2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Interacting with Multi-Robot Systems via Mixed Reality

Florian Kennel-Maushart[1], Roi Poranne[1,2], Stelian Coros[1]

***Abstract*— Mobile robots are becoming safer and more affordable, and their presence in the workspace is increasing. However, many tasks that involve reasoning, long-term planning or human preferences are still hard to automate. While some solutions in specialised areas slowly emerge, an alternative to full autonomy can be to actively leverage intuition and experience of human operators. To do this, suitable interfaces and modes of interaction have to be explored. Inspired by Real-Time Strategy games, we implement a Mixed Reality interface that can be used with either a Microsoft HoloLens 2 headset or a tablet. The interface allows users to interact with multiple mobile robots simultaneously. We conduct a user study to compare the headset and tablet versions of the interface in different scenarios inspired by a real-world construction setting. We show that, while performance and preference of interface are dependent on the task and the complexity of the required interaction, users are able to solve non-trivial tasks on both platforms using our system.**

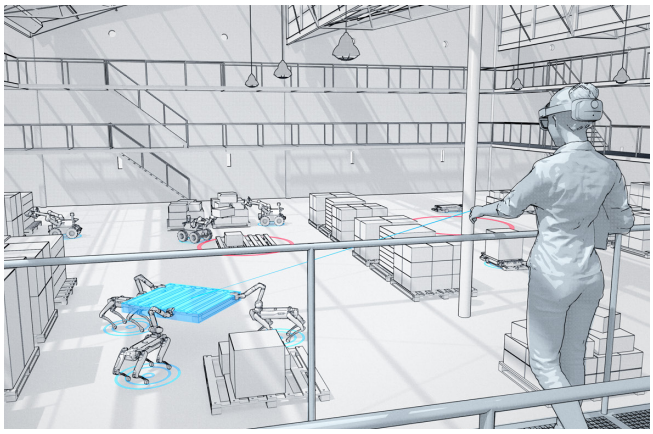


Fig. 1: Artistic depiction of a user guiding robots in MR.

I. Introduction

Robots have come a long way since the beginning of the industrial revolution. Once massive machines, isolated in cages for the safety of their operator, they have become more dexterous, agile, mobile and collaborative. Artificial Intelligence (AI), especially in the context of machine perception and computer vision, led to an unprecedented degree of autonomy in various industries [1]–[3]. However, research challenges such as the DARPA Robotics Challenge [4] show that, while AI and robotics have seen significant improvements in the past decades, a lot of tasks are still very hard for robots to perform on their own, especially when high-level reasoning is required [5]. Also, their ability to react to unforeseen events or ambiguous input is limited.

Based on this and in anticipation of the next industrial revolution [6], an important goal of robotics and AI research is not necessarily how to fully automate every possible task, but rather how to intelligently combine the strengths of robots and human reasoning. Ideally, one operator could then supervise multiple robots and keep them running, occasionally intervening in order to teach robots new skills or correct sub-optimal behaviour, as depicted in the artistic representation in Fig. 1. The kind of resource management described above is the hallmark of a genre of video games known as Real Time Strategy (RTS). In RTS games, the player commonly plays the role of a commander over a group, comprised of many different units. Watching over the map from an aerial view, the player can give units orders to rendezvous, gather resources, assemble a base, and execute different missions. This is done via a set of common interaction modes, usually using a mouse and a keyboard, or, more recently, a touch-screen when playing on a smartphone or tablet computer [7]. While the units usually have some AI, leaving them unattended for longer periods of time is generally not ideal. Instead, the player must plan a strategy, execute it, and adapt it to counteract the opponent's strategy, by efficiently micro-managing the units.

Inspired by RTS games, our goal was to develop a real-life counterpart, to assist operators in managing their robot teams efficiently and effectively. Specifically, we implemented a system and a set of interactions, which mimic and extend those commonly found in RTS games. To be useful and intuitive when working with real robots, the system and the interactions need to address the following considerations:

- *Shared World*: Humans and robots need to have a shared understanding of the workspace and task.
- *Interaction and Interpretation*: The human operator needs to have the ability to intuitively and easily convey intent to the robots.
- *Autonomy vs. Control*: The operator should have access to different interactions with varying degrees of required robot autonomy, depending on the task.

A promising approach in human-robot interaction (HRI), which directly addresses some of these challenges, involves the use of Augmented or Mixed Reality (AR, MR) interfaces. Such AR and MR systems track the environment and can overlay virtual content on top of physical objects. The same tracking system can be used to create a shared frame of

[1] The authors are with the Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland. florian.maushart@inf.ethz.ch; roi.poranne@inf.ethz.ch; scoros@inf.ethz.ch
[2]Roi Poranne is with the Department of Computer Science, University of Haifa, Haifa, Israel. roiporanne@cs.haifa.ac.il
This work was partially supported by the Hilti group and Microsoft.

reference between humans and robots. Mobile AR solutions have been made available to the general public via smartphones and tablets in recent years. MR-headsets, such as the Microsoft HoloLens, are a more recently developed type of interface, which allows for a bigger variety of interactions. MR-headsets use hand-tracking or 6 degree of freedom (6DoF) controllers. This enables a hands-free experience, while interaction with AR content on a smartphone or tablet is usually still limited to touch input.

To test how usable our system is and how intuitive different interactions on the two interfaces are, we conduct a user study, in which a single operator is in charge of controlling multiple miniature differential-drive robots. The envisioned setting is that of an (abstract) construction site and the view from the user's perspective can be seen in Fig. 3b. The operator has to assign robots to different job-sites, guide multiple robots in transportation scenarios or help them with more complex, collaborative tasks. We implement an AR and MR application on a tablet as well as a Microsoft HoloLens 2 and evaluate the performance on both interfaces on three different tasks, testing different aspects of multi-robot HRI. For the evaluation we use objective metrics, as well as the subjective evaluation of participants via a NASA-TLX survey [8] and a follow-up interview. We show that, using our system, users are able to accomplish both simple and complex multi-robot tasks. Our results suggest that, while both the tablet as well as the HoloLens are powerful interfaces for interacting with a multi-robot system, performance and user-preference for the interface as well as interaction modes are strongly task-dependent.

II. Related Work

A. Multi-robot Mixed Reality Interfaces

Mixed Reality for HRI is a topic of great interest. Many recent papers use MR as means of input and visualization for motion planning (see recent reviews [9] and [10]). Here, we highlight some relevant publications in more detail:

An interesting example can be found in [11], where Frank et al. present a study of how tablets with AR capabilities and marker-based tracking can be used to control a group of robots and assign tasks to them. The tasks consist of pushing rods from an initial location to a target location, indicated by the operator. Similarly, in [12], Patel et al. present an AR interface which allows users to group robots around a payload and then to directly interact with a virtual payload instead of individual robots, which in turn controls the collaborative motion of the robots. A similar VR control method for more complex manipulators with a fixed base can be found for example in [13] or [14]. The idea of gesture-based interaction with robot swarms has also been explored in by Alonso-Mora et al. in [15]. In this extension of their earlier paper on multi-robot pattern formation [16], the authors use a Microsoft Kinect RGB-D sensor to capture depth information of an operator. With gesture information extracted from the sensor data, the operator can control cohesion and orientation of the swarm or trace paths for individual robots. They also present a tracking mode in which the whole-body pose of the operator is translated into a formation for the robot swarm. As the intended applications for the algorithms presented are situated in the context of entertainment and robotic games in the form of a distributed display, interactions with the real world were not considered. A gesture-based approach to controlling drones has very recently been presented by Serpiva et al. in [17]. The authors are using gestures and hand-drawn trajectories to start, guide and land a group of unmanned aerial vehicles and change their formation. Another example of hand-drawn interactions is presented in [18], where the authors introduce constraints on robots through virtual sketches on a tablet.

In order to identify different gestures that users would find intuitive when interacting with a swarm of robots, Kim et al. have conducted an exhaustive elicitation study in [19]. They found that both the number of robots as well as the proximity of the user to the robots influence the preferred gesture. While this study is an excellent source for potential interactions, the authors used a *Wizard of Oz* approach, in which robots were controlled by a hidden operator to create the illusion of reacting to user input rather than taking the actual input into account. Another study of gesture-based interaction is presented in [20]. The authors explored an acceleration-based "force push" interaction in VR. They scale the interaction forces with the virtual objects based on the expressiveness, i.e. the scale and speed at which a user performs the force push movement. While the precision of the interaction suffered compared to first-order interactions, i.e. position-based control, users reported the force push to be more "natural", comfortable and fun to use.

B. Swarm Robotic Systems

Many swarm robotic systems have been developed over the past two decades. A recent one, with a focus on HRI is presented in [21]. The swarm robotic platform called "zooids" is introduced and several applications and interactions for its use as an interactive display are presented. The movement is tracked by an overhead DLP projector, and robot modes can be switched between passive and active. As the interaction is based on touch sensors on the robot, the user is limited to direct interaction with a few zooids at time. A similar robotic platform by Sony, called "Toio" (see Fig. 2) has been used in a recent work by Nakagaki et al. in [22]. They are using the small differential-drive robots to interact with different mechanical "shells", which help them perform a diverse set of tasks, both for entertainment and practical purposes. Finally, there exist many more well-established swarm-robotic platforms like the *Kilobots* [23], *e-puck* [24] and *crazyflie* [25], which all have their advantages and disadvantages (also see Sec. III-A).

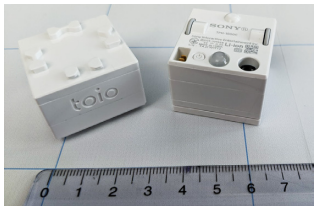


Fig. 2: Toio robots on a sheet with printed dot-pattern

C. Video Games

Much inspiration for human-robot-interaction can be obtained from video games. Game designers have spent decades exploring intuitive interfaces for their titles, both for PC and Consoles, and more recently for mobile devices. Real Time Strategy (RTS) games such as Blizzard Entertainment's "Starcraft 2" are played very competitively, which makes quick and reliable interaction patterns highly important. The goal of these games on a "macro" level is to assign one's units intelligently to different tasks, such as resource collection and base building, in order to gain a competitive advantage over an opponent. Many of these tasks will be executed autonomously by the units once they have been assigned, i.e. they will commute back and forth between a source of resources and a warehouse in the player's base. Ultimately, games are won or lost on a "micro" level, where players need to micromanage their units in direct interactions with the opponent or during exploration tasks. Arguably, this translates well to scenarios of managing a fleet of robots on a construction site. Robots will, for example, be able to autonomously unload a container, but installing the unloaded resources might require human input, e.g. to react to unforeseen circumstances on-site or adjustments of plans (for an artistic representation of this idea see Fig. 1). Another interesting example of an RTS game is Lionhead's "Black and White" (Fig. 3a). This game explores the theme of the player acting as a god that can interact with different units and the environment via a virtual hand, controlled by mouse input. Black and White also supported a "P5 Glove" peripheral [26]. Using bend sensors and external tracking, this glove allowed users to interact with the game using hand gestures instead of mouse input. An important aspect of the user interface are environmental clues, such as flags and highlight effects (see Figure 3a for an example of interactable areas, highlighted with red lighting). These clues elevate the players' perception and can be used to notify them of tasks or areas that need their attention. Mixed Reality devices have very similar capabilities (see Fig. 3), and, together with the localisation and tracking aspects, the ability to overlay information over real objects can be very useful for successful human-robot interfaces.

(a) Black and White 2 (PC) [27] (b) Our interface (HoloLens 2)

Fig. 3: Hand interactions with virtual and robotic agents.

III. Method

A. Robotic Platform

In choosing a platform for our experimentation we considered the following:

1) Precision: Robots need to move and localize within a given coordinate system with sufficient precision.
2) Robustness: The tracking solution (external or on-board) should be robust to obfuscation by the operator. Additionally, robots should have sufficient strength to interact with the environment and withstand minor disturbances.
3) Affordability: The system should be relatively low-cost and easy to set up.
4) Availability: While there are many interesting swarm robotic platforms, many of them have limited availability, or their bill of materials might be outdated.

A platform that meets our criteria are the Toio robots by Sony (see Figure 2). These 3x3x2.5cm differential drive robots feature a small camera at their bottom side, which can read dot patterns. The patterns are printed on a sheet of paper that the robots drive on and which can be combined to allow for potentially very large work spaces. The major advantage of this technology is that any user interaction, including the operator's position or how close their hand is to the robot will not interfere with the robot's position and orientation estimate. While the Toio only has a maximum payload capacity of 200 grams, it reaches linear speeds of up to 350mm/s and rotational speeds of roughly four full rotations per second (1500°/s) [28]. It can connect to a PC via the Bluetooth Low Energy (BLE) protocol, and an SDK for the Unity game engine is available online [29].

B. Robot Controller

A simple method of calculating the desired wheel speeds according to the robot's distance to a goal is via a proportional gain controller. We set the desired forward velocity v_{des} to a fixed value, and get the desired angular velocity as $\omega_{des} = K_p\alpha$, where α represents the angle between the robot's current heading and the goal position. The desired wheel speeds v_R and v_L can then be obtained as $v_R = v_{des} + d_{wheels} \cdot \omega_{des}$ and $v_L = v_{des} - d_{wheels} \cdot \omega_{des}$. To avoid that the robot has to turn by 180 degrees when the goal lies right behind it, we additionally introduce a *reverse condition*: If $\alpha > 90°$, we calculate $\omega_{des} = \text{sign}(\alpha) \cdot (\pi - \text{abs}(\alpha)) \cdot K_p$, and reverse v_R and v_L. This corresponds to the forward direction of the robot being inverted.

While the proportional gain control is easy to implement and works well in the absence of obstacles, robots tend to get stuck on each other, especially because of their square shape. An established algorithm, which is able to avoid collisions for a large number of agents is the Reciprocal Velocity Obstacle (RVO) algorithm [30]. We implemented both proportional control as well as RVO, but chose not to activate RVO during the user study, as the obstacle avoidance could sometimes get in the way of completing the tasks.

C. Input Devices, Software Stack and System

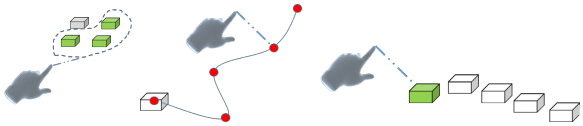
For our experiments we use a Microsoft HoloLens 2 headset as well as a Samsung Galaxy Tab A8, which supports the Android ARCore framework. The HoloLens 2 features a see-through display, RGBD cameras and an IMU [31], which enable features like spatial mapping, articulated hand tracking and projection of virtual content in the form of holograms onto the real world. This is facilitated by the Mixed Reality Toolkit (MRTK), an open-source toolkit which lets users integrate many HoloLens features with Unity. For our system, we used Unity 2021.3.8f1, with MRTK 2.8.2 and the Mixed Reality OpenXR Plugin 1.4.4. Additionally, MRTK integrates with Unity's *ARFoundation* framework (v. 4.2.3), which lets us run the same app with only minor modifications on Android devices like the Samsung Galaxy Tab A8. The Galaxy Tab A8 supports most of ARCore's functionality, except depth estimation, which we do not require for our experiments. Its 10.5" display supports multi-touch. The device weighs about 505g, which is very similar to the HoloLens, which weighs 566g. The BLE connectivity and robot control are handled by a Windows 10 laptop with an Intel Core i7-9750H CPU @ 2.60GHz, 32GB RAM and an Nvidia Geforce RTX 2080 Max-Q GPU. Input data was transmitted from the headset or tablet to the machine via Wi-Fi, using a UDP connection. As described above, robots transmit their positional and heading data directly to the laptop via BLE, which in-turn sends the desired wheel velocities to the robots. Co-localization between the HoloLens and the robots was achieved through an initial, manual alignment of the virtual and the physical playmat. Co-localization between the tablet and robots had to be continuous, as drift would lead to major offsets otherwise, and was achieved through the use of AprilTags [32].

D. Interactions

While there are many ways to distinguish interactions, for the purpose of this paper we focus mainly on two categories:

Discrete Control (Selection Mode) Emulating a mouse pointer, on the HoloLens this interaction uses a ray extended from the operator's hand to point, and a pinch gesture to "click". On the tablet, a simple tap is used instead. When a robot is selected by the gesture, it will be added to a controllable group. When the workspace is hit, the whole controllable group will move towards the hitpoint.

Continuous Control (Fingertip Mode) The fingertip positions are continuously projected onto the plane and serve as the target for the closest robot (Fig. 5c). On tablet, multiple targets can be set via multi-touch and stay in place when the fingers are lifted. On the headset, individual fingers can be hidden by retracting them towards the palm and both hands can be used to control up to 5 targets.



(a) Lasso (b) Waypoints (c) Leader-Follower

Fig. 4: Interactions for discrete control

(a) Formation (b) Force Field (c) Fingertip

Fig. 5: Interactions for continuous control

For the *Selection Mode*, the user can additionally select multiple robots at once by dragging the finger or pointer across the floor and drawing a *lasso* around robots (cf. Fig. 4a). Once the user lets go, the lasso is auto-completed and all robots within the drawn area are selected. While we limit the study for this paper to the interactions described above, our system supports a wider variety of interactions and gestures, some of which can be seen in Figs. 4 and 5. Their functionality can be observed in the accompanying video material, and we would like to explore their applications in future studies. Related work uses similar nomenclature, e.g. both [19] and [33] distinguish *discrete* and *continuous* gestures. In games they are often called *macro* or *micro* interactions (cf. Sec. II-C).

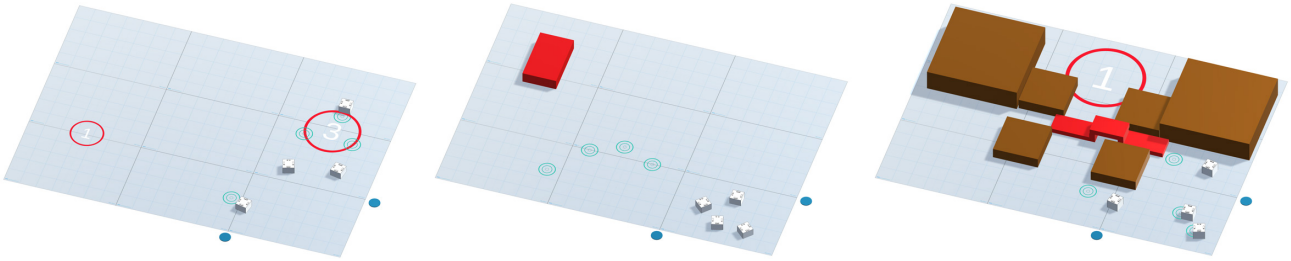
E. Tasks

We implemented three different tasks. An overview of all setups can be seen in Fig. 6.

Task 1: Go to target. The user is shown different target areas in the shape of red circles, each containing a number. The goal is to send as many robots as indicated by the number to each circle. Once enough robots are in the target area, the circle turns green. If each circle contains the correct number of robots, a new set of circles appears. This repeats three times. The locations for the circles as well as the numbers are fixed. Users can choose between direct selection or group selection via the lasso for this task. The goal is to get users acquainted with the system and have them perform a simple task assignment for the robots, where each target area could represent a task on a building site.

Task 2: Move Payload. This task has to be completed using the fingertip mode, i.e. directly controlling each of the robots. The goal is to use the robots to transport a red carboard box to a target area, indicated by two small blue plastic coins at the side of the playmats (cf. Fig. 6b). Users are free to push the box using whatever strategy they see fit, using all robots or a subset. The task is completed as soon as the box enters the target area. This task requires users to take the movement of all robots into account simultaneously, and corresponds to a multi-robot transport task, e.g. to transporting heavy brick palettes to a building site.

Task 3: Escape Room. Users are given the choice of using the selection or fingertip mode for this final task. The goal is for one robot to reach the target area, indicated by a red circle. This time, the path to the target is initially blocked by



(a) Task 1: Go to Target (b) Task 2: Move Payload (c) Task 3: Escape Room

Fig. 6: Overview of the task setups. For each task, the four Toio robots as well as their targets (blue ripples) can be seen. To the side of the playmat, two blue plastic coins indicate the starting area for the robots (and Task 2 target area).

a cardboard contraption, which needs to be pushed aside to open up the path. This task has been designed to test user's ability to simultaneously deal with heterogeneous tasks for multiple robots. In a real scenario, this might correspond to using robots with distinct capabilities in a collaborative manner to achieve a goal, which a single robot cannot achieve on its own.

IV. Results

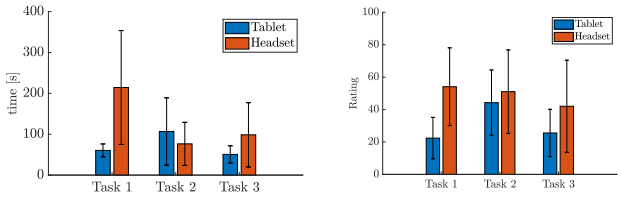
Approval for the study has been obtained from the ETH Zürich Ethics Commission (proposal no. EK 2022-N-168).

A. Demographic

Initially, participants had to complete a quick demographic survey and were given an introduction for each device. The introduction included a 5 minutes training session per device and gesture, in which participants could freely interact with the robots, but did not yet know what the task would be. We completed the study with 14 participants between age 26-38. Six participants wore glasses or contacts, six indicated that they had prior experience with AR or VR devices while eight said they only had little or no experience, and seven indicated that they had played RTS games before. The order in which they used the devices was chosen at random, but all tasks were completed in the same order for each device. After each set of three tasks, participants were given a NASA Task Load Index (TLX) [8] survey to rate their experience. The NASA-TLX has been widely used [34] in HRI as well as more explicitly in recent Mixed-Reality HRI research to assess the perceived workload of a task (see e.g. [12], [17], [35]). For each task except for task 2, the time to completion was measured in-app. For task 2, a stop-watch was used. At the end of the study, a small interview was conducted to ask users more broadly about their experience. One participant did not use the headset and another user did not finish task 1 because of a technical issue. All other participants were able to succesfully complete all tasks on both interfaces.

B. Performance and TLX Rating

Figure 7a shows the average time it took participants to perform each of the tasks, while Fig. 7b shows the averaged TLX scores. We first checked if the data was normally distributed, using a Shapiro-Wilk test [36], which indicated that for the performance comparison a non-parametric test had to be used. A Wilcoxon signed rank test [37] was used to determine if the time to complete the task was significally different between the platforms, with $W_{crit}(N = 12, p < .05) = 12$. Task 1 took participants significantly longer on average when using the HoloLens ($214s \pm 139s$), as compared to the tablet ($60s \pm 16s$), with $W = 0 < W_{crit}$. While the opposite is true for task 2 ($76s \pm 53s$ vs. $107s \pm 82s$), the effect here was not statistically significant ($W = 29 > W_{crit}$). For task 3, some users were able to perform better on the headset than some users using the tablet, but the standard deviation for the headset was much larger and generally tablet performance was better ($98s \pm 79s$ vs. $50s \pm 21s$), with $W = 7 < W_{crit}$. The individual performances can be seen in Fig. 8. A one-sample t-test was used to determine if there was a significant difference between the usage of the tablet or the headset for the same tasks in terms of TLX score. Only task 1 showed a statistically significant difference, with a score of 22.38 ± 12.79 for the tablet vs. 54.1 ± 24 for the headset ($t(12) = 4.03$, $p = 0.0017$). A detailed breakdown of the factors can be seen in Fig. 9, which shows that, especially for task 1, participants reported higher levels of effort and frustration when using the headset as compared to the tablet.



(a) Time (mean and SD) (b) TLX (mean and SD)

Fig. 7: Average completion time and TLX scores per task.

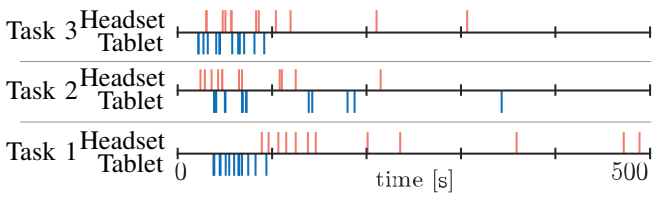


Fig. 8: Dotplot of the individual completion times per task.

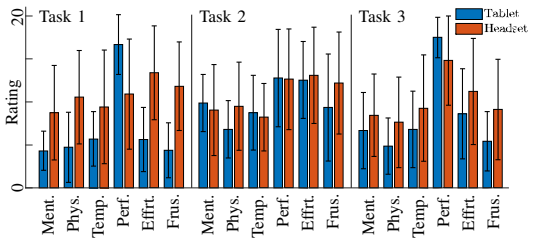


Fig. 9: TLX factors and their standard deviations on a scale from 1-20 for the tablet and headset interfaces respectively.

C. Preferences and post-survey

During the post-survey, participants indicated that their main source of frustration when using the headset was failure of hand tracking, especially for the pinch gesture which needs to be used extensively in task 1. As visible in Tab. I, participants much prefer the tablet for the simple task assignment interaction, while for task 2 there is an even split in preference. When asked what interface they would prefer for HRI in general, there is an even split again, although 3 out of the 5 users who would prefer the tablet indicated that they would probably have answered differently if the hand tracking had been more reliable. Several users highlighted that the headset definitely had an advantage for task 2, as controlling multiple robots via multi-touch on a tablet meant that a large part of the tablet's screen was hidden by the hand and it was hard to see what the robots were doing.

V. Discussion

Today, touch interfaces are well-established. Users are generally familiar with tapping or dragging touch gestures, and for many of them it is their primary input modality [38]. This clearly shows in the results of our first task, as participants were generally very swift in their understanding of the interaction and in reaching the goals when using the tablet, and the standard deviation among participants is low. It is also encouraging that even participants who indicated no prior experience with AR were able to use the interface and reported that it was intuitive to use. In contrast, it took participants much longer to complete task 1 on the headset. Multiple participants reported that they did not think the pinch gesture used to trigger input was very intuitive. This might be due to issues with the hand-tracking, which in turn made participants insecure about how the gesture works. Contrarily, users were more successful in using the headset for task 2, although not significantly more so than with the tablet. We would argue that this is due to a combination of factors, of which the most important are the better immersion and ability to see the real environment clearly, as well as the complexity of the task, for which the free use of the hand on on the headset offered an advantage compared to two-dimensional touch input on the tablet. It can be argued that this gesture does not translate well to the touch screen, which also shows in the higher levels of frustration, mental load and effort participants reported for the tablet compared to task 1 (see Fig. 9). On the other hand, this shows that there are indeed interactions, especially with multiple robots for a more *swarm*-like behaviour, for which the touch screen is limited and other means of input might be required. Finally, it is encouraging to see that some users who had previously leaned towards the tablet for task 1 preferred the headset for task 3. This might in-part be due to a training effect over the course of the session, as well as the encouragement of good performance during the previous task and, finally, the increased complexity of the task itself. Regarding the limitations of our study, it became clear that users might require more in-depth training for the pinch gesture. This is emphasized by an analysis of the word-count of the post-survey, in which "pinch" and "selection" where mentioned a combined 24 times, both ranking in the top 10% of word frequency. Participants who weren't able to use the gesture reliably reported significantly higher frustration had a worse performance, which skewed the results. In future studies we would like to extend the range of possible interactions to some of the options outlined in Figs. 4 and 5. Other interesting aspects would be to use larger robots, 6DoF controllers, or hide objects behind a wall, which would require users to move around and make use of the 3D immersion that AR and MR offer.

Task	Tablet	HoloLens	Both
Go to Target	11	1	2
Move Payload	7	7	0
Escape Room	9	5	0
HRI in general	5	5	4

TABLE I: User preferences for each task.

VI. Conclusion

We have presented an AR and MR system for interaction with multiple robots via touch gestures on a tablet or hand-tracking on a HoloLens 2 headset. Providing several interaction modes, we have conducted a user study in order to compare user preferences and performances on both interfaces across a set of tasks. Our results show, that participants were able to successfully perform both simple and more complex multi-robot tasks using our interface. Furthermore, there is a clear preference for the more familiar touch-based interactions on a tablet for simpler tasks, while users tend to perform better and a significant part of the users actually prefers the headset for more complex interactions. More in-depth studies will have to be conducted in order to explore the possibilities of different hand gestures for HRI and some technical challenges will have to be resolved in order to make hand-gesture interactions more accessible to novice users. Nevertheless, we think that our selection of tasks, interactions and interfaces and the results of our user study serve as an interesting baseline for the comparison of tablet-based AR and hand-tracking-based MR for HRI.

Acknowledgements

The authors would like to thank Velko Vechev for his valuable input and the interesting discussions, as well as Beat Reichenbach for the inspiring render in Figure 1.

REFERENCES

[1] A. Gawel, H. Blum, J. Pankert, K. Krämer, L. Bartolomei, S. Ercan, F. Farshidian, M. Chli, F. Gramazio, R. Siegwart *et al.*, "A fully-integrated sensing and control system for high-accuracy mobile robotic building construction," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 2300–2307.

[2] K. Azadeh, R. De Koster, and D. Roy, "Robotized and automated warehouse systems: Review and recent developments," *Transportation Science*, vol. 53, no. 4, pp. 917–945, 2019.

[3] N. Sousa, A. Almeida, J. Coutinho-Rodrigues, and E. Natividade-Jesus, "Dawn of autonomous vehicles: review and challenges ahead," in *Proceedings of the Institution of Civil Engineers-Municipal Engineer*, vol. 171, no. 1. Thomas Telford Ltd, 2018, pp. 3–14.

[4] C. G. Atkeson, B. Babu, N. Banerjee, D. Berenson, C. Bove, X. Cui, M. DeDonato, R. Du, S. Feng, P. Franklin *et al.*, "What happened at the DARPA Robotics Challenge, and why," *DRC Finals Special Issue of the Journal of Field Robotics*, vol. 1, 2016.

[5] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[6] L. De Nul, M. Breque, A. Petridis. (2021) "Industry 5.0: Towards a sustainable, human-centric and resilient European industry". Directorate-General for Research and Innovation (European Commission). [Online]. Available: https://op.europa.eu/s/pj57

[7] W. Hamilton, A. Kerne, and T. Robbins, "High-performance pen+ touch modality interactions: a real-time strategy game esports context," in *Proceedings of the 25th annual ACM symposium on User interface software and technology*, 2012, pp. 309–318.

[8] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*. Elsevier, 1988, vol. 52, pp. 139–183.

[9] M. Dianatfar, J. Latokartano, and M. Lanz, "Review on existing vr/ar solutions in human–robot collaboration," *Procedia CIRP*, vol. 97, pp. 407–411, 2021, 8th CIRP Conference of Assembly Technology and Systems. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2212827120314815

[10] R. Suzuki, A. Karim, T. Xia, H. Hedayati, and N. Marquardt, "Augmented reality and robotics: A survey and taxonomy for ar-enhanced human-robot interaction and robotic interfaces," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–33.

[11] J. A. Frank, S. P. Krishnamoorthy, and V. Kapila, "Toward mobile mixed-reality interaction with multi-robot systems," *IEEE Robotics and Automation Letters*, vol. 2, no. 4, pp. 1901–1908, 2017.

[12] J. Patel, Y. Xu, and C. Pinciroli, "Mixed-granularity human-swarm interaction," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 1059–1065.

[13] F. Kennel-Maushart, R. Poranne, and S. Coros, "Manipulability optimization for multi-arm teleoperation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

[14] ——, "Multi-arm payload manipulation via mixed reality," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 11 251–11 257.

[15] J. Alonso-Mora, S. H. Lohaus, P. Leemann, R. Siegwart, and P. Beardsley, "Gesture based human-multi-robot swarm interaction and its application to an interactive display," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 5948–5953.

[16] J. Alonso-Mora, A. Breitenmoser, M. Rufli, R. Siegwart, and P. Beardsley, "Multi-robot system for artistic pattern formation," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 4512–4517.

[17] V. Serpiva, E. Karmanova, A. Fedoseev, S. Perminov, and D. Tsetserukou, "Swarmpaint: Human-swarm interaction for trajectory generation and formation control by dnn-based gesture interface," in *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2021, pp. 1055–1062.

[18] H. Kaimoto, K. Monteiro, M. Faridan, J. Li, S. Farajian, Y. Kakehi, K. Nakagaki, and R. Suzuki, "Sketched reality: Sketching bi-directional interactions between virtual and physical worlds with ar and actuated tangible ui," in *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 2022, pp. 1–12.

[19] L. H. Kim, D. S. Drew, V. Domova, and S. Follmer, "User-defined swarm robot control," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[20] R. Yu and D. A. Bowman, "Force push: Exploring expressive gesture-to-force mappings for remote object manipulation in virtual reality," *Frontiers in ICT*, vol. 5, p. 25, 2018.

[21] M. Le Goc, L. H. Kim, A. Parsaei, J.-D. Fekete, P. Dragicevic, and S. Follmer, "Zooids: Building blocks for swarm user interfaces," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, 2016, pp. 97–109.

[22] K. Nakagaki, J. Leong, J. L. Tappa, J. Wilbert, and H. Ishii, "Hermits: Dynamically reconfiguring the interactivity of self-propelled tuis with mechanical shell add-ons," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, 2020, pp. 882–896.

[23] M. Rubenstein, A. Cornejo, and R. Nagpal, "Programmable self-assembly in a thousand-robot swarm," *Science*, vol. 345, no. 6198, pp. 795–799, 2014.

[24] F. Mondada, M. Bonani, X. Raemy, J. Pugh, C. Cianci, A. Klaptocz, S. Magnenat, J.-C. Zufferey, D. Floreano, and A. Martinoli, "The e-puck, a robot designed for education in engineering," in *Proceedings of the 9th conference on autonomous robot systems and competitions*, vol. 1, no. CONF. IPCB: Instituto Politécnico de Castelo Branco, 2009, pp. 59–65.

[25] J. A. Preiss, W. Honig, G. S. Sukhatme, and N. Ayanian, "Crazyswarm: A large nano-quadcopter swarm," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 3299–3304.

[26] Essential Reality, "P5 glove," 2006, [Online; accessed September 08, 2022]. [Online]. Available: http://www.mindflux.com.au/products/essentialreality/p5glove.html

[27] Lionhead Studios Ltd., "Black and white 2," 2005, [Online; accessed September 05, 2022]. [Online]. Available: https://abandonwaregames.net/game/black-and-white-2

[28] Sony Interactive Entertainment. (2021) toio™ core cube technical specifications 2.3.0. [Online]. Available: https://toio.github.io/toio-spec/docs/hardware_other

[29] Morikatron Inc., "toio SDK for Unity," 2022, [Online; accessed September 15, 2022]. [Online]. Available: https://github.com/morikatron/toio-sdk-for-unity

[30] J. Van den Berg, M. Lin, and D. Manocha, "Reciprocal velocity obstacles for real-time multi-agent navigation," in *2008 IEEE International Conference on Robotics and Automation*. IEEE, 2008, pp. 1928–1935.

[31] D. Ungureanu, F. Bogo, S. Galliani, P. Sama, C. Meekhof, J. Stühmer, T. J. Cashman, B. Tekin, J. L. Schönberger, P. Olszta *et al.*, "Hololens 2 research mode as a tool for computer vision research," *arXiv preprint arXiv:2008.11239*, 2020.

[32] K. Takahashi, "Apriltag package for unity," https://github.com/keijiro/jp.keijiro.apriltag, 2022.

[33] J. Berg, A. Lottermoser, C. Richter, and G. Reinhart, "Human-robot-interaction for mobile industrial robot teams," *Procedia CIRP*, vol. 79, pp. 614–619, 2019.

[34] S. G. Hart, "Nasa-task load index (nasa-tlx); 20 years later," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 50, no. 9. Sage publications Sage CA: Los Angeles, CA, 2006, pp. 904–908.

[35] C. Nam, H. Li, S. Li, M. Lewis, and K. Sycara, "Trust of humans in supervisory control of swarm robots with varied levels of autonomy," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018, pp. 825–830.

[36] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3/4, pp. 591–611, 1965.

[37] F. Wilcoxon, *Individual comparisons by ranking methods*. Springer, 1992.

[38] J. Avery, "Enhanced multi-touch gestures for complex tasks," Ph.D. dissertation, University of Waterloo, 2018.