Externalizing Mental Images by Harnessing Size-Describing Gestures: Design Implications for a Visualization System

Sarah Anne Brown
University of Florida
Gainesville, USA
sarah.brown@ufl.edu

Sharon Lynn Chu
University of Florida
Gainesville, USA
slchu@ufl.edu

Neha Rani
University of Florida
Gainesville, USA
neharani@ufl.edu

Figure 1: An illustration of a gesture-based idea visualization system situated within the context of furniture design

ABSTRACT

People use a significant amount of gestures when engaging in creative brainstorming. This is especially typical for creative workers who frequently convey ideas, designs, and stories to team members. These gestures produced during natural conversation contain information that is not necessarily conveyed through speech. This paper investigates the design of a system that uses people's gestures in natural communication contexts to produce external visualizations of their mental imagery, focusing on gestures that describe dimension-related information. While much psycholinguistics research address how gestures relate to the representations of concepts, little HCI work has explored the possibilities of harnessing gestures to support thinking.

We conducted a study to explore how people gesture using a basic gesture-based visualization system in simulated creative gift design scenarios, towards the goal of deriving design implications. Both quantitative and qualitative data were collected from the study, allowing us to ascertain what features (e.g., users' spatial frames of reference and listener types) of a gesture-based visualization system need to be accounted for in design. Results showed that our system managed to visualize users' envisioned gift dimensions, but that visualized object area significantly affected users' perceived accuracy of the system. We extract themes as to what dimensions are important in the design of a gesture-based visualization system, and the possible uses of such a system from the participants' perspectives. We discuss implications for the design of gesture-based visualization systems to support creative work and possibilities for future directions of research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

AVI '20, September 28-October 2, 2020, Salerno, Italy

© 2020 Association for Computing Machinery.

ACM ISBN 978-1-4503-7535-1/20/09...\$15.00

<https://doi.org/10.1145/3399715.3399920>

CCS CONCEPTS

• **Human-centered computing** → **Gestural input**; *Natural language interfaces*; Laboratory experiments.

KEYWORDS

gesture-based visualization, gesture recognition systems, creative brainstorming, storytelling, design

ACM Reference Format:

Sarah Anne Brown, Sharon Lynn Chu, and Neha Rani. 2020. Externalizing Mental Images by Harnessing Size-Describing Gestures: Design Implications for a Visualization System. In *International Conference on Advanced Visual Interfaces (AVI '20)*, September 28-October 2, 2020, Salerno, Italy. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3399715.3399920>

1 INTRODUCTION

When Weiser [34] proffered his vision of ubiquitous computing a few decades ago, he imagined a world where technology recedes into the background while supporting people in daily tasks. In the domain of creativity support systems, much has been done to investigate how systems can support users' creative thinking, but most work relies on the active participation of the user. We propose the idea of a system that passively harnesses gestures to externalize people's mental imagery to support thinking in creative brainstorming scenarios. However, the challenge of designing a system to visualize the information presented by gestures during natural conversation is complex. When we provide descriptions of objects, stories, or settings, we are effectively communicating our mental image for the other person(s) to recreate in their own mind. There are many components to such natural conversation that we as humans are able to understand implicitly, but would require a high level of interpretability for a system to be able to replicate.

Building towards the vision of a system that harnesses gestures to support creative thinking (we henceforth refer to such a system as a *gesture-based visualization system* – see Figure 1 for an illustration of that vision), in this paper, we investigate the design of such a system to produce visualizations from descriptions of specifically an object's dimensions, or size. Our approach for the investigation

consisted of building a prototype system based on high-level guidelines from a preliminary study, and then conducting a study with participants using the prototype system in different conditions to derive implications for design. Based on the gesture literature and our preliminary study, we tested the effects of the user's frame of reference and type of conversational agent or listener type in system use. In the rest of the paper, we present background literature on gestures, related work on gesture-based systems, and then proceed to describing our system, study, data analysis and results.

2 BACKGROUND

2.1 Gestures in Natural Conversation

Although gesture study literature has some debate as to whether the purpose of gestures is for communication to the listener or to aid the speaker's own speech production [18], research more or less consistently shows that gestures, consciously or unconsciously produced in natural conversation, convey significant information [3, 4, 24]. Multiple taxonomies have been proposed to explain different ways in which gestures convey information. One of the most commonly cited is that of McNeill [24]. It includes a separation between gestures that describe concrete objects versus those that describe abstract concepts. Gestures that describe concrete objects are known as *iconic gestures* [23, 28]. They are perhaps the easiest kind of gesture for gesture-based visualization systems, as they provide a visual indication of the speaker's thoughts. Examples of iconic gestures include using your hands to form a triangle to represent the roof of a house, or moving your hand up and down to represent a bunny hopping. They aid in describing the image you are envisioning for the listener as well as for the speaker him/herself, in a way that speech alone cannot always convey [3, 4, 24].

McNeill states that iconic gestures, when paired with speech, "...are the closest look at the ideas of another person that we...can get." [24]. He notes that gestures are not retrieved (such as how we use speech), but rather created during the flow of speech (and hence, are not predetermined) [24, 25]. Instead, the information comes from the speaker's understanding of the mental image they are trying to convey. Furthermore, research has shown that in a discourse, people tend to repeat the same or similar gestures to convey similar ideas. These repeated gesture features have been called *catchments* [27]. For example, if someone uses a triangle-shaped hand to indicate a house with a pointed roof, it is likely that the person will use the same or similar gesture the next time they reference the house.

2.2 Producing and Interpreting Gestures

Many features can affect how an external viewer may interpret the gestures produced by a speaker. One feature highlighted in the gesture literature is the speaker's frame of reference, broadly-speaking, the perspective the speaker takes in the "*phenomena that s/he is considering*" [11]. Frame of reference allows us to understand how the speakers' descriptions are situated within the mental image they are trying to convey. Research on frames of reference is more advanced with respect to speech than gestures, and in relation to describing spatial positioning rather than dimensions. A commonly accepted framework for spatial frames of reference in language is that of Levinson's [21]. Levinson defines three types of frames that can be used. Figure 2 presents an example that can be used to

Figure 2: The three frames of reference example

describe the three frame types, based on the example provided by Danziger [11]. In the scenario, the speaker is trying to describe the position of the ball, which is called the *figure* object. The description is typically done with respect to a referent object called the *ground*. The referent object chosen by the speaker and the relationship that the speaker defines between the figure and the ground determine the type of frame of reference that is adopted.

In an **absolute** frame of reference, the ground is a fixed coordinate system, such as an Earth-centered compass coordinate system. An example would be saying the ball is five inches East of the box. An **intrinsic** frame of reference describes the figure in relation to the inherent features of a ground object that is also in the space of the figure. An example would be saying the ball is by the green side of the box. And in a **relative** frame of reference, the referent is the speaker him/herself. An example would be describing the ball as being right of the box.

Another factor documented in the gesture literature as affecting how gestures are produced is the specifics of the listener or audience [1, 10]. Much research has been conducted towards creating more natural virtual conversational agents. For example, research exists that looks at the degree of attentivity of the listener [18], and the use of communicational signals such as gaze and head movements [15]. Prior work shows that the nature of the conversational agent, such as whether it provides more emotional feedback as opposed to envelope, or primarily nonverbal behaviors (such as gestures) that exist in natural conversation [9] can impact communication. In a study exploring this, subjects in a condition where the conversational agent provided envelope feedback tended to gesture more, and engage in more mimicry with the conversational agent, more so than those who engaged with a conversational agent that only provided emotional feedback. Other minute gestures, for example the ability to produce and recognize nodding, can also influence the quality of feedback between a speaker and listener [30]. The ability to engage in these subtle behaviors of course varies greatly when a conversation is being held between two people, versus when it is held between a person and a virtual agent, which the study context was in [9], or a person and a robot, such as in [30].

3 RELATED WORK

The use of gestures in system design in the HCI literature has mostly focused on gestures as a means of interaction. The large body of work on natural user interfaces falls within this focus (e.g., [12–14, 19, 32]). However, our interest in gesture-based systems is not with respect to system interaction but with respect to the use of gestures to identify and externalize thoughts. Following the theory of distributed cognition, once thought is externalized, the external representations can be used to support further thinking [16].

Previous work brought forth many systems designed to directly or indirectly support thinking through gestures. Many highlight

Figure 3: Example of a dimension gesture (left), and our prototype creating a bounding box with a sample image (right)

the role of gestures in supporting cognition [17, 22, 26]. There are many ways in which gesture can support thinking, such as in the domains of modeling, designing, verbal improvisation and more. All these tasks are cognitively demanding. One study [22] found that not only did people gesture more when improvising, but the amount they gestured was relative to improvisation quality.

Many prior systems incorporate gesture as a supportive medium to make creative thinking more interactive and less dependent on verbal lexicon, such as a system wherein the interpretation of gestures in the spatial context of a 3D shape directly deduces the designer’s intent and subsequent modeling operations [26], or when gesture is used to control a computerized object by associating a navigating command with a specific gesture [2]. For example, gesture can support children’s creative musical activity through an interactive system [17]. In one system, the presenter is able to interact with graphical elements through gestures in real-time. The use of gesture enables them to leverage innate, everyday means of communication and enhance connection with the audience [29]. Vinayak et. al. [26] designed a system wherein the interpretation of gestures in the spatial context of a 3D shape directly deduces the designer’s intent and the subsequent modeling operations.

4 PRELIMINARY STUDY

We conducted a preliminary study to investigate the features of iconic gestures people use during storytelling. Participants were provided with cartoon stimuli, and asked to retell the stories in the context of a natural conversation with the researcher. We analyzed video recordings of the gestures produced during these conversations. Certain meanings conveyed by a speaker’s gestures were found to have some consistent physical gesture features [5]. The most successful example was gestures that described an object’s dimensions (width, height, or depth). These gestures tended to incorporate the use of **both hands moving in a mirrored direction, with typically flat hand shapes**. In all instances of these gestures, the gesture’s motion held no meaning, but **the hands represented the two bounds of the described object’s dimensions**, as though the object were situated between them. Figure 3 shows an example of this from a study participant.

5 SYSTEM DESCRIPTION

Based on the results of our preliminary study, we designed and implemented a prototype system that uses the common hand features for dimension-describing gestures (two flat hands positioned in a mirrored fashion) to adjust the size of a two-dimensional image of an object based on a speaker’s gestures. The system was built using

Figure 4: Interface screens of system used during study

OpenPose [7, 8, 31, 33], a real-time keypoint detection library for the body, face, hands, and feet, which identifies 21 two-dimensional key points for each hand from a continuous frame. Following the guideline from the preliminary study, the system was designed to fit a rectangular bounding-box between the hands using the detected hand data points. Figure 3 also shows the bounding box determined by the system based on a user’s gesture.

We did not implement any gesture recognition in the system - it was unable to identify which gestures are iconic or related to dimensions. Our focus for this paper was on understanding how to design a gesture-based visualization system such that it can produce accurate visualizations of a user’s envisioned object dimensions. Gesture recognition was achieved using a wizard-of-oz method. The system takes one or more video frames as input, identifies the

hand data points, and creates a visualization of a predefined object using the detected height and width dimensions. A researcher interface was built to both control system flow, and capture the gestures needed to form the visualization. As this introduces the possibility for human error, the interface also allows the researcher to capture multiple frames of possible dimension-related gestures, then iterate through them to select the correct frames that could be used to produce the best possible visualization. From our preliminary study, we observed that one gesture tends to convey only one aspect of dimension (e.g., width, height). Thus, the capture of multiple frames allowed us to identify different dimensions, if the user gestured them. The judgment of which aspect of dimension is being addressed by a specific gesture was done by the system, based upon whether the hands were farther apart horizontally or vertically (width) or the inverse (height). As we used two-dimensional data provided by OpenPose, only width and height were used to generate two-dimensional visualizations. If one aspect of dimension needed was not gestured by the user, then the system defaulted to having both width and height of the object's bounding box to be equal to the measurement detected from the gesture.

To enable the investigation of features that can impact gesture production, we also implemented two participant interfaces for the prototype system (see Figure 4B): in one case, the listener was a virtual character named Emily, and in the other, it was a real person embedded in the interface through a live video feed.

In order to provide a context for gesturing, the gesture-based sizing system was couched in an interface that conveyed different gift creation scenarios. Screenshots of the interface are shown in Figure 4. Users had to size gifts for different given characters while interacting with either the virtual agent or the live person. The interface guided the user to select and describe a gift for a particular character that appeared on the screen in the following steps (screenshots in Figure 4A):

Steps One and Two: Pick a gift off the shelf of the shop for the character shown on the right. Then, confirm the chosen gift is correct when it appears on the table.

Step Three: Describe how you would change the gift for the character (features such as color, shape, altering details, etc.).

Step Four: Take a moment to picture the size you would make the gift for the character, then confirm when you're ready. This ensured the user had envisioned the gift's size before describing it.

Step Five: Describe the gift's size. If no gestures are produced, either the researcher or virtual agent would ask the user to show their description more visually. The researcher acting as 'wizard' would engage in frame capturing to feed into the system visualization.

Step Six: With the provided controller, size the gift according to how you mentally pictured it previously. Obviously, we cannot ascertain that the dimensions sized by the user were exactly what they mentally envisioned, but we assume that the user-sized object dimensions were accurate with respect to their mental image for this research, and thus use the term 'assumed mental image' henceforth.

For each gift-sizing task, the system produced a visualization of the object gestured about by the user. Participants would be asked at certain points to rate each system-produced object visualization on a scale of 1 to 7 in terms of how accurately the system captured their idea of the gift's size. Figure 4C shows a screenshot of the rating interface.

6 STUDY DESCRIPTION

6.1 Research Questions

The focus of this paper is to understand how to design a gesture-based visualization system that accurately reflects a user's mental imagery. We implemented a basic dimension-sizing gesture-based visualization system based on preliminary guidelines. The research questions for our main study sought to assess how well the prototype visualized users' assumed mental image of an object's dimension, and if not, why. The specific questions we addressed were:

RQ1: *Does the prototype system accurately visualize a user's mental image of an object's dimensions...*

- **RQ1a:** *...based on the measurements of the system-produced object visualization?*
- **RQ1b:** *...based on the user's perceived accuracy of the system?*

RQ2: *What factors affect the...*

- **RQ2a:** *...accuracy of the system's object visualizations?*
- **RQ2b:** *...the user's perceived accuracy of the system's object visualizations?*

RQ3: *What are the possible scenarios of use for a gesture-based visualization system?*

6.2 Study Design

The study employed a within-subjects design, and followed an ABAB pattern [6] (A being the live person, and B the virtual agent) with regards to listener type. Thus the participant would first interact with the live person (the first A acting as the baseline), followed by the virtual agent (the first B as the treatment condition), and then the live person again (the second A, withdrawal of treatment), and finally the virtual agent again (the second B, re-introduction of treatment). We chose this pattern because interacting with a live person is a process that everyone commonly engages in (thus, being a baseline) as opposed to interacting with a virtual agent.

During all interactions with the system, the researcher interface was screen-captured with audio from both the researcher and participant. We chose to record only the researcher interface, as it provided a view of what the participant saw on their screen, the webcam feed that the system used to capture gestures through the OpenPose data, and the researcher controls.

6.3 Participants

A total of 26 participants, all college students between ages 18 and 29, 14 male and 12 female, participated in the study. All were recruited through the University of Florida's Computer and Information Science and Engineering participant recruitment system, which provided them with extra credit in their courses in exchange for their participation. Due to system errors, we obtained full system data for 20 of these participants. Participant backgrounds varied, with 53.8% of participants identifying as White, 19.2% as Hispanic or Latino, 15.4% as Asian, and 11.5% as African American. We also recorded any languages spoken in addition to English, as differences across culture can impact gesturing [20]. Though the majority (65.4%) did not speak any additional languages, 15.4% of participants also spoke Spanish, 7.7% spoke Chinese, and 3.8% spoke Urdu, Hebrew, and both Hindi and Gujarati, respectively.

6.4 Protocol

The study protocol was as follows:

Informed Consent and Demographics Questionnaire: After each participant consented to participate in the study, we gathered basic demographics using a brief questionnaire.

System Tutorial: The participant practiced using the system guide's interface, doing two gift-description tasks interacting with a live person, and two with the virtual agent. During the tutorial, the researcher could explain how to proceed and answer any questions.

Main System/Study Tasks: Participants engaged in 16 gift-description tasks (Steps 1-6 described in Section 5). For the first 4 tasks, the participant interacted with the system having the live person (first A condition – see Section 6.2), the next 4 tasks were with the system using the virtual agent (first B condition), and then participants did another 4 tasks in the A condition, and another 4 in the B condition. After each AB cycle, participants were asked to do the rating task, rating each of the 8 system-visualized objects thus far in terms of how accurately the system captured their idea of the gifts' sizes (Figure 4C). The 4 characters that participants sized a gift for within a set of tasks varied in size (e.g., a large troll, a cowboy, a fairy). Though characters across the A and B conditions in one cycle were the same, their color schemes were changed to prevent boredom (e.g., see cat character in Figure 4A and 4B).

Perceived Use Interview: After the main study tasks, the researcher explained what an ideal system would be to the participant – one that can take in natural conversation and gestures and produce meaningful visualizations from them. We then asked the participant if they could think of any potential uses for such a system, and if it were directly applicable to them, how often they would use it.

7 DATA ANALYSIS

7.1 RQ1: Accuracy of System Visualizations

7.1.1 Data Pre-Processing. Since the gestures were captured using a wizard-of-oz method, there were instances when gesture frames were wrongly captured and still fed to the visualization system. To eliminate the bias from these human errors in our data, all visualizations created from bad captures were excluded. Bad captures included mostly cases where an appropriate snapshot of the correct gesture could not be obtained (such as the capture landing on the incorrect frame and not the moment the gesture was being performed). In some cases, the participants did not use dimension-describing gestures to describe the object's size. These visualizations were also excluded, as the system was designed specifically for iconic gestures following our preliminary study results (see Section 4).

7.1.2 RQ1A: System Visualization Accuracy based on Measurements. System accuracy for each gift visualization was determined by comparing the dimensions of the visualization generated by our prototype system to the visualization created by the participant during the manual-sizing task. The width and height measurements of each gift visualization were obtained from system logs, and the ratio (width/height) and area calculated. Difference scores in measurements were calculated by subtracting the widths, heights, ratios, and areas of the participant's manually-sized visualizations from each system visualization. Outliers were identified in the difference measurement scores and excluded in later analyses, resulting in a

dataset of 192 cases. We tracked participant session videos associated with these outlier cases and examined the possible causes for these values. We ran one-sample t-tests comparing the measurement differences with 0 (a difference of 0 would mean that there was no difference between the system visualization measurements and the participants' manually-sized visualization measurements).

7.1.3 RQ1B: System Visualization Accuracy based on User Perceptions. Descriptives of the perceived system accuracy scores participants gave were produced. Bivariate correlation tests were ran between these scores and visualization measurement differences.

7.2 RQ2: Factors Affecting Actual and Perceived Visualization Accuracy

Two factors were analyzed in terms of whether they affected the accuracy of the visualizations produced by the system: 'listener type' and 'frame of reference' adopted by the participant gesturer. Three main factors were analyzed in terms of whether they affected users' perceived accuracy of the system visualizations: the objective measurements of the system-produced visualizations; listener type; and frame of reference. A qualitative analysis was also done on the comments uttered by participants during their manual sizing tasks to generate more open-ended understanding of what other factors possibly affected user's perceived accuracy of the system.

7.2.1 Frame of Reference Analysis. Gesture analysis is a very detailed and effortful process. The frame of reference analysis was conducted on 25% of all participants' task videos collected. Each participant had 16 gift-sizing task videos, 4 for each set of the ABAB study design. We randomly selected one task from each cycle (i.e., 1 for the first A, 1 for the first B, 1 for the second A, 1 for the second B), resulting in a total of 4 tasks per participant, or 80 total across all 20 participants. A coder identified every gesture used to describe the gift's size for each task, and transcribed its accompanying speech. Since the identification of gestures can be somewhat subjective in nature, an intercoder agreement between two coders (of 90%) was first established with 10% of the selected data set.

We coded for the three frames of reference previously proposed by Levinson [21], but adapted the framework for dimensions instead of spatial positioning. As mentioned before, most frames of reference frameworks have been designed to analyze spatial gestures instead of dimension gestures. We identified the three frames of reference as follows. A gesture was coded as having...

An absolute frame: if the gesture was accompanied by utterances of specific dimensions, or displayed the exact size of an object without the indication of a reference object or internal viewpoint.

An intrinsic frame: if the speaker uttered a ground object and gestured with respect to it, whether or not the ground was visible.

A relative frame: if the gesture took an internal viewpoint, such as by the performance of an action or reference to oneself.

There were instances of the speaker switching frames of reference throughout a given size-describing task, so only the frame of reference used by the speaker when they explicitly gesture about the gift's dimension was considered. In the identification of reference frames, an intercoder agreement of 85% was established between 2 coders over 2 rounds of coding (10% of the dataset per round), before the rest of the data was split between the 2 coders.

7.2.2 Effects of Listener Type and Frame of Reference. Participants engaged 2 types of listener in the study: a live person and a virtual agent. A repeated measures ANOVA was conducted to determine the effects of listener type on the system measurement differences and the participants' perceived visualization accuracy scores.

We obtained 41 cases where the participant adopted an absolute frame of reference, 11 intrinsic, and 2 with a relative frame. Because of the low number of cases for the relative frame, these cases were dropped from the data. We ran a Kruskal-Wallis H Test to determine the effects of frame of reference on the visualization measurement differences and participants' perceived accuracy scores. A non-parametric test was chosen because of the unbalanced sample sizes in the absolute and intrinsic frames conditions.

To determine whether there was an interaction effect between frame of reference and listener type, a mixed ANOVA was conducted, with listener type as a within-subjects factor, and frame of reference as a between-subjects factor on both the visualization measurement differences and perceived accuracy scores.

7.2.3 Effects of Visualization Measurements. The visualization measurement differences (between the system-produced and user-sized visualizations) were classified into 3 categories: small, medium, and large difference. The thresholds of the categories were determined by the quartile values of the datas. One-way ANOVAs were then ran to determine the effects of visualization measurement differences on users' perceived accuracy of the system.

7.2.4 Analysis of Participants' Comments. We analyzed comments participants made in a think-aloud fashion during the rating task, for insights into their perception of the system. A qualitative open-coding process was performed, identifying codes and themes of interest, on transcripts of 45 identified meaningful comments.

7.3 RQ3: Possible Use Scenarios

Open-coding was performed on the transcribed interviews by two coders to determine patterns and emergent categories from participant responses. Each system use the participant came up with acted as one data point. The final categories determined were *system use*, what general area or industry the system could be used for, and *system purpose*, what overall task the system was suggested to be performing for its use. Inter-coder agreement was established over 30% of this dataset, with 83.33% agreement for system use, and 91.66% agreement for system purpose.

8 STUDY RESULTS

8.1 RQ1: Accuracy of System Visualizations

8.1.1 RQ1a: System Visualization Accuracy based on Measurements. The one-sample t-tests for measurement difference scores (scores of differences between the system-produced and the user-sized visualizations) showed significant differences in terms of all of width, height, area and ratio measurements. Looking at the data carefully, we hypothesized that there were inherent systematic offsets in the visualizations produced by the system. To verify the offsets were indeed systematic, we took the mean width and height difference between the system-produced and user-sized visualizations (0.396 and 0.372, respectively) and calculated mean-adjusted values for our system visualization data. Difference scores were re-calculated, and

Table 1: One-sample t-tests results for system visualization difference measurements

Test Value = 0 (-M indicates mean-adjusted)				
	<i>t</i>	<i>df</i>	<i>Sig. (2-tailed)</i>	<i>Mean Difference</i>
Δ Width	15.800	191	0.000	0.396
Δ Height	13.967	191	0.000	0.372
Δ Area	12.763	191	0.000	0.621
Δ Ratio	1.193	191	0.234	0.025
Δ Width-M	0.012	191	0.991	0.000
Δ Height-M	0.016	191	0.987	0.000
Δ Area-M	1.244	191	0.215	0.044
Δ Ratio-M	-1.048	191	0.296	-1.247

one-sample t-tests were re-ran. With the standardized adjustments, there were no longer significant differences on any of the difference measurements. Results of the one-sample t-tests with the original and adjusted difference measurement values are in Table 1.

The outliers we excluded from this data primarily included: (i) missed bad captures of participant gestures (ii) cases where the participant gestured dimensions for larger objects without being particularly specific with their hands (resulting in abnormally large visualizations), for example trying to gesture for the concept of 'bigger' with a dimension-like gesture and (iii) cases where the larger objects they envisioned required the user to extend their hand out, or nearly out of frame.

8.1.2 RQ1b: System Visualization Accuracy based on User Perceptions. Perceived accuracy scores followed an almost normal distribution, with a mean of 4.076, a mode of 5.0, and a slight skewness of -0.104. There were strong correlations between participant perceived accuracy scores and visualization differences in width, $r(192) = -0.427$, $p = 0.000$; height, $r(192) = -0.495$, $p = 0.000$; area, $r(192) = -0.417$, $p = 0.000$; and ratio, $r(192) = 0.200$, $p = 0.005$. Correlation values were negative, showing that the smaller the measurement difference, the higher the perceived accuracy score.

8.2 RQ2: Factors Affecting Accuracy

8.2.1 RQ2a: Effects on Visualization Accuracy. There was no significant difference between listener types in terms of visualization measurement differences. However, looking at the system visualization measurement differences across cycles in our ABAB design showed interesting trends (see Figure 5). For the live person condition, we can see that the average measurement differences in terms of width, height, area and ratio generally decreased in the second cycle of the ABAB design. For the virtual agent condition, the average measurement differences mostly stayed the same across the 2 cycles, except for area, which increased slightly in the second cycle. There is a clear pattern for ratio whereby the average measurement differences are consistently larger (across the 2 cycles) in the virtual agent conditions than in the live person conditions.

There were no significant differences in the visualization measurement differences in terms of frame of reference on all aspects (width, height, area, ratio). However, there was a marginally significant difference in terms of area ($X^2 = 3.598$, $p = 0.058$) between absolute ($M = 22.49$) and intrinsic ($M = 27.67$) frames of reference.

Figure 5: Measurement differences across ABAB design

Figure 6: Measurement differences and perceived accuracy

There were no interaction effects between frame of reference and listener type on any of the visualization measurement differences.

8.2.2 RQ2b: Effects on User Perceived Accuracy. There were statistically significant effects of level of categorized visualization measurement differences on participants' perceived accuracy scores for all aspects of dimension (width $p = 0.000$, height $p = 0.000$, area $p = 0.000$, ratio $p = 0.008$). The means are shown in Figure 6. Post-hoc analysis showed that the large measurement difference condition was significantly different from the small measurement difference condition for width, height, and area. For ratio, all pairwise comparisons were significantly different, save for the difference between the medium and large measurement differences.

There was no statistically significant difference in perceived accuracy across reference framer or listener, nor was there any interaction effects between the two variables.

8.2.3 Analysis of Participant's Comments. The participants' think-aloud comments were classified into 3 high-level categories: (i) the point-of-view that the participant used for evaluating the system visualization accuracy; (ii) their focus of evaluation; and (iii) their overall impression of the system. Our final coding scheme from the analysis of participant's comments can be found in Figure 7.

8.3 RQ3: Possible Use Scenarios

The final categories from our interview analysis can be found in Table 2, classified in terms of *system use* and *purpose*. System Use showed variety across the established codes, the most prevalent being design and shopping. The codes for system purpose, which reflected possible roles of the system, were: (i) **Translating** a previously non-existent concept the user has in his/her mind into a

Code Lvl 3	Code Lvl 2	Description
Evaluation point-of-View	System functioning	Evaluates based on how the user expects the system to function
	Fit to virtual object	Evaluates based on a target object or character on screen
	Real-world expectations	Evaluates based on how the user expects the object to be in the real-world
	User interaction	Evaluates based on the user's previous interaction with the system
Evaluation focus	Area	Evaluation focus was on visualization area
	Height	Evaluation focus was on visualization height
	Width	Evaluation focus was on visualization width
	Ratio	Evaluation focus was on visualization ratio
System impression	Positive	Positive impression of system
	Negative	Negative impression of system

Figure 7: Coding scheme for participant rating comments

Table 2: Possible Use Scenarios Results

System Use	%	n	System Purpose	%	n
Design	27.5%	11	Translating	65.0%	26
Shopping	20.0%	8	Searching	15.0%	6
Entertainment	15.0%	6	Customizing	12.5%	5
Communication	12.5%	5	N/A	7.5%	3
Education	7.5%	3			
Accessibility	7.5%	3			
Healthcare	5.0%	2			
Etc.	5.0%	2			

visualization for others; (ii) **Searching** for a previously existent object/person the user has in their mind and locate it/them for the user; and (iii) **Customizing** a previously existent object for the user, based on changes the user has in their mind.

9 DISCUSSION

We conducted a study with a basic gesture-based visualization system focusing on dimension descriptions, with the goal of understanding how to move forward in designing this novel type of system. We designed and implemented the basic system using guidelines from a preliminary study, which included that dimension-describing gestures tended to use both hands moving in a mirrored direction, typically flat hand shapes, and that the hands represent the two bounds of the described object's dimensions.

RQ1 asked whether our prototype system accurately visualizes a user's mental image of an object's dimensions. Our findings include the discovery of an inherent offset (0.396 for width, 0.372 for height) in our system's visualizations, which, when corrected, would result in no significant difference between the system visualizations and the user custom-sized visualizations. This indicates, that with the offset corrected, the system visualization portrayed the users' assumed mental image of the object accurately. The reason for the offset is likely a complex combination of factors. One factor could be that we used the shoulders as a fixed reference point for the size of the overall object visualization space. Other factors uncovered in response to RQ2 may also have had a role in causing the offset. Further research is required to unpack the

factors causing the offset, but this suggests that using a one-to-one mapping approach (i.e., dimensions detected from gestures are the exact object dimensions imagined) for a system to produce an object visualization from gestures may not produce entirely accurate visualizations, and it is important to obtain a comprehensive understanding of influencing factors.

In terms of user perceived visualization accuracy, we can infer that participants thought the system had average accuracy. This may easily change in future system iterations, with the inherent offset accounted for. Correlation results showed, non-surprisingly, that the higher the visualization measurement differences in terms of width, height and area, the lower the user's perceived accuracy of the system. However, the correlation for differences in ratio was very slightly positive. We explore possible explanations for this in the discussion of the answers to RQ2 below.

RQ2 asked what factors may affect the accuracy of the system's object visualization, as well as the user's perceived accuracy of that visualization. There was a marginal effect of listener type on visualization measurement differences, with the virtual agent listener leading to larger visualization measurement differences, especially in terms of ratio. This could be due to the level of feedback afforded by the virtual listener. With a more interactive listener, the speaker may be more motivated to better describe what they are imagining. There was also a marginal effect of frame of reference in terms of area specifically. The intrinsic reference frame led to larger visualization measurement differences than the absolute frame. When a speaker adopts an absolute frame of reference in the description of an object, they typically use a fixed coordinate system, such as the metric system or Cartesian coordinates. The coordinate system speakers use can generally be identified easily, and tends to be consistent across speakers. However, in an intrinsic frame, the object used as reference (the ground object in Figure 2) may vary by speaker, likely explaining the larger measurement differences. This indicates the need to recognize speaker frame of reference and account for the use of ground objects in future gesture-based visualization systems. Additionally, our analysis of the reasons for the outlier cases in visualization measurement differences showed that gesturing while describing an object's size may not always have size description as the focus.

In terms of factors affecting users' perceived visualization accuracy, the only factor among those tested that seems to have some effect is the actual visualization measurement differences. There were significant effects of visualization measurement differences on users' perceived visualization accuracy, with smaller differences leading to higher perceived accuracy all aspects of dimension (width, height, area) except for ratio.

The results for ratio were interesting. Ratio was calculated by dividing the width measurements of the visualization by its height measurements. Thus, smaller measurement differences between the system-produced visualization and the user-sized visualization would mean that the system underestimated the width, and overestimated the height of the user's imagined object visualization from his/her gestures. And larger measurement differences would be the reversed. Users' perceived accuracy of the system visualization, though, is significantly lower when measurement differences are smaller (i.e., shorter width, longer height than users imagined), and higher when measurement differences are larger (i.e., longer

width, shorter height). This seems to suggest then that participants' perceptions are affected most when the system shows the object to have a longer width and a shorter height than what they imagined. A possible reason for this bias can be the specific gifts or objects that we loaded into the system for the gift-description tasks. There were 12 objects used in total. Based on one's expectations of the typical dimensions of these objects in real-life, and the exact visuals of these gifts that the system provided (see Figure 4right), 6 of them could have any ratio (e.g., an apple can either have a longer width and shorter height, or shorter width and longer height). Two of the objects are typically longer in width and shorter in height (e.g., a basket), and 4 of them are typically less wide and taller (e.g., a flower vase). So when the system produced a visualization of one of these 4 objects to be wider and shorter than what the participants expected, perceived accuracy was affected. This suggests that the specifics of the objects that are the target of a speaker's description may matter for a gesture-based visualization system.

The analysis of participants' think-aloud comments indicate other factors may influence perceived accuracy, including: (i) users' understanding of the working mechanism of the system; (ii) the referent object that the user used to define his/her object dimensions; (iii) the user's expectations of the actual dimensions of the object in the real world; and (iv) what the user thought he/she gestured with respect to the object. Each of these requires further research and operationalization to inform system design.

For RQ3, we investigated possible uses for a gesture-based visualization system. From our interview analysis, the codes for system use reflected expected application areas to be very wide-ranging, consisting of: Entertainment, Design, Shopping, Accessibility, Healthcare, Communication, Education, and etc. Many participants perceived the use of the system in a way that aligns with our vision: to visualize thoughts from creative, design-based scenarios. This shows promise for our research moving forward as we work towards the design of gesture-based visualization systems.

9.1 Conclusion

Building off of a preliminary study, we designed and developed a gesture-based visualization system for gestures that describe the dimensions of an object. We conducted a study with this system as probe, with our conditions varying the type of listener that the participant conversed with (live or virtual). Our results highlight many detailed aspects that need to be considered in the design of gesture-based visualization systems.

The limitations of our work include the fact that our virtual agent, Emily, was not fully fleshed out for all possible conversational possibilities, such as when a user skipped the 'think of size' task and proceeded to describe the gift's size before she asked them to. And last but not least, we did not analyze the entirety of the dataset for the frame of reference analysis, although the sample analyzed was significantly large enough to derive conclusions.

ACKNOWLEDGMENTS

This research was supported by NSF Grant 1736225. We thank Cameron Kozlin, Honey Tumbaga, and Samantha Su for aiding in the analysis of our study data.

REFERENCES

- [1] Martha W. Alibali, Dana C. Heath, and Heather J. Myers. 2001. Effects of Visibility between Speaker and Listener on Gesture Production: Some Gestures Are Meant to Be Seen. *Journal of Memory and Language* 44, 2 (2001), 169 – 188. <https://doi.org/10.1006/jmla.2000.2752>
- [2] Thomas Baudel and Michel Beaudouin-Lafon. 1993. Charade: remote control of objects using free-hand gestures. *Commun. ACM* 36, 7 (1993), 28–35.
- [3] Geoffrey Beattie and Heather Shovelton. 1999. Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of language and social psychology* 18, 4 (1999), 438–462.
- [4] Geoffrey Beattie and Heather Shovelton. 2002. An experimental investigation of some properties of individual iconic gestures that mediate their communicative power. *British journal of psychology* 93, 2 (2002), 179–192.
- [5] Sarah Anne Brown, Sharon Lynn Chu, Francis Quek, Pomaikai Canaday, Qing Li, Trystan Loustau, Sindy Wu, and Lina Zhang. 2019. Towards a Gesture-Based Story Authoring System: Design Implications from Feature Analysis of Iconic Gestures During Storytelling. In *International Conference on Interactive Digital Storytelling*. Springer, 364–373.
- [6] Breanne J Byiers, Joe Reichle, and Frank J Symons. 2012. Single-subject experimental design for evidence-based practice. *American journal of speech-language pathology* (2012).
- [7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2018. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*.
- [8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In *CVPR*.
- [9] Justine Cassell and Kristinn R Thorisson. 1999. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence* 13, 4-5 (1999), 519–538.
- [10] Susan Wagner Cook and Michael K. Tanenhaus. 2009. Embodied communication: Speakers' gestures affect listeners' actions. *Cognition* 113, 1 (2009), 98 – 104. <https://doi.org/10.1016/j.cognition.2009.06.006>
- [11] Eve Danziger. 2010. Deixis, gesture, and cognition in spatial frame of reference typology. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 34, 1 (2010), 167–185.
- [12] Orlando Erazo and José A Pino. 2015. Predicting task execution time on natural user interfaces based on touchless hand gestures. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. 97–109.
- [13] Rita Francese, Ignazio Passero, and Genoveffa Tortora. 2012. Wiimote and Kinect: gestural user interfaces add a natural third dimension to HCI. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. 116–123.
- [14] Eiji Hayashi, Manuel Maas, and Jason I Hong. 2014. Wave to me: user identification using body lengths and natural gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3453–3462.
- [15] Dirk Heylen. 2006. Head gestures, gaze and the principles of conversational structure. *International Journal of Humanoid Robotics* 3, 03 (2006), 241–267.
- [16] James Hollan, Edwin Hutchins, and David Kirsh. 2000. Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction (TOCHI)* 7, 2 (2000), 174–196.
- [17] Junko Ichino, Aura Pon, Ehud Sharlin, David Eagle, and Sheelagh Carpendale. 2014. Vuzik: the effect of large gesture interaction on children's creative musical expression. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design*. 240–249.
- [18] Naomi Jacobs and Alan Garnham. 2007. The role of conversational hand gestures in a narrative task. *Journal of Memory and Language* 56, 2 (2007), 291–303.
- [19] Christian Jetter, Jens Gerken, and Harald Reiterer. 2010. Natural user interfaces: Why we need better model-worlds, not better gestures. In *CHI 2010*.
- [20] Sotaro Kita. 2009. Cross-cultural variation of speech-accompanying gesture: A review. *Language and cognitive processes* 24, 2 (2009), 145–167.
- [21] Stephen C Levinson. 1996. Frames of reference and Molyneux's question: Crosslinguistic evidence. *Language and space* 109 (1996), 169.
- [22] Carine Lewis, Peter Lovatt, and Elizabeth Kirk. 2015. Many hands make light work: The facilitative role of gesture in verbal improvisation. *Thinking Skills and Creativity* 17 (2015), 149–157.
- [23] Fridanna Maricchiolo, Augusto Gnisci, and Marino Bonaiuto. 2012. Coding hand gestures: A reliable taxonomy and a multi-media support. In *Cognitive behavioural systems*. Springer, 405–416.
- [24] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.
- [25] David McNeill, Justine Cassell, and Karl-Erik McCullough. 1994. Communicative effects of speech-mismatched gestures. *Research on language and social interaction* 27, 3 (1994), 223–237.
- [26] Sundar Murugappan, Hairong Liu, Karthik Ramani, et al. 2013. Shape-It-Up: Hand gesture based creative expression of 3D shapes using intelligent generalized cylinders. *Computer-Aided Design* 45, 2 (2013), 277–287.
- [27] Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E. McCullough, and Rashid Ansari. 2002. Multimodal Human Dis-course: Gesture and Speech. *ACM Trans. Comput.-Hum. Interact.* 9, 3 (Sept. 2002), 171–193. <https://doi.org/10.1145/568513.568514>
- [28] Siddharth S Rautaray and Anupam Agrawal. 2015. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial intelligence review* 43, 1 (2015), 1–54.
- [29] Nazmus Saquib, Rubaiat Habib Kazi, Li-Yi Wei, and Wilmot Li. 2019. Interactive body-driven graphics for augmented video performance. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [30] Candace Sidner and Christopher Lee. 2007. Attentional gestures in dialogues between people and robots. *Engineering approaches to conversational informatics*. Wiley and Sons (2007).
- [31] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand Key-point Detection in Single Images using Multiview Bootstrapping. In *CVPR*.
- [32] Radu-Daniel Vatavu. 2012. User-defined gestures for free-hand TV control. In *Proceedings of the 10th European conference on Interactive tv and video*. 45–48.
- [33] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *CVPR*.
- [34] Mark Weiser. 1998. The future of ubiquitous computing on campus. *Commun. ACM* 41, 1 (1998), 41–42.