A Simulator for Human-Robot Interaction in Virtual Reality

Mark Murnane* Padraig Higgins† Monali Saraf‡ Dr. Francis Ferraro§ Dr. Cynthia Matuszek¶ Dr. Don Engel‖

University of Maryland, Baltimore County

ABSTRACT

We present a suite of tools to model a robot, its sensors, and the surrounding environment in VR, with the goal of collecting training data for real-world robots. The virtual robot observes a rigged avatar created in our photogrammetry facility and embodying a VR user. We are particularly interested in verbal human/robot interactions, which can be combined with the robot's sensor data for grounded language learning. Because virtual scenes, tasks, and robots are easily reconfigured compared to their physical analogs, our approach proves extremely versatile in preparing a wide range of robot scenarios for an array of use cases.

Keywords: ROS, Virtual Reality, Robotics, Unity.

Index Terms: Human-centered computing—Visualization—Visualization systems and tools; Computing methodologies—Artificial intelligence—Computer vision—Computer vision tasks

Figure 1: An experimental participant, instructing a robot

1 MOTIVATION AND INTRODUCTION

When humans learn their first language, visual stimuli and human interaction are a key part of developing a vocabulary. This context provides a "grounding" for the association of words and concepts, and language learned in this way is 'grounded language.' At the same time, learning physically contextualized language is a growing research area in human-robot interaction (HRI), computer vision, and computational linguistics [1]. In robotic grounded language learning, robots acquire mappings between physical concepts and linguistic constructs through sensory inputs that include both human interactions and a dynamic scene.

Developing larger-scale datasets is one of the challenges faced when training robots via language. In addition, current datasets have a tendency to be collected from populations of convenience, rather than being drawn from a diverse pool or one representative of intended users [5]. In this work, we present a suite of tools which allow efficient collection of such datasets by putting humans and virtual robots in virtual environments and collecting high-fidelity sensor data. Our suite of tools mitigates these challenges by providing a portable, virtual alternative to in-person studies.

With these motivations in mind, we present the RIVR simulator (Robot Interaction in VR, pronounced "river"). RIVR is a system we have developed for acquiring a training corpus of human-robot interactions. Our system allows for a robot and its sensors to be represented in a scene where they interact with a human avatar animated by the actions of a human using a commodity virtual reality system. Our approach draws from our team's expertise as researchers in the field of human-centered robotics, virtual reality, and computational linguistics. Our work leverages the open source ROS# [2] project, allowing the robot to be controlled by the same ROS [11] software which would be used for its real-world analog, while the Unity-based VR interface can provide a wide range of scenarios for human participants.

Simulation has been used to advance robotics research from very early in the history of the discipline. Our primary contributions are to (1) enable simulation which includes real-time human-robot interaction, and (2) capture realistic output from the virtual robot's simulated sensors of people in the VR environment. Participants are represented using a photorealistic avatar, currently captured in our photogrammetry facility and rigged to the VR controllers and headset using inverse kinematics. This is a new direction relative to prior work, most of which does not include a human in the scene [3].

This work is directly inspired by similar recent work [7, 8]. The most similar active project is SIGVerse [4], which is being developed concurrently and where similar design choices and application areas were identified. Our work differs in a number of significant ways, including our focus on the linguistics side of grounded learning and in the photo-realism of our photogrammetrically-derived avatars and scenes, which we expect to help optimize authenticity in the behavior of human participants. Our initial use case simulates an indoor apartment scene that is scattered with a number of household objects, drawn from an existing virtual world [6, 10].

In this poster, we present a system that integrates a VR environment with a realistic, controllable virtual robot platform. This system was developed with language-based learning in mind, with the specific goal of gathering robotics sensor data in a setting with sufficient photo-realism and immersion to allow for natural human speech and behavior. This brings virtual reality into an application area in which there is a real, immediate need, potentially supporting a new class of research studies in human-robot interaction.

2 APPROACH & STUDY DESIGN

When originally designed, the simulator was intended to be used in a lab setting, where high performance networking and computing systems were co-located with the test participants. This allowed for a simple architecture and a monolithic system design. However, with the emergence of COVID-19, requiring the physical presence

*e-mail:mark25@umbc.edu
†e-mail:phiggin1@umbc.edu
‡e-mail:monali2@umbc.edu
§e-mail:ferraro@umbc.edu
¶e-mail:cmat@umbc.edu
‖e-mail:donengel@umbc.edu

978-0-7381-1367-8/21/$31.00 ©2021 IEEE
DOI 10.1109/VRW52623.2021.00117

of participants and researchers became untenable. These added constraints have ultimately led to a more flexible, broadly accessible system.

2.1 Architecture

Our overall design uses a distributed approach. The researcher, test participant, and control server are all located in different places, connected via the internet. This makes the system particularly sensitive to internet connections of individuals in the simulation. High latency, low bandwidth, and random packet loss are far more likely than when similar experiments are performed in a lab.

The main components of the system are a VR Client, which was developed in the Unity game engine using ROS# and a REST API to communicate with the server components; a coordination server that exposes a REST API and activates the simulation components and stores test data; a ROS instance that models the robot under test; and a rendering system that uses the Unity game engine to render auxiliary sensor data.

To avoid co-location, participants were recruited who already owned VR systems. This required support for diverse hardware. Based on an informal poll of participants and a review of generally available VR devices, we determined that the broadest base of hardware support would be achieved by supporting SteamVR compatible headsets. An additional percentage of headsets could be made compatible by supporting the Oculus SDK. As such we developed support for both VR ecosystems.[1]

2.2 Pilot Study

In order to validate the design and capability of our system, we ran a series of initial 'pilot' studies with small groups. This allowed us to improve upon both scenario design and applicability to our ultimate HRI goals.

Setup The scene used in the trials is a kitchen scene that was adapted from the AI2Thor [6] environment, containing a kitchen island with a variety of food items on it. For this study, the robot is controlled through the Wizard-of-Oz interface by one of the researchers conducting the human trial. The robot and the human are placed on opposite sides of the table.

Procedure The robot picks up objects from the table and prompts the human for a description of the object. If the initial description gives only the object's name, the wizard can have the robot ask follow-up questions about the physical characteristics of the object, such as color, size, shape, and texture. Once all the objects are described, the robot asks the human to walk it through selecting items for a lunch and placing them on a tray.

Qualitative results After each trial, participants completed a survey to describe their experiences. The responses were used to evaluate the verisimilitude of the simulation, and to determine where future simulator development would be best focused. Responses described limitations in the physics implementation, lack of diversity in responses from the robot, and difficulty distinguishing unfamiliar objects in the scene.

Takeaways Having completed a simple experiment, we have found the simulator itself to be effective. Users were not impeded by latency, the graphics were of a sufficient quality to complete the tasks as required, and the sensor data acquired includes the desired measurements.

In the process of completing the trials we also learned many things about scenario design. Choosing objects to place in the scene required careful thought to elicit the desired responses from the human participants. As an example, models that appear distinct outside of VR may be less distinguishable when viewing them through a headset, particularly if the defining features are text or surface detail rather than overall shape or color.

[1] Although our primary headset target was originally Oculus devices, particularly the Oculus Quest and Quest 2, privacy concerns about changes to Facebook integration [9] rendered these devices unsuitable from an ethical and IRB standpoint.

3 Conclusion

This work forms a framework for conducting human robot interaction experiments in multiple modalities of virtual reality. Much of the future work to be done will be in the form of performing experiments using the simulator, rather than simply continuing work on the simulator itself. However, there are still a number of interesting directions to explore that may enable additional types of research. Having multiple users able to simultaneously interact with the simulated robot could open another avenue for investigation. Our system architecture supports multiple simultaneous users; however, we have not developed any scenarios that would make use of this capability.

During the development of our system ROS2 was released, and the Gazebo simulator began to transition to being part of the Ignition software suite. Once ROS2 becomes mature it will be important to port the network layer of our system as well as the model import and export utilities to be compatible with the new versions. At time of writing, some libraries required have yet to be ported to ROS2, and as such we opted to continue under the existing system.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant Nos. 1428204, 1531491, 1637614, 1657469, 1637937, and 1940931. We would like to thank Adam Berlier for extensive, helpful feedback on this manuscript, and our reviewers for their thoughtful suggestions.

References

[1] M. Bansal, C. Matuszek, J. Andreas, Y. Artzi, and Y. Bisk. Proceedings of the first workshop on language grounding for robotics. In Proceedings of the First Workshop on Language Grounding for Robotics, 2017.

[2] M. Bischoff. ROS#, Dec. 2019.

[3] K. M. Hermann, F. Hill, S. Green, F. Wang, R. Faulkner, H. Soyer, D. Szepesvari, W. M. Czarnecki, M. Jaderberg, D. Teplyashin, et al. Grounded language learning in a simulated 3d world. Technical report, arXiv preprint, 2017.

[4] T. Inamura and Y. Mizuchi. Sigverse: A cloud-based vr platform for research on social and embodied human-robot interaction, 2020.

[5] E. S. Jo and T. Gebru. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 306–316, 2020.

[6] E. Kolve, R. Mottaghi, W. Han, E. VanderBilt, L. Weihs, A. Herrasti, D. Gordon, Y. Zhu, A. Gupta, and A. Farhadi. Ai2-thor: An interactive 3d environment for visual ai. arXiv, 2017.

[7] M. Murnane, M. Breitmeyer, F. Ferraro, C. Matuszek, and D. Engel. Learning from human-robot interactions in modeled scenes. In ACM SIGGRAPH 2019 Posters, SIGGRAPH '19. Association for Computing Machinery, New York, NY, USA, 2019.

[8] M. Murnane, M. Breitmeyer, C. Matuszek, and D. Engel. Virtual reality and photogrammetry for improved reproducibility of human-robot interaction studies. In 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), pp. 1092–1093, 2019. doi: 10.1109/VR.2019.8798186

[9] A. Robertson. Facebook is making oculus' worst feature unavoidable. http://tiny.cc/oculus-privacy. Accessed: 22/1/2021.

[10] M. Shridhar, J. Thomason, D. Gordon, Y. Bisk, W. Han, R. Mottaghi, L. Zettlemoyer, and D. Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[11] Stanford Artificial Intelligence Laboratory et al. Robotic operating system.