Mixed Reality Deictic Gesture for Multi-Modal Robot Communication

Tom Williams
MIRRORLab
Colorado School of Mines
Golden, CO, USA
twilliams@mines.edu

Matthew Bussing
MIRRORLab
Colorado School of Mines
Golden, CO, USA
mbussing@mines.edu

Sebastian Cabrol
MIRRORLab
Colorado School of Mines
Golden, CO, USA
cabrol@mines.edu

Elizabeth Boyle
MIRRORLab
Colorado School of Mines
Golden, CO, USA
eboyle@mines.edu

Nhan Tran
MIRRORLab
Colorado School of Mines
Golden, CO, USA
nttran@mines.edu

***Abstract*—In previous work, researchers have repeatedly demonstrated that robots' use of deictic gestures enables effective and natural human-robot interaction. However, new technologies such as augmented reality head mounted displays enable environments in which mixed-reality becomes possible, and in such environments, physical gestures become but one category among many different types of mixed reality deictic gestures. In this paper, we present the first experimental exploration of the effectiveness of mixed reality deictic gestures beyond physical gestures. Specifically, we investigate human perception of videos simulating the display of allocentric gestures, in which robots circle their targets in users' fields of view. Our results suggest that this is an effective communication strategy, both in terms of objective accuracy and subjective perception, especially when paired with complex natural language references.**

***Index Terms*—Mixed Reality, Augmented Reality, Deixis, Natural Language Generation, Human-Robot Interaction**

I. Introduction

Robots are already being deployed in factories, hospitals, and search-and-rescue operations. In these domains, there is increasing need for robots that are not just taskable, but truly collaborative. Crucially, these robots need to communicate with their human users and teammates in a way that is effective and natural, while minimizing the need for special training. Accordingly, researchers have investigated the use of *natural language* as well as the other communicative behaviors that humans naturally and effortlessly use to facilitate effective human-human communication. For example, humans often use *gesture* to improve the fluency of their speech, communicate abstract concepts, and help them refer to objects, locations, and people in their environment. Perhaps the most important form of gesture in task-based contexts is *deictic gesture* (e.g., pointing), which humans use to draw their interlocutors' attention to different parts of their shared environment, typically to allow them to more readily make natural language reference to nearby objects, people, and locations.

Implementing effective deictic gesture in robots (to reap the same benefits seen in human-human interaction) has many challenges. First, deictic gesture can take a wide variety of forms beyond pointing, including presenting, exhibiting, touching, grouping, and sweeping [28]. Accordingly, researchers have needed to determine not only how robots might physically generate these different types of gestures, but also have needed to investigate the different tradeoffs made by these gesture types in both their effectiveness and how they are perceived, and the effect of context on those tradeoffs [75].

These research challenges are made all the more challenging by the dramatic differences that can be observed in robot morphology. Consider, for example, an autonomous, collaborative, unmanned aerial vehicle (UAV) working in an alpine search and rescue environment. Such a robot may have many reasons to refer to entities in its shared environment, such as reporting the location of disaster victims, or as part of collaborative dialogue about the search through the complex terrain of the disaster zone [96], [97]. However, traditional deictic gesture may not be possible for such a robot, as mounting an arm on such a UAV may not be feasible, and moreover, it may not be effective, due to the scale of the environment and the frequent distance of the UAV from its teammates. For such robots, it is thus critical to investigate whether new communication modalities may enable behaviors that achieve the same functionality as traditional deictic gestures, while respecting these morphological constraints.

We believe that one promising path towards enabling these alternative forms of gesture may be found by leveraging recent Augmented and Mixed Reality technologies [11][1], which allow spatially-grounded visualizations to be rendered over a user's view of their physical environment, typically by way of a Head-Mounted Display (HMD). As a simple example, consider again the case of a UAV communicating with human teammates about the location of a disaster victim. Purely using natural language, a UAV might use an utterance such as "There is an injured person behind the fourth tree to the right of the tall blue pylon." Such an utterance is complex, verbose, may require significant spatial reasoning capabilities to produce and may require sustained attention to interpret. In contrast, if the UAV's teammate were wearing an HMD, the UAV might instead be able to simply draw a circle around the relevant tree and state "There is an injured person behind that tree".

This work was supported by NSF CNS-1823245.

[1]Mixed Reality (MR) refers to any portion of the Reality-Virtuality Continuum containing both real and virtual objects: from Augmented Reality (AR), where virtual objects are displayed in the real world, to Augmented Virtuality (AV), where real objects are displayed in the virtual world [57]. However, AR and MR are often used interchangeably, and are as well in this paper.

This type of *Mixed Reality Deictic Gesture* would leverage the UAV's ability to manipulate its teammate's Mixed Reality Environment to achieve the same communicative goals as a physical gesture would have.

In recent work, we presented the first conceptual framework for categorizing the different types of Mixed Reality Deictic Gestures that may be used in Mixed Reality Human-Robot Interaction, as well as the dimensions along which such categories of gestures are expected to differ [91], [95]. However, to date there has been no systematic empirical examination of the effectiveness and perception of such gestures, in the way that there has been for robots' use of physical deictic gestures [75]. In this paper, we present the first empirical experiment designed to achieve this goal.

In Section II, we provide a brief survey of previous work exploring human and robot use of deictic gesture, as well as of recent work at the intersection of Augmented and Mixed Reality and HRI, including the limited set of work previously exploring Mixed Reality Deictic Gesture for HRI. In Section III, we then describe the design of a human subject experiment designed to provide a preliminary investigation of the effectiveness and human perception of Mixed Reality Deictic Gesture, in which we assess human perceptions of videos simulating the display of such gesture; a study designed to serve as a bridge towards future studies with real AR hardware. We the present the results of that experiment in Section IV. Finally, in Section V we discuss the implications of our experiment and suggest possible design guidelines for robot designers before concluding in Section VI.

II. Related Work

A. Human Deictic Gesture

Deixis is one of the most crucial pieces of human-human communications [52], [61], as well as one of the oldest, both anthropologically and developmentally. Unlike many other aspects of human communication, there are clear analogues of deictic gesture in the animal kingdom (e.g., the signaling capabilities of animals in the presence of predators) [54], [64]. However, deixis itself appears to be uniquely human: even our closest relatives, primates, are unable to point [50], [84] (beyond apes in captivity reaching out for food desired from humans [51] or locations desired access to from humans [76], see also [41], [44]), despite being able to use other kinds of gestures to a limited extent [67]. This divergence in capability may exist in part because deixis requires relatively sophisticated capabilities involving modeling of attentional states and theory of mind [13], [25], [63]. Not only does reasoning about the feasibility and effectiveness of deictic gesture require about perspective taking, but more fundamentally, deictic gesture serves to direct an interlocutor's attention from where it is to where it should be; recognizing that an interlocutor's attention is not where you desire it to be is a complex capability indeed.

In contrast, humans point while speaking even from infancy, with deictic gesture beginning around 9-12 months [8], and general deictic reference mastered around age 4 [20]. Deictic gestures have been shown to be a powerful technique for language learners, as they allow speakers to communicate their intended referents before being able to do so in language, in the same way that other types of gestures help speakers to communicate their intended sense or meaning when they otherwise lack the words to do so. Indeed, developmental changes in deictic gestural capabilities in humans has been demonstrated to be a strong predictor of changes in language development [47]. In addition, long past infancy, humans continue to rely on deictic gesture as a core communicative capability, as its attention-direction presents an efficient and workload-reducing referential strategy in complex environments, far beyond that of purely verbal reference [24], [33], [34], [36], [48], and as deictic gesture allows for communication in environments in which verbal communication would be difficult or impossible, such as in noisy factory environments [38]. Accordingly, it is no surprise that Human-Robot Interaction researchers have sought to enable this effective and natural communication strategy in robots.

B. Robot Deictic Gesture

Within the human-robot interaction literature, there has been widespread evidence for the effectiveness of robots' use of physical deictic gesture[2]. Specifically, studies have shown that robots' use of deictic gesture is effective at shifting attention in the same way as is humans' use of deictic gesture [14], and that robots' use of deictic gesture improves both subsequent human recall and human-robot rapport [12]. This effectiveness has been demonstrated across different contextual scales as well, including gestures to nearby objects on a tabletop [74], gestures to larger regions of space between the robot and its interlocutor [19], and gesture to large-scale spatial locations during direction-giving [62]. Furthermore, this effectiveness has shown to be especially true when gestures are generated in socially appropriate ways [53]. Research has also shown that robots' use of deictic gesture is especially effective when paired with *deictic gaze*, in which a robot (actually or ostensibly) shifts its gaze towards its intended referent [1], [2], [19], and that this is especially effective when gaze and gesture are appropriately coordinated [72]. Also of interest is a recent survey from Cha et al., in which deictic gaze and gesture are discussed within the context of a wide variety of nonverbal signaling mechanisms [15]. These findings have motivated a variety of technical approaches to deictic gesture generation [42], [43], [73], [90], as well as a number of approaches for integrating gesture generation with natural language generation [27] (see also [31], [32], [68], [85]).

Of particular interest to us is the work of Sauppé and Mutlu [75]. Building off the work of Clark, who showed that humans use many deictic gestures beyond pointing [21], Sauppé and Mutlu explored a selection of robotic deictic gestures: pointing, presenting, exhibiting, touching, grouping, and sweeping. Sauppé and Mutlu were especially interested in how these categories differed in both effectiveness and

[2]While there has also been significant work on robot *understanding* of human deictic gesture [56], we focus on robots' *generation* of such gestures.

perceived naturality, and how different contextual factors, such as the density of candidate referents, the number of fully ambiguous distractors for the referent, and the distance of the referent from the referrer. As we will describe, the set of questions we are interested in investigating both in this work and in future work has a number of parallels with those of interest to Sauppé and Mutlu, and accordingly, as we will also describe, the experiment presented in this paper was designed with careful attention to the design used by Sauppé and Mutlu.

C. Augmented Reality for HRI

Although research on augmented and mixed reality have been steadily progressing over the past several decades [5], [6], [11], [86], [98], there has been relatively little work using augmented reality (AR) technologies to facilitate human-robot interactions (despite a number of papers over the past twenty-five years highlighting the advantages of doing so [35], [58]). Recently, however, research at the intersection of these fields has begun to dramatically increase [93], [94]. Recent work in this area includes approaches using AR for robot design [66], calibration [77], and training [80], and for communicating robots' perspectives [39], intentions [4], [16]–[18], [30] and trajectories [29], [70], [88], [99].

Most relevant to this paper are recent works on aligning human and robot perspective to enable more effective robot communication. Amor et al., for example, demonstrate the use of a projector to project instructions and highlight task-relevant objects within a constrained and highly structured task environment shared by robot and human teammates. In that work, however, no natural language generation is used, and projected visualizations are cast as part of the task environment, rather than as part of the robot's communication [3] (see also [4], [30]). Even more closely related, Sibirtseva et al. present an approach in which, as a human teammate describes a target referent to a robot, the robot's maintained distribution over possible intended referents is visualized by circling remaining reference candidates in the user's AR HMD [78] (see also similar work in VR from Perlmutter et al. [65]). This is closer to our area of interest, as the visualizations used in this work are explicitly used to pick out referential candidates, and are explicitly cast as being from the robot's perspective. However, we note that this is *passive* communication, as the robot is generating a backchannel response to the human's communication, whereas we are interested in robots' use of AR as a channel for *active* communication regarding its own intended referents. Moreover, Sibirtseva et al. were principally concerned with the tradeoffs between tablet, projector, and HMD-based AR visualizations, rather than on the impact of contextual factors. Also of interest is recent work from Reardon et al., in which a robot draws the trajectory a human teammate should take onto their field of view, and highlights the intended targets of that trajectory [69]. This work takes a more active communication approach than the work of Sibirtseva et al., but like Sibirtseva et al., Reardon et al. operate outside the context of language-based robot communication.

Finally, this work builds directly off of our own previous work [91], [95] (see also [40]), in which we presented a conceptual framework for categorizing the space of deictic gestures available in Mixed-Reality human-robot interactions, including both traditional physical gestures and purely virtual deictic annotations (categorized into allocentric gestures (e.g., circling a target referent in a user's AR HMD), perspective-free gestures (e.g., projecting a circle around a target referent on the floor of the shared environment), ego-sensitive allocentric gestures (e.g., pointing to a target referent using a simulated arm rendered in a user's AR HMD), and ego-sensitive perspective-free gestures (e.g., projecting a line from the robot to its target on the floor of the shared environment)), as well as combinations of different forms of mixed reality deictic gesture. We then present an initial analysis hypothesizing how these combinations of potential gestures would differ along eleven dimensions, including privacy, cost, and legibility. This framework is especially valuable for our research as, in conjunction with the work of Sauppé and Mutlu [75], it suggests concrete hypotheses regarding the effectiveness and perception of mixed reality deictic gestures in different contexts, allowing us to empirically investigate whether mixed reality deictic gestures have the same communicative benefits as physical gestures, and how those benefits differ according to context. In the next section, we will present a set of such hypotheses, and a human subject experiment designed to investigate them. While in this paper we will only examine allocentric gestures, we have designed our experiment so as to allow all of the gestural categories in our conceptual framework to be examined in future experiments using the same paradigm.

III. Experiment

To better understand the impact of mixed reality deictic gesture as a new modality for robot communication, and its interaction with natural language, we designed a human subject experiment in which participants viewed a robot referring to objects within a visual scene using natural language, mixed reality deictic gesture, or both modalities in combination. This experiment was designed so as to follow the general paradigm used in the seminal evaluation of physical robot gesture presented by Sauppé and Mutlu [75]. All aspects of our experimental design received IRB approval.

A. Experimental Design

Following Sauppé and Mutlu, we used a within-subject design, in which participants watched a robot refer to a series of twelve objects using different communication strategies.

1) Interaction Design: Our first independent variable was *communication style*. For one-third of the objects, the robot used *complex reference* alone, generating an expression of the form "Look at that {color} {shape}" (e.g. "Look at that red cube"). While in future experiments we plan to use fully articulated and minimally articulated baselines similar to those used by Sauppé and Mutlu, in this experiment all complex references followed a common pattern so as to better investigate reaction time. For another third of the objects, the

Fig. 1: Task Environment, with simulated AR visualization

robot used a mixed reality deictic gesture, drawing a circle around the target and stating "Look at that"; a pattern similar to the gestural conditions used by Sauppé and Mutlu. For the final third of the objects, the robot used both complex reference *and* mixed reality deictic gesture, circling the target and then generating a complex reference as described above; a pattern similar to a combination of the gestural and fully articulated conditions used by Sauppé and Mutlu.

2) Environment Design: The experimental environment contained a Kobuki robot positioned behind an array of eighteen blocks, of four shapes (cubes, triangles, cylinders, towers) and four colors (red, yellow, green, blue), evenly spaced in four rows. Specifically, there were six unique blocks and six pairs of non-unique blocks (a difference of *inherent ambiguity*), evenly split between the front and rear rows (a difference of *distance*), and distributed as uniformly as possible according to color and shape. This sought to simultaneously capture multiple environmental dimensions previously determined by Sauppé and Mutlu to affect the accuracy and perceived effectiveness of reference: *ambiguity* and *distance from referrer* while controlling for the other dimensions previously investigated by Sauppé and Mutlu (object clustering, visibility, and noise). Our second and third independent variables were thus referent ambiguity and referent distance[3], yielding a total of twelve (3x2x2) experimental conditions.

B. Procedure

Participants were recruited online using Amazon's Mechanical Turk platform, and directed towards a psiTurk experimental environment [37][4]. After providing informed consent and providing demographic information[5], participants were instructed that they would watch a series of videos in which a robot described and/or visually gestured towards a target object by drawing a circle around it. They were told that they should click on the object that was being described as soon as they had identified it. Participants were then assigned to one of twelve conditions each corresponding to a different video order determined through a counterbalanced Latin Square array. Participants then watched twelve videos, each corresponding with a different experimental condition. When mixed reality deictic gesture was used in a video, gesture onset began 660ms before speech onset, based on the gestural timing model presented by Huang and Mutlu [45] and leveraged by Sauppé and Mutlu [75]. Clicking on any object within a video sent the participant to a survey page in which they were asked to assess the effectiveness of the robot's speech and gesture and the likability of the robot, using the measures described below. Upon answering these survey questions, participants were allowed to proceed to the next video in the series. All videos were six seconds in length, including padding before and after the robot's communicative act.

C. Hypotheses

We examined four core hypotheses:

- H1 We hypothesized that participants would have worse accuracy in identifying the robot's target referent only when ambiguous complex noun phrases were used without an associated mixed reality deictic gesture (i.e., in the complex reference condition for targets with inherent ambiguity)[6].
- H2 We hypothesized **(H2.1)** that the speed at which participants would be able to identify the robot's target referent would be better when mixed reality deictic gesture was used, as it would allow target referents to be disambiguated even before speech began, and **(H2.2)** that this reaction time would increase when a reference was ambiguous.
- H3 We hypothesized **(H3.1)** that participants would perceive the robot to be more effective when mixed reality deictic gesture was used, especially **(H3.2)** when used in combination with complex reference, and **(H3.3)** when the target referent was ambiguous.
- H4 We hypothesized that the extent to which participants liked the robot would correlate with its effectiveness, and accordingly, that **(H4.1)** perceived likability would be higher when mixed reality deictic gesture was used, **(H4.2)** especially in conjunction with complex reference, and **(H4.3)** for ambiguous targets.

D. Measures

To assess these hypotheses, objective and subjective measures were used. All measures were collected once per video.

1) Accuracy: An objective measure of *accuracy* was gathered by recording which item in the scene participants clicked on, and determining whether or not this was in fact the object intended by the robot.

[3]We did not expect to see any effects of distance, but decided to include distance as an independent variable so that we can use an identical experimental design in future experiments in which we will use other types of gestures, e.g., pointing gestures generated with real or simulated arms, for which we *would* expect to see a potential difference.

[4] Mechanical Turk is more successful than traditional university studies at broad demographic sampling [23], though it still has population biases [82].

[5]In online experiments, it is valuable to collect demographic data pre-task to prevent participants who do not meet age requirements from participating.

[6]Performance will obviously be poor in this intersection of conditions, as language used will not be fully discriminating. Our emphasis here is that *with the exception of such cases case*, performance should be uniformly good.

2) Reaction Time: An objective measure of *reaction time* was gathered by recording time stamps at the moment each video phase began (i.e., when the page loaded) and ended (i.e., when an object was clicked on).

3) Effectiveness: A subjective measure of robot *effectiveness* was gathered using a modified version of the Gesture Perception scale presented by Sauppé and Mutlu [75]. Our modified version asked participants to evaluate each of the following statements by clicking a point anywhere along a seven-point Likert-type scale:

1) The robot used its speech and/or mixed reality deictic gesture effectively.
2) The robot's speech and/or mixed reality deictic gesture helped me to identify the object.
3) The robot's speech and/or mixed reality deictic gesture was appropriate for the context.
4) The robot's speech and/or mixed reality deictic gesture was easy to understand.

Each participants' scores for a single video were then transformed to a range of 0-100 and averaged. A reliability analysis indicated that the internal reliability of this scale was very high for our experiment, with Cronbach's α = 0.955.

4) Likability: A subjective measure of robot *likability* was gathered using the Godspeed II Likability scale [7]. Our modified version asked participants to rate their perception of the robot along each dimension by clicking a point anywhere along a five-point Likert-type scale. Each participants' scores for a single video were transformed to a range of 0-100 and averaged. A reliability analysis indicated very high internal reliability (Cronbach's α = 0.963).

E. Participants

50 participants were recruited from Amazon Mechanical Turk (19 F, 31 M). Participants ranged in age from 19 to 69 (M=39.07,SD=11.35). None had participated in any previous studies from our laboratory under the account used.

F. Analysis

Data analysis was performed within a Bayesian analysis framework using the JASP 0.8.5.1 [83] software package, using the default settings as justified by Wagenmakers et al. [87]. All data files are available at tinyurl.com/hri19data. For each measure, a repeated measures analysis of variance (RM-ANOVA) [22], [60], [71] was performed, using communication style, ambiguity, and distance as random factors[7]. Baws factors [55] were then computed for each candidate main effect and interaction, indicating (in the form of a Bayes Factor) for that effect the evidence weight of all candidate models including that effect compared to the evidence weight of all candidate models not including that effect, i.e.

$$\frac{\sum_{m \in M | e \in m} P(m|data)}{\sum_{m \in M | e \notin m} P(m|data)},$$

where e is an effect under consideration, and m is a candidate model in the space of candidate models M.

When sufficient evidence was found in favor of a main effect of communication style (a three-level factor), the results were further analyzed using a post-hoc Bayesian t-test [49], [89] with a default Cauchy prior (center=0, r=$\frac{\sqrt{2}}{2}$=0.707).

While the Bayesian statistical approach has become commonplace in the Cognitive Science and Psychology communities, it is still rare in the Human-Robot Interaction community, and as such we will briefly describe the benefits of this approach. First, the use of a Bayesian approach to statistical analysis provides some robustness to sample size (as it is not grounded in the central limit theorem). Second, the Bayesian approach allows for investigators to examine the evidence for and against hypotheses (whereas the frequentest approach can only quantify evidence towards rejection of the null hypothesis). Third, the Bayesian approach does not require reliance on p-values used in Null Hypothesis Significance Testing (NHST) which have recently come under considerable scrutiny [10], [79], [81]. Finally, we intend for the present study to be the first in a line of studies investigating the effectiveness of mixed reality deictic gesture, and the Bayesian framework facilitates the use of previous study results to construct informative priors so that our experiments may build upon our previous findings rather than starting anew.

IV. Results

A. Accuracy

We hypothesized **(H1)** that accuracy would only drop when ambiguous complex noun phrases were used without an associated mixed reality deictic gesture (i.e., in the complex reference condition). Our results provided extreme evidence in favor of an effect of communication style (Bf 5.626e28)[8] and ambiguity (Bf 2.380e7), and for interactions between communication style and both ambiguity (Bf 1.521e13) and distance (Bf 44577.358). In addition, strong evidence was found in favor of a three-way interaction (22.183).

1) Main effect: Communication style: Post-hoc analysis provided extreme evidence for differences in accuracy, specifically between the complex reference condition (M=0.605,SD=0.49) and both the mixed reality deictic gesture condition (M=0.92,SD=0.272) (Bf 1.129e15) and the complex reference + mixed reality deictic gesture condition (M=0.925,SD=0.264) (Bf 4.728e13). This suggests that the use of complex reference by itself was significantly less effective than mixed reality deictic gesture.

2) Main effect: Ambiguity: Our results also suggest that participants' accuracy was worse when the robot was referring to an ambiguous referent (M=0.743,SD=0.438) than when it was referring to an unambiguous referent (M=0.89,SD=0.313).

[7]Note here that again, we did not expect to see effects of distance based on any of our hypotheses; but we*would* expect to see effects of distance in future experiments in which we plan to compare real and simulated robotic arms. We wanted this experiment to be directly comparable to those future experiments, so we selected a set of analyses that could be performed in both the current and future experiments.

[8]Bayes Factors above 100 indicate extreme evidence in favor of a hypothesis [9]. Here, for example, our Baws Factor Bf of 5.626e28 suggests that our data were 5.626e28 times more likely to be generated under models in which communication style is included than under those in which it is not.

3) Interaction: Communication style and ambiguity: These results are clarified by the interaction found between communication style and ambiguity: performance was only *much* worse when using ambiguous complex references without an associated gesture. This confirms hypothesis **H1**.

4) Interaction: Communication style and distance: Our results demonstrate that when a target referent was close to the robot, using a complex reference alone significantly harmed performance more than when the referent was far away.

5) Interaction: Communication style, ambiguity, and distance: This effect is further clarified through the three-way interaction, which shows that performance drops only occurred when the reference was ambiguous, as shown in Fig. 2.

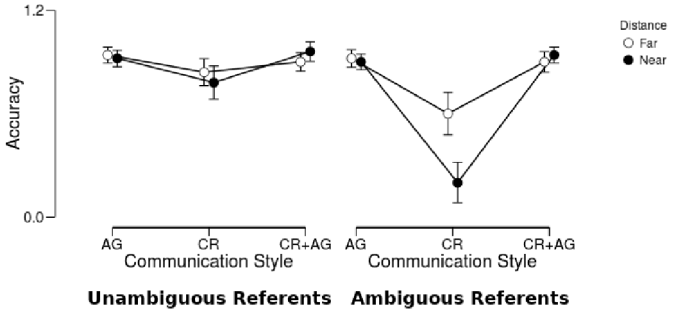

Fig. 2: Effect of communication style (Augmented Gesture (AG), vs Complex Reference (CR) vs both (CR+AG)), referent ambiguity and referent distance on participant accuracy.

B. Reaction Time

We hypothesized **(H2.1)** that reaction time would drop when mixed reality deictic gesture was used, as it would allow target referents to be disambiguated even before speech began, and **(H2.2)** that reaction time would increase when a reference was ambiguous. No results were found in favor of our hypotheses: in fact, our analysis provided strong evidence against a main effect of ambiguity and against any interaction effects. Median reaction time was 7.7 seconds.

C. Effectiveness

We hypothesized **(H3.1)** that perceived effectiveness would be higher when mixed reality deictic gesture was used, especially **(H3.2)** when used in combination with complex reference, and **(H3.3)** when the target referent was ambiguous. Our results provided extreme evidence in favor of main effects of communication style (Bf 1.601e36) and ambiguity (Bf 216.516), and for an interaction between communication style and ambiguity (Bf 1.04e6).

1) Main effect: Communication style: Post-hoc analysis provided extreme evidence in favor of a difference in perceived effectiveness between all three communication styles (mixed reality deictic gesture (M=74.17,SD=23.59) vs. complex reference (M=59.67,SD=27.30) (Bf 2.038e7); mixed reality deictic gesture vs. complex reference + mixed reality deictic gesture (M=87.50,SD=17.08) (Bf 1.462e10); complex reference vs complex reference + mixed reality deictic gesture (Bf 1.581e23)). Specifically, our results show a strong perceived ordering in effectiveness: complex reference < mixed reality deictic gesture < complex reference + mixed reality deictic gesture. This confirms hypotheses **H3.1** and **H3.2**.

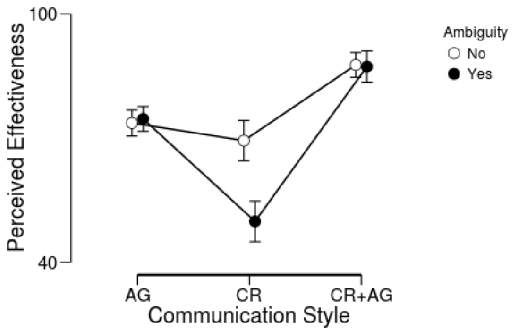

Fig. 3: Effect of communication style (Augmented Gesture (AG), vs Complex Reference (CR) vs both (CR+AG)) and referent ambiguity on perceived effectiveness.

2) Main effect: Ambiguity: In addition, our results showed that robots were perceived as much less effective when describing an ambiguous referent (M=70.63,SD=26.98) than it was when describing an unambiguous referent (M=76.93,SD=23.92).

3) Interaction: Communication style and ambiguity: These results are clarified by examining the observed interaction between communication style and ambiguity, which suggests that while the robot was perceived as less effective when using complex reference alone even when the referent was unambiguous, the robot was perceived as *much* less effective when using complex reference alone to describe ambiguous targets, as seen in Fig. 3. This confirms hypothesis **H3.3**.

D. Likability

We hypothesized that robots' perceived likability would correlate with their effectiveness, and accordingly, that **(H4.1)** perceived likability would be higher when mixed reality deictic gesture was used, **(H4.2)** especially in conjunction with complex reference, and **(H4.3)** when the target referent was ambiguous. Our results provided extreme evidence in favor of a main effect of communication (Bf 5.986e9), and moderate evidence in favor of an effect of ambiguity (Bf 3.088) or an interaction between communication and ambiguity (Bf 7.985).

1) Main effect: Communication style: Post-hoc analysis provided extreme evidence in favor of a difference in likability between the use of complex reference *and* mixed reality deictic gesture (M=69.68, SD=19.27) and the use of *either* complex reference (M=61.35, SD=22.40) (Bf 81289.052) *or* mixed reality deictic gesture (M=60.11, SD=19.64) (Bf 9.940e7). This suggests that participants much more strongly liked the robot when it used both communication styles in combination, confirming hypothesis **H4.1**.

2) Main effect: Ambiguity: Our results suggested that participants liked the robot less when it referred to ambiguous referents. This is clarified by our final interaction effect.

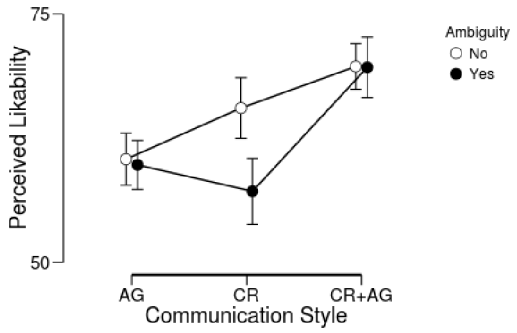

Fig. 4: Effect of communication style (Augmented Gesture (AG), vs Complex Reference (CR) vs both (CR+AG)) and referent ambiguity on likability

3) Interaction: Communication style and ambiguity: This interaction effect suggested that when the robot's target referent was unambiguous, participants exhibited a likability preference ordering of: mixed reality deictic gesture < complex reference < mixed reality deictic gesture + complex reference; but when the robot's target referent ambiguous, participants particularly disliked the use of complex reference alone (which is unsurprising given that in such cases complex reference alone did not allow the target to be properly disambiguated). These findings, as seen in Fig. 4, confirming hypotheses **H4.3** and partially supporting **H4.2**.

V. Discussion

Our results suggest that mixed reality deictic gestures may be an accurate, likable, and effective communication strategy for human-robot interaction, much the same as traditional physical deictic gestures. In this section, we will discuss these results in detail, and leverage them to produce design guidelines for enabling mixed reality deictic gestures.

A. Objective Effectiveness of Mixed Reality Deictic Gesture

Our first and second hypotheses considered the objective effectiveness of mixed reality deictic gestures. Specifically, we hypothesized (**H2.1**) that mixed reality deictic gestures would facilitate faster human reference resolution, especially in the case of ambiguous referents (**H2.2**) – for which referents we also hypothesized that mixed reality deictic gesture would enable increased accuracy (**H1**). Our results did indeed suggest that participants had better accuracy in selecting ambiguous referents when mixed reality deictic gestures were used, and especially when referents were ambiguous (supporting **H1**). This is not particularly surprising, as when complex reference alone was used to refer to otherwise ambiguous referents, the specific descriptions we used were not themselves sufficient to disambiguate those referents. Specifically, to appropriately control language complexity, all instances of complex reference took the form 'Look at that {color} {shape}. When a referent was ambiguous (i.e., there were more than one object of that color and shape), clearly this expression itself was still ambiguous. In future experiments, it will be valuable to use a complex reference condition that more fully aligns with the "fully articulated" baseline used by Sauppé and Mutlu [75], which sacrifices control over linguistic complexity for assurance of complete disambiguation[9].

But while our first hypothesis was supported, no effects on reaction time were observed, thus failing to support **H2**. As median reaction time was 7.7 seconds for videos that were around 5-6 seconds in length, this suggests that participants nearly uniformly waited until videos completed before selecting their targets, and were not hindered by ambiguity. We note, however, that our timestamps may have simply been too noisy a signal, or more likely, that despite our instructions to click on target referents as soon as they were identified, participants may simply not have been aware of the ability or benefit of doing so. In future experiments, it may be valuable to re-examine reaction time, potentially providing participants with "points" based on speed of response, and letting them know after each video whether or not they selected the correct referent. Alternatively, in future work it could be interesting to gain an even more fine-grained measure of how mixed reality deictic reference affects reaction time in complex, multi-entity reference, using eye-tracking techniques such as those employed in Visual World-paradigmatic experiments [46].

In addition, we found a surprising interaction between communication style and distance. We believe that this finding may best be explained by imagining an attentional cone extending in front of the robot. Several theories of qualitative spatial reference (e.g., Ternary Point Configuration Calculus [59]) consider one entity to be "in front" of another if it falls within just such a cone. Our results suggest that when participants had to choose between options that had not been fully disambiguated, they were biased towards options that could be considered to be "in front" of the robot because they fell within that cone. Because of the conic nature of this region, all objects far from the robot may have been considered "in front" of the robot, yielding no bias for any particular distant object, whereas only some of the objects close to the robot would have been considered "in front" of it, yielding a bias towards those objects. This led to poor accuracy in cases of ambiguity where the "true" target referent did not fall within that attentional cone. We would also note that our experimental design uniquely enabled us to identify this interaction; no such interaction was observed by Sauppé and Mutlu because their experimental design did not allow distance and ambiguity to be simultaneously investigated.

B. Subjective Perceptions of Mixed Reality Deictic Gesture

Our third and fourth hypotheses considered the subjective perception of mixed reality deictic gestures. Specifically, we hypothesized (**H3.1**) that participants would perceive the robot to be more effective when mixed reality deictic gesture was

[9]This draws an interesting contrast with Sauppé and Mutlu's experiment, in which the fully articulated baseline was fully disambiguating, but the majority of the deictic gestures examined were *not*; the opposite pattern as observed in our own experimental design.

used, especially when used in conjunction with complex reference (**H3.2**), and when used to refer to an otherwise ambiguous referent (**H3.3**). In addition, we hypothesized (**H4.1**) that participants would perceive the robot to be more likable when mixed reality deictic gesture was used, especially when used in conjunction with complex reference (**H4.2**), and when used to refer to an otherwise ambiguous referent (**H4.3**)

Our results supported all of these hypotheses, with the possible exception of **H4.2**, in that when the target referent would *not* have been otherwise ambiguous, participants actually reported liking the robot more when complex reference alone was used than when the mixed reality deictic gesture alone (that is, accompanied only by a minimally articulated verbal reference) was used. This serves to emphasize that, like physical gesture, mixed reality deictic gesture should be used to supplement rather than replace natural language (excepting extreme circumstances). However, clearly these differences may be exaggerated by the same features of our complex references that potentially exaggerated the accuracy effects.

C. Analysis with respect to Gestural Frameworks

In this paper, we conducted the first exploration of the real and perceived effectiveness of mixed reality deictic gesture. However, as evident within our framework of mixed reality deictic gestures [95], the gestural form examined is but one among many possible forms possible within mixed reality environments. Moreover, while we presented the allocentric gestural form examined in this paper as an alternative to egocentric gesture, of which pointing was presented as the archtypical gesture, it is not clear whether "circling" truly counts as the allocentric equivalent of pointing, especially given the wide variety of physical deictic forms examined by Sauppé and Mutlu [75], who found similar effects of ambiguity and communication style on accuracy in the cases of exhibiting and presenting. Of these, we believe presenting is a closer analogue for allocentric gesture, as, unlike exhibiting, presenting can be used not only to refer to nearby entities, but also to refer to entities out of "arm range", as would allocentric gesture. Finally, Sauppé and Mutlu found significant effects of exhibiting on perceived effectiveness, as we did with allocentric gesture, further strengthening this proposed analogy.

It is also interesting to consider our findings through the lens of our gestural framework [95]. We previously predicted allocentric and perspective-free gestures as being of high static legibility [42] and high dynamic legibility [26], when compared to the other gestural categories. Our findings with respect to accuracy support the first prediction, but our findings with respect to reaction time are not sufficient to support the second. In future work, it will be necessary to explore the other categories of mixed reality deictic gesture beyond allocentric gesture, including physical (egocentric) gestures; while in this study we qualitatively demonstrated similar benefits of allocentric gesture to the benefits of egocentric gesture observed by Sauppé and Mutlu, a direct, empirical, quantitative comparison will be necessary in future work.

D. Design Guidelines

Finally, our results suggest several guidelines for robot designers seeking to enable allocentric mixed reality deictic gesture as a robotic communication strategy.

1) Allocentric mixed reality deictic gestures may be of increased benefit in contexts where disambiguation through language alone is difficult or impossible, such as when describing a specific tree along a treeline.
2) Allocentric mixed reality deictic gestures may be of increased benefit in contexts where the intended target falls outside of what humans may perceive as the robot's real or ostensible attentional cone.
3) Allocentric mixed reality deictic gestures alone are perceived as more effective than language alone, but result in the robot being perceived less well overall. Accordingly, it is beneficial to pair such gestures with complex reference, rather than as a replacement for complex language entirely. Even if this strategy does not glean efficiency benefits traditionally associated with gesture (a point for future investigation, see also below), it may yet glean the other benefits associated with gesture, as discussed in this paper.

VI. Conclusion

In this work we explored the actual and perceived effectiveness of allocentric mixed reality deictic gestures in multi-modal robot communication. Building off the findings presented in this paper, we see several promising directions for future work. We are currently preparing to run a follow-up experiment in which complex references are fully articulated. This will allow us to investigate the effects of allocentric gesture when language alone would be enough for accurate performance. In this experiment, we also hope to collect a more accurate measure of reaction time. If this experiment yields positive results with respect to reaction time, it will then produce a tradeoff that must be examined, i.e., between the efficiency of minimally articulated language and the likability of fully articulated language. If reaction time effects are indeed found in our followup experiment, then we plan to investigate this tradeoff both experimentally and algorithmically. In addition, as described above, it will be important to investigate a wider variety of mixed reality deictic gestures, with respect to both Sauppé and Mutlu [75] and our own [95] frameworks, and to investigate that wider array of gestures with respect to the specific framework dimensions we previously highlighted. We also hope to investigate the effect of different classes of mixed reality deictic gesture when used by robots of differing morphologies, e.g., robots that lack arms vs. robots that have arms they could use instead of (or in conjunction with) allocentric gestures. Finally, we are currently in the process of implementing different mixed reality deictic gestures on the Microsoft Hololens. Once these gestures are integrated with our previous work on natural language generation [92], it will be critical to attempt to replicate the results of this experiment using that integrated system, for increased external validity.

REFERENCES

[1] Henny Admoni, Thomas Weng, Bradley Hayes, and Brian Scassellati. Robot nonverbal behavior improves task performance in difficult collaborations. In *Proceedings of the 11th ACM/IEEE International Conference on Human Robot Interaction (HRI)*, pages 51–58. IEEE Press, 2016.

[2] Henny Admoni, Thomas Weng, and Brian Scassellati. Modeling communicative behaviors for object references in human-robot interaction. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 3352–3359. IEEE, 2016.

[3] Heni Ben Amor, Ramsundar Kalpagam Ganesan, Yash Rathore, and Heather Ross. Intention projection for human-robot collaboration with mixed reality cues. In *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*, 2018.

[4] Rasmus S Andersen, Ole Madsen, Thomas B Moeslund, and Heni Ben Amor. Projecting robot intentions into human environments. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 294–301. IEEE, 2016.

[5] Ronald Azuma, Yohan Baillot, Reinhold Behringer, Steven Feiner, Simon Julier, and Blair MacIntyre. Recent advances in augmented reality. Technical report, Naval Research Lab, Washington, DC, 2001.

[6] Ronald T Azuma. A survey of augmented reality. *Presence: Teleoperators & Virtual Environments*, 6(4):355–385, 1997.

[7] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics (IJSR)*, 1(1):71–81, 2009.

[8] Elizabeth Bates. *Language and context: The acquisition of pragmatics*. Academic Press, 1976.

[9] James O Berger and Luis R Pericchi. The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

[10] James O Berger and Thomas Sellke. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American statistical Association*, 82(397):112–122, 1987.

[11] Mark Billinghurst, Adrian Clark, and Gun Lee. A survey of augmented reality. *Foundations and Trends in Human–Computer Interaction*, 8(2-3):73–272, 2015.

[12] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman, and Matt Berlin. Effects of nonverbal communication on efficiency and robustness in human-robot teamwork. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 708–713. IEEE, 2005.

[13] Inge Bretherton, Sandra McNew, and Marjorie Beeghly-Smith. Early person knowledge as expressed in gestural and verbal communication: When do infants acquire a "theory of mind". *Infant social cognition*, pages 333–373, 1981.

[14] Andrew G Brooks and Cynthia Breazeal. Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-Robot Interaction (HRI)*, pages 297–304. ACM, 2006.

[15] Elizabeth Cha, Yunkyung Kim, Terrence Fong, and Maja J Mataric. A survey of nonverbal signaling methods for non-humanoid robots. *Foundations and Trends in Robotics*, 6(4):211–323, 2018.

[16] Tathagata Chakraborti, Andrew Dudley, and Subbarao Kambhampati. v2v communication for augmenting reality enabled smart huds to increase situational awareness of drivers. In *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*, 2018.

[17] Tathagata Chakraborti, Sarath Sreedharan, Anagha Kulkarni, and Subbarao Kambhampati. Alternative modes of interaction in proximal human-in-the-loop operation of robots. *arXiv preprint arXiv:1703.08930*, 2017.

[18] Mark Cheli, Jivko Sinapov, Ethan E. Danahy, and Chris Rogers. Towards an augmented reality framework for k-12 robotics education. In *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*, 2018.

[19] Aaron St Clair, Ross Mead, and Maja J Matarić. Investigating the effects of visual saliency on deictic gesture production by a humanoid robot. In *Proceedings of the 20th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 210–216. IEEE, 2011.

[20] Eve V Clark and CJ Sengul. Strategies in the acquisition of deixis. *Journal of child language*, 5(3):457–475, 1978.

[21] Herbert H Clark. Coordinating with each other in a material world. *Discourse studies*, 7(4-5):507–525, 2005.

[22] Martin J Crowder. *Analysis of repeated measures*. Routledge, 2017.

[23] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410, 2013.

[24] Antonella De Angeli, Walter Gerbino, Giulia Cassano, and Daniela Petrelli. Visual display, pointing, and natural language: the power of multimodal interaction. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI)*, pages 164–173. ACM, 1998.

[25] Jill De Villiers. The interface of language and theory of mind. *Lingua*, 117(11):1858–1878, 2007.

[26] Anca D Dragan, Kenton CT Lee, and Siddhartha S Srinivasa. Legibility and predictability of robot motion. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 301–308. IEEE Press, 2013.

[27] Rui Fang, Malcolm Doering, and Joyce Y Chai. Embodied collaborative referring expression generation in situated human-robot interaction. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 271–278. ACM, 2015.

[28] Charles J Fillmore. Towards a descriptive framework for spatial deixis. *Speech, place and action: Studies in deixis and related topics*, pages 31–59, 1982.

[29] Jared A Frank, Matthew Moorhead, and Vikram Kapila. Mobile mixed-reality interfaces that enhance human–robot interaction in shared spaces. *Frontiers in Robotics and AI*, 4:20, 2017.

[30] Ramsundar Kalpagam Ganesan, Yash K Rathore, Heather M Ross, and Heni Ben Amor. Better teaming through visual cues. *IEEE Robotics & Automation Magazine*, 2018.

[31] Albert Gatt and Patrizia Paggio. What and where: An empirical investigation of pointing gestures and descriptions in multimodal referring actions. In *Proceedings of the 14th European Workshop on Natural Language Generation*, pages 82–91, 2013.

[32] Albert Gatt and Patrizia Paggio. Learning when to point: A data-driven approach. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 2007–2017, 2014.

[33] Arthur M Glenberg and Mark A McDaniel. Mental models, pictures, and text: Integration of spatial and verbal information. *Memory & Cognition*, 20(5):458–460, 1992.

[34] Susan Goldin-Meadow. The role of gesture in communication and thinking. *Trends in Cognitive Sciences (TiCS)*, 3(11):419–429, 1999.

[35] Scott A Green, Mark Billinghurst, XiaoQi Chen, and J Geoffrey Chase. Human-robot collaboration: A literature review and augmented reality approach in design. *International Journal of Advanced Robotic Systems*, 5(1):1, 2008.

[36] Marianne Gullberg. Deictic gesture and strategy in second language narrative. In *Workshop on the Integration of Gesture in Language and Speech*, pages 155–164. Applied Science and Engineering Laboratories, University of Delaware, 1996.

[37] Todd M Gureckis, Jay Martin, John McDonnell, Alexander S Rich, Doug Markant, Anna Coenen, David Halpern, Jessica B Hamrick, and Patricia Chan. psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, 48(3):829–842, 2016.

[38] Simon Harrison. The creation and implementation of a gesture code for factory communication. *Proceedings of the 2nd International Conference on Gesture in Speech and Interaction (GESPIN)*, 2011, 2011.

[39] Hooman Hedayati, Michael Walker, and Daniel Szafir. Improving collocated robot teleoperation with augmented reality. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 78–86. ACM, 2018.

[40] Leanne Hirshfield, Tom Williams, Natalie Sommer, Trevor Grant, and Senem Velipasalar Gursoy. Workload-driven modulation of mixed-reality robot-human communication. *Workshop on Modeling Cognitive Processes from Multimodal Data at the International Conference on Multimodal Interaction*, 2018.

[41] Catherine Hobaiter, David A Leavens, and Richard W Byrne. Deictic gesturing in wild chimpanzees (pan troglodytes)? some possible cases. *Journal of Comparative Psychology*, 128(1):82, 2014.

[42] Rachel M Holladay, Anca D Dragan, and Siddhartha S Srinivasa. Legible robot pointing. In *Proceedings of the 23rd IEEE International*

Symposium on Robot and Human Interactive Communication (RO-MAN), pages 217–223. IEEE, 2014.

[43] Rachel M Holladay and Siddhartha S Srinivasa. Rogue: Robot gesture engine. In *Proceedings of the AAAI Spring Symposium Series*, 2016.

[44] William D Hopkins, Jared P Taglialatela, and David A Leavens. Chimpanzees differentially produce novel vocalizations to capture the attention of a human. *Animal behaviour*, 73(2):281–286, 2007.

[45] Chien-Ming Huang and Bilge Mutlu. Modeling and evaluating narrative gestures for humanlike robots. In *Robotics: Science and Systems (RSS)*, pages 57–64, 2013.

[46] Falk Huettig, Joost Rommers, and Antje S Meyer. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, 137(2):151–171, 2011.

[47] Jana M Iverson and Susan Goldin-Meadow. Gesture paves the way for language development. *Psychological science*, 16(5):367–371, 2005.

[48] MerryAnn Jancovic, Shannon Devoe, and Morton Wiener. Age-related changes in hand and arm movements as nonverbal communication: Some conceptualizations and an empirical exploration. *Child Development*, pages 922–928, 1975.

[49] Harold Jeffreys. Significance tests when several degrees of freedom arise simultaneously. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, pages 161–198, 1938.

[50] Sotaro Kita. Pointing: A foundational building block of human communication. In *Pointing*, pages 9–16. Psychology Press, 2003.

[51] David A Leavens. Manual deixis in apes and humans. *Interaction Studies*, 5(3):387–408, 2004.

[52] Stephen C Levinson. Deixis. In *The handbook of pragmatics*, pages 97–121. Blackwell, 2004.

[53] Phoebe Liu, Dylan F Glas, Takayuki Kanda, Hiroshi Ishiguro, and Norihiro Hagita. It's not polite to point: generating socially-appropriate deictic behaviors towards people. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 267–274. IEEE Press, 2013.

[54] Peter Marler. Primate vocalization: affective or symbolic? In *Speaking of apes*, pages 221–229. Springer, 1980.

[55] S. Mathôt. Bayes like a baws: Interpreting bayesian repeated measures in JASP [blog post]. https://www.cogsci.nl/blog/interpreting-bayesian-repeated-measures-in-jasp, May 2017.

[56] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *AAAI*, pages 2556–2563, 2014.

[57] Paul Milgram and Fumio Kishino. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77(12):1321–1329, 1994.

[58] Paul Milgram, Shumin Zhai, David Drascic, and Julius Grodski. Applications of augmented reality for human-robot communication. In *Intelligent Robots and Systems' 93, IROS'93. Proceedings of the 1993 IEEE/RSJ International Conference on*, volume 3, pages 1467–1472. IEEE, 1993.

[59] Reinhard Moratz and Marco Ragni. Qualitative spatial reasoning about relative point position. *Journal of Visual Languages & Computing*, 19(1):75–98, 2008.

[60] RD Morey and JN Rouder. Bayesfactor (version 0.9. 9), 2014.

[61] Sigrid Norris. Three hierarchical positions of deictic gesture in relation to spoken language: a multimodal interaction analysis. *Visual Communication*, 10(2):129–147, 2011.

[62] Yusuke Okuno, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. Providing route directions: design of robot's utterance, gesture, and timing. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction*, pages 53–60. ACM, 2009.

[63] Daniela K O'Neill. Two-year-old children's sensitivity to a parent's knowledge state when making requests. *Child development*, 67(2):659–677, 1996.

[64] Daniel Otte. Effects and functions in the evolution of signaling systems. *Annual Review of Ecology and Systematics*, 5(1):385–417, 1974.

[65] Leah Perlmutter, Eric Kernfeld, and Maya Cakmak. Situated language understanding with human-like and visualization-based transparency. In *Robotics: Science and Systems*, 2016.

[66] Christopher Peters, Fangkai Yang, Himangshu Saikia, Chengjie Li, and Gabriel Skantze. Towards the use of mixed reality for hri design via virtual robots. In *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*, 2018.

[67] Simone Pika, Katja Liebal, Josep Call, and Michael Tomasello. Gestural communication of apes. *Gesture*, 5(1):41–56, 2005.

[68] Paul Piwek. Salience in the generation of multimodal referring acts. In *Proceedings of the 2009 international conference on Multimodal interfaces*, pages 207–210. ACM, 2009.

[69] Christopher Reardon, Kevin Lee, and Jonathan Fink. Come see this! augmented reality to enable human-robot cooperative search. In *Proceedings of the IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 1–7. IEEE, 2018.

[70] Eric Rosen, David Whitney, Elizabeth Phillips, Gary Chien, James Tompkin, George Konidaris, and Stefanie Tellex. Communicating robot arm motion intent through mixed reality head-mounted displays. *arXiv preprint arXiv:1708.03655*, 2017.

[71] Jeffrey N Rouder, Richard D Morey, Paul L Speckman, and Jordan M Province. Default bayes factors for anova designs. *Journal of Mathematical Psychology*, 56(5):356–374, 2012.

[72] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. Effects of gesture on the perception of psychological anthropomorphism: a case study with a humanoid robot. In *International Conference on Social Robotics*, pages 31–41. Springer, 2011.

[73] Maha Salem, Stefan Kopp, Ipke Wachsmuth, and Frank Joublin. Towards an integrated model of speech and gesture production for multi-modal robot behavior. In *19th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 614–619. IEEE, 2010.

[74] Maha Salem, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. Generation and evaluation of communicative robot gesture. *International Journal of Social Robotics*, 4(2):201–217, 2012.

[75] Allison Sauppé and Bilge Mutlu. Robot deictics: How gesture and context shape referential communication. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 342–349. ACM, 2014.

[76] E Sue Savage-Rumbaugh. Language as a cause-effect communication system. *Philosophical psychology*, 3(1):55–76, 1990.

[77] Manfred Schönheits and Florian Krebs. Embedding ar in industrial hri applications. In *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*, 2018.

[78] Elena Sibirtseva, Dimosthenis Kontogiorgos, Olov Nykvist, Hakan Karaoguz, Iolanda Leite, Joakim Gustafson, and Danica Kragic. A comparison of visualisation methods for disambiguating verbal requests in human-robot interaction. In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018.

[79] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

[80] Daniele Sportillo, Alexis Paljic, Luciano Ojeda, Giacomo Partipilo, Philippe Fuchs, and Vincent Roussarie. Training semi-autonomous vehicle drivers with extended reality. In *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*, 2018.

[81] Jonathan AC Sterne and George Davey Smith. Sifting the evidence – what's wrong with significance tests? *Physical Therapy*, 81(8):1464–1469, 2001.

[82] Neil Stewart, Jesse Chandler, and Gabriele Paolacci. Crowdsourcing samples in cognitive science. *Trends in Cognitive Sciences (TiCS)*, 2017.

[83] JASP Team. Jasp (version 0.8.5.1)[computer software], 2018.

[84] Michael Tomasello. Why don't apes point? *Trends In Linguistics Studies And Monographs*, 197:375, 2008.

[85] Ielka Francisca Van Der Sluis. *Multimodal Reference, Studies in Automatic Generation of Multimodal Referring Expressions*. PhD thesis, University of Tilburg, 2005.

[86] DWF Van Krevelen and Ronald Poelman. A survey of augmented reality technologies, applications and limitations. *International journal of virtual reality*, 9(2):1, 2010.

[87] EJ Wagenmakers, J Love, M Marsman, T Jamil, A Ly, and J Verhagen. Bayesian inference for psychology, Part II: Example applications with JASP. *Psychonomic Bulletin and Review*, 25(1):35–57, 2018.

[88] Michael Walker, Hooman Hedayati, Jennifer Lee, and Daniel Szafir. Communicating robot motion intent with augmented reality. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 316–324. ACM, 2018.

[89] Peter H Westfall, Wesley O Johnson, and Jessica M Utts. A bayesian perspective on the bonferroni adjustment. *Biometrika*, 84(2):419–427, 1997.

[90] David Whitney, Eric Rosen, James MacGlashan, Lawson LS Wong, and Stefanie Tellex. Reducing errors in object-fetching interactions through social feedback. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, 2017.

[91] Tom Williams. A framework for robot-generated mixed-reality deixis. In *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*, 2018.

[92] Tom Williams and Matthias Scheutz. Referring expression generation under uncertainty: Algorithm and evaluation framework. In *Proceedings of the 10th International Conference on Natural Language Generation (INLG)*, 2017.

[93] Tom Williams, Daniel Szafir, Tathagata Chakraborti, and Heni Ben Amor. Report on the 1st international workshop on virtual, augmented, and mixed reality for human-robot interaction (VAM-HRI). *AI Magazine*, 2018 (forthcoming).

[94] Tom Williams, Daniel Szafir, Tathagata Chakraborti, and Heni Ben Amor. Virtual, augmented, and mixed reality for human-robot interaction. In *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 403–404. ACM, 2018.

[95] Tom Williams, Nhan Tran, Josh Rands, and Neil T. Dantam. Augmented, mixed, and virtual reality enabling of robot deixis. In *Proceedings of the 10th International Conference on Virtual, Augmented, and Mixed Reality (VAMR)*, 2018.

[96] Tom Williams, Fereshta Yazdani, Prasanth Suresh, Matthias Scheutz, and Michael Beetz. Dempster-shafer theoretic resolution of referential ambiguity. *Autonomous Robots*, 2018.

[97] Fereshta Yazdani, Matthias Scheutz, and Michael Beetz. Guidelines for improving task-based natural language understanding in human-robot rescue teams. In *Proceedings of the 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, Debrecen, Hungary, September 2017.

[98] Feng Zhou, Henry Been-Lirn Duh, and Mark Billinghurst. Trends in augmented reality tracking, interaction and display: A review of ten years of ismar. In *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 193–202. IEEE Computer Society, 2008.

[99] Sebastian Meyer zu Borgsen, Patrick Renner, Florian Lier, Thies Pfeiffer, and Sven Wachsmuth. Improving human-robot handover research by mixed reality techniques. In *Proceedings of the 1st International Workshop on Virtual, Augmented, and Mixed Reality for HRI (VAM-HRI)*, 2018.