The International Journal of Robotics Research
Volume 38, Issue 12-13, October-November 2019, Pages 1513-1526

https://doi.org/10.1177/0278364919842925

Article

Communicating and controlling robot arm motion intent through mixed-reality head-mounted displays

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

Abstract

Efficient motion intent communication is necessary for safe and collaborative work environments with co-located humans and robots. Humans efficiently communicate their motion intent to other humans through gestures, gaze, and other non-verbal cues, and can replan their motions in response. However, robots often have difficulty using these methods. Many existing methods for robot motion intent communication rely on 2D displays, which require the human to continually pause their work to check a visualization. We propose a mixed-reality head-mounted display (HMD) visualization of the intended robot motion over the wearer's real-world view of the robot and its environment. In addition, our interface allows users to adjust the intended goal pose of the end effector using hand gestures. We describe its implementation, which connects a ROS-enabled robot to the HoloLens using ROS Reality, using MoveIt for motion planning, and using Unity to render the visualization. To evaluate the effectiveness of this system against a 2D display visualization and against no visualization, we asked 32 participants to label various arm trajectories as either colliding or non-colliding with blocks arranged on a table. We found a 15% increase in accuracy with a 38% decrease in the time it took to complete the task compared with the next best system. These results demonstrate that a mixed-reality HMD allows a human to determine where the robot is going to move more quickly and accurately than existing baselines.

Keywords

Mixed reality, motion planning, human–robot interaction

Humans To Robots Laboratory, Department of Computer Science, Providence, RI, USA

Corresponding author(s):

Eric Rosen, Humans To Robots Laboratory, Department of Computer Science, 115 Waterman Street, Providence, RI 02912, USA. Email: eric_rosen@brown.edu

* The first two authors contributed equally and are listed alphabetically.

1. Introduction

Industrial robots excel at performing precise, accurate, strenuous, and repetitive tasks, which makes them ideal for activities such as car assembly. A 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 move autonomously. 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 (Fong et al., 2003); yet, the motion intention inference problem leads to many safety and efficiency issues for humans working around robots (Han, 2016).

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. In addition, 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 (HMD) (Ruffaldi et al., 2016; Scassellati and Hayes, 2014). 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 test this idea with a system that enables humans to view robot intended motion via 3D graphics on a mixed-reality (MR) HMD: the Microsoft HoloLens. This allows a participant to visualize the robot arm motion in the real workspace before it moves, preventing collisions with the human or with objects (Figure 1). As there is no existing open-source HoloLens ROS integration for 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 Unified Robot Description Format (URDF) parser to quickly import robots into Unity, and network code to send messages between the robot and HoloLens.

Fig. 1. An image captured directly from the MR Headset of a user viewing a robot trajectory.

In addition to visualizing robot motion intent, it is important for the robot to be able to *replan* an intended trajectory based on human response, i.e., when the user notices that the planned robot trajectory will collide with objects in the environment. Using MoveIt (Chitta et al., 2012), we allow a user to command the robot to plan new trajectories with the same start and end points, and so visualize and choose from different robot motion trajectories.

We experimentally compare our system with both a 2D display interface and a control condition with no visualization (Figure 4). 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 38%), increased precision by 11% on average, and increased accuracy by 15% on average, compared with the next best system (2D display). In addition, we improved subjective assessments of system usability (System Usability Scale (SUS)) (Brooke et al., 1996) and mental workload (NASA Task Load Index) (NASA Human Performance Research Group and others, 1987). This experiment shows the promise of MR-HMDs to further human–robot collaboration.

2. Related work

Humans use many non-verbal cues to communicate motion intent. There is much work in approximating these cues in humanoid robots, focusing especially on gestures (Nakata et al., 1998) and gaze (Mutlu et al., 2009), as well as related work on non-verbal communication with non-humanoid robots (Cha et al., 2018). However, robots often lack the faculty or subtlety to physically reproduce human non-verbal cues, especially robots that are not of humanoid form. One alternative is to use animation and animated storytelling techniques, such as forming suggestive poses or generating initial movements (Takayama et al., 2011). This increases legibility: the ability to infer the robot's goal through its directed motion (Dragan et al., 2013). 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 (Scassellati and Hayes, 2014) and co-located teleoperation (Szafir et al., 2014).

Verbal communication has also been shown to be an effective way to have robots communicate their high-level intent Nikolaidis et al. (2017). However, although speech is useful for quickly expressing abstract actions such as "I will rotate the table," it is difficult to communicate low-level actions such as what joint angles the robot will assume throughout the planned motion. Not only is it cumbersome for the robot to explicitly state all of the relevant information for describing a high-degree-of-freedom (high-DoF) arm motion, it is not expected for humans to be able to easily interpret such speech because humans do not typically talk in this manner.

Other related works have used turn and display indicators on the robot to communicate navigational intent (Chadalavada et al., 2016; Schaefer et al., 2017; Szafir et al., 2015). These techniques were found to improve human trust and confidence in robot actions; however, they did not express high detail in the motion plan (Shrestha et al., 2016a,b).

We can also use 2D displays to visualize the robot's future motions within its environment through systems such as RViz (Kam et al., 2015; Leeper et al., 2012). These require the human operator to switch focus from the real-world environment to the visualization display (Milgram et al., 1993). This may lead the operator to expend more time understanding the robot state and environment rather than collaborating with the robot (Burke and Murphy, 2004; Burke et al., 2004).

2.1. Augmented reality and MR 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 (Ahn and Kim, 2016; Andersen et al., 2016; Chadalavada et al., 2015). 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 tablets can allow participants to view a MR of 3D graphics overlaid onto a camera feed of the real world (Rekimoto, 1996). These types of approaches mediate the issue of diverted attention that 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 HMDs 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 (Ruffaldi et al., 2016; Scassellati and Hayes, 2014). 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 MR interfaces have been used in human–human collaboration, such as communicating with remote companions and playing adversarial games (Chen et al., 2015; Kato and Billinghurst, 1999; Ohshima et al., 1998). However, MR as a tool to communicate robot motion intent for human–robot collaboration is nascent. Contemporary work investigates the use of MR for communicating drone paths (Walker et al., 2018), but there is a lack of work dealing with multi-jointed, high-DoF robots. This inspired us to test the hypothesis that a MR-HMD that allows participants to see visual overlays on top of real-world environment in human–robot collaborative tasks is an improvement over 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 MR, provides head-tracked stereo display of 3D graphics to create immersion. In VR, 3D spatial reasoning gains have been tested (Slater and Sanchez-Vives, 2016). Pausch et al. (1993) found that head-tracked displays outperform stationary displays for a visual search task. Ware and Franck (1994) found a head-tracked stereo display three times less erroneous than a 2D display for visually assessing graph connectivity. Slater et al. (1996) measured performance gains in Tri-D chess for first-person perspective VR HMDs over third-person perspective 2D displays (such as RViz). Ruddle et al. (1999) found navigation through a 3D virtual building was faster using HMDs over 2D displays, though with no accuracy increase.

Not all experiments in this area favor large-format VR. Many prior works compared 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 (Demiralp et al., 2006; Kasik et al., 2002). Santos et al. (2009) 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.

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 MR interface for viewing 3D would be better, the evidence from the VR literature tells us that the issue may be more complex.

3. Technical approach

Communicating and controlling robot motion intent requires us to join our robot control system (ROS with ROS Reality) to a motion planner (MoveIt), and to visualize the result on a MR-HMD (HoloLens with Unity) in a shared robot/headset coordinate system.

3.1. ROS and ROS Reality

ROS (Quigley et al., 2009) is a set of tools and libraries to help program robot applications. Designed for Linux systems, ROS connects robot hardware and program processes, or *nodes*, via a local area network (LAN) or wireless local area network (WLAN). Nodes communicate by streaming data over channels, or *topics*. Nodes create publisher objects to send data structures, or *ROS messages*, over different topics, or nodes create subscriber objects that manage incoming publications on those topics. These nodes do not need to be on the same computer as long as they can communicate over the ROS network. For example, the */robot_ state_publisher* node runs on the robot. It subscribes to the joint state topic, performs forward kinematics to calculate the pose of each part of the robot in Cartesian coordinates, and then publishes that information to the transform topic */tf*. Then, any device on the network can subscribe to */tf*.

To visualize robot motion intent, the HoloLens needs to interface with ROS to know upcoming robot poses. This presents a problem, as the HoloLens runs Windows 10, and applications are created in the game engine Unity, neither of which have built in support for ROS. To solve this issue, we have created ROS Reality (Whitney et al., 2017), a software package that enables the HoloLens to subscribe to and publish topics to a ROS network using WebSockets (see Figure 2). ROS Reality uses rosbridge (Toris et al., 2015) to open a WebSocket connection between a server on the ROS network and a client on the HoloLens. Messages are serialized to and deserialized from JSON objects via ROS# (Bischoff, 2017), which allows for easy creation of novel message types.

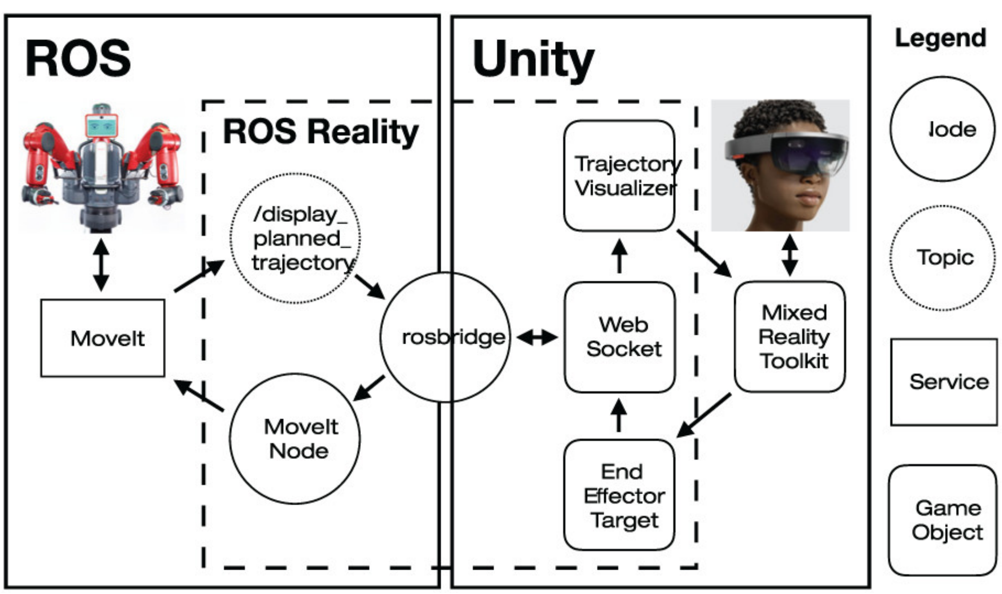


Fig. 2. A schematic of our system. Human operators use the HoloLens to interact with a Unity scene using the MixedRealityToolkit, and specify robot end-effector goal states using hand gestures. Goal poses are wirelessly

communicated over rosbridge to a ROS network using ROS Reality. The MoveIt node receives the goal pose and sends it to MoveIt, which uses the current transform of the robot from /tf as the starting pose, and publishes a plan onto /display_planned_path topic. This motion plan is sent back over rosbridge to Unity, where the TrajectoryVisualizer renders the trajectory fed from the WebSocket client.

3.2. MoveIt

For robot motion planning, we use the MoveIt (Chitta et al., 2012) software package, the most common motion planning software for ROS-enabled robots. Users are able to programmatically specify start and goal robot transforms to MoveIt from a ROS Node. With this, we need to both send desired pose information from the HoloLens to MoveIt and receive back motion plans to visualize.

To receive the planned trajectory from MoveIt, the HoloLens directly subscribes to the */display_planned_path* ROS topic published by MoveIt. This topic contains a list of time-stamped joint angles that determine the trajectory visualization. To send poses to MoveIt, we send the poses to an intermediary node on the ROS network called the MoveIt Node (see Figure 2). This node receives the poses from the HoloLens and uses MoveIt's python API to create a planning service request to MoveIt.

3.3. Microsoft HoloLens and Unity

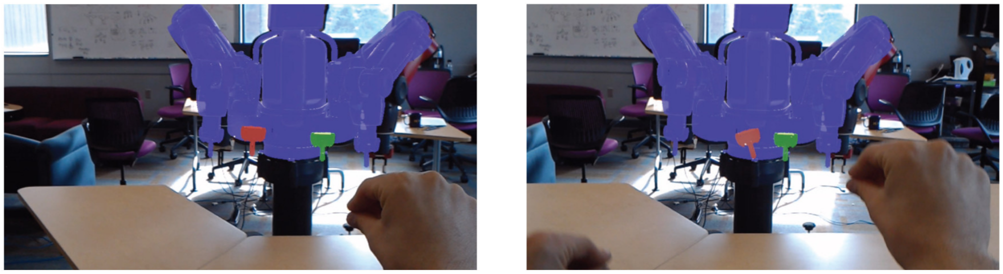
The Microsoft HoloLens is a standalone MR headset that allows users to overlay digital imagery on top of the real world. This is accomplished with an inertial measurement unit, an array of four cameras, and an infrared (IR) depth sensor, which combine to simultaneously map the environment and locate the headset inside of that map. The HoloLens supports the creation of MRs and gesture interfaces using the 3D game engine Unity (Unity Technologies, 2018) in conjunction with Microsoft's MixedReality Toolkit (MRTK) (Microsoft, 2017). Unity applications are composed of scenes for human operators to interact within a virtual space. Operators perceive the scene through the MR headset and interact with it through hand gestures and voice commands.

ROS Reality contains functions for generating realistic Unity models of real ROS robots from URDFs. In addition, ROS Reality allows the virtual robot model to mirror the live robot, and vice versa. This provides natural situational and environmental awareness of the robot, plus robot control.

3.4. Interaction walkthrough

The flow of an interaction using our system can be seen in Figure 3 and the steps are as follows.

0. *Once at startup*: Manually calibrate the MR-HMD coordinate system to the ROS coordinate system.
1. The user specifies a goal pose for each arm in the MR-HMD using gestures (see Figures 3a and 3b).
2. Using speech, the user commands the MR-HMD to send the goal poses to MoveIt via ROS Reality, which computes a motion plan (Figure 3c). Again via ROS Reality, this plan is sent back to the MR-HMD.
3. The human inspects the motion plan, visualized in the MR-HMD via Unity.
4. If the user approves of the trajectory, then the robot performs it (Figure 3d). If not, then the robot repeats from step 2.



(a) The user specifies the goal position for the trajectory. The red and green models represent the right and left goal poses, respectively. The user is currently gesturing with one hand to translate the right arm (red model) towards the center.

(b) The user specifies the rotation of the red goal pose by using a two-handed gesture to rotate the model.

(c) With the goal poses specified, the user says "plan" to visualize the intended motion. (In this case, in the form of an animation; please see our supplemental video at http://h2r.cs.brown.edu/videos/).

(d) After inspecting and agreeing with the motion plan, the user says "move" to cause the real robot arms to execute the motion and move to the desired position.

Fig. 3. An example interaction using our system. A human specifies goal poses of the robot end effector, and the robot visualizes the resulting trajectory generated by its motion planner. Following human approval, the robot executes the motion.

Step 0. To allow MR-HMD users to specify goal poses and visualize plans in the same workspace as the robot, the coordinate spaces of the virtual world in Unity and the real-world robot must align. For this, we manually calibrate. When the MR-HMD app is launched, a life-size virtual version of the robot is displayed to the user. The MR-HMD hand-tracking capabilities enable the user to "grab" the virtual robot and align its position and rotation such that the virtual robot is in the same place as the real robot. This defines a rigid transformation between the two coordinate spaces.

An automatic calibration system is also possible, e.g., using QR tags to calibrate (or constantly recalibrate) the transformation. However, to overcome MR-HMD world-tracking issues, e.g., with complex robot geometry or specular appearance, we settled with the manual approach.

Step 1. After calibration, Unity and ROS have the same coordinate systems. To specify end-effector poses for the robot, our interface uses virtual robot grippers (one for each arm) to represent the goal position and rotation. Using two hands, users move and rotate the virtual robot grippers by gesturing in free space.

Step 2. With the goal poses set, the users says "plan" (or uses a button), triggering the MR-HMD to send this information to MoveIt through the intermediary MoveIt Node. MoveIt calculates a motion plan from the current robot pose to the user-specified goal pose. This plan is sent back to the HoloLens via rosbridge (Figure 2).

Step 3. In our Unity scene, we have a GameObject that acts as a WebSocket client (Figure 2) and interfaces with rosbridge. As trajectories are streamed from MoveIt, the WebSocket client stores them so that they can be used by the TrajectoryVisualizer GameObject for visualization. The two possible visualizations are a looping animation or a sparse static trail. A full discussion of our visualization techniques can be found in Section 3.5.

Step 4. The user decides whether the proposed trajectory from MoveIt is acceptable or not. The user approves by saying "move," and the robot performs the motion. If the user disapproves, then they can repeat step 2 again, and MoveIt will replan a new trajectory. Because we use a stochastic planner, it would be very unlikely to see the same trajectory twice. The user can also go back to step 1 and adjust the goal poses. This process enables both users and robots to communicate motion intent.

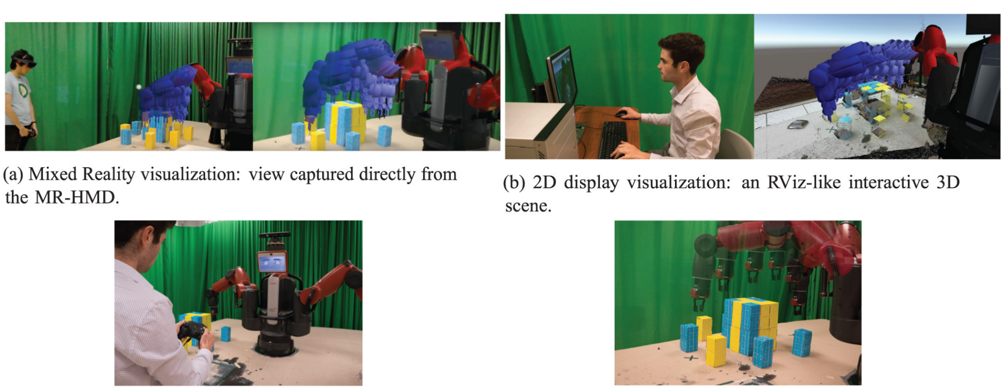
3.5. Visualization design

Any visualization must consider the amount of information conveyed and the ease and efficiency of comprehension. Further, any design must also consider hardware efficiency: the limited computing power of the HoloLens constrains the amount and quality of 3D models visualized, otherwise the rendering and localization loops will slow down and create inaccuracy and virtual/real visual mismatch.

Visualization designs span a large gamut (see RViz for examples (Kam et al., 2015)). We could repeatedly play an animation of the planned motion in real time, which conveys all information but is slow to comprehend. We could show all poses of the motion at once as a continuous trail, which looks cluttered as it is somewhat redundant, and is computationally inefficient. At the other end, we could visualize only the planned end-effector trail, which would be very efficient, but would provide incomplete information on intermediate arm joint locations which may collide with the world.

We drew inspiration from the visualization options provided in the RViz-based GUI to MoveIt. In that interface, users can toggle between an animation of an arm moving through the trajectory and a sparse stroboscopic trail made of multiple arms sampled along the trajectory. For either option, the virtual arm can either be the color of the real robot, or a different, user-specified color.

We implement all of the discussed visualization options in our package. Animation was initially our chosen technique, as it most limits the number of needed draw calls compared to the other options. Unfortunately, this comes at the expense of user comprehension. In our initial testing, we found users needed to watch the animation loop multiple times closely inspect the entire trajectory. For our study, we settled on the sparse trail option from RViz, with two major modifications. First, we had to reduce the polygon count of our virtual arms due to rendering bottlenecks on the HoloLens, and second, we used a light-to-dark color gradient on the trail to emphasize the direction of the motion plan (Figure 4).



(a) Mixed Reality visualization: view captured directly from the MR-HMD.

(b) 2D display visualization: an RViz-like interactive 3D scene.

(c) No visualization.

(d) Robot motion over time.

Fig. 4. Participants must decide whether a robot motion plan collides with the light yellow and blue blocks on the table, across 14 trials and 3 interfaces. Our three interfaces are a MR visualization (a), a 2D display/mouse with an RViz-like visualization (b), and no visualization at all (c). In the first two cases, the experimental setup is shown on the right, and the participant view on the left. In (a), the HoloLens visualizes the robot arm motion plan as a sequence of blue virtual arm graphics overlaid onto the real world. In (b), the 2D display uses the same visualization, but the participant must use the system at a desk. In (c), 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. (d) How a robot motion over time would look in the no-visualization condition.

4. Experiment

With our system, we can now test whether MR-HMDs can aid motion intent communication between humans and robots. We focused on robot-to-human communication, and so goal pose adjustment was not evaluated in this study as it pertains to human-to-robot 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 ($\boldsymbol{M = 26}$, $\mathbf{SD = 6.8}$). We measured 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 Task Load Index (TLX) questionnaires.

4.1. 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 start point to an end point over a table covered in blocks. We did not allow users to specify goal poses and used prerecorded trajectories of the robot's arm rather than use a motion planner in the loop to repeatedly present the same motions to each of the participants. Unknown to the participant, exactly half of the motions collided with the blocks and half did not. Each participant was tasked with labeling the motions as either colliding or non-colliding 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 owing to occlusion from other blocks. Participants could walk around to view the environment from different perspectives. Once a participant had decided how to classify a particular motion, they pressed a button on an Xbox controller to indicate their decision, allowing us to measure the time it took for them to decide.

4.2. Interfaces

We compared three interfaces (Figure 4).

- **No visualization.** This simulated a participant supervising a robot with an emergency stop button. Participants watched the arm, 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; (2) a sparse trail of its future arm poses; and (3) a 3D point cloud of the environment, captured by a Kinect v2 sensor mounted near the robot. 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 (the control scheme was the same as in RViz (Kam et al., 2015)). The visualization remained on the screen for the entire trial. For consistency, participants again recorded their assessment using an Xbox controller.
- **MR.** Through a HoloLens, participants viewed the same visualization of the motion plan 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. Users walked around the room to change their perspective of the robot and visualization. Like the other interfaces, participants decided upon whether the motion collided or not, and

recorded their prediction using an Xbox controller. Like in the monitor interface, the robot arm did not move, and the visualizations remained for the entire trial.

Note that the no visualization interface differs from the monitor and MR 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 evaluate the effectiveness of the MR-based visualization, we consider the no-visualization interface to be a less-direct comparison than the monitor.

4.3. Experimental procedure

We began by reading a consent document to the participant. After consenting, participants completed our motion 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 going 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.

4.4. Measurements

We chose the choice of interface as the within-subjects independent variable. In all three interfaces, our objective dependent variables were the true- and false-positive rates of classifying a path as colliding, and the true- and false-negative rates of classifying a path as non-colliding. By using the mean adjusted accuracy (d') metric, 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). This is discussed further in Section 4.6.1.

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-TLX questionnaire (NASA Human Performance Research Group and others, 1987), system usability as measured by the SUS questionnaire (Brooke et al., 1996), and our own questionnaire measuring perceived predictability and preference for each interface.

- **NASA-TLX.** This is a widely used assessment questionnaire that 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 seven-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 seven-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.

4.5. Hypotheses

We expected that participants would show the best performance in the MR interface condition followed by the monitor interface (i.e., most true positives/negatives, fewest false positives/negatives, lowest levels of mental workload, highest usability, predictability, and system preference scores). In addition, we hypothesize that participants would have a faster labeling speed with the MR interface compared with the monitor interface.

- **H1.** MR will be the easiest interface for completing the motion labeling task, as demonstrated by participants achieving the best performance out of the three conditions, across (a) most true positives/negatives, (b) fewest 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) more true positives/negatives, (b) fewer false positives/negatives, (c) lower levels of workload, (d) higher usability scores, and (e) higher predictability and preference scores.
- **H3.** The MR interface will have faster 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 planned robot motion, whereas in the no-visualization condition, labeling times are generated by watching the robot enact the planned motion. As such, only the monitor and MR conditions are directly comparable.

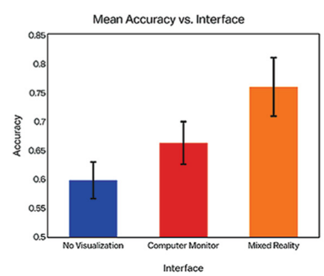
4.6. Results

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

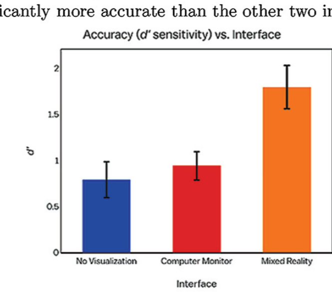
SDT describes accuracy in human perception and decision-making tasks by taking into account preferences for responses (Macmillan, 2002; Tanner and Swets, 1954). For instance, in our task, always responding that a motion plan will collide would yield high true positive scores ("hits"), and also high false-positive scores ("false alarms"). In decision-making tasks with innocuous false alarms, adopting this strategy would not affect overall performance. However, for tasks with high false-alarm cost, a strategy that results in low false-alarm rates while retaining high hit rates is better. For human–robot interaction tasks such as 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 (1999).

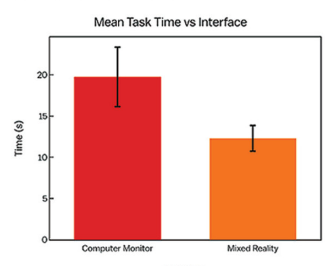
4.6.2. Accuracy. We counted the number of participant true positives, false positives, true negatives, and false negatives in each condition. From this, we report accuracy as the proportion of true positives plus true negatives out of the total number of motion plans (Figure 5a). MR was the most accurate ($M = 0.76$, $\mathrm{SD} = 0.19$), followed by the monitor ($M = 0.66$, $\mathrm{SDSD} = 0.14$), followed by the no-visualization condition ($M = 0.60$, $\mathrm{SD} = 0.12$). These differences were statistically significant (Wilks $\Lambda = 0.619$, $F(2, 30) = 9.244$, $p = 0.001$, $\eta^2 = 0.381$), and accuracy in the MR condition was significantly better than in the monitor condition ($p = 0.001$) and the no-visualization condition ($p < 0.001$). Performance in the monitor condition was not significantly better than in the no-visualization condition ($p = 0.065$).



(a) Mean accuracy across interfaces. The MR interface is significantly more accurate than the other two interfaces.



(b) Mean adjusted accuracy (d') across interfaces. The MR interface has a significantly better d' compared with the other two baseline interfaces.



(c) Mean task times for comparable interfaces (see hypothesis **H3**). The MR interface is significantly faster than the monitor.

Fig. 5. Objective measure user study results. Error bars represent standard error.

We also report d' scores for each participant in each of the three conditions (Figure 5b). There was a significant difference in d' performance scores between the conditions (Wilks $\Lambda = 0.523 F(2, 30) = 13.675$, $p < 0.001$,

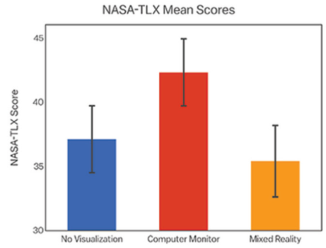
$\eta^2 = 0.477$). Further, the performance in the MR condition ($M = 1.79$, $\mathrm{SD} = 0.88$) was significantly better than the monitor condition ($M = 0.94$, $\mathrm{SD} = 0.58$) and the no-visualization condition ($M = 0.79$, $\mathrm{SD} = 0.72$), all with $p < 0.001$. The difference between performance in the monitor condition was not significantly better than performance in the no-visualization condition ($p = 0.38$). 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 neither of performance indicators in the monitor condition were significantly better than the no-visualization condition. Thus, hypotheses H1(a) and H1(b) were supported, but hypotheses H2(a) and H2(b) were not supported.

Finally, as a manipulation check, we verified that participants who completed the no-visualization condition followed by the monitor condition and then the MR condition (*Order 1*:, $M = 0.67$, $\mathrm{SD} = 0.16$) 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*:, $M = 0.68$, $\mathrm{SD} = 0.17$), $t(94) = 0.220$, $p = 0.826$. The same was true for the d' scores (*Order 1*:, $M = 1.14$, $\mathrm{SD} = 0.82$; *Order 2*: $M = 1.21$, $\mathrm{SD} = 0.89$), $t(94) = 0.428$, $p = 0.669$.

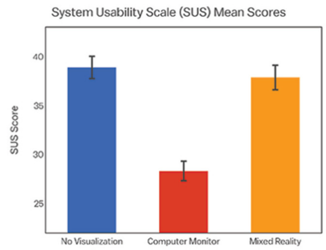
4.6.3. Task time. Hypothesis H3 stated that motion labeling times would be faster in the MR condition than in the monitor condition. A paired samples t-test showed significant differences in mean motion labeling times between the two conditions ($t(31) = 3.415,\ p < 0.001$). Mean labeling times trended in the hypothesized direction (Figure 5c). Labeling times in the MR condition were significantly shorter ($M = 11.95$, $\mathrm{SD} = 8.42$) than in the monitor condition ($M = 19.39$, $\mathrm{SD} = 19.28$). Hypothesis H3 was thus supported.

4.6.4. Subjective workload. Hypotheses H1(c) and H2(c) stated that workload would increase from MR to monitor, and from monitor to no visualization. We used one-way repeated measures ANOVA to test for statistical significance in workload scores across the three interface conditions (Wilks $\Lambda = 0.802$, $F(2, 30) = 3.693$, $p = 0.037$, $\eta^2 = 0.198$).

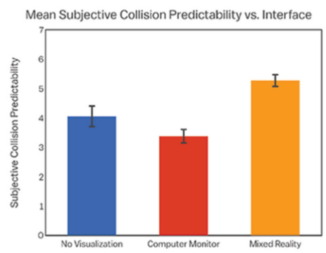
The MR condition was associated with the lowest workload scores ($M = 35.39$, $\mathrm{SD} = 15.73$), followed by the no-visualization condition ($M = 37.11$, $\mathrm{SD} = 14.78$), and then the monitor condition ($M = 42.32$, $\mathrm{SD} = 14.71$; Figure 6a). Post hoc comparisons showed that mean scores in the MR condition were significantly lower than in the monitor condition ($p = 0.040$). 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. Hypothesis H1(c), which stated that MR would have the lowest workload scores, was partially supported. Hypothesis H2(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. Participants reported the lowest levels of subjective workload in the MR condition, significantly lower than in the monitor condition.



(b) Mean SUS scores across all interfaces. The monitor interface had significantly lower usability scores than the other interfaces. All interfaces were significantly different from one another.



(c) Mean subjective collision predictability scores across all interfaces. The MR interface score was significantly higher than the other two interfaces. The monitor and no-visualization conditions were not significantly different.

Fig. 6. Subjective questionnaire user study results. Error bars represent standard error.

4.6.5. Subjective usability. Hypotheses H1(d) and H2(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 $\mathbf{\Lambda = 0.151}$, $\mathbf{F(2, 30) = 84.342, p < 0.001, \eta^2 = 0.849}$). However, the no-visualization condition was associated with the highest SUS scores ($\mathbf{M = 38.91}$, $\mathbf{SD = 6.52}$), followed by the MR condition ($\mathbf{M = 37.88}$, $\mathbf{SD = 7.10}$), and the monitor condition ($\mathbf{M = 28.31}$, $\mathbf{SD = 5.62}$; Figure 6b). Mean SUS scores in the MR condition were significantly higher than the monitor condition ($\mathbf{p < 0.001}$), and mean SUS scores in the no-visualization condition were significantly higher than the monitor condition ($\mathbf{p < 0.001}$). The difference between the MR condition and the no-visualization condition was not significant. Hypotheses H1(d) and H2(d) were not supported.

4.6.6. Subjective collision predictability. Hypotheses H1(e) and H2(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 $\mathbf{\Lambda = 0.246 F(2, 30) = 45.891}$, $\mathbf{p < 0.001}$, $\mathbf{\eta^2 = 0.754}$). Participants showed the highest agreement that MR helped them to predict collisions ($\mathbf{M = 5.28}$, $\mathbf{SD = 1.11}$), followed by the no-visualization condition ($\mathbf{M = 4.06}$, $\mathbf{SD = 1.95}$), and then the monitor condition ($\mathbf{M = 3.38}$, $\mathbf{SD = 1.31}$; Figure 6c). The difference between mean scores in the MR condition were significantly higher than in the monitor condition and the no-visualization condition (both $p < 0.05$), supporting hypothesis H1(e). Means scores in the monitor condition were lower than the no-visualization condition but not significantly so ($p = 0.15$). Hypothesis H2(e) was not supported.

4.6.7. 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.

5. 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 15% increase in collision prediction accuracy and a 38% decrease in time taken. MR 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.

As MR-HMDs are a novel technology, it would be unsurprising for there to be a corresponding novelty effect in our subjective enjoyment measures. However, considering the objective benefits of the MR interface, we feel it is unlikely to be the only cause of the reported subjective enjoyment.

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: the motion plan overlay was not always perfectly aligned on top of the robot owing to the manual calibration and owing to inaccuracy in HoloLens tracking (the authors noticed a drift of several centimeters over a long period of use), the setup took a long time, and 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, 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 as 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. 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. The similarity of SUS and NASA-TLX scores between the no visualization and the MR condition was somewhat surprising, as the interfaces are extremely different. It is possible that the increased cognitive load of interpreting the MR visualizations was offset by the increase in ease of task resulting from those visualizations.

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 H2 was supported.

As a consideration for future work, our system only considers one method of human–robot motion intention communication, and alternative methods may prove effective. MR can also be used to communicate other things in addition to motion intent, such as shared goals, needed objects, or other aspects of robot state.

In addition, further user studies could be conducted to evaluate the effectiveness of using the MR-HMD for communicating motion intent in scenarios with varying cost of mistakes. In our study, we instructed users to label the trajectories as quickly and accurately as possible, but did not directly penalize the users for mislabeling the trajectories. Real-world situations that have actual costs associated with making mistakes may more heavily rely on having an interface that has higher usability for conveying information.

Although this study evaluated the effectiveness of using MR-HMDs for communicating robot motion intent, we did not evaluate the use of our system for enabling users to adjust trajectories, measuring the effectiveness of human-to-robot communication. Future work will address different methodologies of allowing end-users to interact directly with the planned trajectories, such as the end-effector goal pose specification described in Section 3.4, or perhaps a broader system that allows for fine-grain and high-level adjustment.

6. Conclusion

If robots and humans are to form fluid cooperative work partnerships, we will need efficient communication and control of robot motion. We describe a system to allow MR visualizations of robotic motion intent, an interface to control robot motion using MR, and a user study investigating the hypothesis that MR would be a natural interface for robot motion intent communication. In this study, we 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 MR is one way to bridge the robot–human motion communication gap.

Acknowledgments

We thank David Laidlaw for fruitful discussion on VR literature.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by DARPA (grant number D15AP00102) and by the AFRL (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

Ahn Jg, Kim GJ (2016) Remote collaboration using a tele-presence mobile projector robot tele-operated by a smartphone. In: IEEE/SICE International Symposium on System Integration (SII). IEEE, pp. 236–241. Crossref.

Andersen RS, Madsen O, Moeslund TB, Amor HB (2016) Projecting robot intentions into human environments. In: Robot and Human Interactive Communication (RO-MAN). IEEE, pp. 294–301. Crossref.

Bischoff M (2017) ROS Sharp. https://github.com/siemens/ros-sharp.

Brooke J, et al. (1996) SUS - a quick and dirty usability scale. *Usability Evaluation in Industry* 189(194): 4–7.

Burke JL, Murphy RR (2004) Situation awareness and task performance in robot-assisted technical search: Bujold goes to Bridgeport. Technical Report CRASAR-TR2004-23, Center for Robot-Assisted Search and Rescue, Tampa, FL.

Burke JL, Murphy RR, Coovert MD, Riddle DL (2004) 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. Crossref. ISI.

Cha E, Kim Y, Fong T, Mataric MJ, et al. (2018) A survey of nonverbal signaling methods for non-humanoid robots. *Foundations and Trends in Robotics* 6(4): 211–323. Crossref.

Chadalavada RT, Andreasson H, Krug R, Lilienthal AJ (2015) 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). IEEE, pp. 1–6. Crossref.

Chadalavada RT, Lilienthal A, Andreasson H, Krug R (2016) Empirical evaluation of human trust in an expressive mobile robot. In: *RSS Workshop on Social Trust in Autonomous Robots*.

Chen H, Lee AS, Swift M, Tang JC (2015) 3D collaboration method over HoloLens and Skype end points. In: *Proc. 3rd International Workshop on Immersive Media Experiences*. New York: ACM Press, pp. 27–30. Crossref.

Chitta S, Sucan I, Cousins S (2012) MoveIt! *IEEE Robotics and Automation Magazine* 19(1): 18–19. Crossref. ISI.

Demiralp C, Jackson CD, Karelitz DB, Zhang S, Laidlaw DH (2006) Cave and fishtank virtual-reality displays: A qualitative and quantitative comparison. *IEEE Transactions on Visualization and Computer Graphics* 12(3): 323–330. Crossref. PubMed.

Dragan AD, Lee KC, Srinivasa SS (2013) Legibility and predictability of robot motion. In: 2013 8th ACM/IEEE International Conference on Human–Robot Interaction (HRI). IEEE, pp. 301–308. Crossref.

Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. *Robotics and Autonomous Systems* 42(3): 143–166. Crossref. ISI.

Han Y (2016) *The Social Behavior Guide for Confused Autonomous Machines*. Master's Thesis, Rhode Island School of Design. https://issuu.com/horatiohan/docs/the_soical_behavior_guide_for_confu

Kam HR, Lee SH, Park T, Kim CH (2015) RViz: A toolkit for real domain data visualization. *Telecommunication Systems* 60(2): 337–345. Crossref.

Kasik DJ, Troy JJ, Amorosi SR, Murray MO, Swamy SN (2002) Evaluating graphics displays for complex 3D models. *IEEE Computer Graphics and Applications* 22(3): 56–64. Crossref.

Kato H, Billinghurst M (1999) Marker tracking and HMD calibration for a video-based augmented reality conferencing system. In: IEEE and ACM International Workshop on Augmented Reality (IWAR). IEEE, pp. 85–94. Crossref.

Leeper AE, Hsiao K, Ciocarlie M, Takayama L, Gossow D (2012) Strategies for human-in-the-loop robotic grasping. In: *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. New York: ACM Press, pp. 1–8. Crossref.

Macmillan NA (2002) Signal detection theory. *Stevens' Handbook of Experimental Psychology*. New York: John Wiley & Sons, Inc. Crossref.

Microsoft (2017) MixedRealityToolKit. https://github.com/Microsoft/MixedRealityToolkit-Unity

Milgram P, Zhai S, Drascic D, Grodski J (1993) Applications of augmented reality for human–robot communication. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vol. 3. IEEE, pp. 1467–1472. Crossref.

Mutlu B, Yamaoka F, Kanda T, Ishiguro H, Hagita N (2009) Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. In: *ACM/IEEE International Conference on Human Robot interaction*. New York: ACM Press, pp. 69–76. Crossref.

Nakata T, Sato T, Mori T, Mizoguchi H (1998) Expression of emotion and intention by robot body movement. In: *International Conference on Autonomous Systems*.

NASA Human Performance Research Group and others (1987) Task Load Index (NASA-TLX) v1.0 Computerised Version. NASA Ames Research Centre.

Nikolaidis S, Kwon M, Forlizzi J, Srinivasa S (2017) Planning with verbal communication for human–robot collaboration. *arXiv preprint arXiv:1706.04694*.

Ohshima T, Satoh K, Yamamoto H, Tamura H (1998) AR^2 Hockey: A case study of collaborative augmented reality. In: Proceedings IEEE 1998 Virtual Reality Annual International Symposium, 1998. IEEE, pp. 268–275.

Pausch R, Shackelford MA, Proffitt D (1993) A user study comparing head-mounted and stationary displays. In: Proceedings IEEE 1993 Symposium on Research Frontiers in Virtual Reality, 1993. IEEE, pp. 41–45. Crossref.

Quigley M, Conley K, Gerkey B, et al. (2009) ROS: An open-source robot operating system. In: *ICRA Workshop on Open Source Software*, Vol. 3, Kobe, Japan, p. 5.

Rekimoto J (1996) Transvision: A hand-held augmented reality system for collaborative design. *Virtual Systems and Multimedia* 96: 18–20.

Ruddle RA, Payne SJ, Jones DM (1999) Navigating large-scale virtual environments: What differences occur between helmet-mounted and desk-top displays? *Presence: Teleoperators and Virtual Environments* 8(2): 157–168. Crossref. ISI.

Ruffaldi E, Brizzi F, Tecchia F, Bacinelli S (2016) Third point of view augmented reality for robot intentions visualization. In: *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*. Berlin: Springer, pp. 471–478. Crossref.

Santos BS, Dias P, Pimentel A, et al. (2009) Head-mounted display versus desktop for 3D navigation in virtual reality: A user study. *Multimedia Tools and Applications* 41(1): 161. Crossref.

Scassellati B, Hayes B (2014) Human–robot collaboration. *AI Matters* 1(2): 22–23. Crossref.

Schaefer KE, Straub ER, Chen JY, Putney J, Evans A (2017) Communicating intent to develop shared situation awareness and engender trust in human–agent teams. *Cognitive Systems Research* 46: 26–39. Crossref.

Shrestha MC, Kobayashi A, Onishi T, et al. (2016a) Intent communication in navigation through the use of light and screen indicators. In: ACM/IEEE International Conference on Human Robot Interaction. IEEE Press, pp. 523–524. Crossref.

Shrestha MC, Kobayashi A, Onishi T, et al. (2016b) Exploring the use of light and display indicators for communicating directional intent. In: Advanced Intelligent Mechatronics. IEEE, pp. 1651–1656. Crossref.

Slater M, Linakis V, Usoh M, Kooper R (1996) Immersion, presence, and performance in virtual environments: An experiment with tri-dimensional chess. *ACM Virtual Reality Software and Technology* 163: 72.

Slater M, Sanchez-Vives MV (2016) Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI* 3: 74. Crossref.

Stanislaw H, Todorov N (1999) Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, and Computers* 31(1): 137–149. Crossref. PubMed.

Szafir D, Mutlu B, Fong T (2014) Communication of intent in assistive free flyers. In: *ACM/IEEE International Conference on Human–Robot Interaction*. New York: ACM Press, pp. 358–365. Crossref.

Szafir D, Mutlu B, Fong T (2015) Communicating directionality in flying robots. In: *ACM/IEEE International Conference on Human–Robot Interaction*. New York: ACM Press, pp. 19–26. Crossref.

Takayama L, Dooley D, Ju W (2011) Expressing thought: Improving robot readability with animation principles. In: *International Conference on Human–Robot Interaction*. New York: ACM Press, pp. 69–76. Crossref.

Tanner WPJr, Swets JA (1954) A decision-making theory of visual detection. *Psychological Review* 61(6): 401. Crossref. PubMed. ISI.

Toris R, Kammerl J, Lu DV, et al. (2015) Robot web tools: Efficient messaging for cloud robotics. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, pp. 4530–4537. Crossref.

Unity Technologies (2018) Unity Software. https://unity.com

Walker M, Hedayati H, Lee J, Szafir D (2018) Communicating robot motion intent with augmented reality. In: *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. New York: ACM Press, pp. 316–324. Crossref.

Ware C, Franck G (1994) Viewing a graph in a virtual reality display is three times as good as a 2D diagram. In: Proceedings IEEE Symposium on Visual Languages, 1994. IEEE, pp. 182–183. Crossref.

Whitney D, Rosen E, Phillips E, Konidaris G, Tellex S (2017) Comparing robot grasping teleoperation across desktop and virtual reality with ROS Reality. In: *International Symposium on Robotics Research*.