Virtual Reality and Photogrammetry for Improved Reproducibility of Human-Robot Interaction Studies

Mark Murnane*
University of Maryland, Baltimore County

Max Breitmeyer†
University of Maryland, Baltimore County

Cynthia Matuszek‡
University of Maryland, Baltimore County

Don Engel§
University of Maryland, Baltimore County

ABSTRACT

Collecting data in robotics, especially human-robot interactions, traditionally requires a physical robot in a prepared environment, which presents substantial scalability challenges. First, robots provide many possible points of system failure, while the availability of human participants is limited. Second, for tasks such as language learning, it is important to create environments which provide interesting, varied use cases. Traditionally, this requires prepared physical spaces for each scenario being studied. Finally, the expense associated with acquiring robots and preparing spaces places serious limitations on the reproducible quality of experiments. We therefore propose a novel mechanism for using virtual reality to simulate robotic sensor data in a series of prepared scenarios. This allows for a reproducible data set which other labs can recreate using commodity VR hardware. The authors demonstrate the effectiveness of this approach with an implementation that includes a simulated physical context, a reconstruction of a human actor, and a reconstruction of a robot. This evaluation shows that even a simple "sandbox" environment allows us to simulate robot sensor data, as well as the movement (e.g. view-port) and speech of humans interacting with the robot in a prescribed scenario.

Keywords: Virtual Reality, VR, Photogrammetry, Human-Robot Interaction, Virtual Presence

Index Terms: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, Augmented, and Virtual Realities I.2.9 [Artificial Intelligence]: Robotics—Operator Interfaces I.2.9 [Artificial Intelligence]: Robotics—Sensors

1 INTRODUCTION

A major topic of research in the field of human robot interaction has been grounded language acquisition, which is the process of learning language while being informed by the surrounding environment [1–3,9]. This methodology resolves a number of classic linguistic issues such as syntactic ambiguity [4, 10] by providing additional sensor data and an environment that correlates with the presented speech. However, grounded language acquisition presents new challenges to researchers in the form of logistical issues and the addition of a number of new variables to experiments.

By tying linguistic experiments to physical robots, real environments, and humans that populate them, researchers are able to capture rich data-sets. However, these data-sets may be contaminated by mechanical failures, difficult or impossible replication, and greatly increased cost of experimentation. The authors sought to find an effective middle ground between isolated language acquisition and grounded acquisition that might support fully repeatable experiments with sufficient isolation, while not excluding the necessary environment and context needed for effective learning.

In order to achieve such a middle ground the authors have created a VR simulation that includes elements of context from a sample scenario, a reconstruction of a human actor, and a reconstruction of a robot. While pure simulation has been a mainstay in robotics since the fields' inception [5–7], the main contribution of the author's work is that by using VR the authors are able to tie the simulation to a real-world interaction in real-time, combining real sensor data and human interaction with the simulated data both from the robot's perspective and that of one or more human participants. This is distinct from the most similar prior work, which lacked a human presence in the scene [8].

The authors' software incorporates virtual sensors (the virtual world as seen by the virtual robot, e.g. point clouds and audio recordings), as well as virtual actuators and mechanics in a scene that may be experienced by a human subject in real-time. The opposite of this is also the true, in that a real robot may use the authors' system to encounter a simulated or recorded human actor.

This system was designed and implemented with direct input from researchers in the field of human-centered robotics, which supported the creation of both a test methodology and an apparatus grounded in real-world challenges specific to human-robot interactions (HRI). The representative nature of the authors' team ensures this work is directly responsive to the needs of HRI researchers in carrying out their work.

The authors' collaborators identified several key features that such an instrument and experimental design must support:

- Such an instrument must simulate sensors that match real-world effects
- The system must allow the simulated results to be directly compared to measured data from the real-world analog
- The system must enable researchers to easily test a large variety of scenarios and contexts with minimal effort

2 DESIGN

Researchers have need for a large corpus of data sampled from individual human-robot interactions that characterize the wide variety of communications a human may attempt to convey to a robot. Currently, this data requires bringing individual participants into a lab in order to expose them to a physical robot and demonstrate their communication in the given environment. This requires a physical artifact, correct function of the robot, and the physical presence of the human. By moving this interaction to a virtual environment and simulating the robot the authors allow a much wider variety of participants to engage in this research as long as such a simulation is able to achieve the same results that would be attained through a physical encounter with a robot.

Current virtual reality technology can present a nearly convincing environment to the human participant while also capturing the portion of their performance that the physical robot would itself capture. The telepresence robot this experiment was designed around is able to capture video, audio, a depth image, and odometric data about its

*mark25@umbc.edu
†mb17@umbc.edu
‡cmat@umbc.edu
§donengel@umbc.edu

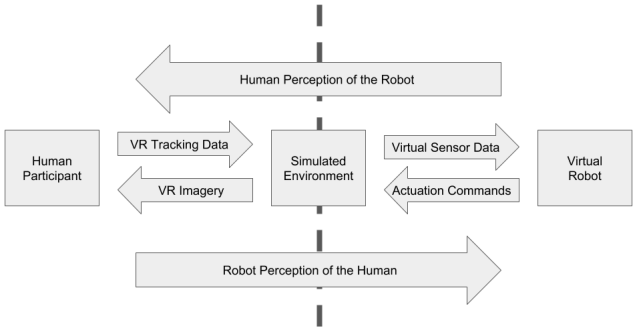

Figure 1: Information Flow In The Authors' System

environment. The authors combine a 3D scan of the environment and the participant within it to supplement the data already acquired by the virtual reality system in order to fully span the sensor inputs needed by the robot.

By combining this acquired sensor data with a simulation of the robot's internal logic, the authors are able to generate the actuator commands that would be given by the physical robot in its model. By mapping these actuation commands to a set of virtual actuators in real-time, the authors are then able to close the loop from the human perspective by providing a convincing simulation of how the robot would react to the human given their actions and communications.

The verisimilitude of the authors' simulation from the robot's perspective is dependent on the model of both the human participant and the accuracy of the simulated sensors. Therefore the authors needed a high-detail, high-accuracy 3D model of the user, as well as tracking data of their motion, and the audio that they produce. The authors rely on a 3D scan of the user combined with the VR tracking data and the microphones in a VR headset to acquire data about the participant, then model the sensor noise that would occur due to the structure and implementation of the robot itself.

3 System Development

The various sensors available on the physical robot were duplicated within the game engine using software analogs. The authors implemented a virtual kinect sensor, cameras, microphones, collision sensors, distance sensors, and wheel encoders using the game engine.

For the participant, the authors used the HTC Vive, which includes two hand trackers, along with two Vive Trackers and the headset itself to estimate the position and orientation of the participant in the VR system, as derived from the Inertial Measurement Unit (IMU) and Lighthouse information from the Vive API.

To optimize for realism - both in terms of human behavior and in terms of how closely the virtual sensors on the virtual robot will emulate their real-world analogs - it is necessary to have a physically realistic scene, including a physically realistic human avatar. To that end, the authors have employed the UMBC Photogrammetry Facility. The rig used for this project is optimized for scanning human participants, and therefore utilizes the simultaneous triggering of many cameras to acquire a static view of potentially moving objects such as live animals or human participants. The output of this process is a polygon mesh on the order of 20 million triangles, and an 8k texture.

The resulting decimated model contains approximated 500,000 triangles and uses a 4k texture. This model was rigged for animation with a model allowing inverse kinematic control of the limbs and head. While the face of the model has been captured at a high resolution, at this point no effort has been made to animate the face separately from the head. The inverse kinematic controls of the model were tied to five real world control points: The center-point of the Vive headset, the left and right hand-held controllers, and two Vive Trackers attached to the participant's feet. In combination, these provide plausible poses for the simulated model of the participant that closely mimic the real-world pose for a large variety of conditions. In the future, the authors intend to integrate a complete skeletal tracking system to increase the accuracy of the model and to investigate facial tracking and modelling.

The authors scanned the BeamPro robot using a depth camera, producing a polygon mesh and an RGB texture. As the robot lacks any articulated limbs or separable components, this model required only a rigid body rig to fully express the capabilities of the physical counterpart. In addition to the motion of the robot's frame, the authors sought to simulate the appearance of the robots video display. This was accomplished by applying a video texture to the surface of the proxy that could be fed by either live footage of a real world camera, or recorded files from previous sessions.

4 Future Work

With this system now available, the natural next step is for the authors to begin using this system to begin capturing sample interactions, and to extend the system to support additional sensors and robots. One significant feature missing from the current implementation is integration with ROS, the Robot Operating System. ROS provides additional simulation capabilities and has a large library of existing robot models. By combining this system with ROS many researchers will immediately be able to test their existing human interactive robots against the authors' library of scenarios.

References

[1] D. Arumugam, S. Karamcheti, N. Gopalan, L. L. Wong, and S. Tellex. *Accurately and efficiently interpreting human-robot instructions of varying granularities*. Science and Systems, In Proceedings of Robotics, 2017.

[2] 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.

[3] J. Y. Chai, Q. Gao, L. She, S. Yang, S. Saba-Sadiya, and G. Xu. *Language to action: Towards interactive task learning with physical agents*. In IJCAI, pages 29, 2018.

[4] D. Chen and R. Mooney. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI-2011)*. pages 859865, 2011.

[5] S. Chernova, N. DePalma, E. Morant, and C. Breazeal. *Crowdsourcing human-robot interaction: Application from virtual to physical worlds*. In RO-MAN, 2011 IEEE, pages 2126. IEEE, 2011.

[6] M. Forbes, M. J.-y. Chung, M. Cakmak, and R. P. Rao. Robot programming by demonstration with crowdsourced action fixes. In *Second AAAI Conference on Human Computation and Crowd-sourcing*, 2014.

[7] M. Forbes, R. P. Rao, L. Zettlemoyer, and M. Cakmak. *Robot programming by demonstration with situated spatial language understanding*. In Robotics and Automation (ICRA), 2015 IEEE International Conference on, pages 20142020. IEEE, 2015.

[8] 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.

[9] C. Matuszek. *Grounded language learning: Where robotics and nlp meet*. In IJCAI, pages 56875691, 2018.

[10] J. Thomason, A. Padmakumar, J. Sinapov, J. Hart, P. Stone, and R. J. Mooney. Opportunistic active learning for grounding natural language descriptions. In *Conference on Robot Learning*. pages 6776, 2017.