Communicating Robot Arm Motion Intent Through Mixed Reality Head-mounted Displays

Eric Rosen*, David Whitney*, Elizabeth Phillips, Gary Chien,
James Tompkin, George Konidaris, Stefanie Tellex

Abstract Efficient motion intent communication is necessary for safe and collaborative work environments with collocated humans and robots. Humans efficiently communicate their motion intent to other humans through gestures, gaze, and social cues. However, robots often have difficulty efficiently communicating their motion intent to humans via these methods. Many existing methods for robot motion intent communication rely on 2D displays, which require the human to continually pause their work and check a visualization. We propose a mixed reality head-mounted display visualization of the proposed robot motion over the wearer's real-world view of the robot and its environment. To evaluate the effectiveness of this system against a 2D display visualization and against no visualization, we asked 32 participants to labeled different robot arm motions as either colliding or non-colliding with blocks on a table. We found a 16% increase in accuracy with a 62% decrease in the time it took to complete the task compared to the next best system. This demonstrates that a mixed-reality HMD allows a human to more quickly and accurately tell where the robot is going to move than the compared baselines.

Eric Rosen, David Whitney, and Stefanie Tellex
Humans To Robots Lab, Brown University. {eric_rosen,david_whitney,stefanie_tellex}@brown.edu

Elizabeth Phillips
Humanity Centered Robotics Initiative, Brown University, elizabeth_phillips1@brown.edu

Gary Chien and James Tompkin
Brown University, {gary_chien,james_tompkin}@brown.edu

George Konidaris
Intelligent Robot Lab, Brown University, george_konidaris@brown.edu

* First two authors contributed equally

1 Introduction

Industrial robots excel at performing precise, accurate, fast, and repetitive tasks. This makes them ideal for activities like car assembly. One major drawback of these robots is that humans are unable to easily predict their motions, which forces most industrial robots to be isolated from human workers and restricts human-robot collaboration. This is especially true in a fluid working environment without rigidly-defined tasks, or where robots have autonomy. Although the intended robot motion is defined ahead of time through motion planning, efficiently conveying the intended motion to a human is difficult. Human-robot collaboration requires robots to communicate to humans in ways that are intuitive and efficient [11]; yet, the motion intention inference problem leads to many safety and efficiency issues for humans working around robots [12].

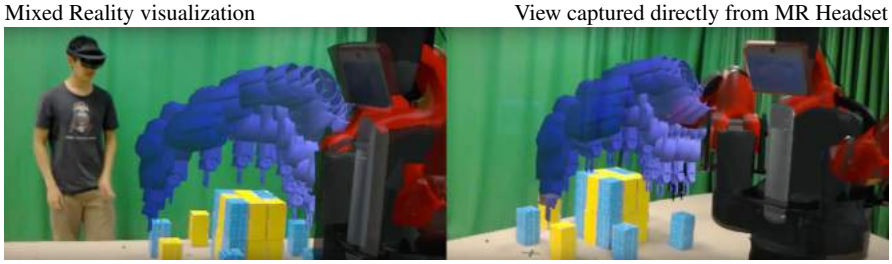
This problem has inspired research into how robots might effectively communicate intent to humans. Current interfaces for communicating robot intent have limitations in expressing motion plans within a shared workspace. Humanoid robots can try to mimic the gestures and social cues that humans use with each other, but many robots are not and cannot be humanoid by design. The motion robots intend to make can also be visualized on a 2D display near the robot. This requires the human to take their attention away from the robot's physical space to observe the display, which could be dangerous. Additionally, a 2D projection of a 3D motion plan can take time for a human to understand, requiring interaction to inspect different points of view.

Natural communication might be achieved when humans can see a robot's future motion in the real world from their own point of view, via a head-mounted display [27, 26]. This could increase safety and efficiency as the human no longer needs to divert their attention. Further, as the 3D motion plan would be overlaid in 3D space, human users would not need to make sense of 2D projections of 3D objects.

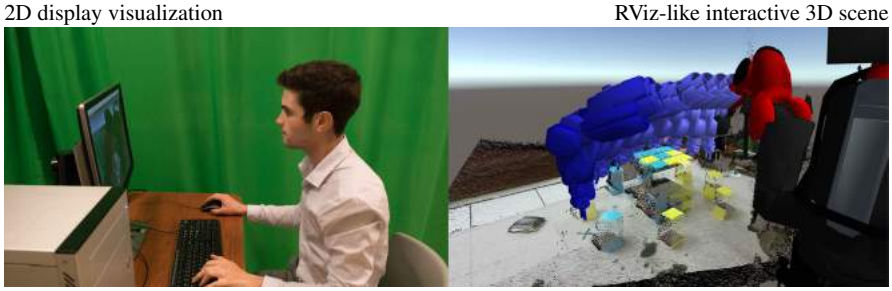
We experimentally test this idea with a system that enables humans to view a robot's intended motion via 3D graphics on a mixed reality (MR) head-mounted display (HMD)—the Microsoft HoloLens. This allows a participant to visualize the motion of the robot's arm in the real workspace before it moves, preventing collisions with the human or with objects. As there is no existing open source HoloLens ROS integration within the robotics community, we have released our code: `https://github.com/h2r/Holobot`. This integrates HoloLens with the widely-used Unity game engine, provides a URDF parser to quickly import robots into Unity, and networking code to send messages between the robot and HoloLens.

We evaluated our MR system by comparing it to both a 2D display interface and a control condition with no visualization (Fig. 1). In a within-subjects-design study, 32 participants used all three system variants to classify arm motion plans of a Rethink Robotics Baxter as either colliding or not colliding with blocks on a table. Our MR system reduced task completion time by 7.4 seconds on average (a reduction of 62%), increased precision by 11% percent on average, and increased accuracy by 16% percent on average, compared to the next best system (2D display). Additionally, we improved subjective assessments of system usability (System Usability Scale) and mental workload (NASA Task Load Index). This experiment shows the promise of mixed-reality HMDs to further human-robot collaboration.

Mixed Reality visualization View captured directly from MR Headset



2D display visualization RViz-like interactive 3D scene



No visualization Stroboscopic photo of robot arm motion

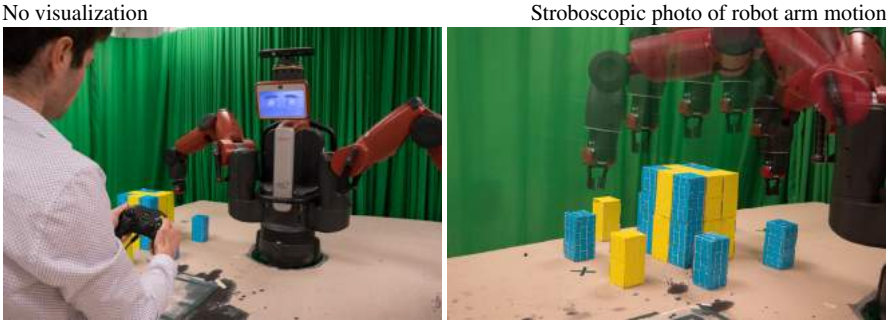


Fig. 1: Participants must decide whether a robot arm motion plan either collides or not with the light yellow and blue blocks on the table, across 14 trials and three interfaces. *Top to bottom:* Our three interfaces: an MR visualization, a 2D display/mouse with an RViz-like visualization, and no visualization at all. In each case, the left shows the experimental setup, and the right shows the participant view. *Top:* The HoloLens visualizes the robot arm motion plan as a sequence of blue virtual arm graphics overlaid onto the real world. *Middle:* The 2D display uses the same visualization, but the participant must use the system at a desk. *Bottom:* In the no visualization condition, the participant directly observes the robot arm move and pushes a 'stop' button on an Xbox controller if they think collision will occur.

2 Related Work

Humans use many non-verbal cues to communicate motion intent. There have been some successes at approximating these cues in humanoid robots, such as with gestures [20] and gaze [19], including via robot anthropomorphism [17]. However, often robots lack the faculty or subtlety to physically reproduce human non-verbal cues—especially robots that are not of human form. One alternative is to use animation and animated storytelling techniques, such as forming suggestive poses or generating initial movements [37]. This increases legibility: the ability to infer the robot's goal through its directed motion [10]. However, these methods still lack the ability to transparently communicate complex paths and motions. Further, tasks involving close proximity teamwork may require more detailed knowledge of how the robot will act both before and during the motion, such as in collaborative furniture assembly [27] and co-located teleoperation [35].

Other related works have used turn and display indicators on the robot to communicate navigational intent [36, 7, 28]. These techniques were found to improve human trust and confidence in robot actions; however, they did not express high detail in the motion plan [30, 29].

We can also use 2D displays to visualize the robot's future motions within its environment through systems like RViz [13]. These require the human operator to switch focus from the real world environment to the visualization display [18]. This may lead the operator to expend more time understanding the robot state and environment rather than collaborating with the robot [5, 4].

2.1 Augmented and Mixed Reality for Human-robot Collaboration

We can adapt the real-world environment around the human-robot collaboration to help indicate robot intent. One way is to combine light projectors with object tracking software to build a general-purpose augmented environment. This has been used to convey shared work spaces, robot navigational intention, and safety information [6, 1, 2]. However, building special purpose environments is time consuming and expensive, with a requirement for controlled lighting conditions. Further, they exhibit occlusions of the augmenting light from objects in the environment, and limit the number of people able to see perspective-correct graphics.

Hand-held tablet technology can allow participants to view a mixed reality of 3D graphics overlaid onto a camera feed of the real world [24]. These types of approaches mediate the issue of diverted attention which 2D displays suffer. However, they limit the ability of the operator to use their hands while working, and there is a mismatch in perspective between the eyes of the human and the camera in the tablet.

Optical head-mounted displays can overlay 3D graphics on top of the real world from the point of view of the human. This has been hypothesized to be a natural and transparent means of robot intent communication, for instance, with the overlay of future robot poses [26, 27]. Hopefully such a system would reduce human-robot

collaborative task time and produce fewer errors. The recent introduction of the Microsoft HoloLens has made off-the-shelf implementations of such a visualization possible. Previously, the HoloLens and other similar MR interfaces have been used in human-human collaboration, such as communicating with remote companions and playing adversarial games [15, 22, 8]. However, mixed reality as a tool to communicate robot motion intent for human-robot collaboration is nascent. This inspired us to test the hypothesis that an MR HMD which allows participants to see visual overlays on top of real-world environment in human-robot collaborative tasks is more performant than existing approaches.

2.2 3D Spatial Reasoning in Virtual Reality Displays

As HoloLens and its contemporaries are new as pieces of integrated technology, there is little direct evidence to support their efficacy in robot intent communication. However, hypotheses may be informed from literature in the parallel technology of virtual reality (VR) which, in a similar way to mixed reality (MR), provides head tracked stereo display of 3D graphics to create immersion. In VR, 3D spatial reasoning gains have been tested [32]. Pausch et al. found that head-tracked displays outperform stationary displays for a visual search task [23]. Ware and Franck found a head-tracked stereo display $3\times$ less erroneous than a 2D display for visually assessing graph connectivity [39]. Slater et al. measured performance gains in Tri-D chess for first-person perspective VR HMDs over third-person perspective 2D displays (like RViz) [31]. Ruddle et al. found navigation through a 3D virtual building was faster using HMDs over 2D displays, though with no accuracy increase [25].

Not all experiments in this area favor large-format VR. Many prior works compare immersive head-tracked CAVE displays against desktop and 'fishtank VR' displays, and often smaller higher-resolution displays induce greater performance thanks to faster visual scanning [14, 9]. Sousa Santos et al. reviewed all HMD to 2D display comparisons in the literature until 2009, and found their results broadly conflicting. Then, they conducted their own comparison for 3D navigation: on average, the desktop setup was better than the VR HMDs [33].

In general, the relationship between VR display and task performance is one with many confounding factors. The benefits over traditional 2D desktop displays are task dependent, and no clear prescriptive guidelines exist for which techniques to employ to gain what benefit. As such, while we may assume that a mixed reality interface for viewing 3D would be better, the evidence from the VR literature tells us that the issue may be more complex.

3 Experiment

To test whether mixed reality HMDs can aid robot motion intent communication, we asked novice participants to decide whether or not a robot arm motion plan would collide with blocks on a table using three interfaces: no visualization, an RViz-like 2D display visualization, and our MR visualization. Our evaluation used 32 participants (15 male, 17 female) with ages ranging from 20 to 55 ($M = 26, SD = 6.8$). We measure task completion time and true/false positive/negative rates as objective metrics, as well as the subjective assessments of system usability, likability, and workload via the SUS and NASA TLX questionnaires.

3.1 Task

In each interface, we presented each participant with the same set of 14 robot arm motions in a random order. These motions each moved from a fixed start point to a fixed end point over a table covered in blocks. Unknown to the user, exactly half of the motions collided and half did not. Each participant was tasked with labeling the motions as either colliding or non-colliding with the blocks as quickly and accurately as possible. The blocks were assembled such that it would be difficult to obtain a complete view of all blocks from just one perspective due to occlusion from other blocks. Once a participant had decided how to classify a particular motion, they pressed a button on an Xbox controller to indicate their decision. Participants could freely walk around to view the environment from different perspectives.

3.2 Interfaces

We compared three interfaces (Fig. 1):

- **No visualization:** This condition simulated a participant supervising a robot with an emergency stop button. Participants watched the arm move, and pressed an Xbox controller button to stop the arm if they thought it would collide.
- **Monitor:** Participants viewed and interacted with a 2D monitor on a desk. The visualization consisted of: 1) a 3D model of the robot and its future arm poses, obtained by importing a description of the robot in URDF format and preplanned path joint poses, and 2) a 3D point-cloud of the environment, captured by a Kinect v2 sensor mounted near the robot. The robot's future arm poses changed color from a light to dark blue to represent what point in time the robot would be in that arm pose, allowing participants to see the entire planned path. In this interface, the robot arm did not move. Participants could move the virtual camera in the visualization to gain different perspectives using a keyboard-and-mouse-based

control scheme (as in RViz [13]). For consistency, participants recorded their assessment using an Xbox controller.

- **Mixed Reality (MR):** Through a HoloLens, participants viewed a similar 3D visualization of the robot with the motion overlaid on top of the real world. In this case, there is no need to visualize the environment via a point cloud because the participant can see it directly. Based on the visualization and their physical motion in the real world, participants decided upon whether the motion collided or not and recorded their prediction using an Xbox controller. Likewise, in this interface, the robot arm did not move.

The no visualization component of our study is different from the monitor and mixed reality components because the arm moves. We move the arm because asking the participant to judge whether the arm will collide in the future with no clues whatsoever is pure guesswork. However, moving the arm makes it less comparable to the visualization components, especially in the case of measuring task time. Given that our main interest was to compare the effectiveness of the mixed reality and monitor based visualization techniques, we consider the no visualization technique to be an indirect baseline only.

3.3 Experimental Procedure

We began by reading a consent document to the participant. After consenting, participants were asked to complete our navigational intent task using all three interfaces. The no visualization condition was always completed before the other two interfaces. Participants received instruction to hit the stop button if and only if they thought the arm was about to collide with a tower. Then, after a 3-2-1 countdown, we started the arm moving.

The monitor and MR interfaces then followed. We counterbalanced the order in which participants completed the monitor and MR conditions after completing the no visualization condition (i.e., half of participants completed the no visualization condition followed by the MR condition and then the monitor condition, the other half of participants completed the no visualization condition followed by the monitor condition and then the MR condition). Participants were randomly assigned to complete one of the two counterbalancing conditions. For the MR and monitor conditions, participants received instructions to label the robot's planned motion as quickly and accuracy as possible. Then, after a 3-2-1 countdown, we displayed the visualization. After completing the task for all 14 robot arm motions with each interface, the participant completed three questionnaires.

3.4 Measurements

We chose the choice of interface as the within-subjects independent variable. In all three interfaces, our objective dependent variables were the true positive rate of correctly classifying a path as colliding, and the true negative rate of correctly classifying a path as non-colliding. We also accounted for participant strategy in labeling each motion as colliding or non-colliding (e.g., showing a tendency to always label a motion plan as colliding).

In the monitor and MR interface conditions, we also measured the average speed of labeling each motion plan by recording the time elapsed from first seeing the visualization of the planned path to labeling the path. This allowed us to measure the accuracy and precision with which each interface allowed participants to label the robot's intended motion.

Our subjective dependent variables were participant workload as measured by the NASA Task Load Index (NASA-TLX) questionnaire [21], system usability as measured by the System Usability Scale (SUS) questionnaire [3], and our own questionnaire measuring perceived predictability and preference for each interface.

- **NASA-TLX:** This is a widely-used assessment questionnaire which asks participants to provide a rating of their perceived workload during a task across six sub-scales: mental demand, physical demand, temporal demand, effort, frustration, and performance. We measured the first five on scales from 0 (Low) to 100 (High), with performance measured from 0 (Perfect) to 100 (Failure). For this evaluation, the weighted measure of paired comparisons among the sub-scales was not included. The workload score is calculated as the average of the six sub-scales.
- **SUS:** This questionnaire assesses overall system usability by asking participants to rate ten statements on a 7-point Likert scale ranging from "strongly disagree" to "strongly agree." The statements cover different aspects of the system, such as complexity, consistency, and cumbersomeness. SUS is measured on a scale from 0 to 100, where 0 is the worst score and 100 is the best.
- **Ours:** This assessed how participants felt each interface helped them to accurately predict collisions. Participants were asked to rate three statements, one for each condition, on a 7-point Likert scale ranging from "strongly disagree" to "strongly agree." For instance, "When using the monitor and keyboard, I felt I could accurately predict collisions." We also asked participants to select which interface they enjoyed the most, which interface made understanding the robot's motion the easiest, and which interface they preferred for completing the task.

3.5 Hypotheses

We expected that participants would show the best performance in the Mixed Reality interface condition followed by the monitor interface (i.e., highest true positives/negatives, least false positives/negatives, lowest levels of mental workload,

highest usability, predictability, and system preference scores). Additionally, we hypothesize that participants would have a faster labeling speed with the MR interface compared to the monitor interface.

- **H1:** MR will be the easiest interface for completing the motion labeling task. This will be demonstrated by participants achieving the best performance out of the three conditions, across (a) highest true positives/negatives, (b) lowest false positives/negatives, (c) lowest levels of workload, (d) highest usability scores, and (e) highest predictability and preference scores.
- **H2:** The monitor interface will be easier for completing this task than using no visualization at all. This will be demonstrated by participants achieving better performance than with no visualization, across (a) higher true positives/negatives, (b) lower false positives/negatives, (c) lower levels of workload, (d) higher usability scores, and (e) higher predictability and preference scores.
- **H3:** The MR interface will be associated with quicker labeling times than the monitor interface, as demonstrated by the average time it took for participants to label each motion as colliding or not colliding. Labeling times in the monitor and MR conditions are a function of evaluating the visualization of the robot's planned motion, whereas in the no visualization condition, labeling times are generated by watching the robot enact the planned motion. As a result, only the monitor and MR conditions are directly comparable.

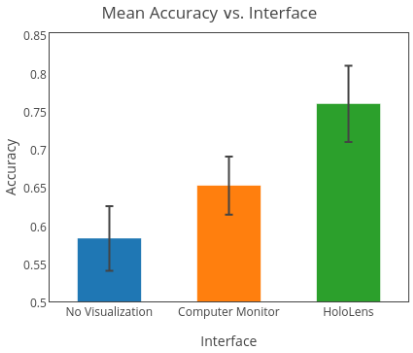
3.6 Results

Analysis Techniques

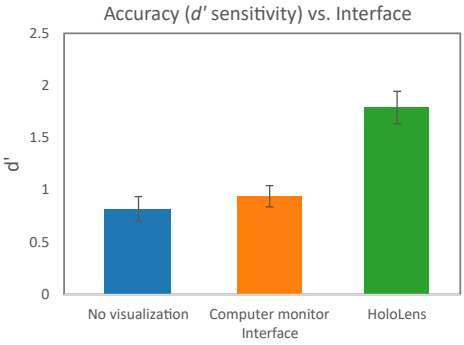
We used repeated measures analysis of variance (ANOVA) and signal detection theory (SDT) to determine if differences between measures in the three conditions were significant at the 95% confidence level. While ANOVA is likely to be familiar to the reader, SDT is less likely to be familiar, and so we will describe its use.

Signal detection theory (SDT) describes accuracy in human perception and decision making tasks by taking into account preferences for certain types of responses [16, 38]. For instance, in our task, always responding that a motion plan will collide would yield high true positive scores ("hits"), and also high false positive scores ("false alarms"). In decision making tasks with innocuous false alarms, adopting such a strategy would not affect overall performance. However, for tasks with high false alarm cost, then a strategy that results in low false alarm rates while retaining high hit rates is better. For HRI tasks like ours, false alarms would slow the collaboration considerably and so we consider them high cost.

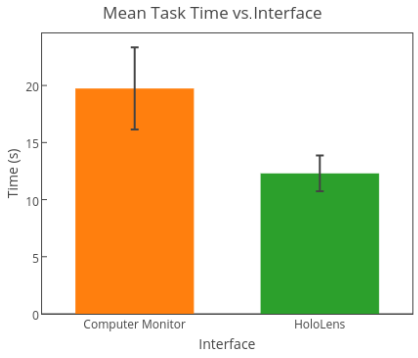
In SDT tasks, d' (also called sensitivity) is a common measure which considers decision making strategy. It is the standardized difference between the hit rate and the false alarm rate. To handle perfect scores (i.e., correctly labeling all the colliding and non-colliding paths), zero false alarm scores, and zero hit scores, we adopted the technique outlined by Stanislaw and Todorov [34].



(a) Mean accuracies across interfaces.



(b) Mean adjusted accuracy (d') across interfaces.



(c) Mean task times for comparable interfaces (please see **H3** definition).

Fig. 2: The MR interface significantly increases accuracy (*top left*) and strategy-compensated accuracy (d') (*top right*) over the baseline systems. *Bottom:* The MR interface is also significantly faster for assessing motion plans.

Accuracy

We counted the number of participant true positives, false positives, true negatives, and false negatives in each condition. From this, we report the familiar accuracy measure as the proportion of true positives plus true negatives out of the total number of motion plans (Fig. 2b). MR was the most accurate (M= 0.76, SD= 0.19), followed by the monitor (M= 0.66, SD= 0.14), followed by the no visualization condition (M= 0.60, SD= 0.12). These differences were statistically significant (Wilks $\Lambda = 0.619$, $F(2,30) = 9.244$, $p = .001$, $\eta^2 = 0.381$), and accuracy in the MR condition was significantly better than in the monitor condition ($p = .001$) and the no visualization condition ($p < .001$). Performance in the monitor condition was not significantly better than in the no visualization condition ($p = .065$).

We also report d' scores for each participant in each of the three conditions (Fig. 2c). There was a significant difference in d' performance scores between the conditions (Wilks $\Lambda = 0.449$ $F(2,30) = 18.378$, $p < .001$, $\eta^2 = 0.551$). Further, the performance in the MR condition (M= 1.79, SD= 0.88) was significantly better than the monitor condition (M= 0.94, SD= 0.58) and the no visualization condition

($M = 0.82, SD = 0.66$), all with $p < .001$. The difference between performance in the monitor condition was not significantly better than performance in the no visualization condition (p= .42). A look at the mean accuracy and mean d' scores showed that performance in the MR, monitor, and no visualization conditions trended in the hypothesized direction although both performance indicators in the monitor condition were not significantly better the no visualization condition. Thus, hypotheses 1 (a) and (b) were supported, but hypotheses and 2 (a) and 2 (b) were not supported.

Finally, participants who completed the no visualization condition followed by the monitor condition and then the MR condition (*Order 1:*, $M = 0.67$, $SD = 0.16$) did not have significantly different accuracy scores than participants who completed the no visualization condition followed by the MR condition then the monitor condition (*Order 2:*, $M = 0.68$, $SD = 0.17$), $t(94) = .220$, $p = .826$. The same was true for the d' scores (*Order 1:*, $M = 1.13$, $= 0.85$; *Order 2:* $M = 1.24$, $SD = 0.82$), $t(94) = 0.687$, $p = 0.494$.

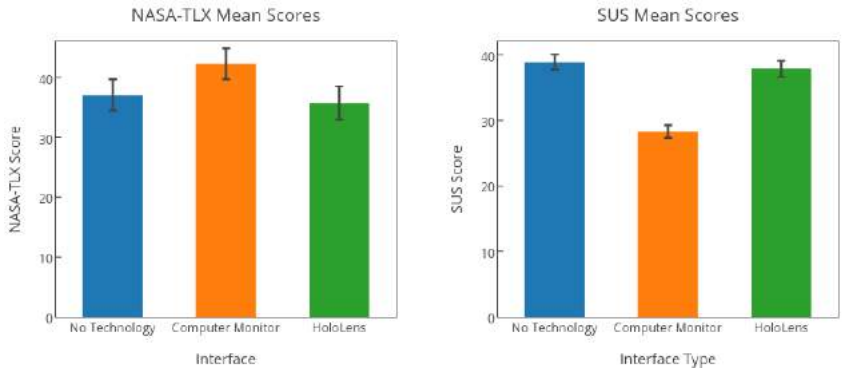
Task Time

Hypothesis 3 stated that motion labeling times would be faster in the MR condition than in the monitor condition. A paired samples t-test showed significant differences in mean motion labeling times between the two conditions ($t(31) = 3.415, p < .001$). Mean labeling times trended in the hypothesized direction (Fig. 2c). Labeling times in the MR condition were significantly shorter ($M = 11.95, SD = 8.42$) than in the monitor condition ($M = 19.39, SD = 19.28$). Hypothesis 3 was supported.

Subjective Workload

Hypotheses 1 (c) and 2 (c) stated that workload would increase from MR to monitor, and from monitor to no visualization. One-way repeated measures ANOVA was used to test for statistically significant differences in workload scores across the three interface conditions (Wilks $\Lambda = 0.802$, $F(2,30) = 3.693$, $p = 0.037$, $\eta^2 = 0.198$).

The MR condition was associated with the lowest workload scores ($M = 35.39, SD = 15.73$), followed by the no visualization condition ($M = 37.11, SD = 14.78$), and then the monitor condition ($M = 42.32, SD = 14.71$; Fig. 3a.). Post hoc comparisons showed that mean scores in the MR condition were significantly lower than in the monitor condition ($p = .040$). There was not a significant difference in workload scores between the MR condition and the no visualization condition. The difference between workload scores in the monitor condition and the no visualization condition were not significantly different. Hypotheses 1 (c), which stated that MR would have the lowest workload scores, was partially supported. Hypothesis 2 (c) was not supported as the workload scores in the monitor condition were higher than in the no visualization condition.



(a) Mean NASA-TLX scores across all interfaces. (b) Mean SUS scores across all interfaces.

Fig. 3: Participants reported the lowest levels of subjective workload in the MR condition and significantly lower workload than in the monitor condition (NASA-LTX). Participants reported the highest assessments of system usability in the no visualization condition (SUS). However, the difference between the no visualization condition and the MR condition was not significant.

Subjective Usability

Hypotheses 1 (d) and 2 (d) stated that MR would have the highest usability scores, followed by monitor, followed by no visualization. A one-way repeated measures ANOVA, showed that there was a significant difference in mean usability scores across the three conditions (Wilks $\Lambda = 0.151$, $F(2,30) = 84.342$, $p < 0.001$, $\eta^2 = 0.849$). However, the no visualization condition was associated with the highest SUS scores ($M = 38.91, SD = 1.15$), followed by the MR condition ($M = 37.88, SD = 1.26$), and the monitor condition ($M = 28.31, SD = 0.99$; Fig. 3b.). Mean SUS scores in the MR condition were significantly higher than the monitor condition ($p < 0.001$), and mean SUS scores in the no visualization condition were significantly higher than the monitor condition ($p < 0.001$). The difference between the MR condition and the no visualization condition was not significant. Hypotheses 1 (d) and 2 (d) were not supported.

Subjective Collision Predictability

Hypotheses 1 (e) and 2 (e) stated that the ordering of highest collision predictability scores would be MR, then monitor, then no visualization. We used one-way repeated measures ANOVA to test for significant differences in participants' assessments of whether or not they felt the interfaces could help them predict collisions. There were significant differences between the interfaces on this measure (Wilks $\Lambda = 0.246$ $F(2,30) = 45.891$, $p < 0.001$, $\eta^2 = 0.754$). Participants showed the highest agree-

ment that MR helped them to predict collisions ($M = 5.28, SD = 0.20$), followed by the no visualization condition ($M = 4.06, SD = 0.35$), and then the monitor condition ($M = 3.38, SD = 0.23$). The difference between mean scores in the MR condition were significantly higher than in the monitor condition and the no visualization condition (both $p's < .05$), supporting Hypothesis 1 (e). Means scores in the monitor condition were lower than the no visualization condition but not significantly so ($p = .44$). Hypothesis 2 (e) was not supported.

Subjective Enjoyment

We compared the frequencies with which participants selected each interface as the one they enjoyed the most, the one they preferred for completing the task, and the one they felt made understanding the robot's motion the easiest. All participants selected MR as the interface they enjoyed the most ($N = 32$). For the interface participants felt made understanding the robot's motion the easiest, almost all of the participants selected MR ($N = 29, 90.6\%$), while only three participants (9.4%) selected the monitor. Finally, when asked about preference for completing that task, almost all participants selected MR ($N = 30, 93.8\%$). Only two participants (6.3%) selected the monitor interface as their preferred interface for completing the task. No participants selected the no interface condition.

4 Discussion

Overall, our results demonstrate the potential benefit of MR to communicate robot motion intent to humans. Participants in the MR condition significantly outperformed the monitor condition, showing a 16% increase in collision prediction accuracy and a 62% decrease in time taken. Mixed reality also allowed participants to outperform the control condition of no visualization. Almost universally, participants selected MR as the most enjoyable interface, the easiest for completing the task, and the one they preferred for assessing the robot motion plans. Taken together, these findings strongly support our hypotheses that MR would be associated with the best objective performance measures.

An examination of participant free responses regarding why they preferred MR over monitor offers some insight into these findings. Many participants reported that using the monitor and mouse to virtually move around the robot was cumbersome, unintuitive, difficult to manipulate, distracting, and confusing. Participants reported that MR was not perfect, e.g., the motion plan overlay was not always perfectly aligned on top of the robot[2], the setup took a long time, and that physically moving around the robot was difficult at times. Even so, 34% of participants reported that they liked that they could freely move around the robot to see the planned motion,

[2] Due to imperfections in the HoloLens' SLAM, the authors noticed a drift of several centimeters could occur over a long period of use.

and that this made determining whether or not collisions would occur faster, easier, and more intuitive than when using the monitor and mouse.

The subjective questionnaire responses offered mixed but promising support for the MR condition. Although participants working with the MR condition reported lower workload than in the no visualization condition, it was not significantly lower, which offered only partial support for hypothesis H1 (c). The mean workload scores did trend in the hypothesized direction (i.e., the MR condition had the lowest workload scores overall) and the results suggests that participants did not find the MR interface more taxing than using no interface at all. Similarly, although participants rated the no visualization condition as slightly more usable than the MR condition (counter to hypothesis H1 (d)), the no visualization condition was not rated significantly more usable.

Perhaps surprisingly, the monitor condition did not significantly outperform the no visualization condition for both objective and subjective measures. Participant accuracy (and accuracy accounting for decision making strategy) was not significantly better, and when working with the computer monitor, participants reported higher workload and lower assessments of usability than when working with the no visualization condition. Put another way, looking at a robot with an emergency stop button in your hand is about as simple an interface as you could build. Finally, participants also reported the least agreement that the monitor interface could help them to accurately predict robot collisions. Thus, no part of hypothesis 2 was supported.

There are several limitations of our system. At a high level, our system only considers one method of robot-to-human motion intention communication, and there is much to explore still in enabling human-to-robot communication. Alternative methods of communicating robot motion intent may prove better. Mixed reality can also be used to communicate other things beside motion intent, such as shared goals, needed objects, or other aspects of the robot's state.

5 Conclusion

If robots and humans are to form fluid cooperative work partnerships, then they need to be able to communicate their motion intent to each other effectively. We investigated the hypothesis that mixed reality would be a natural interface for robot motion intent communication, and found that both participant performance and participant perceptions were improved with an MR visualization over the more traditional monitor interface for visualization and over no visualization at all. Our results provide evidence that mixed reality is one way to bridge the communication gap and allow robots to communicate their motion intent to humans.

Acknowledgements We thank David Laidlaw for fruitful discussion on VR literature. This work was supported by DARPA under grant number D15AP00102 and by the AFRL under grant number FA9550-17-1-0124. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA or AFRL.

References

1. J.-g. Ahn and G. J. Kim. Remote Collaboration Using a Tele-Presence Mobile Projector Robot Tele-Operated by a Smartphone. In *IEEE/SICE International Symposium on System Integration (SII)*, pages 236–241. IEEE, 2016.
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)*, pages 294–301. IEEE, 2016.
3. J. Brooke et al. SUS-A Quick and Dirty Usability Scale. *Usability Evaluation in Industry*, 189(194):4–7, 1996.
4. J. L. Burke and R. R. Murphy. Situation Awareness and Task Performance in Robot-Assisted Technical Search: Bujold Goes to Bridgeport, 2004.
5. J. L. Burke, R. R. Murphy, M. D. Coovert, and D. L. Riddle. Moonlight in Miami: Field study of human-robot interaction in the context of an urban search and rescue disaster response training exercise. *Human-Computer Interaction*, 19(1-2):85–116, 2004.
6. 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 *European Conference on Mobile Robots (ECMR)*, pages 1–6. IEEE, 2015.
7. R. T. Chadalavada, A. Lilienthal, H. Andreasson, and R. Krug. Empirical Evaluation of Human Trust in an Expressive Mobile Robot. In *RSS Workshop on Social Trust in Auton. Robots*, 2016.
8. H. Chen, A. S. Lee, M. Swift, and J. C. Tang. 3d collaboration method over hololens and skype end points. In *Proc. 3rd International Workshop on Immersive Media Experiences*, pages 27–30. ACM, 2015.
9. C. Demiralp, C. D. Jackson, D. B. Karelitz, S. Zhang, and D. H. Laidlaw. Cave and Fishtank Virtual-Reality Displays: A Qualitative and Quantitative Comparison. *IEEE Transactions on Visualization and Computer Graphics*, 12(3):323–330, 2006.
10. A. D. Dragan, K. C. Lee, and S. S. Srinivasa. Legibility and predictability of robot motion. In *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, pages 301–308. IEEE, 2013.
11. T. Fong, I. Nourbakhsh, and K. Dautenhahn. A Survey of Socially Interactive Robots. *Robotics and Autonomous Systems*, 42(3):143–166, 2003.
12. Y. Han. The Social Behavior Guide for Confused Autonomous Machines. Master's thesis, Rhode Island School of Design, 2016.
13. H. R. Kam, S.-H. Lee, T. Park, and C.-H. Kim. RViz: A Toolkit for Real Domain Data Visualization. *Telecommunications Systems*, 60(2):337–345, Oct. 2015.
14. D. J. Kasik, J. J. Troy, S. R. Amorosi, M. O. Murray, and S. N. Swamy. Evaluating Graphics Displays for Complex 3D Models. *IEEE Computer Graphics and Applications*, 22(3):56–64, May 2002.
15. H. Kato and M. Billinghurst. Marker Tracking and HMD Calibration for a Video-based Augmented Reality Conferencing System. In *IEEE and ACM International Workshop on Augmented Reality (IWAR)*, pages 85–94. IEEE, 1999.
16. N. A. Macmillan. Signal Detection Theory. *Stevens' Handbook of Experimental Psychology*, 2002.
17. A. D. May, C. Dondrup, and M. Hanheide. Show Me Your Moves! Conveying Navigation Intention of a Mobile Robot to Humans. In *European Conference on Mobile Robots (ECMR)*, pages 1–6. IEEE, 2015.
18. P. Milgram, S. Zhai, D. Drascic, and J. Grodski. Applications of Augmented Reality for Human-Robot Communication. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pages 1467–1472. IEEE, 1993.
19. B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita. Nonverbal Leakage in Robots: Communication of Intentions Through Seemingly Unintentional Behavior. In *ACM/IEEE International Conference on Human Robot interaction*, pages 69–76. ACM, 2009.
20. T. Nakata, T. Sato, T. Mori, and H. Mizoguchi. Expression of Emotion and Intention by Robot Body Movement. In *International Conference on Autonomous Systems*, 1998.

21. NASA Human Performance Research Group and others. Task Load Index (NASA-TLX) v1. 0 computerised version. *NASA Ames Research Centre*, 1987.
22. T. Ohshima, K. Satoh, H. Yamamoto, and H. Tamura. AR2 Hockey: A Case Study of Collaborative Augmented Reality. *Proceedings of the Virtual Reality Annual International Symposium*, (s 268), 1998.
23. R. Pausch, M. A. Shackelford, and D. Proffitt. A User Study Comparing Head-Mounted and Stationary Displays. In *Research Properties in Virtual Reality Symposium*, 1993.
24. J. Rekimoto. Transvision: A Hand-Held Augmented Reality System for Collaborative Design. In *Virtual Systems and Multimedia*, volume 96, pages 18–20, 1996.
25. R. A. Ruddle, S. J. Payne, and D. M. Jones. Navigating Large-Scale Virtual Environments: What Differences Occur Between Helmet-Mounted and Desk-Top Displays? *Presence*, 8(2):157–168, April 1999.
26. E. Ruffaldi, F. Brizzi, F. Tecchia, and S. Bacinelli. *Third Point of View Augmented Reality for Robot Intentions Visualization*, pages 471–478. Springer International Publishing, Cham, 2016.
27. B. Scassellati and B. Hayes. Human-Robot Collaboration. *AI Matters*, 1(2):22–23, 2014.
28. K. E. Schaefer, E. R. Straub, J. Y. Chen, J. Putney, and A. Evans. Communicating Intent to Develop Shared Situation Awareness and Engender Trust in Human-Agent Teams. *Cognitive Systems Research*, 2017.
29. M. C. Shrestha, A. Kobayashi, T. Onishi, E. Uno, H. Yanagawa, Y. Yokoyama, M. Kamezaki, A. Schmitz, and S. Sugano. Intent Communication in Navigation Through the Use of Light and Screen Indicators. In *ACM/IEEE International Conference on Human Robot Interaction*, pages 523–524. IEEE Press, 2016.
30. M. C. Shrestha, A. Kobayashi, T. Onishi, H. Yanagawa, Y. Yokoyama, E. Uno, A. Schmitz, M. Kamezaki, and S. Sugano. Exploring the Use of Light and Display Indicators for Communicating Directional Intent. In *Advanced Intelligent Mechatronics*, pages 1651–1656. IEEE, 2016.
31. M. Slater, V. Linakis, M. Usoh, and R. Kooper. Immersion, presence, and performance in virtual environments: An experiment with tri-dimensional chess. In *ACM Virtual Reality Software and Technology (VRST)*, volume 163, page 72. ACM Press New York, NY, 1996.
32. M. Slater and M. V. Sanchez-Vives. Enhancing Our Lives with Immersive Virtual Reality. *Frontiers in Robotics and AI*, 3:74, 2016.
33. B. Sousa Santos, P. Dias, A. Pimentel, J.-W. Baggerman, C. Ferreira, S. Silva, and J. Madeira. Head-mounted Display Versus Desktop for 3D Navigation in Virtual Reality: a User Study. *Multimedia Tools and Applications*, 41(1):161, 2009.
34. H. Stanislaw and N. Todorov. Calculation of Signal Detection Theory Measures. *Behavior Research Methods, Instruments, & Computers*, 31(1):137–149, 1999.
35. D. Szafir, B. Mutlu, and T. Fong. Communication of Intent in Assistive Free Flyers. In *ACM/IEEE International Conference on Human-robot Interaction*, pages 358–365. ACM, 2014.
36. D. Szafir, B. Mutlu, and T. Fong. Communicating Directionality in Flying Robots. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 19–26. ACM, 2015.
37. L. Takayama, D. Dooley, and W. Ju. Expressing Thought: Improving Robot Readability with Animation Principles. In *International Conference on Human-robot Interaction*, pages 69–76. ACM, 2011.
38. W. P. Tanner Jr and J. A. Swets. A Decision-Making Theory of Visual Detection. *Psychological Review*, 61(6):401, 1954.
39. C. Ware and G. Franck. Viewing a Graph in a Virtual Reality Display is Three Times as Good as a 2D Diagram. In *IEEE Symposium on Visual Languages*, pages 182–183, Oct 1994.