Robot-Generated Mixed Reality Gestures Improve Human-Robot Interaction

Nhan Tran[1(✉)], Trevor Grant[2], Thao Phung[1], Leanne Hirshfield[2], Christopher Wickens[3], and Tom Williams[1(✉)]

[1] Department of Computer Science, Colorado School of Mines, Golden, CO 80401, USA
nttran@alumni.mines.edu,twilliams@mines.edu
[2] Institute for Cognitive Science, University of Colorado Boulder, Boulder, CO 80309, USA
[3] Department of Psychology, Colorado State University, Fort Collins, CO 80523, USA

Abstract. We investigate the effectiveness of robot-generated mixed reality gestures. Our findings demonstrate how these gestures increase user effectiveness by decreasing user response time, and that robots can pair long referring expressions with mixed reality gestures without cognitively overloading users.

1 Introduction

HRI researchers have sought to enable robots to understand [4] and generate [5,6] deictic gestures as humans do. But even for armed robots, traditional deictic gestures have limitations. In search and rescue, for example, robots may need to communicate about hard-to-describe and/or highly ambiguous referents. We present a *mixed reality* solution that enables robots to generate effective *mixed reality deictic gestures* (MRDGs) without morphological requirements.

Per Hirshfield et al. [2], the tradeoffs between language and visual gesture may be highly sensitive to teammates' level and type of cognitive load. It may not be advantageous to rely on visual communication in contexts with high visual load, or to rely 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 Wickens' Multiple Resource Theory (MRT) [7,8]. In this paper, we thus also present the first exploration of mixed reality communication under different levels and types of cognitive load.

2 Experiment

We experimentally assessed whether different robot communication styles improve user task performance under four conditions: high visual perceptual load, high auditory perceptual load, high working memory load, and low overall

This work was funded by NSF grants IIS-1909864 and CNS-1823245.

H. Li et al. (Eds.): ICSR 2021, LNAI 13086, pp. 768–773, 2021.
https://doi.org/10.1007/978-3-030-90525-5_69

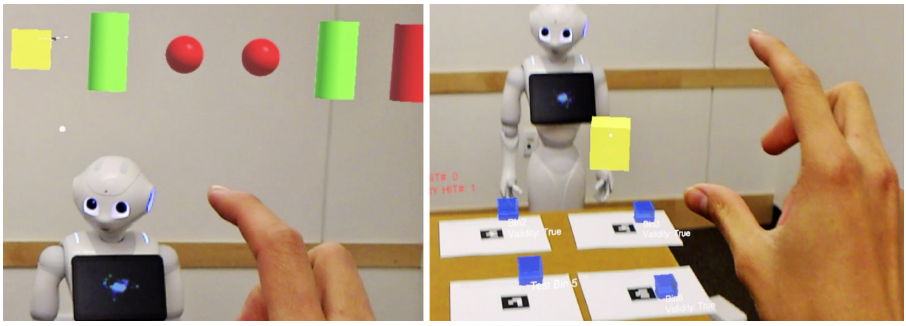

Fig. 1. Participants play a mixed reality game using the Microsoft HoloLens. The Pepper robot interacts with them from behind a table. (Color figure online)

load. On the assumption that there are different perceptual resources, and that MRDGs employ visual-spatial resources in accordance to MRT, we specifically tested four hypotheses, which formalize the intuitions of Hirshfield et al. [2].

H1 Users under high **visual perceptual load** will perform quickest and most accurately when robots use complex natural language without MRDGs.

H2 Users under high **auditory perceptual load** will perform quickest and most accurately when robots use MRDGs without using complex natural language.

H3 Users under high **working memory load** will perform quickest and most accurately when robots use MRDGs without using complex natural language.

H4 Users under **low overall load** will perform quickest and most accurately when robots use MRDGs paired with complex natural language.

2.1 Experimental Context

Participants interacted with a language-capable robot while wearing the Microsoft HoloLens over a series of trials, with robot communication style and user cognitive load varied between trials. We employed a dual-task paradigm in a tabletop pick-and-place task. Participants view the primary task through the Microsoft HoloLens, allowing them to see virtual bins overlaid over mixed reality fiducial markers, and a panel of blocks that changes every few seconds (Fig. 1). The Pepper robot is positioned behind the table, ready to interact.

2.2 Experimental Task

Primary Task: The user's *primary task* is to watch the block panel for a target block: a *red cube*, *red sphere*, *red cylinder*, *yellow cube*, *yellow sphere*, *yellow cylinder*, *green cube*, *green sphere*, or *green cylinder*. These blocks were formed by combining three colors with three shapes. When participants see the target block, their task is to place it into any of a particular set of bins. For example, the robot might tell a user that whenever they see a *red cube* they

should place it in bins *two or three*. Two factors increase the complexity of this primary task. First, at every point during the task, one random bin is unavailable and greyed out. This forces users to remember all target bins. Second, to create a demanding auditory component to the primary task, 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 and non-target bins are switched.

Secondary Task: Three times per experiment trial, the participant encounters a secondary task, in which the robot interrupts with a new request to move a block to a bin. Depending on trial condition, the robot's spoken request may be accompanied by a mixed reality gesture.

2.3 Experimental Design

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

Cognitive Load
Cognitive load was manipulated through our primary task. Following Beck and Lavie [3], we manipulated cognitive load by jointly manipulating memory constraints and target/distractor discriminability, producing four load profiles: (1) all load low, (2) high working memory load, (3) high visual perceptual load, and (4) high auditory perceptual load.

Working Memory Load: In the high working memory load condition, participants had to remember the identities of three out of six visible bins, producing a memory load of seven items: three target bins, target block color and shape, and target syllable consonant and vowel. In all other conditions, participants only had to remember the identities of two out of four visible bins, producing a total memory load of six items.

Visual Perceptual Load: In the high visual perceptual load condition, the target block was always difficult to discriminate, sharing one common property with all distractors. For example, if the target block was a red cube, all distractors were red or cubes (but not both). In the low visual perceptual load condition, the target block was always easy to discriminate, sharing no common properties with any distractors. For example, if the target block was a red cube, no distractors were red or cubes.

Auditory Perceptual Load: Auditory perceptual load conditions followed a similar structure to visual perceptual load conditions. For example, if the target syllable was *kah*, in the high load condition all distractors started with *k* or end with *ah* (but not both), and in the low load condition no distractors started with *k* or end with *ah*.

Communication Style
Communication style was manipulated through our secondary task, following Williams et al. [9]: (1) In blocks using **complex language (CL)**, the robot referred to objects using full referring expressions needed to disambiguate those objects (e.g., "the red sphere"). (2) In blocks using **MR + CL**, the robot referred to objects using full referring expressions paired with a MRDG (e.g., an arrow drawn over the red sphere). (3) In blocks using **MR + simple language (SL)**, the robot referred to objects using minimal referring expressions (e.g., "that block"), paired with a MRDG. We didn't examine SL without MR, as that communication style typically does not enable referent disambiguation, requiring the user to ask for clarification or guess at random.

2.4 Measures

Accuracy was measured for both tasks by logging which objects participants clicked on, determining whether these were intended by the task or robot, and whether they were placed in the correct bins.

Response time (RT) was measured by logging when participants interacted with blocks and bins. In a primary task, when participants see a target block, their task is to pick-and-place it into a particular set of bins. Thus, RT was measured as delay between when the target block is displayed and when placement is completed. In the secondary task, RT was measured as time between start of Pepper's utterance and placement of the secondary target block.

Perceived mental workload was measured using the NASA TLX [1].

Perceived communicative effectiveness was measured using the modified Gesture Perception Scale [6] employed by Williams et al. [9], which assesses effectiveness, helpfulness, and appropriateness of communication.

2.5 Participants and Procedure

36 participants were recruited from Mines (31 M, 5 F), aged 18–32. After providing informed consent and completing demographic and visual capability surveys, participants were introduced to the task through verbal instruction and an interactive tutorial. Participants then engaged in the twelve (Latin square counterbalanced) trials formed by combining the four cognitive load conditions and the three communication style conditions, with surveys after each block.

3 Results

Bayesian repeated measures analyses of variance (RM-ANOVA) with Bayes Inclusion Factor analyses were performed, using communication style and cognitive load as random factors. A log transformation was applied to all RT data.

Response Time: We found strong evidence against effects on primary task RT (BFs < 0.028), but strong evidence for an effect of communication style (BF_{Incl}

= 17.86) on secondary task RT. Post-hoc analysis revealed extreme evidence (BF = 601.46) for a difference in RT between CL ($\mu = 2.10$, $\sigma = 0.33$; untransformed $\mu = 8.88$ s, $\sigma = 4.07$ s) and MR + CL ($\mu = 1.96$, $\sigma = 0.32$; untransformed $\mu = 7.78$, $\sigma = 3.88$), weak evidence (BF = 1.55) for a difference in RT between CL and MR + SL ($\mu = 2.01$, $\sigma = 0.44$; untransformed $\mu = 8.76$, $\sigma = 6.20$), and moderate evidence (BF = 0.20) *against* a difference between MR + CL and MR + SL.

Accuracy: Strong evidence was found *against* effects on primary or secondary task accuracy (All $BFs_{Incl} < 0.033$ for an effect). Mean primary task accuracy was 0.71 ($\sigma = 0.26$). Mean secondary task accuracy was 0.98 ($\sigma = 0.07$).

Perceived Mental Workload: Strong evidence was found *against* effects on perceived mental workload (BF_{Incl} between 0.006 and 0.040 in favor of an effect). Most participants' perceived workload indicated "medium load".

Perceived Communicative Effectiveness: Anecdotal to strong evidence was found *against* any effects on perceived communicative effectiveness (BF_{Incl} between 0.05 and 0.12 in favor of an effect on all questions). Participants' perceived communicative effectiveness had a mean of 5.61 out of 7 ($\sigma = 1.21$).

4 Discussion and Conclusion

We examined the effectiveness of different combinations of language and MRDG under different types of mental workload, through a mixed-reality robotics laboratory experiment. Our results suggest the primary benefit of MRDGs in robot communication is increasing secondary task speed by reducing visual search time (especially when paired with complex language) regardless of mental workload. However, our results failed to support our hypotheses. While we expected differences between communication styles based on workload, we observed that visual augmentations may *always* be helpful for a secondary task, regardless of workload. Furthermore, we found no effects on perceived workload or perceived effectiveness. The differences in participants' own secondary RTs might have been too small for participants to notice, or participants may have only considered their primary task when reporting their perceptions.

References

1. Hart, S., Staveland, L.: Development of NASA-TLX (task load index): results of empirical and theoretical research. In: Human Mental Workload (1988)
2. Hirshfield, L., Williams, T., Sommer, N., Grant, T., Gursoy, S.V.: Workload-driven modulation of mixed-reality robot-human communication. In: Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data (2018)
3. Lavie, N.: The role of perceptual load in visual awareness. Brain Res. **1080**, 91–100 (2006)
4. Matuszek, C., Bo, L., Zettlemoyer, L., Fox, D.: Learning from unscripted deictic gesture and language for human-robot interactions. In: Proceedings of AAAI (2014)

5. Salem, M., Kopp, S., Wachsmuth, I., Rohlfing, K., Joublin, F.: Generation and evaluation of communicative robot gesture. Int. J. Soc. Robot. **4**(2), 201–221 (2012)
6. Sauppé, A., Mutlu, B.: Robot deictics: how gesture and context shape referential communication. In: Proceeding of HRI (2014)
7. Wickens, C.D.: Processing resources and attention. In: Multiple-task Performance (1991)
8. Wickens, C.D.: Multiple resources and mental workload. Hum. Factors **50**(3), 449–555 (2008)
9. Williams, T., Bussing, M., Cabrol, S., Lau, I., Boyle, E., Tran, N.: Investigating the potential effectiveness of allocentric mixed reality deictic gesture. In: Chen, J.Y.C., Fragomeni, G. (eds.) HCII 2019. LNCS, vol. 11575, pp. 178–198. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-21565-1_12