Learning from Human-Robot Interactions in Modeled Scenes

Mark Murnane
Max Breitmeyer
mark25@umbc.edu
mb17@umbc.edu
University of Maryland, Baltimore County
Baltimore, MD

Francis Ferraro
Cynthia Matuszek
Don Engel
ferraro@umbc.edu
cmat@umbc.edu
donengel@umbc.edu
University of Maryland, Baltimore County
Baltimore, MD

Figure 1: Using a monitor wall to talk to a virtual robot, modeled using a combination of Unity, ROS, and Gazebo.

ABSTRACT

There is increasing interest in using robots in simulation to understand and improve human-robot interaction (HRI). At the same time, the use of simulated settings to gather training data promises to help address a major data bottleneck in allowing robots to take advantage of powerful machine learning approaches. In this paper, we describe a prototype system that combines the robot operating system (ROS), the simulator Gazebo, and the Unity game engine to create human-robot interaction scenarios. A person can engage with the scenario using a monitor wall, allowing simultaneous collection of realistic sensor data and traces of human actions.

CCS CONCEPTS

• **Computer systems organization** → **External interfaces for robotics**; • **Computing methodologies** → *Physical simulation.*

SIGGRAPH '19 Posters, July 28 - August 01, 2019, Los Angeles, CA, USA

ACM ISBN 978-1-4503-6314-3/19/07…$15.00
https://doi.org/10.1145/3306214.3338546

KEYWORDS

Robotics, Virtual Reality, Machine Learning

ACM Reference Format:
Mark Murnane, Max Breitmeyer, Francis Ferraro, Cynthia Matuszek, and Don Engel. 2019. Learning from Human-Robot Interactions in Modeled Scenes. In *SIGGRAPH '19: ACM Special Interest Group on Computer Graphics and Interactive Techniques, July 2019, Los Angeles, CA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3306214.3338546

1 INTRODUCTION

Gathering enough data to perform large-scale machine learning is a significant bottleneck in robotics. Robots are complex and sometimes fragile systems that are not easy to move among a variety of settings, and can only collect sensor data at the time of interaction. For tasks in the human-robot interaction (HRI) space, the problem is multiplied by the complexity of involving human participants in data-gathering tasks. As well as being time-consuming, it is difficult to guarantee that data is collected consistently across participants [Murnane et al. 2019] and settings. As a direct result, there has been increasing interest in using robots in simulation for tasks such as teleoperation [Whitney et al. 2018] and robot control.

We describe a prototype of a system designed to gather robot sensor data, human actions, and speech in a virtual setting, allowing for a variety of robots, environments, and human activities. In our

system, complex world generation and display is provided by the Unity game engine, robot control and perception is provided by the Robot Operating System (ROS) and simulation environment Gazebo, and interaction is via a curved display wall that can track head and optionally controller movements.[1] We envision this system as an aid to grounded language acquisition systems—machine learning systems that learn the situated meaning of words from a nonspecialist describing tasks and objects—by facilitating language-based interactions between users and simulated robots.

2 APPROACH

Though current robotic simulators provide excellent tools for accurately modeling rigid jointed robots, they lack many of the tools used by the entertainment industry for modeling humans and human worlds [Whitney et al. 2018]. In choosing our approach, we sought to combine the strengths of two existing engines in order to create a simulation that models both the human's experience of the robot and the robot's experience of the human with sufficient verisimilitude to build a corpus of training data.

Based on our evaluation of the available engines, we created a parallel system that can model a scenario in both the Unity game engine and the Gazebo simulation simultaneously. Using the ROS Remote Procedure Call (RPC) Application Programming Interface (API) as a common abstraction layer allows portions of the robot's sensor inputs to be rendered in both engines. This hybrid approach allows each sensor to be modeled by the engine that best fits the test scenario, and allows the VR display to be rendered in the more powerful Unity engine [Codd-Downey et al. 2014].

In our motivating demonstration, we show a REEM-C robot interacting with a human in a hospital setting. This bipedal robot requires the use of numerous joint encoders, an Inertial Measurement Unit (IMU), and load cells in order to walk and stand upright. Gazebo is able to model each of these sensors, and has a plugin interface supporting multiple noise and error models for these sensors. However, this robot also provides a stereo camera that is more difficult to accurately model in Gazebo. When interacting with a human, it is important for a robot to be able to perceive gestures and body language made by the human. In order to generate the video input received by the simulated robot, we captured a model of a human subject using the UMBC Photogrammetry Facility, then rigged the model in Maya to be imported into Unity, where we animate the skeleton of the model in real-time using a variety of capture techniques. In virtual reality, there is existing work modeling full body motion from the three point tracking available through most virtual reality platforms [DeepMotion 2018].

The UMBC Pi[2] Facility[2] provides an immersive curved-wall display as well as head and hand tracking via an Advanced Research Tracking system and hand-held controllers. For tests requiring more precise full-body tracking in the future, we plan to add support for Vicon capture data from the UMBC Interactive Systems Research Center User Studies Lab, optionally coupled with a head-mounted VR system.

[1]The system can also render a full VR environment using a headset, if desired, as different users tend to prefer one or the other [Philpot et al. 2017].
[2]http://pisquared.umbc.edu

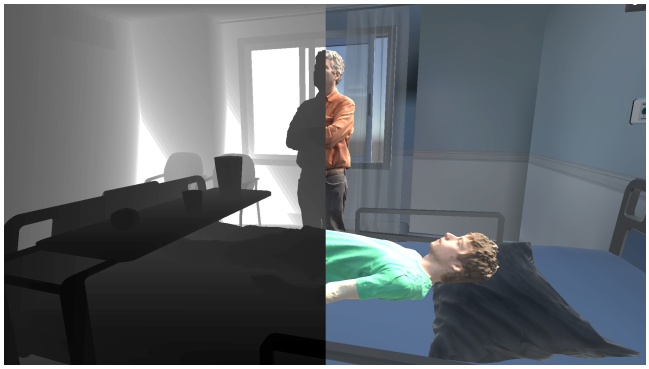

Figure 2: A split view of the robot's perspective, with an RGB image on the right and depth sensor data on the left.

3 LANGUAGE-BASED HRI IN SIMULATION

Our system allows for the creation of collections of rich human-robot interactions. By capturing and recreating the human model rather than directly capturing a performance using traditional video cameras or an RGB-D camera (such as the Kinect), we are able to create a corpus of data that allows for the testing and development of new sensor arrays without requiring repeated performance from human participants. If a particular sensor placement or layout fails to reliably capture a gesture, additional iterations of a robot may be tested against the entire corpus of data automatically.

Human language technology has significantly advanced in recent years; though not perfect, automatic speech recognition and transcription has become generally available for downstream applications. With the inclusion of a microphone among the monitor wall's sensors and off-the-shelf speech recognition tools, participants will be able to directly communicate with the simulated robot and provide training data for grounded language systems. We view our prototype system as a critical milestone toward this integration.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grants No. 1531491 and 1428204. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Support was also provided for this work by the Next Century Corporation.

REFERENCES

Robert Codd-Downey, P Mojiri Forooshani, Andrew Speers, Hui Wang, and Michael Jenkin. 2014. From ROS to Unity: Leveraging robot and virtual environment middleware for immersive teleoperation. In *IEEE ICIA*.

DeepMotion. 2018. How To Make 3 Point Tracked Full-Body Avatars in VR. http://tiny.cc/3pt-deepmotion

Mark Murnane, Max Breitmeyer, Cynthia Matuszek, and Don Engel. 2019. Virtual Reality and Photogrammetry for Improved Reproducibility of Human-Robot Interaction Studies. In *IEEEVR*. IEEE Press, Osaka, Japan.

Adam Philpot, Maxine Glancy, Peter J Passmore, Andrew Wood, and Bob Fields. 2017. User Experience of Panoramic Video in CAVE-like and Head Mounted dDisplay Viewing Conditions. In *ACM TVX*. ACM, Hilversum, The Netherlands, 65–75.

David Whitney, Eric Rosen, Daniel Ullman, Elizabeth Phillips, and Stefanie Tellex. 2018. ROS Reality: A Virtual Reality Framework Using Consumer-Grade Hardware for ROS-Enabled Robots. In *IROS*. IEEE, 1–9.