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

Xiaoyan Zhou
Colorado State University
Fort Collins, Colorado, USA
Xiaoyan.Zhou@colostate.edu

Adam S. Williams
Colorado State University
Fort Collins, Colorado, USA
adam.sinclair.williams@colostate.edu

Francisco R. Ortega
Colorado State University
Fort Collins, Colorado, USA
fortega@colostate.edu

ABSTRACT

As augmented reality (AR) technology and hardware become more mature and affordable, researchers have been exploring more intuitive and discoverable interaction techniques for immersive environments. This paper investigates multimodal interaction for 3D object manipulation in a multi-object AR environment. To identify the user-defined gestures, we conducted an elicitation study involving 24 participants and 22 referents using 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 present the gestures and speech results, and the differences compared to similar studies in a single object AR 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.

CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Mixed / augmented reality**; **Interaction techniques**; **Empirical studies in HCI**.

KEYWORDS

elicitation, multimodal interaction, augmented reality, gesture and speech interaction, multi-object AR environment

ACM Reference Format:
Xiaoyan Zhou, Adam S. Williams, and Francisco R. Ortega. 2022. Eliciting Multimodal Gesture+Speech Interactions in a Multi-Object Augmented Reality Environment. In *28th ACM Symposium on Virtual Reality Software and Technology (VRST '22), November 29-December 1, 2022, Tsukuba, Japan.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3562939.3565637

1 INTRODUCTION

Easy-to-remember gestures produce high usability interfaces [16]. 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 [36], facilitating intuitive gesture design without implementing perfect interaction recognizers in advance. Prior findings indicated that users prefer to choose input modalities based on their needs during the interaction [1, 8, 12]. Previous elicitation studies mostly focused on gesture set design for different devices and interfaces [4, 12, 20, 28, 36], and several studies have explored multimodal interactions with speech and hand gestures [10, 29, 32]. However, few to no researches have involved multimodal interactions in a multi-object augmented reality (AR) or virtual reality (VR) environment.

These prior works raised multiple questions that were explored during this study: Does multimodal interaction look different in multi-object AR environments? How does a multi-object AR environment impact an elicitation study's gesture and speech proposals? What gestures do users prefer for multiple object manipulation, and are there any differences from single object manipulation? What speech commands do users choose for multiple object manipulation, and what are the differences compared to single object manipulation? The raised questions drive our motivation to understand if previous single object studies can transfer to more realistic environments. For this work, an elicitation study was conducted for multimodal interaction in AR with a Wizard of Oz (WoZ) experiment design (i.e., a researcher emulating a live system) [26, 36]. It involved 24 participants completing 22 referents (i.e., command) each in an AR head-mounted display (HMD). It yielded 528 proposals, which were used to generate a winning gesture set with 25 gestures after binning and ranking. We compared our single virtual object manipulation proposals to the findings from prior studies in a single object AR environment [18, 29, 32]. 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 [36], which aims to assist in designing more intuitive [36], guessable [35], learnable, and memorable [15] 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 [14]. Based on recent literature review results [27], over two hundred studies have adopted the use of an elicitation methodology

VRST '22, November 29-December 1, 2022, Tsukuba, Japan

ACM ISBN 978-1-4503-9889-3/22/11…$15.00
https://doi.org/10.1145/3562939.3565637

in their work. Elicitation studies can provide valuable insights for developing new interaction techniques.

Prior findings proved that users prefer to choose input modalities based on their needs during the interaction, such as choosing gestures over speech in a quiet environment [1, 8, 12]. Wobbrock et al. [36] 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 [12], which can offer users different modalities to achieve the same effect and increase the acceptance of new technology.

Nevertheless, although the realistic scenarios in AR interaction include more than one virtual object, most elicitation studies involving mid-air gestures in AR only considered single object manipulation in a single object AR environment [18, 19, 29, 32]. Pham et al. conducted an elicitation study with an AR headset that included a scenario of single building manipulation among multiple buildings [18]. 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 on 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, yet the elicited gestures were surface gestures [19]. Wittorf et al. adopted an elicitation methodology for exploring mid-air gestures, yet the participants were interacting with a wall display [34]. Danielescu and Piorkowski conducted an elicitation study to explore free-space gestures with a projector display that included multiple target selections among a set of photos [5]. However, the referent showed that photos were selected one by one, which could bias the participants' gesture proposals. Furthermore, Wobbrock et al. found that users preferred one-handed over two-handed interactions for tabletop interaction [17, 36]. As a result, we were interested in whether users preferred two hands for two-object manipulation in an AR environment. To fill the gap in multimodal interaction in a multi-object AR environment, this work conducted a study to elicit speech and mid-air gestures in an augmented reality environment that contains four virtual objects (Figure 1) and compared results to elicited proposals with a single-object AR environment.

3 STUDY DESIGN

This study conducted the elicitation experiment using a similar process as previous work [24, 26, 36]. Twenty-two tasks (i.e., referents) were used for each modality in this work. Of those, 17 basic referents were selected based on their inclusion in prior works [29, 32], while the other five were developed to be multi-object versions of basic referents. Specifically, participants were required to manipulate two objects simultaneously when multi-object referents were presented. Referents included six translations (along x, y, and z axes), six rotations (about x, y, and z axes), three abstract actions (create, destroy, and select), and two scaling actions (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.

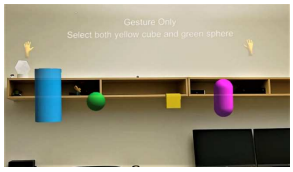


Figure 1: Participant view

Figure 2: Experiment setup

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 sensing 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 [30], as shown in Figure 1. If either hand was 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 could use any interaction they felt was appropriate to execute the command based on the presented text referent and input modality. Participants were told to perform gestures inside the headset's view, which they could tell by the hand icons displayed. 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 [29]: H_1) for the same single object manipulation referent, winning gestures in a multi-object AR environment will be different from ones in a single object AR environment; H_2) participants would prefer to use both hands for two object manipulation referents. Moreover, through elicited multimodal interaction, further connections between speech commands and hand gestures are expected to be found in this study.

4 RESULTS

With the experiment, 528 proposals were collected from each modality. To eliminate the effect of the referent text biasing the speech

proposal [29], prior to the analysis, speech proposals that were identical 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 ($\mathcal{AR}$), co-agreement rate ($\mathcal{CR}$) and (V_{rd}) significance test were used to determine consensus among gesture proposals [25]. $\mathcal{AR}$ is used to quantify the consensus of the binned proposals for interactions by referent [33], as shown in Eq. 1. $\mathcal{CR}$ is used to measure the amount of agreement shared between referents [25]. This study adopted Fliess's Kappa coefficient (k_F) and the related chance agreement term (p_e) [24] 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 [24]. 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 [12]. For a complete treatment on elicitation studies and methods, see Williams et al. [33].

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

The agreement rate $\mathcal{AR}$ 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 [25], **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 [24]. The GS block obtained .193 agreement rate with 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 [29], 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 ($\mathcal{AR}$ > .10), and they showed significant difference between agreement rates ($V_{rd(12,N=312)}$ = 191.492, p <.001). Accordingly, 9 referents have agreement rates below 0.10, 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 (Figure 3). **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.** As shown in Figure 3, the referents Shrink Both and Roll Counter Clockwise (RCC) are also achieved high agreements. Among abstract referents, Destroy and Destroy Both got the two lowest agreement rates. In rotation referents, Pitch up and Pitch down exhibited the two lowest agreement rates. For the translation referent, referent Move Up has the lowest agreement rate ($\mathcal{AR}_{MoveUp}$ = .072), yet Move Down has a much higher agreement rate ($\mathcal{AR}_{MoveDown}$ = .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 ($\mathcal{AR}_{Select}$ = .457, $\mathcal{AR}_{SelectBoth}$ = .457, $\mathcal{CR}$ = .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 ($\mathcal{AR}_{Shrink}$ = .286, $\mathcal{AR}_{ShrinkBoth}$ = .341, $\mathcal{CR}$ = .217).

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

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

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 [12]. 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 [29]. 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 [29]. 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 that 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

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

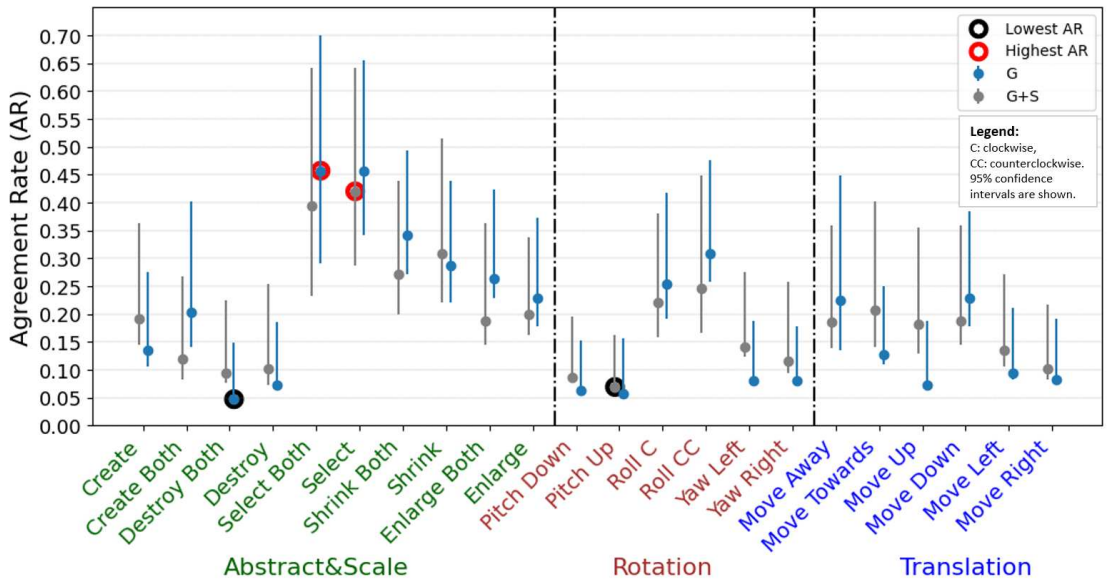


Figure 3: Agreement rates for the gesture (G) and gesture+speech (G+S) conditions grouped by referent category

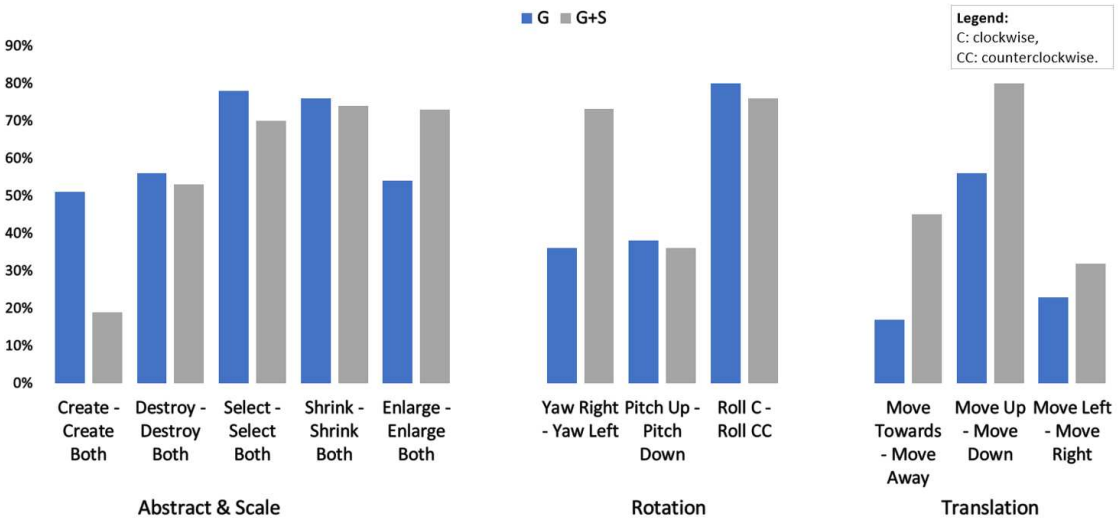


Figure 4: Gestures co-agreement between referents in gesture only (G) block and gesture+speech (G+S) block

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 ($\mathcal{AR}$ > 0.10), and presented significant difference between agreement rates ($V_{rd(18,N=456)}$ = 262.325, p <.001). Only referents Destroy Both, Pitch Up, and Pitch Down have low agreement rates that less than 0.10 (Figure 3), and further significant differences among those agreement rates were not found ($V_{rd(2,N=72)}$ = 1.368, p <1.000). As shown in Figure 3, the highest agreement rate in the GS block was from referent Select, and referent Select Both was not far behind in rank. The lowest agreement rates in the GS block came from referents

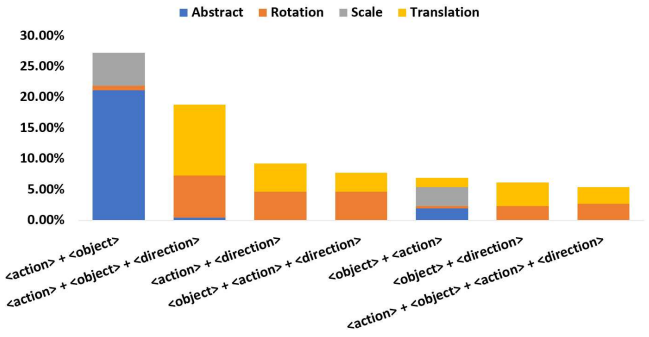


Figure 5: Usage of syntax format by referent type in the speech only (S) block

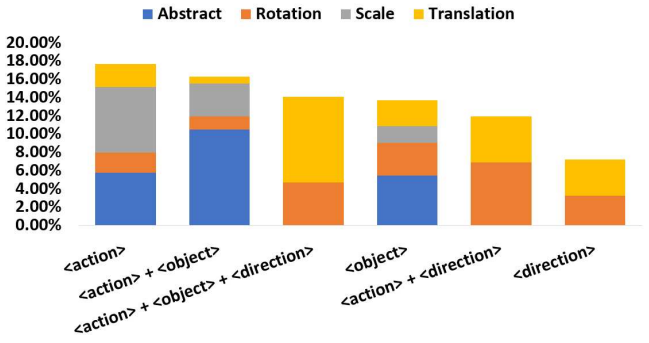


Figure 6: Usage of syntax format by referent type in the gesture and speech (GS) block

Pitch Up, and Pitch Down was very close by. As in the G block, referent Shrink also obtained a high agreement rate while combining with speech ($\mathcal{AR}_{Shrink}$ = 0.308). Moreover, referents Destroy and Destroy Both showed a similar low agreement as in the G block, compared to mostly other referents ($\mathcal{AR}_{Destroy}$ = 0.101, $\mathcal{AR}_{DestroyBoth}$ = 0.094).

Regarding 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 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 ($\mathcal{AR}_{Create}$ = .192, $\mathcal{AR}_{CreateBoth}$ = .120, $\mathcal{CR}$ = .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 in 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 ⟨*action*⟩ only, ⟨*object*⟩ only, ⟨*direction*⟩ only, and ⟨*status*⟩ 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 ⟨*action*⟩ ⟨*object*⟩ syntax proposed in the S block turned into ⟨*action*⟩ plus gesture proposals in the GS block [29]. In the GS block, the most used syntax format was ⟨*action*⟩ 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 AR environment. As anticipated, based on the results, 87.5% of proposals with ⟨*action*⟩ only syntax format have involved gestures of "pointing", "tapping", or "grabbing". Furthermore, with ⟨*direction*⟩ only syntax proposals, 84.21% of gestures showed "pointing" or "tapping" to indicate the target object. In contrast, proposals consisting of the ⟨*object*⟩ only syntax format had merely 28.95% of the proposals involving the 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 [11], 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 far 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 counterclockwise showed more association with nominal expressions.

5 DISCUSSION

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 AR environments was statistically significant (X^2 = 255.33, p <.001). **This means that using a multi-object environment increased the usage of both hands.** The results support H_1 for some tasks, because there were 11 out of 17 single object manipulation referents

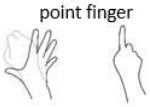
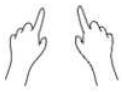
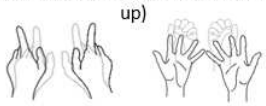
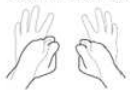
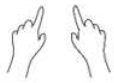
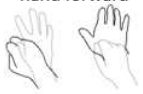
Referent	Category	Winning Gesture	Second Place	Third Place	Coverage
Create	Abstract	bloom gesture (palm face forward) / point finger 	tap finger 	zoom out 	77.09%
Create Both	Abstract	bloom gesture (palm face forward) 	point finger 	tap finger/bloom gesture (palm face up) 	72.34%
Destroy Both	Abstract	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 for abstract referent

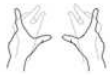
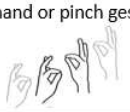
Referent	Category	Winning Gesture	Second Place	Third Place	Coverage
Enlarge Both	Scale	zoom out 	bloom gesture (palm face forward) 	bloom gesture (palm face forward) while move hand backward 	79.16%
Enlarge	Scale	zoom out 	bloom gesture (palm face forward) 	extend hands distance diagonally (open hand 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 for 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 for translation referent

resulting in different winning gestures, compared to ones from a single object AR environment. In general, there are similarities and dissimilarities in multimodal interaction within a multi-object AR environment and a single-object AR environment. **The multi-object environment has inspired more physical interaction which came from experience with the real world.**

Based on our resulting top three proposal variants for each referent, among 17 single object manipulation tasks, six referents have the same winning gestures as prior findings in a single object AR environment [18, 29]. Another five referents' second place proposals were identical to previous results in a single object AR environment [18, 29]. Within the six referents that have the same winning gestures in both AR 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 existing interfaces [13]. 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 [18, 29]. Unsurprisingly, abstract referents have less similarity in 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 [23], 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.

Besides using the agreement rate to measure the consensus among gesture proposals and facilitate gesture sets generation, other criteria such as reversibility can also be considered validation after gesture sets are generated [21]. As shown in the gesture figures (Figure 7, Figure 8, Figure 9, Figure 10), all winning gestures for opposite scale referents and rotation referents met the reversibility criteria, then winning gestures for opposite translation referents were mostly reversible except for Move Left and Move Right. The winning gestures for opposite abstract referents were not fully reversible, which might be due to the difficulty of suggesting gestures for abstract interaction. Nevertheless, winning gestures for Create Both and Destroy Both perfectly met the reversibility criteria.

This paper has eliminated the presentations of variations of similar hand poses, which could be due to the physical features or object sizes or individual differences between users [37]. Through the binning procedure, most interchangeable variants of a hand pose with the same movement path are grouped into the same gesture. For example, pinching with two and three fingers is all grouped into the same pose. And if they all move clockwise, they are binned into the same gesture class.

In terms of speech proposals, our results showed more variety in syntax formats. We have two more single-word syntax formats

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	point 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 for rotation referent

in GS block compared to the ones found in Williams et al. [32], and they were ⟨*object*⟩ only and ⟨*status*⟩ only. For speech-only interaction, our study presented ⟨*action*⟩ + ⟨*object*⟩ as the top rank syntax format, compared to the ⟨*action*⟩ only syntax format which has a similar proportion in a single object environment [32]. 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 actions 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 grounding 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 AR environment by sensing the user's intent 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 previous work only focused on pointing gestures [3], our study discovered the association between common manipulation gestures and speech commands for interaction in a multi-object AR environment.

6 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 AR 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 including aliasing for gestures and speech [12, 29, 36], we propose that including aliasing could significantly improve the performance of the recognizer. For example, using the commands "spin" or "rotate" plus a gesture that indicates direction should be equal to using commands "roll/yaw/pitch". With gestures, performing pinching or grasping and then moving the hand for virtual object translation should be

equivalent to pointing at the target and 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. For example, to perform an interaction with a virtual object such as "move the clock to left", the system provides options that the user can either say "clock" plus move a finger to the left, or use a finger to point at the clock plus say "(go) left". Moreover, as Williams and Ortega mentioned in their work, legacy bias could be a benefit to new technology because it is memorable and discoverable [31]. 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].

7 LIMITATIONS AND FUTURE WORK

The text referents could bias participants' speech proposals in our experiment. We also know that using animations as referents would bias the gesture proposal in the elicitation study [9, 29]. 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 unnatural speech proposals from participants. Therefore, we will use a more efficient but flexible way to elicit proposals from participants in our future elicitation studies. For example, we could adopt the "before" and "after" approach to present the desired effect of a referent for our future study [18, 22]. 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.

This study has focused on interaction in an AR environment, still, the elicited speech proposals can also be used for interaction in a VR environment due to the more occlusive VR environment. In terms of hand gesture proposals, further evaluation and comparison studies are necessary since gesture usability can be affected by the field of view of the HMD and the clutter of the scene.

This work used two object manipulation as multi-object interaction cases. Future works are expected to verify the current elicited proposals with referents that involve three or more object manipulation. Also, further studies can answer the questions of how more virtual object manipulation influences the speech and hand gesture proposals for interaction in AR.

8 CONCLUSION

This study investigated multimodal interaction in a multi-object AR 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 AR 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 AR environment, indicating more physical interaction where preferred. Further, in the study, more speech syntax formats were proposed in multimodal interaction in a multi-object AR 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 AR environment.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (NSF) awards NSF IIS-1948254, DARPA HR00112110011, NSF 2106590, NSF 2016714, NSF 2037417, and NSF 1948254. We would like to acknowledge Brandon Kelly for his hand illustrations. We also like to acknowledge Lauren Mangus and Lexi Sanchez for their help with grammar corrections.

REFERENCES

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

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

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

[4] Aurélie Cohé and Martin Hachet. 2012. Understanding User Gestures for Manipulating 3D Objects from Touchscreen Inputs. In *Proceedings of Graphics Interface 2012* (Toronto, Ontario, Canada) *(GI '12)*. Canadian Information Processing Society, Toronto, Ont., Canada, Canada, 157–164. http://dl.acm.org/citation.cfm?id=2305276.2305303

[5] Andreea Danielescu and David Piorkowski. 2022. Iterative design of gestures during elicitation: Understanding the role of increased production. https://arxiv.org/abs/2104.04685

[6] Niloofar Dezfuli, Mohammadreza Khalilbeigi, Max Mühlhäuser, and David Geerts. 2011. A Study on Interpersonal Relationships for Social Interactive Television. In *Proceedings of the 9th European Conference on Interactive TV and Video* (Lisbon, Portugal) *(EuroITV '11)*. Association for Computing Machinery, New York, NY, USA, 21–24. https://doi.org/10.1145/2000119.2000123

[7] Robert J.K. Jacob, Audrey Girouard, Leanne M. Hirshfield, Michael S. Horn, Orit Shaer, Erin Treacy Solovey, and Jamie Zigelbaum. 2008. Reality-Based Interaction: A Framework for Post-WIMP Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 201–210. https://doi.org/10.1145/1357054.1357089

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

[9] Sumbul Khan and Bige Tunçer. 2019. Gesture and speech elicitation for 3D CAD modeling in conceptual design. *Automation in Construction* 106 (2019), 102847.

[10] Minkyung Lee and Mark Billinghurst. 2008. A Wizard of Oz Study for an AR Multimodal Interface. In *Proceedings of the 10th International Conference on Multimodal Interfaces* (Chania, Crete, Greece) *(ICMI '08)*. Association for Computing Machinery, New York, NY, USA, 249–256. https://doi.org/10.1145/1452392.1452444

[11] David Mcneill. 2005. *Gesture and Thought*. the University of Chicago Press, USA. https://doi.org/10.7208/chicago/9780226514642.001.0001

[12] Meredith Ringel Morris. 2012. Web on the Wall: Insights from a Multimodal Interaction Elicitation Study. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces* (Cambridge, Massachusetts, USA) *(ITS '12)*. ACM, New York, NY, USA, 95–104. https://doi.org/10.1145/2396636.2396651

[13] Meredith Ringel Morris, Andreea Danielescu, Steven Drucker, Danyel Fisher, Bongshin Lee, M c Schraefel, and Jacob O Wobbrock. 2014. Reducing Legacy Bias in Gesture Elicitation Studies. *Interactions* 21, 3 (May 2014), 40–45.

[14] Meredith Ringel Morris, Jacob O Wobbrock, and Andrew D Wilson. 2010. Understanding users' preferences for surface gestures. In *Proceedings of graphics interface 2010*. Canadian Information Processing Society, 261–268.

[15] Miguel A. Nacenta, Yemliha Kamber, Yizhou Qiang, and Per Ola Kristensson. 2013. Memorability of Pre-Designed and User-Defined Gesture Sets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. Association for Computing Machinery, New York, NY, USA, 1099–1108. https://doi.org/10.1145/2470654.2466142

[16] Michael Nielsen, Moritz Störring, Thomas B. Moeslund, and Erik Granum. 2004. A Procedure for Developing Intuitive and Ergonomic Gesture Interfaces for HCI. In *Gesture-Based Communication in Human-Computer Interaction*, Antonio Camurri and Gualtiero Volpe (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 409–420.

[17] Francisco R. Ortega, Alain Galvan, Katherine Tarre, Armando Barreto, Naphtali Rishe, Jonathan Bernal, Ruben Balcazar, and Jason-Lee Thomas. 2017. Gesture elicitation for 3D travel via multi-touch and mid-Air systems for procedurally generated pseudo-universe. In *2017 IEEE Symposium on 3D User Interfaces (3DUI)*. 144–153. https://doi.org/10.1109/3DUI.2017.7893331

[18] Tran Pham, Jo Vermeulen, Anthony Tang, and Lindsay MacDonald Vermeulen. 2018. Scale Impacts Elicited Gestures for Manipulating Holograms: Implications for AR Gesture Design. In *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM, 227–240.

[19] Thammathip Piumsomboon, Adrian Clark, Mark Billinghurst, and Andy Cockburn. 2013. User-Defined Gestures for Augmented Reality. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (Paris, France) *(CHI EA '13)*. Association for Computing Machinery, New York, NY, USA, 955–960. https://doi.org/10.1145/2468356.2468527

[20] Jaime Ruiz, Yang Li, and Edward Lank. 2011. User-Defined Motion Gestures for Mobile Interaction *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 197–206. https://doi.org/10.1145/1978942.1978971

[21] Giulia Wally Scurati, Michele Gattullo, Michele Fiorentino, Francesco Ferrise, Monica Bordegoni, and Antonio Emmanuele Uva. 2018. Converting maintenance actions into standard symbols for Augmented Reality applications in Industry 4.0. *Computers in Industry* 98 (2018), 68–79. https://doi.org/10.1016/j.compind.2018.02.001

[22] Teddy Seyed, Chris Burns, Mario Costa Sousa, Frank Maurer, and Anthony Tang. 2012. Eliciting Usable Gestures for Multi-Display Environments. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces* (Cambridge, Massachusetts, USA) *(ITS '12)*. Association for Computing Machinery, New York, NY, USA, 41–50. https://doi.org/10.1145/2396636.2396643

[23] SHENG KAI TANG and David Coulter. 2022. Start gesture - mixed reality. https://docs.microsoft.com/en-us/windows/mixed-reality/design/system-gesture

[24] Theophanis Tsandilas. 2018. Fallacies of Agreement: A Critical Review of Consensus Assessment Methods for Gesture Elicitation. *ACM Trans. Comput. Hum. Interact.* 25, 3 (June 2018), 18.

[25] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2015. 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* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1325–1334. https://doi.org/10.1145/2702123.2702223

[26] Radu-Daniel Vatavu and Jacob O. Wobbrock. 2022. Clarifying Agreement Calculations and Analysis for End-User Elicitation Studies. *ACM Trans. Comput.-Hum. Interact.* 29, 1, Article 5 (jan 2022), 70 pages. https://doi.org/10.1145/3476101

[27] Santiago Villarreal-Narvaez, Jean Vanderdonckt, Radu-Daniel Vatavu, and Jacob A Wobbrock. 2020. 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)*. ACM Press, Eindhoven, NA.

[28] Panagiotis Vogiatzidakis and Panayiotis Koutsabasis. 2022. 'Address and command': Two-handed mid-air interactions with multiple home devices. *International Journal of Human-Computer Studies* 159 (2022), 102755. https://doi.org/10.1016/j.ijhcs.2021.102755

[29] Adam S. Williams, Jason Garcia, and Francisco Ortega. 2020. Understanding Multimodal User Gesture and Speech Behavior for Object Manipulation in Augmented Reality Using Elicitation. *IEEE Transactions on Visualization and Computer Graphics* 26, 12 (2020), 3479–3489. https://doi.org/10.1109/TVCG.2020.3023566

[30] Adam S Williams and Francisco Ortega. 2020. 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)*. 19–22. https://doi.org/10.1109/VRW50115.2020.00286

[31] Adam S. Williams and Francisco R. Ortega. 2020. Evolutionary Gestures: When a Gesture is Not Quite Legacy Biased. *Interactions* 27, 5 (sep 2020), 50–53. https://doi.org/10.1145/3412499

[32] Adam S. Williams and Francisco R. Ortega. 2020. Understanding Gesture and Speech Multimodal Interactions for Manipulation Tasks in Augmented Reality Using Unconstrained Elicitation. *Proc. ACM Hum.-Comput. Interact.* 4, ISS, Article 202 (nov 2020), 21 pages. https://doi.org/10.1145/3427330

[33] Adam S. Williams and Francisco R. Ortega. 2021. A concise guide to elicitation methodology. https://arxiv.org/abs/2105.12865

[34] Markus L Wittorf and Mikkel R Jakobsen. 2016. Eliciting Mid-Air Gestures for Wall-Display Interaction. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction* (Gothenburg, Sweden) *(NordiCHI '16)*. ACM, New York, NY, USA, 3:1–3:4.

[35] Jacob O. Wobbrock, Htet Htet Aung, Brandon Rothrock, and Brad A. Myers. 2005. Maximizing the Guessability of Symbolic Input. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (Portland, OR, USA) *(CHI EA '05)*. Association for Computing Machinery, New York, NY, USA, 1869–1872. https://doi.org/10.1145/1056808.1057043

[36] Jacob O Wobbrock, Meredith Ringel Morris, and Andrew D Wilson. 2009. User-defined Gestures for Surface Computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI '09)*. ACM, New York, NY, USA, 1083–1092.

[37] Xiaoyan Zhou, Adam S. Williams, and Francisco R. Ortega. 2022. Towards Establishing Consistent Proposal Binning Methods for Unimodal and Multimodal Interaction Elicitation Studies. In *Human-Computer Interaction. Theoretical Approaches and Design Methods*, Masaaki Kurosu (Ed.). Springer International Publishing, Cham, 356–368.