Proceedings of the 27th IEEE International Symposium
on Robot and Human Interactive Communication,
Nanjing, China, August 27-31, 2018

A Comparison of Visualisation Methods for Disambiguating Verbal Requests in Human-Robot Interaction

Elena Sibirtseva, Dimosthenis Kontogiorgos, Olov Nykvist, Hakan Karaoguz,
Iolanda Leite, Joakim Gustafson and Danica Kragic

***Abstract*— Picking up objects requested by a human user is a common task in human-robot interaction. When multiple objects match the user's verbal description, the robot needs to clarify which object the user is referring to before executing the action. Previous research has focused on perceiving user's multimodal behaviour to complement verbal commands or minimising the number of follow up questions to reduce task time. In this paper, we propose a system for reference disambiguation based on visualisation and compare three methods to disambiguate natural language instructions. In a controlled experiment with a YuMi robot, we investigated real-time augmentations of the workspace in three conditions – head-mounted display, projector, and a monitor as the baseline – using objective measures such as time and accuracy, and subjective measures like engagement, immersion, and display interference. Significant differences were found in accuracy and engagement between the conditions, but no differences were found in task time. Despite the higher error rates in the head-mounted display condition, participants found that modality more engaging than the other two, but overall showed preference for the projector condition over the monitor and head-mounted display conditions.**

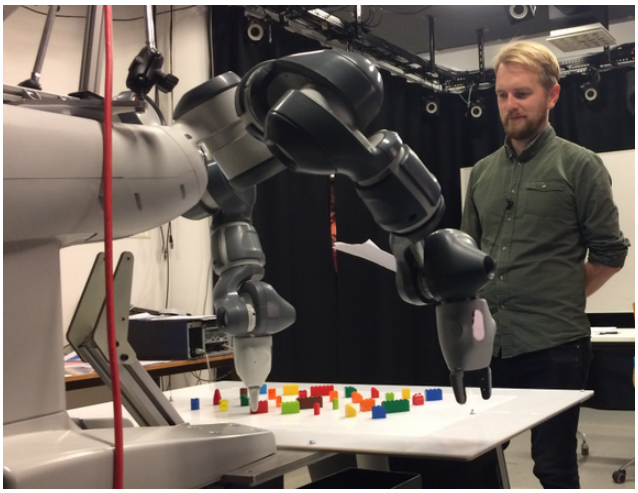


Fig. 1: A participant interacting with the YuMi robot in our experiment using verbal requests to exchange Lego blocks.

I. Introduction

Picking up objects is a common task for robots that work alongside people in home and workplace environments. A typical human-robot interaction task consists of a robot assisting a worker as a third hand, retrieving requested items out of a variety of similar objects.

It is intuitive for humans to use natural language when their hands are busy and they cannot point at the target object. However, such interactions can often lead to ambiguous requests because of speech recognition and language understanding errors, limitations in the robot's understanding of the scene or the presence of similar objects in the workspace.

Previous research has tackled the problem of disambiguating requests from two different perspectives. One perspective aims to reduce ambiguity by asking follow-up questions. However, the more clarification questions the robot asks, the longer the task takes and the risk of speech recognition errors is likely to increase. Previous work that focuses on minimizing the number of follow-up questions has shown that verbal interactions increase task time and can influence accuracy [24]. An alternative approach consists of employing visualisation techniques such as projectors [2] or head-mounted displays [8] to augment the scene with the robot's or human's intentions. While the first few works in this direction have started to appear [4], [16], [19], the effects of augmenting the workspace to disambiguate user verbal requests are still unknown.

In this paper, we performed an experiment to investigate different real-time visualisation modalities for disambiguating verbal requests in an object retrieval task. We developed a system that, in the presence of ambiguous verbal requests, highlights candidate objects and updates this selection as the user refines the target object description with new verbal requests. Using this system, we tested three modalities for providing visual information about the candidate objects that the robot is considering in the workspace: projector, head-mounted display (using Microsoft HoloLens), and a side monitor as the baseline condition.

Our experimental setup consisted of an ABB YuMi robot and a table with Lego blocks (Figure 1). The robot and the human took turns while requesting Lego blocks to pick up. Participants had to verbally explain which Lego block they wanted, using shape and colour information, and were able to perceive by looking at the real-time visualisation the robot's hypothesis about which objects match that description. We intentionally designed this setup to include blocks that would originate ambiguous requests.

In a within-subjects experiment, we collected task times

Elena Sibirtseva, Olov Nykvist, Hakan Karaoguz, Iolanda Leite and Danica Kragic are with the Robotics, Perception and Learning Lab, EECS at KTH Royal Institute of Technology, Stockholm, Sweden. {elenasi, onykvist, hkarao, iolanda, dani}@kth.se

Dimosthenis Kontogiorgos and Joakim Gustafson are with the Speech, Music and Hearing Lab, EECS at KTH Royal Institute of Technology, Stockholm, Sweden. diko@kth.se, jocke@speech.kth.se

978-1-5386-7980-7/18/$31.00 ©2018 IEEE

and accuracies, as well as subjective metrics such as engagement, task observability, display interference and personal preferences. The results of the study showed no significant difference in task time between three conditions. Furthermore, accuracy significantly decreased in the head-mounted display condition; however, participants regarded this condition as the most engaging compared to the other two. As anticipated, the projector condition provided better observability of robot's behaviour and was considered less disruptive. Finally, the projector interface was preferred by most participants and viewed as the most natural and easy to understand visualisation method.

II. Related work

In object retrieval tasks natural language is commonly interpreted into semantically informed representations of the physical space between humans. In human communication, language grounding refers to establishing a "common ground" and understanding that both parts refer to the same object or concept [7]. There have been early attempts in linguistics research in the '70s [25], where users interact with a machine that can understand simple references to objects. Further attempts were made to solve the problem using multimodal features [3], and disambiguate verbal references to objects in a virtual space.

Humans use various methods to establish common ground when they instruct each other in collaborative object retrieval tasks. Common problems occur when object ambiguity is encountered. This makes it more challenging to establish grounding. Li et al. [12] experimented with natural language instructions to investigate the effect of object descriptors, perspective and spatial references and found that ambiguous sentences take more time to process.

Establishing language grounding, particularly in situated human-robot dialogue, can be challenging. Robots need to perceive human behaviour and build internal representations and spatial semantic understanding based on human intentions [23]. Recent research has approached the problem linguistically and through incremental reference resolution [5], [11], [22], spatial references [9], [15], modelling uncertainty [10], but also through past visual observations [18].

Other approaches have considered multimodal features to disambiguate verbal references to the physical space. Several studies have investigated methods such as eye gaze and pointing gestures to disambiguate referring expressions to objects in the shared space between humans and robots [1], [14], [17], [21], and explored non-verbal communicative behaviours to achieve grounding.

Whitney et al. [24] used language and pointing gestures at specific objects when there was ambiguity in the human request. A POMDP based framework was developed in order to balance out the trade off between gaining additional information and the risk of facing speech-to-text failures. However, such an interaction can take a lot of time and would be infeasible with a larger amount of objects. One of the ways to solve this is to visualise the current state of the robot's understanding of the request.

Fig. 2: Human-human interaction pilot study to investigate the most common verbal references used by participants in the task.

Several works have shown effectiveness of using projector based approaches to augment robot's intentions into the shared workspace [2], [4], [16], [20]. In particular, Andersen et al. [2] proposed an object-aware projection technique which takes into account the 3D nature of the environment. As a possible use-case they proposed a car assembly line, where car doors are transported on a conveyor belt and both human and robot have to engage as co-workers on the door. Augmented reality is used to mark the parts that the robot is currently working on. A user study was performed, in which the task was to either rotate or move a white box, based on the instructions provided by one of the three interfaces: projector, monitor display, and text description. The evaluation of this study showed that the projector approach scored higher in user effectiveness and user satisfaction compared to a baseline condition.

Moreover, another successful application of projectors for showing robot's intent was demonstrated in [4], where Chadalavada et al. equipped a robotic fork-lift with a projector to visualise its future trajectory a few meters ahead. The results of the human study showed that by visualising the robot's intent, they achieved significant increase in predictability and transparency; the attributes most crucial for the acceptance of the robots in the workspace.

The application of head-mounted displays to human-robot interaction is an emerging field of research and shows promising results. For instance, Rosen et al. [19] proposed a mixed reality head-mounted display framework to visualise future trajectories of the robot motion. To evaluate the performance of the proposed framework, they conducted a study where participants were asked to detect collisions of robot arm motions using three interfaces: no visualisation, monitor 3D point cloud view from a Kinect sensor, and mixed reality with Hololens. The authors found that the head-mounted display condition for this specific task is faster, more accurate, and subjectively more enjoyable.

III. Human-human pilot study

In order to inform the design of our reference disambiguation visualisation system, we first carried out a human-human

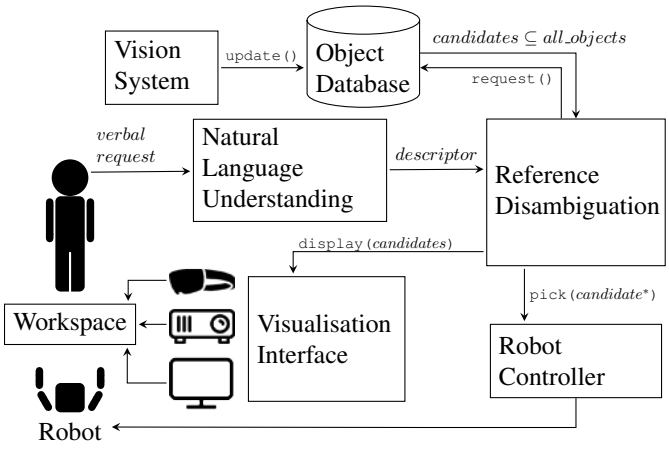


Fig. 3: The architecture of the proposed system for visualising ambiguous fetching requests.

interaction pilot study on a collaborative task involving object retrieval. We recruited 10 participants (5 pairs) that took turns in asking for and fetching Lego blocks of various colours and shapes to build a model. Since we were interested in verbal references, we asked participants to avoid pointing and instead use only verbal instructions (see Figure 2).

We found that most participants used the terms colour and shape to describe the blocks, which informed the design of the system described in the next section. Using an off-the-shelf speech recognition system to transcribe the collected audio data resulted in many incorrect object descriptors, possibly augmented by the fact that none of the participants were native English speakers. We therefore decided to make the language understanding module of our system controlled by a wizard.

IV. System description

We designed and implemented a system that takes user verbal requests as an input and processes them through the modules depicted in Figure 3. If there is ambiguity in the request (i.e., more than one object matches the colour or shape described by the user) the system highlights the candidate objects using the visualisation interface while awaiting for further verbal commands that refine the request. This process continues until there is no more ambiguity and the robot is able to pick up the target object.

When the user makes a verbal request explaining which block the robot should fetch, a human wizard performs the **natural language understanding** to extract colour and shape object descriptors supported by the system. This module is the only wizarded component of the system.

Given the object descriptors, the **reference disambiguation** module queries the object database to get the candidate objects that match the provided descriptors. The **object database** stores colour and shape attributes, 3D positions and rotation with respect to the robot of all the objects present in the workspace. The object attributes are continuously updated by the **vision system**, which uses a Microsoft Kinect

Algorithm 1 Visualisation of object highlighting extracted by object descriptors from human instructions.

```
1:  candidates ← []
2:  procedure VISUALISE(OBJECTDESCRIPTORS)
3:      candidates ← queryObjectDB(objectDescriptors)
4:      display(candidates)
5:      if len(candidates) = 1 then
6:          pick(candidate*)
7:          updateObjectDB()
8:      if singleShape(candidates)
9:          and singleColour(candidates) then
10:         displayIDs(candidates)
```

sensor. The vision system works as follows. A region-of-interest (ROI) that represents the robot's workspace is defined on the image with the objects. Individual objects within the workspace are continuously segmented using colour segmentation and morphological operations. Finally, 3D position and rotation estimates of the Lego blocks with regard to the robot are calculated using the depth information from the Kinect sensor.

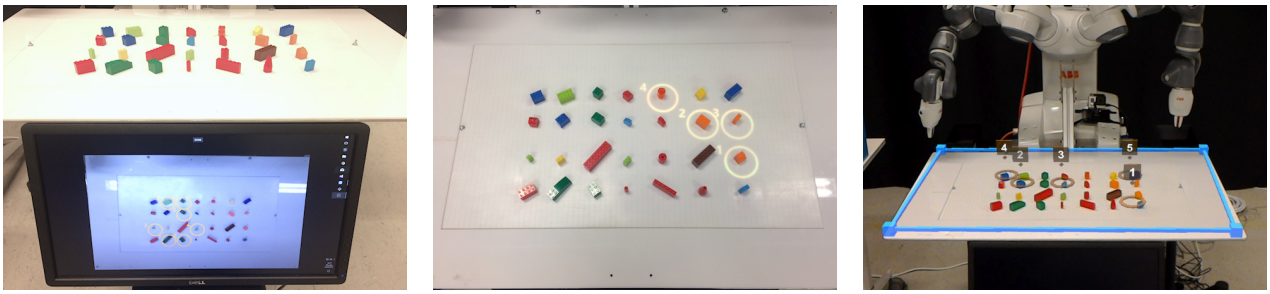
The **robot controller** module is responsible for receiving the object coordinates from the reference disambiguation module and performing motion planning to pick up the target object and place it on a side bin. In our implementation with the YuMi robot, the low-level arm motions are planned and executed using an open source ROS-based motion planner [6]. The corresponding arm for the action is selected based on the target's location.

Finally, one of the **visualisation interfaces** highlights the candidate objects that, in the robot's perceptive, match the human's verbal request of one or several objects. We developed interfaces that support three different visualisation methods: a side monitor, projector, and head-mounted display. These methods will be described in more detail in the next section.

V. Evaluation

Our evaluation scenario consisted of a YuMi robot capable of retrieving Lego blocks following participants' verbal instructions. Using this scenario, we evaluated the reference disambiguation system described in the previous section by comparing three different visualisation modalities using a within-subjects design:

- **Monitor (M)**. A monitor near the workspace streaming the video from a web-camera directed at the table from the top (Figure 4a). The candidate object highlights were overlayed on the video stream. The monitor was placed in a best possible position we encountered without interfering with the robot's manipulations of the objects.
  The monitor was positioned in a place where the cognitive mapping of the physical objects to the monitor is realised in an optimal way considering the available positions in the setup.



(a) Monitor (M) (b) Projector (P) (c) Head-mounted display (HMD)

Fig. 4: The three visualisation methods evaluated in our experiment.

- **Projector (P)**. In this condition, we used a projector which augmented the candidate highlights directly on the physical workspace (Figure 4b).
- **Head-mounted display (HMD)** we used a commercial head-mounted display[1] to show the candidate objects by merging the virtual 3D highlights into the real world (Figure 4c). The virtual workspace was initially calibrated to align with the real workspace using a fiducial marker, but the continuous tracking was performed based on the spatial mapping provided by the head-mounted display.

In our design, we did not want to favour any particular modality by creating a scenario that is not possible to replicate in the other conditions. For example, if the shared workspace was not a flat surface or had many occlusions, it would be expected that the head-mounted display condition would outperform the other two. To provide a fair comparison between the three conditions and avoid potential bias of the interface design in our measures of interest, we developed user interfaces as similar to each other as possible.

The interface of each modality consisted of overlaying circles around objects that the robot is considering to pick up at the moment, and the numbers near them appear when further descriptive disambiguation is impossible. For example, a request "give me a blue cylinder" cannot be further disambiguated if there are two suitable blocks left on the table. Thus, we augment numbers near the objects so that the request could be disambiguated. In the head-mounted display condition, to make the boundaries of the field of view more clear for participants, we show an additional bounding box around the table.

A. Hypotheses

We formulated the following hypotheses for this experiment:

- **H1:** Participants will take longer to complete trials in the M condition than in the P and HMD conditions.
- **H2:** Participants will commit fewer mistakes in the P and HMD conditions than in the M condition.
- **H3:** Participants will consider the HMD condition more engaging than the M and P conditions.
- **H4:** Participants will consider the P condition less disruptive compared to the other two conditions.
- **H5:** Participants will prefer the P and HMD conditions to the M condition.

We base **H1** and **H2** on the premise that if participants need to perform spatial mapping from the shared workspace to the the monitor, this will potentially contribute to a higher cognitive load. Similarly, because participants need to look away from the workspace and back at the monitor in M, this will likely increase the number of errors. Our reasoning for establishing **H3** and **H5** is drawn from previous research showing that augmented reality applications can improve user experience [16], [19]. **H4** is argued for by reasoning that the augmented reality condition will enable participants to dedicate full attention to the workspace.

B. Participants

A total of 29 subjects (12 female, 17 male), with ages between 22 and 50 ($M = 28.8$), were recruited for this experiment using mailing lists and flyers. To be able to participate in the experiment, subjects needed to be fluent in English, not have any colour vision deficiency and not wear glasses (due to difficulties wearing the mixed reality device).

On a scale from 1 to 5 (with 1 representing high), participants' familiarity with digital technology was 1,8. Additionally, 21 out of the 29 participants had tried Augmented or Virtual Reality before, while 9 out of 29 had interacted with a robot before.

C. Procedure

Upon arrival, participants were given a consent form and instructions about the experimental process. They were instructed to ask a robot to pick up a set of Lego blocks using only colour and shape descriptors without pointing or using spatial references (e.g. "the block next to the red one").

After that, participants went through a training phase with the experimenter where they picked up Lego blocks in turns as if they were interacting with the robot to get familiar with the task. Before each experimental trial, participants were given a piece of paper listing images of the blocks they would have to request from YuMi. Each trial consisted of 15 turns where the human participant and the robot took turns while requesting Lego blocks from each other from the

[1]https://www.microsoft.com/en-us/hololens

shared workspace. The participant started first and requested in each trial 8 objects and the robot 7. While participants had to make their requests using verbal descriptors, YuMi's requests consisted of highlighting the target block using the active visualisation modality in that trial. This type of request was included in the experiment to ensure that participants took actions in the physical workspace and avoid, for example, that in the M condition they followed the video feed shown in the monitor. Each trial took 8 minutes on average. Participants filled task questionnaires after each trial and a final questionnaire at the end of the experiment.

We used a balanced Latin square array to counterbalance the order of conditions being tested by each participant and avoid order effects. The initial arrangement of Lego blocks on the table was randomised in each session, meaning that participants did not use the same arrangement twice. An experimenter was always present in the room to ensure blocks were removed from the table in cases of occasional grasping errors and intervene if necessary.

We recorded audio and video in all sessions and logged time measurements and object requests for further analysis.

D. Measurements

To investigate the presented hypotheses, we collected both objective and subjective measures. From the interaction logs, we extracted the average **request time** per object considering the portions of the task where the participant describes a Lego block for the robot to pick up to the moment the Reference Disambiguation module sends a pick request to the robot controller (note that this excludes the robot's action completion time). The first two human block requests were excluded from each trial because their duration might have been biased by the fact that participants were still getting used to the modality/device (especially in the HMD condition). A human annotator analysed the video recordings and counted the number of incorrect task executions per trial, i.e. when participants either described the wrong block to the robot or picked a block different than the requested one. This frequency was normalised by the total number of turns of each trial and will be referred to as the **error rate** per trial.

After participating in each trial, participants answered subjective questions extracted from The Presence Inventory [13] and the Presence Questionnaire [26] about their perceived **engagement**, **observability** (i.e. how well they could observe the robot's behaviour) and **display interference** (i.e. the degree to which the visual display quality interfered with or distracted from task performance). Participants answered these questions using a 7-point Likert scale where 1 meant "Not at all" and 7 meant "Very much". At the end of the experiment, they answered additional questions regarding their **preferences** such as which condition they preferred, which condition they found easiest to perform the task and which condition would they pick to work with in the future. The final survey also included open ended questions about the advantages and disadvantages of each modality, as well as generic questions about participants' previous experience with robots, video games and AR/VR devices.

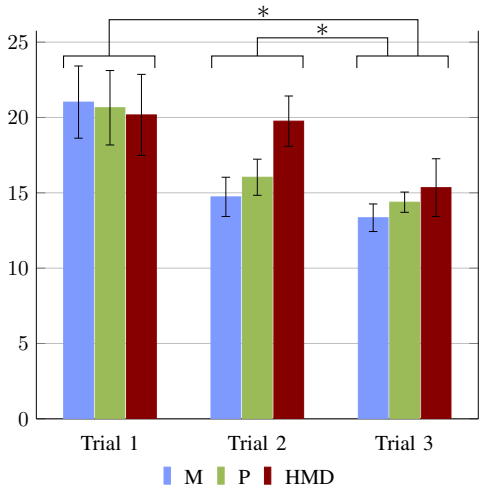


Fig. 5: Average duration (in seconds) of participants' verbal request by trial number and condition. (*) denotes p $<$ .05.

VI. Results

This section presents the results of the objective and subjective measures collected in the experiment.

A. Objective Measures

The objective measures were analysed using one-way repeated measures ANOVA. Because the first trial of each participant took longer than the other two trials in general, for analysing the request time variable, we included the order of the trials as a within-subjects factor. Post hoc tests were performed using the Bonferroni correction. Figure 5 complements the results presented below.

1) Request Time: For the portions of the task where participants described a Lego block for the robot to pick, we found no significant main effect of condition, $F(2, 14) = 1.19, p = .33, \eta^2 = .15$. A significant order effect was found, $F(2, 14) = 11.43, p < .05, \eta^2 = .62$, such that the average duration of request turns was higher in the first trial ($M =$ 20.61, $SE =$ 1.48) than in the second ($M =$ 16.84, $SE =$ .53) and third ($M =$ 14.36, $SE =$ .44) trials, regardless of condition. Post hoc tests revealed no significant differences between the first and second trials ($p = .18$), but a significant difference between the second and third trials ($p < .05$), as well between the first and third trials ($p < .05$). No significant interaction effect was found between condition and trial, $F(4, 28) = .57, p = .69, \eta^2 = .08$.

2) Error Rates: We found a significant effect of condition, $F(2, 56) = 3.22, p < .05, \eta^2 = .10$, such that in the P condition the participants had the lowest error rates ($M =$ .01, $SE =$.01), followed by the M condition ($M =$ .02, $SE =$.01) and then the HMD condition ($M =$ .04, $SE =$.01). Post hoc tests revealed that the P condition had significantly lower error rates than the HMD condition ($p = 1.0$), but no significant differences were found between error rates between M and HMD, nor M and P.

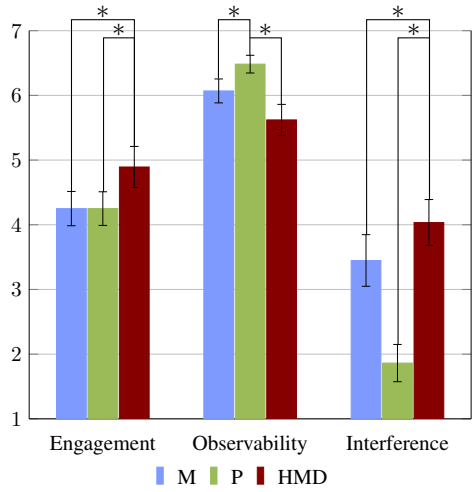


Fig. 6: Questionnaire responses for perceived Engagement, Observability and Display Interference. Ratings were provided on a 7-point Likert scale. (*) denotes p $<$ .05

B. Subjective Measures

The subjective measures collected after each trial (engagement, observability and display interference) were analysed using one-way repeated measures ANOVA, and the multiple choice questions of the final survey we analysed using Chi Squared tests. When post hoc comparisons were done, we used the Bonferroni correction. The results reported here are summarised in Figure 6.

1) Engagement: We found a statistically significant effect of condition, $F(2, 54) = 4.93, p < .05, \eta^2 = .15$, such that participants found the HMD condition to be the more engaging ($M = 4.89$, $SE = .32$) than both M ($M = 4.25$, $SE = .27$) and P ($M = 4.25$, $SE = .26$). Post hoc tests revealed that engagement ratings were significantly higher in the HMD condition than both the M and P conditions ($p < .05$ in both comparisons), but no significant differences were found in perceived engagement between the M and P conditions ($p = 1.0$).

2) Observability: There was a statistically significant effect of condition, $F(2, 56) = 8.74, p < .01, \eta^2 = .24$, such that participants considered that they were best able to observe the robot's behaviour in the P condition ($M = 6.48$, $SE = .14$), followed by the M condition ($M = 6.07$, $SE = .19$) and finally the HMD condition ($M = 5.62$, $SE = .24$). Post hoc tests showed that the P condition was considered better to observe the robot's behaviour compared to the M and HMD conditions ($p < .05$ for both comparisons), but no significant differences were found between the M and HMD conditions ($p = .19$)

3) Display interference: A statistically significant effect was found of condition, $F(2, 56) = 14.11, p < .001, \eta^2 = .34$. The condition in which the display less interfered with the task was the P ($M = 1.86$, $SE = .29$), followed by the M ($M = 3.45$, $SE = .40$) and then the HMD ($M = 4.03$, $SE = .36$). Post hoc comparisons revealed that these differences were statistically significant between M and HMD ($p < .05$), P and HMD ($p < .001$), but not between M and HMD ($p = .64$).

TABLE I: Preference Results (one participant did not answer one of the questions).

Question	M	P	HMD
Prefer	1	25	3
Easiest	4	20	4
Use Again	2	23	4

4) Overall Preferences: There was a significant difference in the answers to "In which condition did you prefer to use the robot?" ($\chi^2 = 36.69, p < 0.001$) such that the highest number of participants preferred the P condition. Similarly, in the responses to the question "In which condition did you find it easiest to perform this task?", participants found the P condition significantly easier than the other two conditions ($\chi^2 = 18.29, p < 0.001$). Finally, we found a statistically significant difference in answers to the question "Which condition would you pick to work with?" ($\chi^2 = 27.79, p < 0.001$), such that the P condition was the one participants would prefer to work with in the future.

VII. Discussion

Our first hypothesis stated that participants would take longer to complete the task in the M condition compared to the P and HMD conditions. This hypothesis was not supported, as there were no significant differences between the request times between conditions. The significant difference between the average request duration in the first trial compared to the other two trials was likely caused by a learning curve on how to interact with the system: even though participants were told that YuMi was only capable of understanding shapes and colour descriptions, in the first trial participants tended to use other ways to describe the objects such as spacial references (e.g., "the one closer to you") that were not supported by the system.

H2 stated that participants would commit fewer mistakes in the P and HMD conditions than in the M condition, a hypothesis that was partially supported. Although the smallest error rates occurred in the P condition, participants committed more task mistakes in the HMD than in the M condition. We believe that the errors in the HMD condition were mainly a consequence of limitations of the mixed reality device such as limited field of view, which lead participants to sometimes lose their perspective of the entire workspace. Nevertheless, the average error rate was fairly low in all conditions.

Despite the higher error rates in the HMD condition, participants did find this condition more engaging than the other two conditions, a finding aligned with previous research on augmented reality in HRI [16]. One of the mentioned advantages of the HMD condition which might have contributed to higher engagement was the increased freedom to move around in the environment; regardless of their point of view, they were able to visualise the highlighted objects. Therefore, H3 (participants will consider the HMD condition the most engaging) was supported.

In H4, we stated that the P condition would be considered less disruptive than the other two conditions. This hypothesis was supported by our results for observability and display interference. Not surprisingly, in the open ended questions participants mentioned that because of the wearable device in the HMD condition, and the fact that they had to switch their attention between the monitor and the workspace in the M condition, these two conditions were more disruptive than the P condition.

The questions regarding modality preferences followed the same trend as H4 and participants clearly chose the P condition over the other two conditions. Many participants used words like "natural', "easy to understand" and "simple" to characterize the P condition. Some participants considered this modality to require the least cognitive load of all the conditions they interacted with. On the other hand, participants considered the HMD condition to be more intrusive, with a limited field of view for the visualisation projection and somewhat uncomfortable to wear after some time because of its weight. While some of these disadvantages will become less evident with advances in hardware, see-through head-mounted displays will likely remain more intrusive than the other two types of modalities we investigated. Regardless of these limitations, participants appreciated the "portability" aspect in the HMD condition, especially when compared to the projector in the P condition. The most common disadvantage identified in the M condition was the need to map the scene back and forth between the monitor and the physical workspace. Participants who preferred the M condition often did so for considering this modality to be the most familiar to them.

Our main goal was to investigate the impact of projector and head-mounted display visualisation methods when compared with typical ways of visualising information such as a monitor. As such, we deliberately decided not to include a control condition where the robot used pointing or follow up questions to disambiguate requests. Furthermore, it is important to note that without any sort of disambiguation requests, participants would not be able to complete parts of the task, since in each trial there was at least one situation where two objects had the same shape and colour.

A. Limitations

As one of the initial explorations in this domain, our experiment has several limitations that need to be addressed in future work. For example, we did not account for task difficulty (all the trials had similar levels of ambiguity), the objects were arranged in such a way that from most participants' viewpoints there were no occlusions, and the shared workspace consisted of a flat surface. As such, further research is needed to see whether the same results apply to more difficult tasks that would increase participants' cognitive load, as well as to more complex scenes where either because of the object placement or the nature of the projection surface, the 3D projections (only possible in the mixed reality condition) would play a more important role in the visualisations.

Finally, in the trial phase participants were able to practice the flow of the task with the experimenter, but we did not give them the opportunity to wear the head-mounted display until they actually had to use it in the trial. While most participants reported to have used other AR and VR devices before, the lack of experience with such interfaces might have an impact on participants' performance. In the attempt to account for this effect, we excluded the first two request turns of each trial, but a larger participant sample would have helped us to better understand whether previous experience with such devices influenced the results.

B. Design implications

Our findings indicate that the three investigated visualisation methods (monitor, projector and head-mounted display) are equally effective for displaying the robot's intentions in the presence of ambiguous requests. Nevertheless, other factors such as user experience, the nature of the task and practical considerations about cost and flexibility of the setup might affect the choice of one modality over another. This section discusses the advantages and disadvantages of each modality along these factors to inform future decisions of employing these methods in HRI scenarios.

User experience. While users found the head-mounted display modality more engaging, not surprisingly they also considered it the most intrusive. Since engagement and attention are related concepts [26], head-mounted displays can be useful in tasks requiring the user to remain extremely focused. However, given the current hardware limitations in weight and field of view of these devices, head-mounted displays might not be suitable for very long tasks. As discussed in the limitations, the cognitive load in the monitor condition is likely to increase as task complexity increases, which might negatively affect users' engagement and task performance. As such, projector or head-mounted display modalities might be suitable for more complex tasks.

Technical Considerations. The head-mounted display modality is better at dealing with occlusions and non-flat surfaces, but its limited field of view can become an issue in very large workspaces. These considerations are therefore relevant when considering the target application domain where the projections will be used. It is important to note, however, that with hardware improvements (which are likely to happen given the increasing research in this area) these considerations will tend to change over time.

Practical Issues. Although the monitor and the projector are more familiar and in general less expensive solutions, it should be noted that they are less flexible for requiring a permanent installation on top of the workspace. While this is not a problem for stationary workspaces, when considering, for example, fetching tasks with mobile robots, the lack of mobility in the setup can become an issue. In this case, a head-mounted display solution becomes a clear choice.

VIII. Conclusions

In this paper, we investigated different visualisation methods for conveying to users which objects a robot is consider-

ing given verbal requests. We conducted a controlled experiment to compare three visualisation interfaces: head-mounted display, projector and a monitor as a control condition. Both objective (request time and error rate) and subjective measures (engagement, observability, display interference and preferences) were taken into account.

Our assumption was that head-mounted display and projector interfaces will decrease task time and increase accuracy compared to the control condition. However, the results of our findings showed no significant difference in task time related to condition. On the other hand, the head-mounted display interface increased error rates compared to the other two conditions (although these were generally low). Despite this fact, participants found the head-mounted display condition more engaging. Most participants preferred the projector modality because they found it the easiest to use and less intrusive for this specific setting.

In future work, we will explore two different research directions. One of them is to explore benefits of the head-mounted displays in the tasks with irregular surfaces and object occlusion. Another relevant topic to investigate is the integration of other human perception modalities, such as pointing and gaze direction, to complement verbal requests and investigate the effects of visualisation methods for even more effective disambiguation.

IX. Acknowledgements

This work is supported by the SSF (Swedish Foundation for Strategic Research) projects COIN and FACT.

References

[1] H. Admoni, T. Weng, and B. Scassellati, "Modeling communicative behaviors for object references in human-robot interaction," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 3352–3359.

[2] R. S. Andersen, O. Madsen, T. B. Moeslund, and H. B. Amor, "Projecting robot intentions into human environments," in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 2016, pp. 294–301.

[3] R. A. Bolt, *"Put-that-there": Voice and gesture at the graphics interface*. ACM, 1980, vol. 14, no. 3.

[4] R. T. Chadalavada, H. Andreasson, R. Krug, and A. J. Lilienthal, "That's on my mind! robot to human intention communication through on-board projection on shared floor space," in *Mobile Robots (ECMR), 2015 European Conference on*. IEEE, 2015, pp. 1–6.

[5] J. Y. Chai, L. She, R. Fang, S. Ottarson, C. Littley, C. Liu, and K. Hanson, "Collaborative effort towards common ground in situated human-robot dialogue," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 2014, pp. 33–40.

[6] S. Chitta, I. Sucan, and S. Cousins, "Moveit![ros topics]," *IEEE Robotics & Automation Magazine*, vol. 19, no. 1, pp. 18–19, 2012.

[7] H. H. Clark, *Using language*. Cambridge university press, 1996.

[8] J. A. Frank, M. Moorhead, and V. Kapila, "Mobile mixed-reality interfaces that enhance human-robot interaction in shared spaces," *Frontiers in Robotics and AI*, vol. 4, p. 20, 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2017.00020

[9] S. Guadarrama, L. Riano, D. Golland, D. Go, Y. Jia, D. Klein, P. Abbeel, T. Darrell *et al.*, "Grounding spatial relations for human-robot interaction," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*. IEEE, 2013, pp. 1640–1647.

[10] J. Hough and D. Schlangen, "It's not what you do, it's how you do it: Grounding uncertainty for a simple robot," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 274–282.

[11] C. Kennington and D. Schlangen, "A simple generative model of incremental reference resolution for situated dialogue," *Computer Speech & Language*, vol. 41, pp. 43–67, 2017.

[12] S. Li, R. Scalise, H. Admoni, S. Rosenthal, and S. 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*. IEEE, 2016, pp. 44–51.

[13] M. Lombard, T. B. Ditton, and L. Weinstein, "Measuring presence: the temple presence inventory," in *Proceedings of the 12th Annual International Workshop on Presence*, 2009, pp. 1–15.

[14] G. Mehlmann, M. Häring, K. Janowski, T. Baur, P. Gebhard, and E. André, "Exploring a model of gaze for grounding in multimodal hri," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 247–254.

[15] R. Paul, J. Arkin, N. Roy, and T. Howard, "Grounding abstract spatial concepts for language interaction with robots."

[16] A. Pereira, E. J. Carter, I. Leite, J. Mars, and J. F. Lehman, "Augmented reality dialog interface for multimodal teleoperation," in *Robot and Human Interactive Communication (RO-MAN), 2017 26th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2017.

[17] P. Renner, T. Pfeiffer, and I. Wachsmuth, "Spatial references with gaze and pointing in shared space of humans and robots," in *International Conference on Spatial Cognition*. Springer, 2014, pp. 121–136.

[18] S. F. B. K. N. R. Rohan Paul, Andrei Barbu, "Temporal grounding graphs for language understanding with accrued visual-linguistic context," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 4506–4514. [Online]. Available: https://doi.org/10.24963/ijcai.2017/629

[19] E. Rosen, D. Whitney, E. Phillips, G. Chien, J. Tompkin, G. Konidaris, and S. Tellex, "Communicating robot arm motion intent through mixed reality head-mounted displays," *arXiv preprint arXiv:1708.03655*, 2017.

[20] E. Ruffaldi, F. Brizzi, F. Tecchia, and S. Bacinelli, "Third point of view augmented reality for robot intentions visualization," in *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*. Springer, 2016, pp. 471–478.

[21] A. Sauppé and B. Mutlu, "Robot deictics: How gesture and context shape referential communication," in *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*. ACM, 2014, pp. 342–349.

[22] G. Skantze, "Jindigo: a java-based framework for incremental dialogue systems," *Proceedings of Interspeech. submitted, www. jidingo. net*, 2010.

[23] L. Steels and M. Hild, *Language grounding in robots*. Springer Science & Business Media, 2012.

[24] D. Whitney, E. Rosen, J. MacGlashan, L. L. Wong, and S. Tellex, "Reducing errors in object-fetching interactions through social feedback," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1006–1013.

[25] T. Winograd, "Understanding natural language," *Cognitive psychology*, vol. 3, no. 1, pp. 1–191, 1972.

[26] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence: Teleoperators and virtual environments*, vol. 7, no. 3, pp. 225–240, 1998.