Human-Robot Interaction in a Shared Augmented Reality Workspace

Shuwen Qiu*  Hangxin Liu*  Zeyu Zhang  Yixin Zhu  Song-Chun Zhu

Abstract—We design and develop a new shared Augmented Reality (AR) workspace for Human-Robot Interaction (HRI), which establishes a bi-directional communication between human agents and robots. In a prototype system, the shared AR workspace enables a shared perception, so that a physical robot not only perceives the virtual elements in its own view but also infers the utility of the human agent—the cost needed to perceive and interact in AR—by sensing the human agent’s gaze and pose. Such a new HRI design also affords a shared manipulation, wherein the physical robot can control and alter virtual objects in AR as an active agent; crucially, a robot can proactively interact with human agents, instead of purely passively executing received commands. In experiments, we design a resource collection game that qualitatively demonstrates how a robot perceives, processes, and manipulates in AR and quantitatively evaluates the efficacy of HRI using the shared AR workspace. We further discuss how the system can potentially benefit future HRI studies that are otherwise challenging.

I. INTRODUCTION

Recent advance in Virtual Reality (VR) and Augmented Reality (AR) has blurred the boundaries between the virtual and the physical world, introducing a new dimension for Human-Robot Interaction (HRI). With new dedicated hardware [1], [2], [3], VR affords easy modifications of the environment and its physical laws for HRI; it has already facilitated various applications that are otherwise difficult to conduct in the physical world, such as psychology studies [4], [5], [6] and AI agent training [7], [8], [9], [10].

In comparison, AR is not designed to alter the physical laws. By overlaying symbolic/semantic information and visual aids as holograms, its existing applications primarily focus on assistance in HRI, e.g., interfacing [11], [12], [13], data visualization [14], [15], [16], robot control [17], [18], and programming [19], [20]. Such a confined range of applications hinders its functions in broader fields. We argue such a deficiency is due to the current setting adopted in prior AR work; we call it a active human, passive robot paradigm, as illustrated by the red arrows in Fig. 1. In such a paradigm, the virtual holograms displayed in AR introduce asymmetric perceptions to humans and robots; from two different views, the robot and human agents may possess a different amount of information. This form of information asymmetry prevents the robot from properly assisting humans during collaborations. This paradigm also heavily relies on a one-way communication channel, which intrinsically comes with a significant limit: only human agents can initiate the communication channel, whereas a robot can only passively execute the commands sent by humans, incapable of proactively manipulating and interacting with the augmented and physical environment.

To overcome these issues, we introduce a new active human, active robot paradigm and propose a shared AR workspace, which affords shared perception and manipulation for both human agents and robots; see Fig. 1:

1) Shared perception among human agents and robots. In contrast to existing work in AR that only enhances human agents’ understanding of robotic systems, the shared AR workspace dispatches perceptual information of the augmented environment to both human agents and robots equivalently. By sharing the same augmented knowledge, a robot can properly assist its human partner during HRI; the robot can accomplish a Level 1 Visual Perspective Taking (VPT1) by inferring if a human agent perceives certain holograms and estimating associated costs.

2) Shared manipulation on AR holograms. In addition to manipulating physical objects, shared AR workspace endows a robot with the capability to manipulate holograms proactively, in the same way as a human agent does, which would instantly trigger the update of shared perception. As a result, HRI in the shared AR workspace permits a more seamless and harmonious collaboration.

* Shuwen Qiu and Hangxin Liu contributed equally to this work.
UCLA Center for Vision, Cognition, Learning, and Autonomy (VCLA) at Statistics Department. Emails: {s.qiu, hk.liu, zeyuzhang, yixin.zhu}@ucla.edu, sczhu@stat.ucla.edu.

The work reported herein was supported by ONR N00014-19-1-2153, ONR MURI N00014-16-1-2007, and DARPA XAI N66001-17-2-4029.

Fig. 1: A comparison between the existing AR systems and the proposed shared AR workspace. Existing AR systems limit to an active human, passive robot, one-way communication, wherein a physical robot would only react to human commands via AR devices without taking its own initiatives; see the red arrows. The proposed shared AR workspace constructs an active human, active robot, bi-directional communication channel that allows a robot to perceive and proactively manipulate holograms as human agents do; see green arrows. By offering shared perception and manipulation, the proposed shared AR workspace affords more seamless HRI.
We develop a prototype system using a Microsoft Hololens and TurtleBot2, and demonstrates the efficacy of the shared AR workspace in a case study of a resource collection game.

The remainder of the paper is organized as follows. Section II introduces the system setup and details some critical system components. The two essential functions, shared perception and shared manipulation of the proposed shared AR workspace, are described in Section III. Section IV demonstrates the efficacy of the proposed system by a case study, and Section V concludes the paper with discussions on some related fields the system could potentially promote.

II. SYSTEM SETUP

In this section, we describe the prototype system that demonstrates the concept of the shared AR workspace; Fig. 2 depicts the system architecture. Our prototype system assumes (i) a human agent wearing an AR device and (ii) a robot with perception sensors; however, the system should be able to scale up to multi-human, multi-robot settings.

Robot Platform: We choose TurtleBot2 mobile robot with a ROS compatible laptop as the robot platform; see Fig. 2a. The robot’s perception module includes a Kinect2 RGB-D sensor and a Hokuyo Lidar, which constructs the environment’s 3D structure using RTAB-Map [21]. Once the map is built, the robot only needs to localize itself within the map by fusing visual and wheel odometry.

AR Headset: Human agents in the present study wear a Microsoft Hololens as the AR device; see Fig. 2b. HoloLens headset integrates a 32-bit Intel Atom processor and runs Windows 10 operating system onboard. Using Microsoft’s Holographic Processing Unit, the users can realistically view the augmented contents as holograms. The AR environment is created using the Unity3D game engine.

Communication: Real-time interactions in the shared AR workspace demands timely communication between HoloLens (human agents) and robots, established using ROS# [22]. Between the two parties, HoloLens serves as the client, who publishes the poses of holograms, whereas the robot serves as the server, which receives these messages and integrates them into ROS. In addition to the perceptual information obtained by its sensors, the robot also has access to the 3D models of holograms so that they can be rendered appropriately and augmented to the shared perception.

Overall Framework: The shared perception in the shared AR workspace allows a robot to perceive virtual holograms in three different levels with increasing depth: (i) know the existence of holograms in the environment, (ii) see the holograms from the robot’s current coordinate obtained by localizing itself using physical sensors, and (iii) infer human agent’s utility/cost of seeing holograms. Take an example shown in Fig. 2e: human agents can directly see objects in the yellow region as it is within their Field-of-View (FoV), but they need to change the views to perceive the objects marked in light blue; objects in dark blue are fully occluded. Only having with such a multi-resolution inference could the robot properly initiate interactions or collaboration with the human, forming a bi-directional communication. For instance, in Fig. 2f, the robot estimates a hologram is occluded from the human agent’s current view and plans and carries this occluded hologram to assist a human agent to accomplish a task. Since the robot proactively helps the human agent form collaborations, such a new AR paradigm contrasts the prior one-directional communication.

III. SHARED AR WORKSPACE

Below we describe the shared perception and shared manipulation implemented in the shared AR workspace.

A. Detection and Transformation

A key feature in the shared AR workspace is the ability to know where the holograms are at all time, which requires to localize human agents, robots, and holograms, and construct transformations among them. Using an AR headset, the human agent’s location is directly obtained. Given the corresponding transformations between a human agent and a hologram $h_i^T h_i T$, the AR headset with the human agent’s egocentric view can render the holograms.
Fig. 3: The cost of a human agent seeing an object is defined by visibility/occlusion and the angle between two vectors—her current facing direction and the looking direction of an object. Suppose she is facing to the doorknob (green arrow), the cost to see the clock is higher than seeing the sprayer as the angle \( \beta \) is larger than \( \alpha \). Although the angle \( \gamma \) to see the plant under the desk is smaller than \( \beta \), the plant is currently occluded from the human agent’s view, resulting in a higher cost despite a smaller angle.

By estimating the human pose from a single RGB-D image [23], the robot establishes a transformation to the human agent \( T_h \). Fig. 2c shows one example. Specifically, the frame of a human agent is attached to the head, whose \( x \) axis is aligned with the human face’s orientation estimated by three key points—two eyes and the neck. When the human agent is partially or completely outside of the robot’s FoV, the frame of the human agent is directly estimated by leveraging the visual odometry provided by the Hololens.

By combining the above two transformations, the transformations from the robot to a hologram can be computed by \( T_{r \rightarrow h} = T_{h \rightarrow r} \circ T_h \). The transformations and the coordination of human agents, robots, and virtual holograms are represented in the same coordinate for easy retrieval by the robot.

B. Augmenting Holograms

Only knowing the existence of holograms is insufficient; the robot ought to “see” the holograms in a way that can be naturally processed for its internal modules (e.g., planning, reconstruction). We design a rendering schema to “augment” holograms to the robot and incorporate them into the robot’s ROS data messages, such as 3D point clouds and 2D images.

3D Point Clouds: The holograms rendered for human agents are stored in a mesh format. To render them in 3D for robots, we use a sampling-based method [24] to convert holograms to point clouds. With the established transformations, these holograms are augmented to the robot’s point clouds input with both position and color information; see Fig. 5a for examples of rendered holograms for the robot.

2D Image Projection: We render the holograms by projecting them onto the robot’s received images. Following a general rendering pipeline [25], we retrieve the hologram’s coordinate \( P_r \) with respect to camera frame by the established transformation \( T_r \) and calculate the 2D pixel position \( P_s = M_p \times P_r \), given the camera’s intrinsic matrix \( M_p \).

C. Visual Perspective Taking

Simply “knowing” and “seeing” holograms would not be sufficient for a robot to help the human agent in the shared AR workspace proactively. Instead, to collaborate, plan, and execute properly, the robot would need to possess the ability to infer whether others can see an object. Such an ability to attribute others’ perspective is known as Level 1 Visual Perspective Taking (VPT1) [26], [27]. Specifically, we hope to endow the robot in the shared AR workspace with capabilities of inferring (i) whether the human agent can see certain objects, and (ii) how difficult it is.

VPT1 of a robot is devised and implemented at both the object level and scene level. At the object level, we define the human agent’s cost to see an object as a function proportional to the angle between the human agent’s current facing direction and looking direction of the object; see an illustration in Fig. 3. The facing direction is jointly determined by the pose detection from the robot’s view and the IMU embedded in Hololens. The system also accounts for the visibility of objects as they may be occluded by other real/virtual objects in the environment. To identify an occluded object, multiple virtual rays are emitted from AR headset’s FoV to the points in a standard plane whose pose would be updated along with the human agent’s pose. The object would be identified as occluded if any of those rays intersect with (i) other holograms whose poses are known in the system, or (ii) real objects or structures whose surfaces are detected by Hololens’s spatial mapping.

At the scene level, we categorize the augmented environment into three regions: (i) Focusing region, highlighted in yellow in Fig. 2c, is considered within the human’s FoV excluding occluded regions, determined by the FoV of Hololens—a 30° by 17.5° area centered at the human’s eye. (ii) Transition region, highlighted in light blue, does not directly appear in the human’s FoV, but it can be perceived with minimal efforts (e.g., by turning head). (iii) Blocked region, highlighted in the dark blue, is occluded and cannot be seen by merely rotating view angles; the human agent has to traverse the space with large body movements, e.g., spaces under tables are typical Blocked regions.

D. Interacting with Holograms

By “seeing,” “knowing,” and even “inferring” human agents about holograms in the shared AR workspace, the robot could subsequently plan and manipulate these holograms as an active user in the very same way as a human agent does. However, the holograms are not yet tangible for the robot to “interact.” In our prototype system, we devise a simple rule-based algorithm to determine the conditions to be triggered for a robot to interact with holograms.

Fig. 4 illustrates the core idea. After obtaining a hologram’s 3D mesh, the algorithm fits a circumscribed sphere to the mesh and to itself with 20% enlargement. Once the robot’s sphere is sufficiently close to the hologram’s (i.e., there is an intersection between two spheres), it triggers a manipulation mode, and the hologram is attached to the robot and move together. The movements are also synced in
the shared perception to the human agent in real-time; see Fig. 2f. Since the present study adopts a ground mobile robot, we project the spheres to circles on the floor plane to simplify the intersection check. More sophisticated interactions, such as a mobile manipulator grasping a hologram in 3D space, is achievable using standard collision checking methods.

E. Planning

The last component of the system is the planner. In fact, the shared AR workspace poses no constraints on task and motion planning algorithms; the decision should be made mainly based on robot platforms (e.g., ground mobile robot, mobile manipulator, humanoid) and executed tasks (e.g., HRI, navigation, prediction) during the interactions; see the next section for the planning schema adopted in this paper.

IV. EXPERIMENT

A. Experimental Setup

We design a resource collection game in the shared AR workspace to demonstrate the efficacy of the system. Fig. 5a depicts the environment. Six holograms, rendered as point clouds and highlighted in circles with zoomed-in views, are placed around the human agent (marked by a red skeleton at the center of the room), whose facing direction is indicated in yellow. Some holograms can be easily seen, whereas others are harder due to their tricky locations in 3D or occlusion (e.g., object 6). A human agent’s task is to collect all holograms and move them to the table as fast as possible. The robot stationed in the green dot would help the human in collecting the resources.

As described in Section III-C, the robot first estimates the cost for a human agent to see the holograms and whether they are occluded; the result is shown in Fig. 5b. In our prototype system, the robot prioritizes to help the occluded holograms and then switch to the one with the highest cost. In future, it is possible to integrate prediction models (e.g., [28], [29], [30]) that anticipate human behaviors.

B. Qualitative Results

Intuitively, we should see a better overall performance during HRI via shared AR workspace due to its shared perception and manipulation that enables a robot to help the human agent for task completion collaboratively.

Fig. 6 gives an example of a complete process, demonstrating a natural interaction between the human agent and the robot to accomplish a given task collaboratively. The top row shows the human agent’s egocentric views through the Hololens that overlays the holograms to the image captured by its PV camera. The middle row is a sequence of the interactions between the robot and holograms from a third-person view. The bottom row reveals the robot’s knowledge of the workspace and its plans. In this particular trial, the human agent first collected the roman tomato and the bottle as they appear to have a lower cost. In parallel, the robot collaboratively carries holograms—the occluded cabbage and the tomato with the highest cost—to the human agent.

C. Quantitative Results

We conduct a pilot study to evaluate shared AR workspace quantitatively. Twenty participants were recruited to assess the robot performance in a between-subject setting ($N = 10$ for each group). The participants in the Human group are asked to find and collect all six holograms by themselves. The participants in the Human+Robot group use the shared AR workspace system, where the robot proactively helps the participants to accomplish the task. Each subject has no familiarization with the physical environments, but they received simple training about how to use the AR device right before the experiments started.

Fig. 7 compares the results between the two groups. The difference of the completion time is statistically significant; $t(19) = 1.0, p = 0.028$. Participants with robot’s help take significantly less time (mean: 135 seconds, median: 134 seconds) to complete the given task. In contrast, the baseline group requires much more time with a larger variance (mean: 202 seconds, median: 206 seconds). This finding indicates a new role that a robot can play in the shared AR workspace by assisting human agents to accomplish a task collaboratively.
V. RELATED WORK AND DISCUSSION

We design, implement, and demonstrate how the shared perception and manipulation provided by the shared AR workspace improve HRI with a proof-of-concept system using a resource collection game. In future, more complex and diverse HRI studies are needed to further examine and benfits and limits of the shared AR workspace by (i) varying the degree of human agent’s and/or robot’s perception and manipulation capability; e.g., only the robot can see and act on holograms while the human agent cannot, as an opposite to current AR setup, and (ii) introducing virtual components to avoid certain costly and dangerous setups in the physical world. Below, we briefly review related work and scenarios that shared AR workspace could potentially facilitate.

The idea of creating a shared workspace for human agents and robots has been implemented in VR, where they can re-target views to each other to interact with virtual objects [31]. Prior studies have demonstrated advantages in teleoperation [32] and robot policy learning [33]. More recently, a system [34], [35] that allows multiple users to interact with the same AR elements is devised. In comparison, the shared AR workspace deals with the perceptual noise in the physical world and promotes robots to become active users in AR to work on tasks with humans collaboratively.

In recent years, Human-Robot Interaction and Collaboration have been developing with increasing breadth and depth. One core challenge of the field is to seek how the robot or the human should act to promote understanding and trust, usually in terms of predictability, with the other. From a robot’s angle, it models humans by inferring goals [36], [37], tracking mental states [38], [39], predicting actions [40], and recognizing intention and attention [41], [42]. From a human agent’s perspective, the robot needs to be more expressed [43], to promote human trust [44], to assist properly [45], [46], and to generate proper explanations of its behavior [44]. We believe the proposed shared AR workspace is an ideal platform for evaluating and benchmarking existing and new algorithms and models.

Human-robot teaming [47], [48] poses new challenges to computational models aiming to endow robots with the Theory of Mind abilities, which are usually in a dyadic scenario [49]. With the adaptability to multi-party settings and the fine-grained controllability of users’ situational awareness, the proposed shared AR workspace offers a unique solution to test the robot’s ability to maintain belief, intention, and desires [50], [51], [39] of other agents. Crucially, the robot would play the role of a collaborator to help and as a moderator [52] to accommodate each agent. The ultimate goal is to forge a shared agency [53], [54], [55], [56] between robots and human agents for seamless collaboration.

How human’s cognition emerges and develops is a fundamental question. Researchers have looked into the behaviors of primates’ collaboration and communication [57], imitation [58], and crows’ high-level reasoning [59], planning and tool making [60] for deeper insights. Cognitive robots are still in their infancy in developing such advanced cognitive capabilities, despite various research efforts [61], [62]. These experimental settings can be relatively easier to replicate in the shared AR workspace, which would open up new avenues to study how a robot would emerge similar behaviors.


