment. Moreover, as Williams and Ortega mentioned in their work, legacy bias could be a benefit to new technology because it is memorable and discoverable [28]. We suggest that emerging technology such as AR-HMD should consider both legacy bias from the touchscreen and physical interaction based on body awareness and environmental skills [7].

6 Limitation and Future Work

The text referents could bias participants' speech proposals in our experiment. We also know that using animation as referents would bias the gesture proposal in the elicitation study [9, 26]. It is still a research question that how to eliminate the bias from referent presentation. Our experiment design requires participants to give both speech and gesture in GS block, which could end with the unnatural speech proposals from participants. Therefore, we will use a more efficient but flexible way to elicit proposals from participants in our future elicitation study. For example, we could adopt the "before" and "after" approach to present the desired effect of a referent for our future study [16, 19]. Reducing the fatigue caused by mid-air interactions is another necessary vein of future work. One way to mitigate this issue is to use other modalities such as eye-gazing combined with speech to replace mid-air gestures. Another option for reducing fatigue could be developing microgestures that require less psychical effort than mid-air gestures.

7 Conclusion

This study investigated multimodal interaction in a multi-object virtual environment. We chose 22 referents for the elicitation study that included canonical referents for scale, translation, and rotation tasks and three abstract referents. We generated a consensus set of gestures for interaction in a multi-object virtual environment and found that participants used the same gesture for one and two objects but with both hands for two object manipulation. The results further demonstrated that participants tended to act on the target objects in a multi-object virtual environment, indicating more physical interaction where preferred. Further, in the study, more speech syntax formats were proposed in multimodal interaction in a multi-object virtual environment. We discovered the association between expressions and stroke, which can improve the accuracy and efficiency of the recognition system. We also provided design guidelines based on our findings and comparison with prior works in a simple virtual environment.

Acknowledgments

The authors wish to thank A, B, and C. This work was supported in part by a grant from XYZ.

References

[1] M. Z. Baig and M. Kavakli. Qualitative analysis of a multimodal interface system using speech/gesture. In *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 2811–2816. IEEE, IEEE, Wuhan, China, 2018.

[2] R. Balakrishnan and K. Hinckley. Symmetric bimanual interaction. *GROUP ACM SIGCHI Int. Conf. Support. Group Work*, 2000.

[3] M.-L. Bourguet and A. Ando. Synchronization of speech and hand gestures during multimodal human-computer interaction. In *CHI 98 Conference Summary on Human Factors in Computing Systems*, pp. 241–242. ACM, Apr. 1998.

[4] A. Cohé and M. Hachet. Understanding user gestures for manipulating 3d objects from touchscreen inputs. In *Proceedings of Graphics Interface 2012*, GI '12, pp. 157–164. Canadian Information Processing Society, Toronto, Ont., Canada, Canada, 2012.

[5] A. Danielescu and D. Piorkowski. Iterative design of gestures during elicitation: Understanding the role of increased production, Jan 2022.

[6] N. Dezfuli, M. Khalilbeigi, M. Mühlhäuser, and D. Geerts. A study on interpersonal relationships for social interactive television. In *Proceedings of the 9th European Conference on Interactive TV and Video*, EuroITV '11, p. 21–24. Association for Computing Machinery, New York, NY, USA, 2011. doi: 10.1145/2000119.2000123

[7] R. J. Jacob, A. Girouard, L. M. Hirshfield, M. S. Horn, O. Shaer, E. T. Solovey, and J. Zigelbaum. Reality-based interaction: A framework for post-wimp interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, p. 201–210. Association for Computing Machinery, New York, NY, USA, 2008. doi: 10.1145/1357054.1357089

[8] A. A. Karpov and R. M. Yusupov. Multimodal interfaces of Human–Computer interaction. *Her. Russ. Acad. Sci.*, 88(1):67–74, Jan. 2018.

[9] S. Khan and B. Tunçer. Gesture and speech elicitation for 3d cad modeling in conceptual design. *Automation in Construction*, 106:102847, 2019.

[10] D. Mcneill. *Gesture and Thought*. the University of Chicago Press, USA, 01 2005. doi: 10.7208/chicago/9780226514642.001.0001

[11] M. R. Morris. Web on the wall: Insights from a multimodal interaction elicitation study. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces*, ITS '12, pp. 95–104. ACM, New York, NY, USA, 2012. doi: 10.1145/2396636.2396651

[12] M. R. Morris, A. Danielescu, S. Drucker, D. Fisher, B. Lee, M. c. Schraefel, and J. O. Wobbrock. Reducing legacy bias in gesture elicitation studies. *Interactions*, 21(3):40–45, May 2014.

[13] M. R. Morris, J. O. Wobbrock, and A. D. Wilson. Understanding users' preferences for surface gestures. In *Proceedings of graphics interface 2010*, pp. 261–268. Canadian Information Processing Society, 2010.

[14] M. A. Nacenta, Y. Kamber, Y. Qiang, and P. O. Kristensson. Memorability of pre-designed and user-defined gesture sets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, p. 1099–1108. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2470654.2466142

[15] M. Nielsen, M. Störring, T. B. Moeslund, and E. Granum. A procedure for developing intuitive and ergonomic gesture interfaces for hci. In A. Camurri and G. Volpe, eds., *Gesture-Based Communication in Human-Computer Interaction*, pp. 409–420. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[16] T. Pham, J. Vermeulen, A. Tang, and L. MacDonald Vermeulen. Scale impacts elicited gestures for manipulating holograms: Implications for AR gesture design. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pp. 227–240. ACM, June 2018.

[17] T. Piumsomboon, A. Clark, M. Billinghurst, and A. Cockburn. User-defined gestures for augmented reality. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '13, p. 955–960. Association for Computing Machinery, New York, NY, USA, 2013. doi: 10.1145/2468356.2468527

[18] J. Ruiz, Y. Li, and E. Lank. User-defined motion gestures for mobile interaction. CHI '11, p. 197–206. Association for Computing Machinery, New York, NY, USA, 2011. doi: 10.1145/1978942.1978971

[19] T. Seyed, C. Burns, M. Costa Sousa, F. Maurer, and A. Tang. Eliciting usable gestures for multi-display environments. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces*, ITS '12, p. 41–50. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2396636.2396643

[20] S. K. TANG and D. Coulter. Start gesture - mixed reality, 2022.

[21] T. Tsandilas. Fallacies of agreement: A critical review of consensus assessment methods for gesture elicitation. *ACM Trans. Comput. Hum. Interact.*, 25(3):18, June 2018.

[22] R.-D. Vatavu and J. O. Wobbrock. Formalizing agreement analysis for elicitation studies: New measures, significance test, and toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, p. 1325–1334. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2702123.2702223

[23] R.-D. Vatavu and J. O. Wobbrock. Clarifying agreement calculations and analysis for end-user elicitation studies. *ACM Trans. Comput.-Hum. Interact.*, 29(1), jan 2022. doi: 10.1145/3476101

[24] S. Villarreal-Narvaez, J. Vanderdonckt, R.-D. Vatavu, and J. A. Wobbrock. A systematic review of gesture elicitation studies: What can we learn from 216 studies. In *Proceedings of ACM Int. Conf. on Designing Interactive Systems (DIS'20)*, p. NA. ACM Press, Eindhoven, 2020.

[25] P. Vogiatzidakis and P. Koutsabasis. 'address and command': Two-handed mid-air interactions with multiple home devices. *International Journal of Human-Computer Studies*, 159:102755, 2022. doi: 10.1016/j.ijhcs.2021.102755

[26] A. S. Williams, J. Garcia, and F. Ortega. Understanding multimodal user gesture and speech behavior for object manipulation in augmented reality using elicitation. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3479–3489, 2020. doi: 10.1109/TVCG.2020.3023566

[27] A. S. Williams and F. Ortega. Insights on visual aid and study design for gesture interaction in limited sensor range augmented reality devices. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 19–22, 2020. doi: 10.1109/VRW50115.2020.00286

[28] A. S. Williams and F. R. Ortega. Evolutionary gestures: When a gesture is not quite legacy biased. *Interactions*, 27(5):50–53, sep 2020. doi: 10.1145/3412499

[29] A. S. Williams and F. R. Ortega. Understanding gesture and speech multimodal interactions for manipulation tasks in augmented reality using unconstrained elicitation. *Proc. ACM Hum.-Comput. Interact.*, 4(ISS), nov 2020. doi: 10.1145/3427330

[30] A. S. Williams and F. R. Ortega. A concise guide to elicitation methodology, May 2021.

[31] M. L. Wittorf and M. R. Jakobsen. Eliciting Mid-Air gestures for Wall-Display interaction. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, NordiCHI '16, pp. 3:1–3:4. ACM, New York, NY, USA, 2016.

[32] J. O. Wobbrock, H. H. Aung, B. Rothrock, and B. A. Myers. Maximizing the guessability of symbolic input. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, p. 1869–1872. Association for Computing Machinery, New York, NY, USA, 2005. doi: 10.1145/1056808.1057043

[33] J. O. Wobbrock, M. R. Morris, and A. D. Wilson. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1083–1092. ACM, New York, NY, USA, 2009.