Knowledge Acquisition for Robots Through Mixed Reality Head-Mounted Displays

Nishanth Kumar* Eric Rosen* Stefanie Tellex

***Abstract*—A robot can clean a room by performing a series of pick-and-place tasks on relevant items. To accomplish this sequence, the robot must know the original poses of the relevant items, how to grasp these items, and the final poses to place each item at. Language and gestures have been shown to successfully enable humans to help a robot acquire this knowledge, but are limited because they do not allow people to express poses precisely. Mixed Reality Head-Mounted Displays (MR-HMDs) have recently gained traction as a means to facilitate human-robot communication because they can visualize 3D graphics on the environment from the perspective of of the user. We hypothesize that MR-HMDs provide an intuitive interface for enabling end-users to help robots acquire grounded knowledge about how to clean up a shared workspace. We propose an MR system that allows a user to communicate to a mobile manipulator what items need to be placed away, how to grab them, where they are located and where they need to be placed. After this information is provided, the robot can autonomously clean a room.**

***Index Terms*—Mixed Reality, Pose Detection, Grasp Annotation, Human-Robot Interaction**

I. Introduction

Having a robot clean up a room is a large desiderata for home robots. This clean up task can be decomposed into a sequence of pick and place tasks performed on specific items. However, there are many things the robot must know about the environment in order to successfully accomplish a clean up task. The robot must know 1) the relevant items in the room that it must identify and place elsewhere, 2) the locations of these items so that it may navigate to them, 3) the affordance points where it may grasp these items, and 4) the locations it must place these items at such that the room becomes cleaned.

In an ideal world, a robot would autonomously learn or identify everything it needs given language-based commands from a human. However, despite current advances, robots are still unable to accurately identify and localize items in real-time due to perception errors. Moreover, even if this were possible, robots are currently unable to successfully grasp and manipulate arbitrary objects autonomously. These issues are exacerbated by the often partially-observable nature of clean up in the real world: there is a good chance the robot will be unable to see all the relevant items immediately (i.e, it may have to search for them). Furthermore, the robot's sensor may be imperfect, introducing noise into the observations that make inference even more difficult. Having a human help a robot build a knowledge base about the shared environment has been shown to be useful in circumventing these issues. [1, 2]. These approaches use modalities such as language and gestures to enable symbol grounding through human-robot interactions. MR has started to gain attention as a useful interface for facilitating human-robot interactions. Existing approaches have used MR to enable robots to communicate motion intent to users [3], as well as program robot motions [4]. However, these approaches parameterize motions through robot poses. To our knowledge, MR has not been used to allow users to ground item information for mobile manipulators to perform motion behaviors in a shared environment.

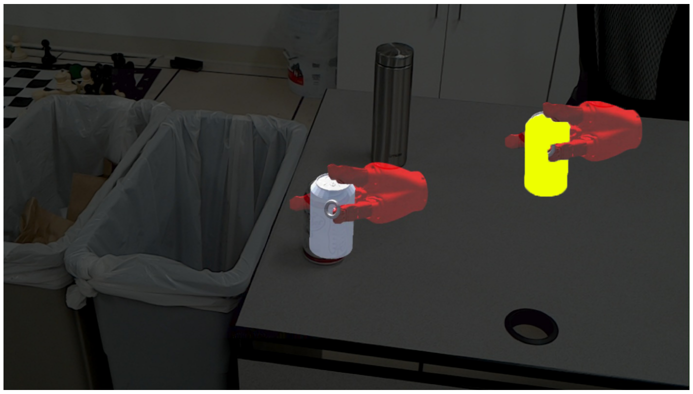


Fig. 1: A picture from the perspective of a user with our proposed MR interface grounding the pose of the item, how to grab it, and where to place it. A white Dr. Pepper soda can mesh model has been positioned over a real soda can and a grasp annotation has been added. In addition, a yellow version of the mesh has been positioned somewhere else on the table to represent a goal pose for the item to be placed, and how its hand should be positioned.

We hypothesize that MR can be used to create an intuitive interface for enabling end-users to help robots acquire knowledge about items in their environment (Fig. 1). We outline our approach to an MR interface that allows a user to add 3D meshes of items into the shared environment that they can place and orient over real items to tell the robot what items are in the room and what their poses are. Furthermore, the user is able to generate 3D meshes of the robot's end effector and place them relative to the item to provide grasp annotations that instruct the robot on how to grab the items. Finally, the user is able to specify goal poses for the items by reposing yellow 3D meshes that represent the user's desired placement

*These authors contributed equally.

of the item.

Future work will consist of connecting all these components together under one interface, and enabling a robot to use grounded information from a user with our MR interface to clean a room. This work also has the potential to be integrated with a learning agent, helping relieve issues via autonomous approaches. Furthermore, future work can integrate this system into a non-deterministic planner to allow a robot to deal with noise and failure cases.

II. Related Work

Related works have proposed methods for enabling robots to acquire information about their environment through interaction with a user [1, 2]. Bastianelli et al. [1] propose a system that integrates a) a Simultaneous Localization and Mapping (SLAM) subsystem, to provide a metric map of the environment, and b) a multi-modal interface that allows a user to use speech and gestures to point to items in the environment and assign symbolic information to them. The user is able to use a laser pointer to highlight items, and the agent is able to locate the point and use a RGB-D sensor to segment out the element. After the item has been segmented, the sensor information can be used to perform pose detection and contribute to the robot's information about the environment. Bastianelli et al. [1] shows that a multi-modal interface that uses language and gesture can be used to build an effective knowledge base representation for an agent performing symbol grounding in a human robot interaction.

Randelli et al. [5] also found that multi-modal human-robot interactions facilitated by speech and gestures can be useful for helping robots build relevant knowledge bases about environments they have never been in before. Randelli et al. [5] propose a tangible user interface that incorporates mechanisms for having users select items in the environment similar to Bastianelli et al. [1], as well as ground landmarks. In addition, Bastianelli et al. [1] perform a case study in a home environment, where a robot may need to perform tasks such as cleaning. Related works have given service robots access to semantic information about the environment to help them perform tasks, but constructing these knowledge bases is still difficult [5].

Similarly, Kemp et al. [6] developed an interface for a laser-pointer to be used to specify the 3D object location of an item. They showed that this interface allowed users to enable a mobile manipulator to pick up objects from the floor reliably. This system was especially novel because it does not require any instrumentation of the environment, and is also robust for real-world domains. Kemp et al. [6] motivates our MR interface to also work in general, unstructured environments, such as a home. However, a laser pointer is limited in what it can "highlight" in the environment, and information about the object's affordance points or goal poses may not be so precisely communicated through the modalities used by Kemp et al. [6] and Bastianelli et al. [1].

Other related work has shown that human knowledge can be leveraged to enable robots to grasp previously unseen objects. Leeper et al. [7] developed a monitor and mouse based interface to allow users to specify objects and grasp locations and showed that this enabled the robot to grasp objects in cluttered environments and minimize collisions. Lin and Chiang [8] show that human gestures can be used to specify information to a robot about a pick-and-place task.

Nguyen and Kemp [9] propose a method to autonomously compute feasible locations on items to perform manipulation behaviors by using RGB images and 3D point clouds of an item. Nguyen and Kemp [9] evaluate different 3D locations around the item, and record whether the behavior would be successful or not. Our proposed MR interface can be used to generate seed locations for these learning algorithms by leveraging the annotated grasp poses the users give.

Mixed Reality (MR) has gained popularity in human-robot interactions. Rosen et al. [3] developed a MR interface for enabling a robot to communicate its motion intent to a user. In addition, Rosen et al. [3] conducted a user study to compare the MR interface to a 2D monitor approach, and found 16 percent increase in accuracy with a 62 percent decrease in the time it took to label robot motions as either safe or unsafe, compared to the next best system. Gadre et al. [4] developed an MR based interface and showed that this is more intuitive and easier to use than 2D interfaces for users to program simple tasks for a stationary robot arm to complete. Frank et al. [10] showed that MR based interfaces are more intuitive and easier-to-use than 2D interfaces for users to command multiple-robots to perform simple tasks. However, no MR system to our knowledge has enabled users to ground item information to enable a mobile manipulator to perform a complex behavior such as cleaning an unknown room.

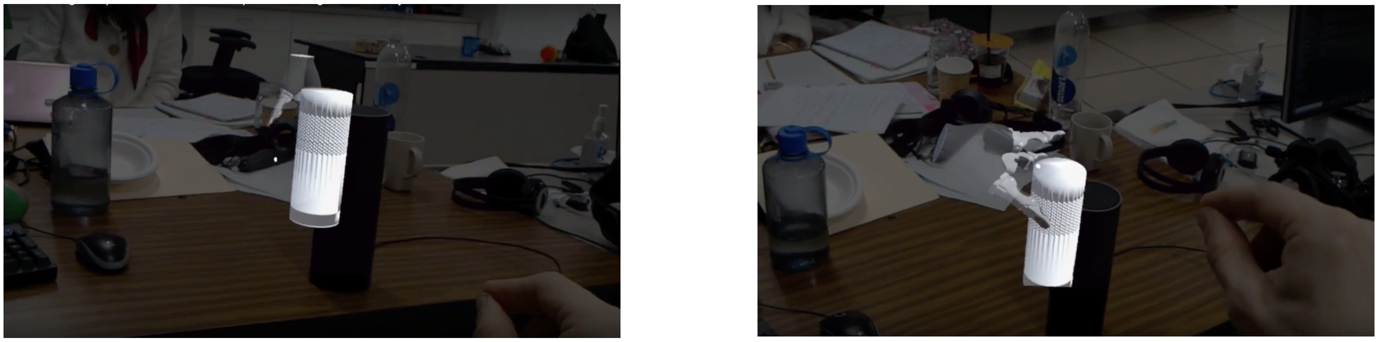
III. System Description

Our proposed MR interface enables a user to directly ground information about items in the environment (Fig. 2). To do this, the MR-HMD map must be calibrated to the mobile manipulator's map. We accomplish this task by presenting the user with a 3D mesh model of the robot, which they move to align with the real one using the MR-HMD. This calibration step aligns the MR-HMD coordinate frame with the robot's.

Now that the robot and MR-HMD are calibrated, the MR-HMD can display 3D information about items it knows about, overlaid on the physical items. At the start, the mobile manipulator may know nothing but how to move to locations, but the user may now ground information about the items using MR.

A. Pose Detection

To specify the object's current location and orientation, a user simply selects the relevant 3D mesh model, and places it over the real object using hand gestures (Fig. 2). This is done in a similar manner to the coordinate frame calibration step, but instead of dragging a 3D mesh model of the robot over the real robot, we drag a 3D mesh model of the item over the real item. Because the MR-HMD coordinate frame is aligned with the robot's, the robot is now aware of the pose of the



a) A user repositioning the 3D mesh over the real item to ground the current pose.

b) A user repositioning the 3D mesh of a robot end-effector to ground the grasp pose relative to the item.

Fig. 2: Two images of our currently implemented system (https://www.youtube.com/watch?v=Qipza556sWQt). A user is grounding a) the pose of an Amazon Alexa to be cleaned up, and b) how to grab the Alexa. Once the robot has this information, it can use a motion planner to move to the item, and grab it.

object in 3D space with respect to its coordinate system. This enables the robot to know where the item is, and can be used to help the robot navigate to the item.

This assumes that a 3D model of the item already exists and is available. However, in unstructured environments, the robot may encounter items it has never seen before. If the system does not have access to the object model, the user can use MR to help the robot generate it through an autonomous approach. For example, the user can select a 3D primitive shape, like a sphere, and place it over the desired object. A robot with a RGBD sensor could then go the item and move around it, using the sphere as a mask for what parts of the perception data are relevant for mesh generation.

B. Grasp Annotation

The user must also provide at least one grasp annotation around the mesh model so that the robot can grab the item. These annotations are in the form of MR visualizations of the robot's end-effectors that the user can place around a mesh model (Fig. 2). In our proposed system, users will be able to pair the end effectors with the 3D mesh model, so that the end effector pose is saved relative to the item pose. This allows the robot to know where the affordance poses of the item are for the 3D mesh, regardless of current pose. For example, if the item moved and the mesh model was updated to reflected that, we would still know the affordance poses for the item. The robot is now aware of various relative poses it could move its end-effector to in order to grasp the object.

C. Goal Pose Specification

Finally, the user must specify the pose at which each object should be placed. The user may specify the goal pose by using hand gestures to move a copy of the object model to a goal location (Fig 1). This is very similar to the pose detection component, except instead of dragging the 3D mesh over the real item, the user drags it to a desired pose for where the item should be put away. In addition, the user could use MR to highlight a 3D region of the environment to indicate a cleanup area for the item, letting the agent place the item in any stable pose inside the region.

D. Execution

The robot will begin executing the sequence in the same order the user specified. First, it will localize itself, and navigate to a randomly-chosen position near the object such that it is close enough to attempt grasps, using a move base planner such as GMapper [11]. It will then orient itself to face the object. It will choose a random grasp annotation from those the user has specified and attempt to grasp the object at this pose. It will use torque sensors in its grippers to reason whether it has successfully grasped the item or not. If it has not, it will attempt to grasp using a different grasp annotation. If it is unsuccessful at every annotation provided, it will move the base to a new location to find a different plan. Once it succeeds at grasping, it will use SLAM to navigate to a region near the user-specified pose to place the object. The robot knows the goal pose of the item and how it is grabbing the item, so the agent can calculate the pose it must move its end-effector to such that the object is placed correctly. The robot will then repeat this sequence of steps for the next object.

E. Proposed Work

We intend to complete the system described in Section III above (video in Fig. 2) and implement it with an MR-HMD and mobile manipulator. We will then setup a representative kitchen cleanup task that will involve the robot performing a series of pick-and-place tasks in a cluttered environment. We will perform a user-study and evaluate both quantitative and qualitative metrics. Specifically, we will measure task completion rate and speed, as well as evaluate our system's usability via the NASA-TLX [12] and System Usability Score (SUS) [13]. We will evaluate our results against the baseline of users performing the same task using a 2D monitor and mouse based interface.

In the future, we intend to extend our system to plan execution of the robot's task more efficiently and also recover from failure autonomously. We expect to accomplish this by framing the cleanup task as a Partially-Observable Markov Decision Process (POMDP) [14]. We will then use a known POMDP solver to generate plans where the robot will execute the pick-and-place sequences in the optimal order and also be able to autonomously recover from failures such as inability to grasp the object at the provided annotations, inability to navigate to a location near the object, etc. Using the POMDP framework also opens the doors for the robot to use Reinforcement Learning (RL) to improve itself at the cleanup task.

We also plan to extend our system allow users to help robots learn new skills in addition to cleanup. We intend to use MR to allow users to directly specify relevant information about the skill to the robot, and then use Learning From Demonstration (LFD) [15] to teach the robot the skill. We intend to investigate the effectiveness of MR as a method of teaching robots arbitrary skills and also how well these skills will generalize.

In addition, parts of our proposed MR system could be used to help aid the autonomous approaches to different parts of the system. For example, after the user selects the 3D mesh of an item in the room they want the robot to interact with, it can be paired with a 3D point cloud captured by the robot to be fed into a mesh alignment algorithm to autonomously perform pose detection. Furthermore, a user could additionally use MR to highlight a 3D region of the environment to crop out what parts of the 3D point cloud should be inputted to the mesh alignment algorithm.

IV. Conclusion

Our proposed MR interface allows users to specify information necessary for a mobile manipulator to complete complex pick-and-place tasks in novel environments. We intend to complete development of this system and implement it using an MR-HMD and mobile manipulator. We will then setup a cleanup task involving multiple pick-and-place actions executed on different items and perform a substantial user-study based on this task.

In the future, we intend to incorporate POMDP's into our interface to allow the robot to autonomously recover from failure. We also plan to enable the robot to learn and improve at its pick-and-place tasks via Reinforcement Learning. Finally, we intend to extend our system to allow the robot to be taught arbitrary actions and the conditions under which to execute them via Learning From Demonstration. We then plan to evaluate the ease-of-use and effectiveness of this system.

References

[1] E. Bastianelli, D. Bloisi, R. Capobianco, G. Gemignani, L. Iocchi, and D. Nardi, "Knowledge representation for robots through human-robot interaction," *arXiv preprint arXiv:1307.7351*, 2013.

[2] S. Rosenthal, J. Biswas, and M. Veloso, "An effective personal mobile robot agent through symbiotic human-robot interaction," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 2010, pp. 915–922.

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

[4] S. Y. Gadre, E. Rosen, G. Chien, E. Phillips, S. Tellex, and G. Konidaris, "End-user robot programming using mixed reality," *ICRA*, 2019.

[5] G. Randelli, T. M. Bonanni, L. Iocchi, and D. Nardi, "Knowledge acquisition through human–robot multimodal interaction," *Intelligent Service Robotics*, vol. 6, no. 1, pp. 19–31, 2013.

[6] C. C. Kemp, C. D. Anderson, H. Nguyen, A. J. Trevor, and Z. Xu, "A point-and-click interface for the real world: Laser designation of objects for mobile manipulation," in *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2008, pp. 241–248.

[7] A. E. Leeper, K. Hsiao, M. Ciocarlie, L. Takayama, and D. Gossow, "Strategies for human-in-the-loop robotic grasping," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. ACM, 2012, pp. 1–8.

[8] H.-I. Lin and Y. P. Chiang, "Understanding human hand gestures for learning robot pick-and-place tasks," *International Journal of Advanced Robotic Systems*, vol. 12, no. 5, p. 49, 2015. [Online]. Available: https://doi.org/10.5772/60093

[9] H. Nguyen and C. C. Kemp, "Autonomously learning to visually detect where manipulation will succeed," *Autonomous Robots*, vol. 36, no. 1-2, pp. 137–152, 2014.

[10] 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, Oct 2017.

[11] B. Gerkey. (2017) GMapping. [Online]. Available: http://wiki.ros.org/gmapping

[12] NASA Human Performance Research Group and others, "Task Load Index (NASA-TLX) v1.0 Computerised Version," *NASA Ames Research Centre*, 1987.

[13] J. Brooke *et al.*, "SUS-A Quick and Dirty Usability Scale," *Usability Evaluation in Industry*, vol. 189, no. 194, pp. 4–7, 1996.

[14] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.

[15] D. Whitney, E. Rosen, and S. Tellex, "Learning from crowdsourced virtual reality demonstrations," *Virtual Augmented Mixed Reality Human Robot Interactions (VAM-HRI)*, 2018.