Get This! ⇓
Mixed Reality Improves Robot Communication Regardless of Mental Workload

Nhan Tran
Colorado School of Mines
Department of Computer Science
nttran@mines.edu

Trevor Grant
University of Colorado
Department of Computer Science
trevor.grant@colorado.edu

Thao Phung
Colorado School of Mines
Department of Computer Science
thaophung@mines.edu

Leanne Hirshfield
University of Colorado
Department of Computer Science
leanne.hirshfield@colorado.edu

Christopher Wickens
Colorado State University
Department of Psychology

Tom Williams
Colorado School of Mines
Department of Computer Science
twilliams@mines.edu

ABSTRACT

We present the first experiment analyzing the effectiveness of robot-generated mixed reality gestures using real robotic and mixed reality hardware. Our findings demonstrate how these gestures increase user effectiveness by decreasing user response time during visual search tasks, and show that robots can safely pair longer, more natural referring expressions with mixed reality gestures without worrying about cognitively overloading their interlocutors.

KEYWORDS

augmented reality, mixed reality, cognitive load, deictic gesture

ACM Reference Format:
Nhan Tran, Trevor Grant, Thao Phung, Leanne Hirshfield, Christopher Wickens, and Tom Williams. 2021. Get This! ⇓ Mixed Reality Improves Robot Communication Regardless of Mental Workload. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3434074.3447203

1 INTRODUCTION

Successful human-robot interaction in many domains relies on successful communication. Accordingly, there has been a wealth of research on enabling human-robot communication through natural language [7, 13]. However, just like human-human dialogue, human-robot dialogue is inherently multi-modal, and necessarily involves communication channels other than speech, with human interlocutors regularly using gaze and gesture cues to augment, modify, or replace their natural language utterances. Speakers regularly use deictic gestures such as pointing, for example, to direct interlocutors' attention to objects in the environment, both to reduce the number of words that the speaker must use to refer to their target referents, as well as to lower the cognitive burden imposed on listeners to interpret those utterances.

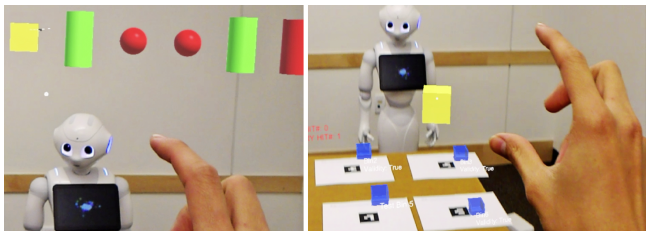

Figure 1: During the experiment, participants play a mixed reality game using the Microsoft HoloLens. The Pepper robot is positioned behind the table, ready to interact.

Due to the prevalence and utility of deictic gestures in situated communication, human-robot interaction researchers have sought to enable robots to understand [6] and generate [10–12] deictic gestures just as humans do. However, the ability to understand and generate deictic gestures comes with hardware requirements that can be onerous or unsatisfiable in certain use cases. While perceiving deictic gestures only requires a camera or depth sensor, generating deictic gestures requires a specific robotic morphology (e.g., expressive robotic arms). This fundamentally limits the gestural capabilities, and thus overall communicative capabilities, of the *majority* of robotic platforms in use today, such as mobile bases used in warehouses, assistive wheelchairs, and unmanned aerial vehicles (UAVs). Moreover, even for robots that do have arms, traditional deictic gestures have fundamental limitations. In contexts such as urban or alpine search and rescue, for example, robots may need to communicate about hard-to-describe and/or highly ambiguous referents in novel, uncertain, and unknown environments.

A scenario that illustrates all of these problems is an aerial robot in a search and rescue context that needs to generate an utterance such as "I found a victim behind *that tree*" [cf. 21]. In this case, the ability to precisely pick out the target tree using some sort of gestural cue would be of great value, as the referring expressions the robot would need to generate without using gesture would likely be convoluted (e.g., "the fourth tree from the left in the clump

HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA

ACM ISBN 978-1-4503-8290-8/21/03…$15.00
https://doi.org/10.1145/3434074.3447203

of trees to the right of the large boulder") or not readily human-understandable (e.g., "the tree 48.2 meters to the northwest").

Unfortunately, it is unlikely that such a UAV would have an arm mounted on it, meaning that physical gesture is not a realistic possibility, no matter its utility. Moreover, even in the unlikely case that the robot had an arm mounted on it, it is unlikely that a traditional pointing gesture generated by such an arm would be able to pinpoint a specific far-off tree.

In this work, we present a *mixed reality* (MR) solution that enables robots to generate effective deictic gestures without imposing any morphological requirements. Specifically, we present the first implementation of the *mixed reality deictic gestures* proposed by Williams et al. [20] on real robotic and mixed reality hardware.

Mixed reality deictic gestures are visualizations that can serve the same purpose as traditional deictic gestures, and which fall within the broad category of *view-augmenting* mixed reality interaction design elements in the Reality-Virtuality Interaction Cube framework [19]. Williams et al. [20] divide these new forms of visual gestures into *perspective-free* gestures that can be projected onto the environment, and *allocentric* gestures (visualized in the perspective of the listener) that can be displayed in teammates' augmented reality (AR) head-mounted displays.

Recent work on perspective-free gestures has focused on the *legibility* of projected gestures [14], while recent work on allocentric gestures has focused on gesture effectiveness when paired with different kinds of language (in virtual online testbeds) [17, 18] and on effectiveness of *ego-sensitive allocentric* gestures such as virtual arms [2, 3]. In this work we focus on this first, (non-egosensitive) allocentric category of mixed reality deictic gesture.

In previous work in this space, Williams et al. [18] [see also 17], demonstrated that (non-ego-sensitive) allocentric mixed reality deictic gestures, at least when tested in a simulated video-based experiment, could increase communication accuracy and efficiency, and, when paired with complex referring expressions, were viewed as more effective and likable than purely linguistic communication. However, to date, mixed reality deictic gestures have only been tested in video-based simulations. In this article, we present the first demonstration of mixed reality deictic gestures generated on actual AR Head-Mounted Displays (the Microsoft Hololens) in the context of task-based human-robot interactions.

Moreover, as previously pointed out by Hirshfield et al. [4], the tradeoffs between language and visual gesture may be highly sensitive to teammates' level and type of cognitive load. For example, Hirshfield et al. [4] suggest that it may not be advantageous to rely heavily on visual communication in contexts with high visual load, or to rely heavily on linguistic communication in contexts with high auditory or working memory load. These intuitions are motivated by prior theoretical work on human information processing, including the Multiple Resource Theory (MRT) by Wickens [15, 16]. In this article, we thus also present the first exploration of the tradeoffs between different forms of mixed reality communication under different levels and types of cognitive load.

2 EXPERIMENT

In this section, we present the design of a human-subject experiment to assess whether different robot communication styles improve participants' task performance under four conditions: high visual perceptual load, high auditory perceptual load, high working memory load, and low overall load.

2.1 Hypotheses

Based on the assumptions that there are different perceptual resources, and that mixed reality deictic gestures employ visual-spatial resources in accordance to MRT, this experiment was designed to test the following hypotheses, which formalize the intuitions of Hirshfield et al. [4].

- **H1** Users under high **visual perceptual load** will perform quickest and most accurately when robots rely on complex natural language without the use of mixed reality deictic gestures.
- **H2** Users under high **auditory perceptual load** will perform quickest and most accurately when robots rely on mixed reality deictic gestures without the use of complex natural language.
- **H3** Users under high **working memory load** will perform quickest and most accurately when robots rely on mixed reality deictic gestures without the use of complex natural language.
- **H4** Users under **low overall load** will perform quickest and most accurately when robots rely on mixed reality deictic gestures paired with complex natural language.

2.2 Experimental Context

In this experiment, participants interacted with a language-capable robot while wearing the Microsoft HoloLens, over a series of trials, with the robot's communication style and the user's cognitive load systematically varying between trials.

Our experiment employed a dual-task paradigm oriented around a tabletop pick-and-place task. Participants view the primary task through the Microsoft HoloLens, allowing them to see virtual bins overlaid over the mixed reality fiducial markers on the table, as well as a panel of blocks above the table that changes every few seconds (Fig. 1). As shown in Fig. 2, the Pepper robot is positioned behind the table, ready to interact with the participant.

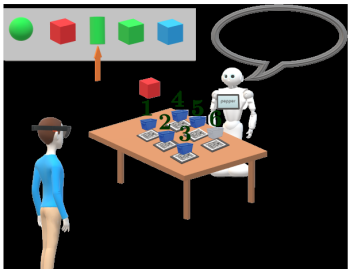

Figure 2: Participants wearing the HoloLens are asked to correctly pick-and-place virtual blocks into virtual bins.

2.3 Experimental Task

Primary Task

The user's *primary task* is to look out for a particular block in the block panel (selected from among *red cube*, *red sphere*, *red cylinder*, *yellow cube*, *yellow sphere*, *yellow cylinder*, *green cube*, *green sphere*, *green cylinder*[1]). These nine blocks were formed by combining three colors (red, yellow, green) with three shapes (cube, sphere, cylinder). Whenever participants see this target block, their task is to pick-and-place it into any one of a particular set of bins. For example, as the game starts, the robot might tell a user that whenever they see a *red cube* they should place it in bins *two or three*.

Two additional factors increase the complexity of this primary task. First, in order to have participants remember the full set of candidate bins, rather than just one particular bin from that set, at every point during the task one random bin is marked as unavailable and greyed out (with the disabled bin changing each time a block is placed in a bin). Second, to create a demanding auditory component to the primary task ensemble, the user hears a series of syllables playing in the task background, is given a target syllable to look out for, and is told that whenever they hear this syllable, the target bins and non-target bins are switched. In other words, the bins that they should consider to place blocks in should be exchanged with those they were previously told to avoid. For example, if the user's target bins from among four bins are bins two and three, and they hear the target syllable, then future blocks will need to be placed instead into bins one and four. The syllables heard are selected from among (*bah*, *beh*, *boh*, *tah*, *teh*, *toh*, *kah*, *keh*, *koh*). These nine syllables were formed by combining three consonant sounds (b,t,k) with three vowel sounds (ah,eh,oh).

Secondary Task

Three times per experiment trial, the participant encounters a secondary task, in which the robot interrupts and utters a new request, asking the participant to move a particular, currently visible block, to a particular, currently accessible bin. Depending on the condition of the experiment trial, the robot's spoken request may be accompanied by a mixed reality gesture. Unlike the long-term primary task that requires participants to remember the initial target block and keep track of the continuously changing target bins during the 90 second round, in the secondary task the robot asks participants to pick a different target block and place it in a different target bin, after which participants can continue with the primary task.

2.4 Experimental Design

We used a Latin square counterbalanced within-subjects experimental design with two within-subjects factors: Cognitive Load (4 loads) and Communication Style (3 styles) (see Fig. 3).

Cognitive Load

Our first independent variable, cognitive load, was manipulated through our primary task. Following Beck and Lavie [5], we manipulated cognitive load by jointly manipulating memory constraints and target/distractor discriminability (cp. [5]), producing four different load profiles: (1) one in which all load was considered low, (2) only working memory load was considered high, (3) only visual perceptual load was considered high, and (4) only auditory perceptual load was considered high.

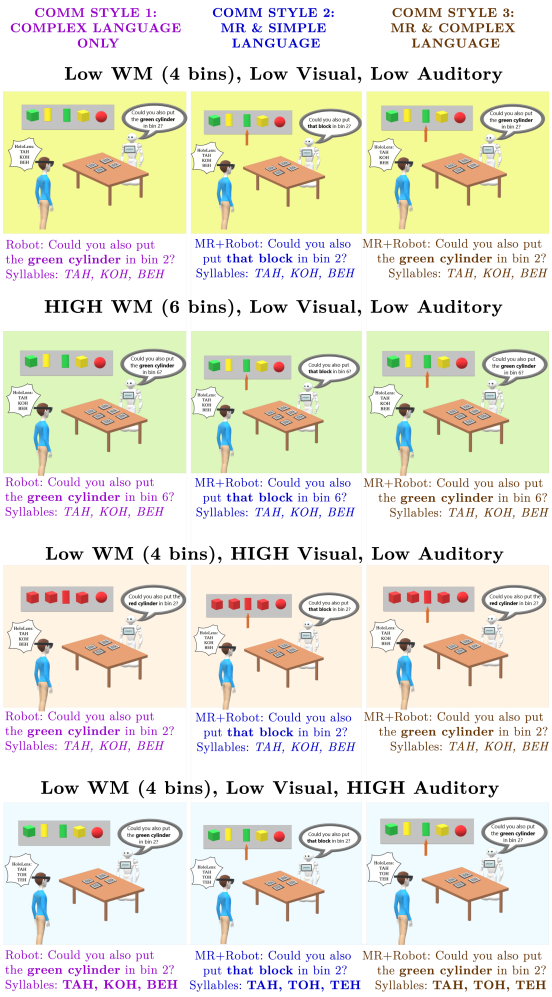

Figure 3: Participants engage in 12 (Latin square counterbalanced) experimental trials formed by combining 4 cognitive load conditions and 3 communication style conditions.

Working memory load was manipulated as follows: In the high working memory load condition, participants were required to remember the identities of three target bins out of a total of six visible bins, producing a total memory load of seven items (an "item" is defined by either a bin and/or an attribute) when including the two properties of the target block (shape and color) and the two properties of the target syllable (consonant and vowel). In all other conditions, participants were only required to remember the identities of two target bins out of a total of four visible bins, producing a total memory load of six items.

[1]These colors were chosen for consistent visual processing, as blue is processed differently within the eye due to spatial and frequency differences of red/green and blue cones. This did mean our task was not accessible to red/green colorblind participants, requiring us to remove from our dataset the data of several colorblind participants.

Visual perceptual load was manipulated as follows: In the high visual perceptual load condition, the target block was always difficult to discriminate from distractors due to sharing one common property with all distractors. For example, if the target block was a red cube, all distractors would be either red or cubes (but not both). In the low visual perceptual load condition, the target block was always easy to discriminate from distractors due to sharing no common properties with any distractors. For example, if the target block was a red cube, no distractors would be red or cubes.

Auditory perceptual load was manipulated as follows: In the high auditory perceptual load condition, the target syllable was always difficult to discriminate from distractors due to sharing one common property with all distractors. For example, if the target syllable was *kah*, all distractors would either start with *k* or end with *ah* (but not both). In the low auditory perceptual load condition, the target syllable was always easy to discriminate from distractors due to sharing no common properties with any distractors. For example, if the target syllable was *kah*, no distractors would either start with *k* or end with *ah*.

Communication Style

Our second independent variable, communication style, was manipulated through our secondary task. Following Williams et al. [17] and Williams et al. [18], we manipulated communication style by having the robot exhibit one of three behaviors:

During experiment blocks associated with the **complex language** communication style condition, the robot referred to objects using full referring expressions needed to disambiguate those objects (e.g., "the red sphere").

During blocks associated with the **MR + complex language** communication style condition, the robot referred to objects using full referring expressions (e.g., "the red sphere"), paired with a mixed reality deictic gesture (an arrow drawn over the red sphere).

During blocks associated with the **MR + simple language** communication style condition, the robot referred to objects using minimal referring expressions (e.g., "that block"), paired with a mixed reality deictic gesture (an arrow drawn over the object to which the robot was referring).

Following Williams et al. [17] and Williams et al. [18], we did not examine the use of simple language without MR, as that communication style typically does not enable referent disambiguation, resulting in the user needing to ask for clarification or guess at random between ambiguous options.

3 RESULTS AND DISCUSSION

36 participants were recruited from our university (31 M, 5 F), ranging in age from 18 to 32. For both primary and secondary tasks, we measured perceived mental workload, perceived communicative effectiveness, task accuracy, and task response time.

For each measure, a repeated measures analysis of variance (RM-ANOVA) [1, 8, 9] was performed, using communication style and cognitive load as random factors. Anecdotal to strong evidence was found *against* any effects of communication style or imposed cognitive load on perceived mental workload, perceived communicate effectiveness, task accuracy, and the primary task response time. However, we did find moderate evidence in favor of an effect of mixed reality communication style on secondary task response time, but no effect of or interaction with workload was found.

Our results suggest that the primary benefit of mixed reality deictic gestures in robot communication lies in their ability to increase users' speed at performing a secondary task by reducing the time taken to perform constituent visual searches (especially when paired with complex referring expressions), regardless of the level and type of workload users are experiencing.

These results align with previous work *not* performed in realistic task environments [18], which found that participants demonstrated slower response times when complex language alone was used, with no clear differences between simple and complex language when pairing language with mixed reality deictic gestures, and suggested that people found a robot more likable when it used longer more natural referring expressions. When combined with the results of our experiment, this suggests that robots likely can pair complex referring expression with mixed reality gestures without worrying about cognitively overloading their interlocutors.

Despite these positive findings, our results failed to support our four workload-driven hypotheses. While we originally expected differences between communication styles under different cognitive load profiles, especially based on whether communication style was overall visual or auditory, in fact what we observed is that visual augmentations, especially when paired with complex referring expressions, may *always* be helpful for a secondary task, regardless of level and type of imposed workload.

While this study shows the effect of mixed reality deictic gestures on human's task response time, it has a number of limitations. We found that some participants failed early into the game and completely lost track of what block to place in what bin. Providing some sort of real-time, directive cues might help participant recover from errors. However, the purpose of a challenging primary task is to impose high workload on the participants and to observe how the robot's different communication styles can help enhance its human teammate's task performance while being cognitive overloaded. Additional consideration is needed to design ways that recovery hints can be presented (either visual and/or auditory) without interfering with the imposed workload profiles during the experiment.

Additionally, we received feedback from some participants during the debriefing that they felt the series of syllables playing in the task background (e.g., *bah*, *beh*, *boh*, *tah*, *teh*, *toh*, *kah*, *keh*, *koh*) could easily be misheard. After missing the auditory cue that signals the switch of the target and non-target bins, they started to guess the target bins to attempt to proceed with the primary task. We recommend future studies to use distinguishable sounds instead of these syllables in order to improve auditory discrimination.

Lastly, future research could use a mixed reality device that supports both eye-tracking and hand-tracking such as the newer Microsoft HoloLens 2. This will enable researchers to more precisely capture response time and allow users to interact with holograms through completely natural hand gestures rather than the simple gaze-and-commit (e.g., air tap) interaction of the Hololens 1.

ACKNOWLEDGMENTS

This research was funded in part by NSF grants IIS-1909864 and CNS-1810631.

REFERENCES

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

[2] Thomas Groechel, Zhonghao Shi, Roxanna Pakkar, and Maja J Matarić. 2019. Using Socially Expressive Mixed Reality Arms for Enhancing Low-Expressivity Robots. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 1–8.

[3] Jared Hamilton, Nhan Tran, and Tom Williams. 2020. Tradeoffs Between Effectiveness and Social Perception When Using Mixed Reality to Supplement Gesturally Limited Robots. In *Proceedings of the 3rd International Workshop on Virtual, Augmented, and Mixed Reality for HRI*.

[4] Leanne Hirshfield, Tom Williams, Natalie Sommer, Trevor Grant, and Senem Velipasalar Gursoy. 2018. Workload-driven modulation of mixed-reality robot-human communication. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*. ACM, 3.

[5] Nilli Lavie. 1995. Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human perception and performance* 21, 3 (1995), 451.

[6] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. 2014. Learning from unscripted deictic gesture and language for human-robot interactions. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

[7] Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems* 63 (2015), 22–35.

[8] RD Morey and JN Rouder. 2014. BayesFactor (Version 0.9. 9).

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

[10] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323.

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

[12] Allison Sauppé and Bilge Mutlu. 2014. Robot deictics: How gesture and context shape referential communication. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 342–349.

[13] Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. 2020. Robots That Use Language. *Annual Review of Control, Robotics, and Autonomous Systems* 3 (2020).

[14] Thomas Weng, Leah Perlmutter, Stefanos Nikolaidis, Siddhartha Srinivasa, and Maya Cakmak. 2019. Robot Object Referencing through Legible Situated Projections. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 8004–8010.

[15] Christopher D Wickens. 1991. Processing resources and attention. *Multiple-task performance* 1991 (1991), 3–34.

[16] Christopher D Wickens. 2002. Multiple resources and performance prediction. *Theoretical issues in ergonomics science* 3, 2 (2002), 159–177.

[17] Tom Williams, Matthew Bussing, Sebastian Cabrol, Elizabeth Boyle, and Nhan Tran. 2019. Mixed Reality Deictic Gesture for Multi-Modal Robot Communication. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*.

[18] Tom Williams, Matthew Bussing, Sebastian Cabrol, Ian Lau, Elizabeth Boyle, and Nhan Tran. 2019. Investigating the Potential Effectiveness of Allocentric Mixed Reality Deictic Gesture. In *Proceedings of the 11th International Conference on Virtual, Augmented, and Mixed Reality*.

[19] Tom Williams, Daniel Szafir, and Tathagata Chakraborti. 2019. The Reality-Virtuality Interaction Cube. In *Proceedings of the 2nd International Workshop on Virtual, Augmented, and Mixed Reality for HRI*.

[20] Tom Williams, Nhan Tran, Josh Rands, and Neil T Dantam. 2018. Augmented, mixed, and virtual reality enabling of robot deixis. In *International Conference on Virtual, Augmented and Mixed Reality*. Springer, 257–275.

[21] Tom Williams, Fereshta Yazdani, Prasanth Suresh, Matthias Scheutz, and Michael Beetz. 2019. Dempster-shafer theoretic resolution of referential ambiguity. *Autonomous Robots* 43, 2 (2019), 389–414.