Structuring Human-Robot Interactions via Interaction Conventions

Ji Han*[1], Gopika Ajaykumar*[1], Ze Li[2], and Chien-Ming Huang[1]

***Abstract*— Interaction conventions (e.g., using pinch gestures to zoom in and out) are designed to structure how users effectively work with an interactive technology. We contend in this paper that successful human-robot interactions may be achieved through an appropriate use of interaction conventions. We present a simple, natural interaction convention—"Put That Here"—for instructing a robot partner to perform pick-and-place tasks. This convention allows people to use common gestures and verbal commands to select objects of interest and to specify their intended location of placement. We implement an autonomous robot system capable of parsing and operating through this convention. Through a user study, we show that participants were easily able to adopt and use the convention to provide task specifications. Our results show that participants using this convention were able to complete tasks faster and experienced significantly lower cognitive load than when using only verbal commands to give instructions. Furthermore, when asked to give natural pick-and-place instructions to a human collaborator, the participants intuitively used task specification methods that paralleled our convention, incorporating both gestures and verbal commands to provide precise task-relevant information. We discuss the potential of interaction conventions in enabling productive human-robot interactions.**

I. INTRODUCTION

Robots hold promising potential in assisting people in various domains, ranging from flexible manufacturing and collaborative construction to healthy aging at home. To maximize the utility of robotic assistance in the envisioned domains, humans must be able to easily and effectively interact with their robotic assistants and collaborators in these diverse environments and situations. A common approach taken by Human-Robot Interaction (HRI) researchers to design effective human-robot interactions has been to model the interactions off of human-human interactions, with the robot playing the role of a human-like, social agent [1].

While it is important for robots to have a semantic understanding of human language, behaviors, and environments in order to interact with people seamlessly, modeling human-robot interaction off of human-human interactions introduces several technical challenges, such as understanding spontaneous human behavior, considering a diversity of modalities in human expressions, and understanding interaction contexts [2]. These technical challenges can be computationally demanding and algorithmically intractable, especially in uncontrolled settings or long-term deployments.

As an alternate approach to requiring robots to acquire the level of semantic understanding necessary for human-human interactions, we argue that users themselves can learn and use simple, intuitive *interaction conventions* to effectively interact with robots. Interaction conventions have commonly been leveraged within Human-Computer Interaction (HCI) to drive and, as needed, constrain user behaviors in using computing devices [3]. Examples of interaction conventions employed by users when they interact with modern computing devices include dragging and dropping virtual objects within an interface, pinching two fingers to zoom out of a screen, or long-pressing to activate a hidden interaction mode. Such learned interaction conventions have been shown to produce natural human-computer interactions [4].

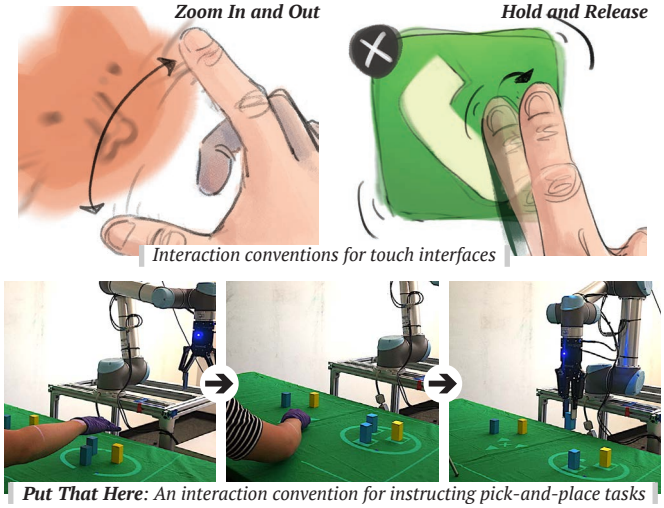


Fig. 1. Similar to how *interaction conventions* can be used to structure human-computer interactions (Top), we contend that appropriate interaction conventions can enable productive human-robot interactions. We present a real-time, interactive system that implements an interaction convention through which users use multimodal behaviors to instruct a robot manipulator in performing pick-and-place tasks (Bottom).

In this work, we explore the application of learned interaction conventions to human-robot interactions to enable productive, natural interactions (Fig. 1). We study an interaction convention, "Put That Here," that human users can use to interact with robots during tabletop pick-and-place tasks, which are fundamental to many manipulation tasks that robots are envisioned to perform [5]. This convention ("put that (*pointing toward an object*) here (*pointing at a target location*)" [6]) builds on implicit multimodal human communication and proved to be natural and robust for directing a robot to perform a variety of pick-and-place tasks.

Next, we provide relevant background that motivates this work. We then describe our interaction convention design and system implementation; we report a user evaluation and its results in Sections IV and V, respectively; we conclude this paper with a discussion of design implications.

*Both authors contributed equally to this work.

[1]Johns Hopkins University, Baltimore, MD 21218, USA. Email: cmhuang@cs.jhu.edu

[2]Tsinghua University, Beijing, 100084, China.

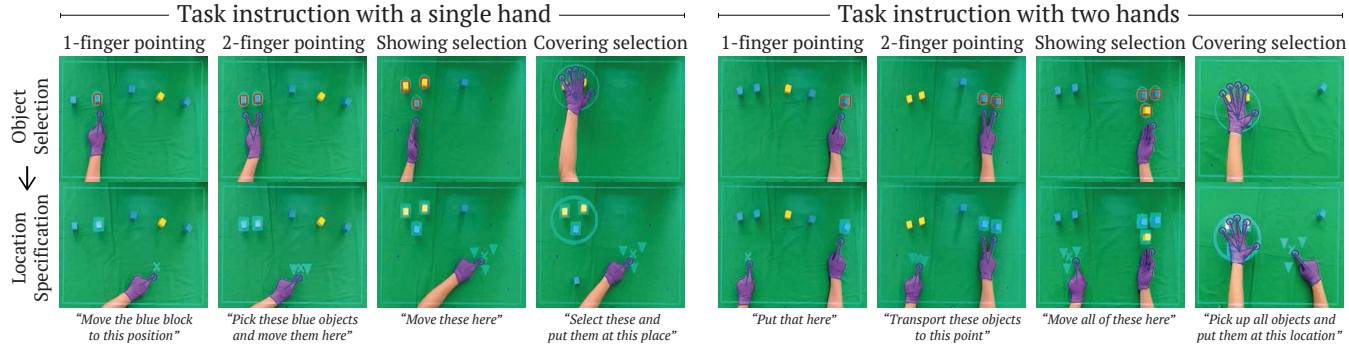


Fig. 2. Our interaction convention for multimodal instruction of pick-and-place tasks. It supports the use of indicative gestures with one or two hands for object selection and for location specification, as well as the use of verbal commands to describe task actions and object features.

II. BACKGROUND

A. Interaction Conventions

Interaction conventions have been widely used by designers in Human-Computer Interaction to convey to users the range of interactions that are feasible for a particular computing system [3]. In this way, interaction conventions are not limited to originating in natural human behavior, but may instead be constructed to minimize technological failures resulting from the limitations of interactive technologies. For example, navigating a screen-based interface using a scrollbar is a convention that was designed so that users are able to access the full extent of content available on the interface while still working within the technological limitations of their computing device (i.e., the limited screen size). In this work, we similarly use an interaction convention to constrain user behaviors such that they may interlock with a limited set of robot capabilities.

Prior works have suggested that the ease of use and learnability of an interface relying on interaction conventions can depend on the user's prior experience with the convention [7], [8]. As a result, although interaction conventions do not have to be tied to human behavior, basing interaction conventions off of existing human behaviors that users are already familiar with can facilitate users in quickly and effectively using the convention. Several works in HCI have modeled interaction conventions off of implicit human behaviors [9], [10] or natural multimodal inputs from humans [11]–[15]. Our interaction convention, "Put That Here," similarly takes root in implicit multimodal human behaviors.

B. Multimodal Human-Robot Interaction

People instinctively employ verbal and non-verbal behaviors in communication and instruction. Among various non-verbal behaviors, gestures play a particularly integral role in human communication [16], [17]. Gestures are used to disambiguate references and supplement spoken content. Moreover, they help facilitate expressions that are hard to convey simply through speech (e.g., precise location in space) and thus reduce the communicator's cognitive load [18], [19] and enhance the level of communication between communicators and recipients. Therefore, for robots to understand human pick-and-place task instructions in situ, they need to have situational awareness of human multimodal behaviors, and they must utilize these behaviors effectively for reference resolution (e.g., knowing what objects are being referred to and where the precise target location is) [20]–[23].

Multimodal behaviors for robot instruction, particularly gestures and speech, have demonstrated their potential in enabling effective interactions between humans and robots [24]–[26]. Previous approaches to multimodal interactions with robots have included verbal instruction for mobile robot navigation [27], deictic gestures for directing a robot collaborator's attention [28], and pointing gestures for pick-and-place tasks [29]. In this work, we describe an interaction convention composed of task instructions consisting of verbal and gestural commands for tabletop pick-and-place tasks.

III. SITUATED MULTIMODAL INTERACTION CONVENTION

In this section, we first present a simple interaction convention for intuitive instruction of robot pick-and-place tasks. We then describe an interactive system implemented to allow end users to follow this convention in collaborating with a robot manipulator through multimodal instructions.

A. "Put That Here": An Interaction Convention for Multimodal Instruction

People naturally—and almost necessarily—employ a combination of verbal commands and gestures to provide unambiguous instruction of pick-and-place tasks to a collaborative partner. Drawing on our experience and observations of human interactions, as well as prior research on human multimodal interactions with technology (e.g., [6]) and robot instruction (e.g., [20]), we designed a straightforward interaction convention, "Put That Here," for specifying robot pick-and-place tasks. Our convention involves the use of hand gestures to select objects and to specify intended move-to locations, while allowing users to naturally utter task actions (e.g., "move" and "pick up") and to verbally point out relevant features of objects and the environment (e.g., "blue blocks"). Figure 2 illustrates a set of common gestural specifications and verbal commands that were implemented in our system.

We implemented four types of indicative gestures—single-finger pointing, two-finger pointing, showing, and covering

Algorithm 1 Multimodal Put-That-Here Convention

Gesture-Type ← Gesture-Recognition()
Fingertip-location ← Fingertip-Detection()
Speech-Command ← Speech-Recognition()
Pointed-Objects ← Object-Detection(*Gesture-Type*)
if *Speech-Command* = *pick* **then**
  Pick-Object ← *Pointed-Objects*
  Feedback(*Pick-Object*, *Gesture-Type*, *pick*)
else if *Speech-Command* = *place* **then**
  Place-Location ← *Fingertip-location*
  Feedback(*Place-Object*, *Gesture-Type*, *place*)
  Manipulation(*Pick-Object*, *Place-Location*)
else if *Speech-Command* = *cancel* or *stop* **then**
  Reset(*Speech-Command*)
else
  Feedback(*Pointed-Object*, *Gesture-Type*, *point*)
end if

(or "brushing")—that have been observed in the manner by which people talk and interact with objects [20], [28]. In our own analysis of people's natural instruction of pick-and-place tasks (see Section V-A), we also observed the various uses of pointing and covering gestures by the participants (Figure 4 (b)). In fact, several participants' natural instructions (prior to learning the "Put That Here" convention) were following our designed convention and could already be recognized by our system. In addition to simple task indication with gestures and speech, people can use the aforementioned behaviors to refine and clearly identify objects for selection. For instance, users can use a covering gesture to select a group of objects and then use verbal commands, such as "pick blue objects," to specify only blue objects within the covering selection. Next, we describe our autonomous, interactive system that enables users to instruct a robot manipulator to perform pick-and-place tasks via the "Put That Here" convention.

B. System Overview

Our interactive pick-and-place system consisted of a UR5 robot manipulator, a mini-projector, a ceiling-mounted RGB camera facing downward, and several software modules for gesture recognition, speech recognition, object detection, robot manipulation, and system feedback. Algorithm 1 details how these modules are integrated to enable multimodal specifications of pick-and-place tasks.

C. Gesture Recognition

We implemented two methods for recognizing indicative gestures: (1) contour-based fingertip detection and (2) deep learning-based gesture recognition. For both methods, we used purple gloves and a green screen to simplify hand segmentation and recognition. The segmented ROI image was then processed by the following methods in parallel:

1) Fingertip Detection: Our implementation of fingertip detection followed an image processing pipeline similar to that proposed by Gurav and Kadbe [30]. Our pipeline began with image blurring using an average filter. The filtered

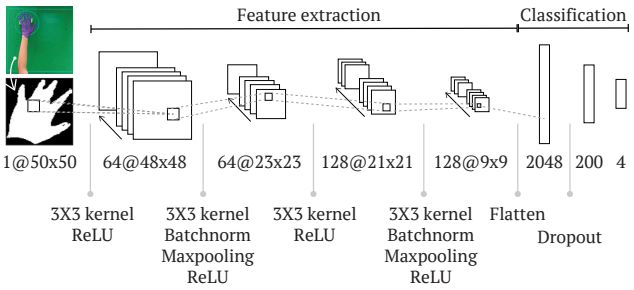


Fig. 3. The design of our neural network for gesture classification.

image was then converted into HSV space for hand segmentation and contour extraction. We further smoothed the contour of the hand [31] to filter out noisy convex points. We then empirically determined additional thresholds to qualify the remaining convex points as fingertips.

2) Learning-Based Gesture Recognition: In addition to geometry-based fingertip detection, we also applied deep learning for gesture classification. Our designed neural network is depicted in Fig. 3. The input to this network is a binary image of a segmented hand. The output is one of four classifications: one-finger pointing, two-finger pointing, showing gesture, and covering gesture. We used cross entropy for loss computation and Adam [32] for optimization. Our training data was collected locally. To improve generalization of our model, we augmented our training images via image rotations, affine transformations, and flips [33]. We also added L2 regularization to avoid model overfitting [34].

D. Object Detection and Reference Estimation

We used color segmentation to identify objects and our contour-based approach to locate the center of a hand and its fingertips. This information allowed us to estimate the pointing direction and the objects being referenced. To recognize the direction of a "showing" gesture, we used principle component analysis (PCA) to compute the primary direction of the gesture. For the "covering" gesture, we considered objects within the area around the hand shape.

E. Speech Recognition

We used the Google Speech-to-Text service to recognize users' verbal commands. We passed audio streams from a microphone placed in front of the user to the Google Cloud service, which in turn returned the recognized texts. Locally, we used a language model representing the linguistic structure of a task instruction. A valid instruction consisted of two parts: object selection ("pick") and location specification ("place") (see examples in Fig. 2). The recognized texts from the Google Cloud service were parsed using our language model. Parsed verbal commands and recognized gestures and objects were associated temporally. The fused information was then used to identify the selected objects and intended location. Our method of associating verbal commands and gestures was informed by models of human communication, which indicate that a gesture-relevant linguistic cue occurs within the timespan of the corresponding gesture [35].

F. Robot Manipulation

We used simple wooden blocks in our pick-and-place task and preprogrammed the robot's grasp configuration for the blocks. The positions of blocks in image coordinates were transformed to real-world coordinates relative to the base link of the robot via homography transformations. The desirable grasp pose of the block was then used to solve inverse kinematics to plan the robot's motion. The system chose the inverse kinematics solution that was closest to the robot's current joint configuration while avoiding introducing awkward movements into the motion plan.

G. Signaling System Feedback

In addition to instruction recognition and robot manipulation, we implemented a projection-based feedback system that signals system states through projected highlights (Fig. 2). Projected feedback indicated the system's interpretation of user task specifications. In our implementation, projected feedback highlighted the object(s) that a user was pointing at with a circle, the object(s) that was/were selected with a square, and the location specified by the user with a cross. When an instruction included multiple objects, the cross represented the center of the object cluster, and triangles represented the place location of each object. These signals provided real-time feedback to users about their instructions. Prior research has shown that effective feedback from robotic systems can significantly improve collaborative task performance and user experience [36], [37]. In our study, users were able to use verbal commands to "cancel" their instructions or "stop" the robot action based on projected system feedback.

IV. EVALUATION

We sought to assess whether people are able to learn and use the interaction convention of "Put That Here" to effectively instruct a robot to perform a variety of pick-and-place tasks. More broadly, this evaluation aimed to understand the possibility of structuring interactions between humans and robots via interaction conventions.

A. Study Design and Experimental Conditions

Our user study consisted of two parts. In the first, participants were asked to provide pick-and-place instructions as naturally as they would when interacting with a human partner. The robot was not revealed to the participants during this part of study. To avoid linguistic or gestural priming that could potentially influence participants' behaviors, we projected task information directly onto the workspace (Fig. 4 (a)) and told participants that their partner could not perceive the projected information. We recorded participants' task instruction performances. For the second part of the study, we designed a within-subjects experiment in which each participant was asked to provide task instructions to a UR5 robot using three methods. The order of the methods used was counterbalanced using a Latin square method to reduce potential order effects. The three methods of instruction are described below:

1) Gesture-Only Instruction: In this condition, participants were instructed to use only gestures to provide task instructions. There were no experimental instructions or constraints on how to perform these gestures. When the participants completed their gestural commands, an experimenter manually provided information on objects and target location to the robot via a Wizard-of-Oz (WoZ) interface for robot action execution. Conceptually, this condition represented *gesture-based interactions* that are commonly used with modern touch interfaces.

2) Speech-Only Instruction: In this condition, participants were asked to instruct the robot with only verbal commands. Similar to the gesture-only condition, an experimenter provided necessary task information based on the user's verbal specification to the robot via a WoZ interface. Conceptually, this condition represented *speech-based interactions* that are commonly used with smart speakers (e.g., Amazon Echo).

3) Multimodal Interaction Convention: Participants followed the "Put That Here" interaction convention (Section III) to instruct the robot using both gestures and verbal commands. The system was fully autonomous in this condition.

B. Experimental Task & Procedure

Upon consenting to participate in the study, participants started with the first part of the study, in which they provided task instructions as naturally as they could. After this part of the data collection, the participants were randomly assigned to one of the condition orders to interact with the robot. In the multimodal interaction convention condition, participants were allowed to practice the instruction method as long as they wanted until they felt confident about using it.

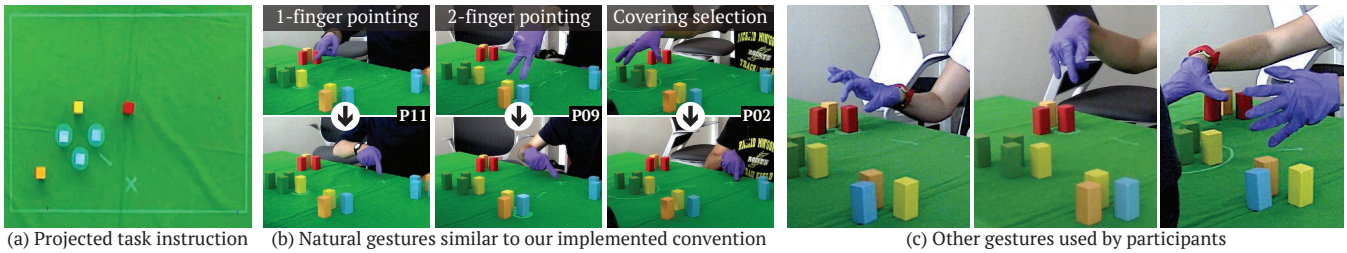
In each condition, participants were asked to give five task instructions with different object configurations (e.g., locations and number of objects) to the robot, one at a time. The participants were asked to look away from the workspace while each task configuration was being set up. Once the participants were instructed to return to the workspace, a task instruction was projected for three seconds (Fig. 4 (a)). The participants were allowed to provide a task instruction to the robot only after the projected information disappeared.

After each experimental condition, the participants filled out a questionnaire regarding their experience using the instruction method. After completing all three conditions, an experimenter interviewed the participants for additional comments. The whole study was about an hour long and the participants were compensated with $10 USD.

C. Measures

To measure the effectiveness of our interaction convention for pick-and-place instructions, we used a combination of objective, subjective, and behavioral measures.

1) Training Time: We defined the training time for the method of multimodal instruction as the length of the practice trial. We note that participants could practice as long as they wanted and were asked to end the practice trial only when they felt comfortable using the method for task instruction.



(a) Projected task instruction (b) Natural gestures similar to our implemented convention (c) Other gestures used by participants

Fig. 4. (a): Each task instruction was presented through projection to avoid possible linguistic or gestural priming. (b): Participants naturally used multimodal behaviors similar to our designed interaction convention for instruction. (c): Besides the common behaviors shown in (b), participants also used various gestures for task specification.

2) System Performance: To evaluate system performance, we focused on whether our system was able to recognize participants' gestures accurately and whether participants were able to use our convention to instruct the robot successfully.

3) Task Efficiency and Cognitive Load: We measured task efficiency using the time needed for a user to complete a task instruction. In our experiment, the time needed for the robot to execute a given instruction was the same across conditions; thus, we used the amount of time taken for instruction to represent task efficiency.

Task efficiency is largely related to cognitive load experienced during a task; we sought to assess a participant's cognitive load using a variety of measures. These measures included: (1) preparation time—the time needed before issuing an instruction (presumably for thinking and planning the instruction); (2) number of linguistic pauses within an instruction; (3) length of linguistic pauses; and (4) number of speech disfluencies (e.g., uh, uhm, and like) within an instruction. In addition to these behavioral measures of cognitive load, we adapted the NASA TLX [38] for our study context. In particular, we used three items—mental demand, effort, and frustration—from the TLX scale to assess task load (Cronbach's $\alpha = .74$).

4) Usability: We used a questionnaire to assess participants' subjective perceptions of the instruction methods, focusing on the aspects of *ease of use* and *flexibility* of the method for specifying pick-and-place instructions. The scale of ease of use consisted of four items (Cronbach's $\alpha = .84$). We used a single item to measure flexibility.

D. Participants

We recruited 14 participants for this study. Four participants were excluded from our analysis: one participant did not finish the study, and three participants did not follow our experimental protocol. The resulting 10 participants (two of which were female) were included in our data analysis. The participants were all native English speakers and aged 18.8 years on average ($SD = 1.03$).

V. RESULTS

We used one-way repeated-measures analysis of variance (ANOVA) in which instruction method—speech-only, gesture-only, or multimodal—was set as a fixed effect, and participant was set as a random effect. Our analyses focused on how the "Put That Here" multimodal convention compared to speech-only and gesture-only task instruction. Therefore, two *a priori* pairwise comparisons, using a Bonferroni-adjusted α level of .025 (.05/2) for significance, were carried out to measure differences across conditions for each quantitative measure. For readability, main effects and all pairwise comparisons are reported in Fig. 5.

A. Natural Instructions for Pick-and-Place

We first present our observations of participants' natural behaviors when giving pick-and-place instructions (the first part of our study). Nine out of the ten participants used both gestures and verbal commands during their instruction. To our surprise, one participant gave instructions through only verbal commands. We note that several participants' natural instruction, including their uses of gestures and verbal commands, was very similar to our designed convention (e.g., "Move this to here" and "Move these objects to here" (P9)) (Fig. 4 (b)). Moreover, we observed a wide range of gestures that participants used. See examples in Fig. 4 (c). These variations posed technical challenges in recognizing the intended indications successfully.

B. Training Time

On average, the participants spent less than 90 seconds ($M = 83.37$, $SD = 41.26$) on practicing our interaction convention before they felt confident in using it for task instruction, suggesting the learnability of the convention.

C. System Performance

Overall, our system was able to recognize participants' multimodal task instructions effectively. Out of 50 trials (five trials per participant), our system failed to recognize a participant's fingertips twice and a pointed-at block once. Only one instruction mistake was made by a participant, who used a covering gesture but missed one target object.

D. Task Efficiency and Cognitive Load

Task efficiency was measured in terms of time needed to complete an instruction. The analysis of variance found a significant main effect of the instruction method on task efficiency, $F(2, 137) = 23.92$, $p < .001$. Pairwise comparisons further revealed that the participants needed a significantly longer time to finish an instruction if they could only use verbal commands, $F(1, 137) = 14.82$, $p < .001$, or gestural instructions, $F(1, 137.1) = 9.43$, $p = .003$.

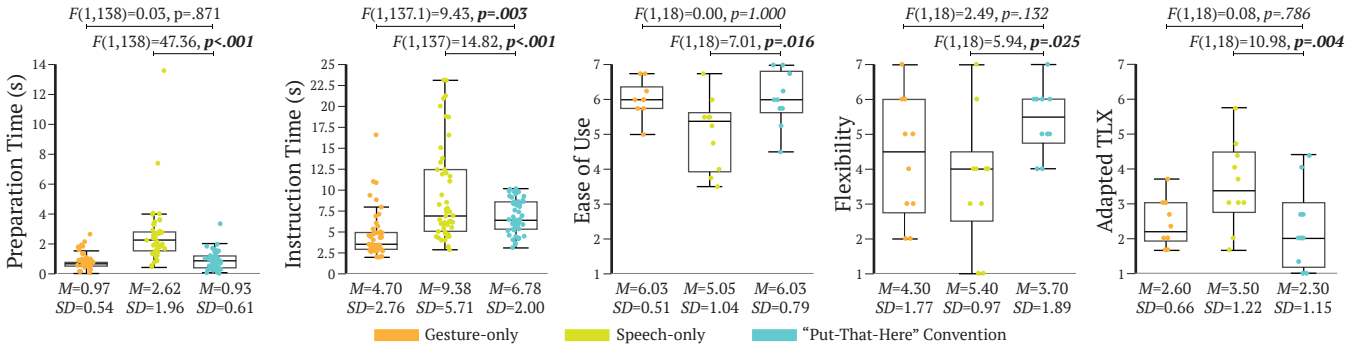


Fig. 5. Results of cognitive load, task efficiency, and usability. Pairwise comparisons using a Bonferroni-adjusted α level of .025 for significance were carried out to evaluate the effectiveness of using the multimodal method for pick-and-place instruction.

To assess participants' cognitive load, we used several behavioral and subjective measures, including preparation time for instruction, linguistic dysfluency, and an adapted TLX scale. The analysis of variance found a significant main effect of the instruction method on preparation time, $F(2, 138) = 30.85$, $p < .001$. On average, the participants were able to begin their instruction within less than a second after the projected task disappeared when using the gesture-only and multimodal methods; no significant difference was found between gesture-only and multimodal methods, $F(1, 138) = 0.03$, $p = .871$. In contrast, the participants spent more than 2.5 seconds when using the speech-only method, $F(1, 138) = 47.36$, $p < .001$. Participants' comments indicated that it was difficult to only use verbal commands to give instructions and that they needed a bit more time to think to compensate for the difficulty:

> "*I had to think a little before giving certain commands.*" (P11)
>
> "*[The speech-only method] was a little more difficult because it was harder to specify where exactly within the rectangle [workspace] I wanted to move the box and also which box I wanted to move.*" (P7)

Our data also showed that the participants were likely to pause during their instructions when using the speech-only method (Fig. 6). Moreover, their instructions contained dysfluent tokens such as "uh" in this condition ($M = 0.66$, $SD = 1.19$). Furthermore, the analysis of variance found a significant main effect of the instruction method on our adapted TLX scale, $F(2, 18) = 6.76$, $p = .006$. Participants reported significantly higher task load when using the speech-only method compared to the multimodal method, $F(1, 18) = 10.98$, $p = .004$.

E. Usability

We focused on two aspects of usability in this evaluation: ease of use and flexibility of method. The analysis of variance found a significant main effect of the instruction method on subjective perceptions of ease of use, $F(2, 18) = 4.67$, $p = .023$. Participants rated the multimodal convention as significantly easier to use than the speech-only method, $F(1, 18) = 7.01$, $p = .016$. Regarding the perceived flexibility, the analysis of variance found a marginal main effect of

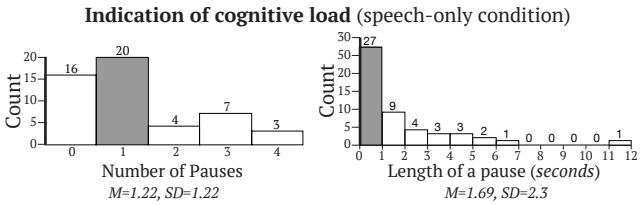


Fig. 6. The participants reported higher cognitive load, as measured through an adapted TLX scale and linguistic dysfluency, when using the speech-only method for task instruction compared to the other two methods.

the instruction method on subjective perceptions of method flexibility, $F(2, 18) = 3.05$, $p = .072$. The results showed that the multimodal convention provided higher flexibility for the participants in specifying task instructions compared to the speech-only method, $F(1, 18) = 5.94$, $p = .025$.

Overall, the multimodal convention was the most preferred method. Seven out of the ten participants preferred the multimodal method over the other two methods:

> ". . . *with verbal and gestures, um, was I guess the most convenient because I could both specify, like, which objects I wanted to move and exactly where I wanted them to go, and it was easier to select them overall.*" (P7)

VI. DISCUSSION

Just as a new interactive technology (e.g., touch interfaces) requires a new set of interaction conventions (e.g., pinch to zoom in and out, hold and release, and swipe), we argue that robots necessitate new interaction conventions for effective use by people. In this paper, we explored a natural, simple convention for giving pick-and-place instructions to a robot partner. Our observations of people's natural behaviors when giving pick-and-place instructions supported the "Put That Here" convention that allows people to use common indicative gestures and verbal commands for task specification. The results of our experiment show that participants were able to learn how to use the convention quickly and could use it with ease. The results further indicate that the participants using the multimodal convention were able to complete their instructions faster and experienced significantly lower cognitive load during the task than when using the speech-only method. While we did not observe significant differences be-

tween the multimodal and gesture-only methods, we note that the participants did not use gestures without accompanying speech when asked to conduct natural instruction.

A. Universal Conventions for HRI

In this work, we demonstrated that structuring user behaviors to conform to learned interaction conventions can create natural, effective, and easy human-robot interactions without requiring the extensive technical robot capabilities that are typically required in human-robot interactions modeled off of human-human interactions. Besides simplifying the problem of robot understanding of the unbounded space of human behaviors, abstracting human interactive behaviors into interaction conventions also has the potential to create *universal* human-robot interactions.

Since human behaviors and human-robot interactions can differ depending on human characteristics, such as culture, gender, or educational background [39]–[41], robots must tailor their behavior to individuals during interactions, while following the norms corresponding to the current cultural context and avoiding potential pitfalls such as stereotyping. Since most robots have not reached this level of adaptability, humans may have to adapt their behaviors and actions to effectively interact with a robot that has not been personalized to their characteristics, or, in the worst case, may be unable to interact with the robot.

If a universal interaction convention is instead used to structure the human-robot interaction, both the human and robot interactants will have common ground on the interaction structure, facilitating communication and collaboration despite varying human characteristics or contexts. In a similar vein, human-robot interactions built upon universal conventions could transcend individual robot platforms or manufacturers, so that humans do not need to learn new interaction patterns for each new robot that they encounter. This idea can already be found in computers and mobile devices, where conventions such as drag and drop and touchscreen swipe gestures are universal regardless of the user's geographic location or the device's manufacturer. By extending the concept of universal interaction conventions to HRI, we can move one step closer towards making human-robot interaction accessible for all.

B. Variations of "Put That Here"

The "Put That Here" convention used in this work enables users to perform tabletop pick-and-place tasks. We envision several design variations that can extend this convention to make it applicable for a wider range of task contexts. The convention currently allows users to instruct the robot to move multiple objects when they are proximally located to one another. For future design iterations of this convention, we plan to explore methods for users to specify multiple objects that are distant from each other during task instruction, analogous to how the 'Shift' key on a keyboard may be used to select multiple items on a screen-based interface. Similarly, we will investigate how to improve upon our existing convention so that users may specify distant locations for object grasping or placement that are not limited to the constrained tabletop workspace used in this study. We will consider incorporating deictic gestures or instructional tools for robot manipulation that have been used in previous studies, such as laser pointers [42], [43], into our convention to make it more applicable outside the tabletop context.

Lastly, our convention currently supports four classes of gestures (Fig. 2). However, our user evaluation demonstrated that participants may use gestures that are not covered by the four gesture types used in our convention during task specification (Fig. 4 (c)). Therefore, we would like to extend our convention to incorporate a wider range of gestures, as well as enable users to extend the convention themselves to include their own custom gestures. Finally, in this study, our user evaluation included a limited number of participants. For future design iterations, we aim to conduct evaluations with more users to gain a better understanding of user behaviors and needs in using our multimodal convention.

VII. CONCLUSION

Appropriate interaction conventions that are straightforward to learn and use can effectively structure communication between end users and technology. In this paper, we argue that there exists a need for interaction conventions for enabling productive human-robot interactions. We studied a particular multimodal convention—"Put That Here"—for situated instruction of pick-and-place tasks. Our empirical evaluation indicated the usability and learnability of the multimodal convention. This work highlights the promise of using learned interaction conventions as a tool for designing human-robot interactions, particularly for new, technologically limited, or cross-cultural robotic technologies.

ACKNOWLEDGMENT

We thank Yuxiang Gao, Jaimie Patterson, Eden Metzger, Anthonia Duru, and Erica Hwang for their help with this work. This work was partially supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE-1746891, the Nursing/Engineering joint fellowship from the Johns Hopkins University, and the National Science Foundation award #1840088.

REFERENCES

[1] Nicole C Krämer, Astrid von der Pütten, and Sabrina Eimler. Human-agent and human-robot interaction theory: similarities to and differences from human-human interaction. In *Human-computer interaction: The agency perspective*, pages 215–240. Springer, 2012.

[2] Ruth S Aylett, Ginevra Castellano, Bogdan Raducanu, Ana Paiva, and Mark Hanheide. Long-term socially perceptive and interactive robot companions: challenges and future perspectives. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 323–326, 2011.

[3] Donald A Norman. Affordance, conventions, and design. *interactions*, 6(3):38–43, 1999.

[4] Jef Raskin. Intuitive equals familiar. *Communications of the ACM*, 37(9):17–19, 1994.

[5] Tomás Lozano-Pérez, Joseph L. Jones, Emmanuel Mazer, and Patrick A. O'Donnell. Task-level planning of pick-and-place robot motions. *Computer*, 22(3):21–29, 1989.

[6] Richard A Bolt. *Put-that-there: Voice and gesture at the graphics interface*, volume 14. ACM, 1980.

[7] Susan Wiedenbeck and Sid Davis. The influence of interaction style and experience on user perceptions of software packages. *International Journal of Human-Computer Studies*, 46(5):563–588, 1997.

[8] S Shyam Sundar, Saraswathi Bellur, Jeeyun Oh, Qian Xu, and Haiyan Jia. User experience of on-screen interaction techniques: An experimental investigation of clicking, sliding, zooming, hovering, dragging, and flipping. *Human–Computer Interaction*, 29(2):109–152, 2014.

[9] Matthias Rauterberg, Morten Fjeld, Helmut Krueger, Martin Bichsel, Ulf Leonhardt, and Markus Meier. Build-it: a computer vision-based interaction technique for a planning tool. In *People and Computers XII*, pages 303–314. Springer, 1997.

[10] Albrecht Schmidt. Implicit human computer interaction through context. *Personal technologies*, 4(2-3):191–199, 2000.

[11] Sharon Oviatt, Antonella DeAngeli, and Karen Kuhn. Integration and synchronization of input modes during multimodal human-computer interaction. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 415–422, 1997.

[12] Veikko Surakka, Marko Illi, and Poika Isokoski. Gazing and frowning as a new human–computer interaction technique. *ACM Transactions on Applied Perception (TAP)*, 1(1):40–56, 2004.

[13] Robert JK Jacob. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems (TOIS)*, 9(2):152–169, 1991.

[14] Robert JK Jacob. Eye movement-based human-computer interaction techniques: Toward non-command interfaces. *Advances in human-computer interaction*, 4:151–190, 1993.

[15] David Rempel, Matt J Camilleri, and David L Lee. The design of hand gestures for human–computer interaction: Lessons from sign language interpreters. *International journal of human-computer studies*, 72(10-11):728–735, 2014.

[16] Adam Kendon. Do gestures communicate? a review. *Research on language and social interaction*, 27(3):175–200, 1994.

[17] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[18] Jean Ann Graham and Simon Heywood. The effects of elimination of hand gestures and of verbal codability on speech performance. *European Journal of Social Psychology*, 5(2):189–195, 1975.

[19] Susan Goldin-Meadow. *Hearing gesture: How our hands help us think*. Harvard University Press, 2005.

[20] 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.

[21] Tom Williams, Saurav Acharya, Stephanie Schreitter, and Matthias Scheutz. Situated open world reference resolution for human-robot dialogue. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 311–318. IEEE Press, 2016.

[22] Hendrik Zender, Geert-Jan M Kruijff, and Ivana Kruijff-Korbayová. Situated resolution and generation of spatial referring expressions for robotic assistants. In *IJCAI*, pages 1604–1609, 2009.

[23] Shen Li, Rosario Scalise, Henny Admoni, Stephanie Rosenthal, and Siddhartha S Srinivasa. Spatial references and perspective in natural language instructions for collaborative manipulation. In *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pages 44–51. IEEE, 2016.

[24] O Rogalla, M Ehrenmann, R Zollner, R Becher, and R Dillmann. Using gesture and speech control for commanding a robot assistant. In *Robot and Human Interactive Communication, 2002. Proceedings. 11th IEEE International Workshop on*, pages 454–459. IEEE, 2002.

[25] Geraint Jones, Nadia Berthouze, Roman Bielski, and Simon Julier. Towards a situated, multimodal interface for multiple uav control. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, pages 1739–1744. IEEE, 2010.

[26] Takuya Takahashi, Satoru Nakanishi, Yoshinori Kuno, and Yoshiaki Shirai. Human-robot interface by verbal and nonverbal behaviors. In *Intelligent Robots and Systems, 1998. Proceedings., 1998 IEEE/RSJ International Conference on*, volume 2, pages 924–929. IEEE, 1998.

[27] Stanislao Lauria, Guido Bugmann, Theocharis Kyriacou, and Ewan Klein. Mobile robot programming using natural language. *Robotics and Autonomous Systems*, 38(3-4):171–181, 2002.

[28] 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.

[29] Roberto Cipolla and Nicholas J Hollinghurst. Human-robot interface by pointing with uncalibrated stereo vision. *Image and Vision Computing*, 14(3):171–178, 1996.

[30] Ruchi Manish Gurav and Premanand K Kadbe. Real time finger tracking and contour detection for gesture recognition using opencv. In *Industrial Instrumentation and Control (ICIC), 2015 International Conference on*, pages 974–977. IEEE, 2015.

[31] Non-parametric Ramer-Douglas-Peucker. Ramer-douglas-peucker algorithm. 1972.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. Understanding data augmentation for classification: when to warp? *arXiv preprint arXiv:1609.08764*, 2016.

[34] Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM, 2004.

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

[36] 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 *Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on*, pages 708–713. IEEE, 2005.

[37] Chien-Ming Huang and Andrea L Thomaz. Effects of responding to, initiating and ensuring joint attention in human-robot interaction. In *RO-MAN, 2011 IEEE*, pages 65–71. IEEE, 2011.

[38] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 904–908. Sage Publications Sage CA: Los Angeles, CA, 2006.

[39] Tatsuya Nomura and Satoru Takagi. Exploring effects of educational backgrounds and gender in human-robot interaction. In *2011 International conference on user science and engineering (i-user)*, pages 24–29. IEEE, 2011.

[40] Frédéric Kaplan. Who is afraid of the humanoid? investigating cultural differences in the acceptance of robots. *International journal of humanoid robotics*, 1(03):465–480, 2004.

[41] Dane Archer. Unspoken diversity: Cultural differences in gestures. *Qualitative sociology*, 20(1):79–105, 1997.

[42] Charles C Kemp, Cressel D Anderson, Hai Nguyen, Alexander J Trevor, and Zhe Xu. A point-and-click interface for the real world: laser designation of objects for mobile manipulation. In *Human-robot interaction (HRI), 2008 3rd ACM/IEEE international conference on*, pages 241–248. IEEE, 2008.

[43] Kentaro Ishii, Shengdong Zhao, Masahiko Inami, Takeo Igarashi, and Michita Imai. Designing laser gesture interface for robot control. In *IFIP Conference on Human-Computer Interaction*, pages 479–492. Springer, 2009.