The Thesis Committee for Steve Han
certifies that this is the approved version of the following thesis:

Demonstrations for Robots in Immersive Virtual Reality

APPROVED BY

SUPERVISING COMMITTEE:

Yuke Zhu, Supervisor

Etienne Vouga

Demonstrations for Robots in Immersive Virtual Reality

by

Steve Han

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2023

Acknowledgments

Thanks to Mingyo Seo and Yuke Zhu for the project idea. Thanks to Mingyo Seo for creating the simulation, training the baselines, and adapting the whole body control code. Thanks to Kyutae Sim for helping with the data collection. Thanks to Carlos Isaac Gonzalez and Seung Hyeon Bang for help with the real robot experiments.

Demonstrations for Robots in Immersive Virtual Reality

Steve Han, MSCompSci
The University of Texas at Austin, 2023

Supervisor: Yuke Zhu

Our world is designed by humans, for humans. This makes humanoid robots the perfect general-purpose platform to automate repetitive or dangerous tasks done by humans. However, due to the complexity of humanoid robots and the shortage of demonstration data, research in robot learning for humanoids is scarce. To address these challenges, I present a VR interface named DRIVR (Demonstrations for Robots in Immersive Virtual Reality) for collecting human demonstrations for humanoid robots in both simulation and reality. The demonstrations are then used to train an imitation learning algorithm that uses an underlying controller to abstract away the complexity of whole-body control. I further propose that by embedding this data collection mechanism in VR video games, we can amass a large-scale dataset of high quality human demonstrations that can drive the development of future autonomous humanoids. To illustrate the feasibility of this idea, I collect a small dataset on toy tasks in simulation and a real robot using the VR interface. I then show that the trained policy can be deployed with a reasonable success rate.

Table of Contents

Acknowledgments **iii**

Abstract **iv**

List of Tables **vii**

List of Figures **viii**

Chapter 1. Introduction **1**

1.1 Humanoid Robots 4

1.2 VR Teleoperation Interface 5

1.3 Scaling up Demonstration Dataset for Humanoids 6

Chapter 2. Background **9**

2.1 Imitation Learning 9

2.1.1 RNN 9

2.2 Whole-body Control 9

2.3 Related Work 9

2.3.1 VR for Imitation Learning 9

2.3.2 Humanoid Learning Locomotion 10

2.3.3 Humanoid Teleoperation 10

2.3.4 Demonstration Interface 10

2.3.5 Videos 12

Chapter 3. VR Interface **14**

Chapter 4. Future Works **15**

Appendices **16**

Appendix A. Lerma's Appendix 17

Appendix B. My Appendix #2 18

B.1 The First Section 18

B.2 The Second Section 18

B.2.1 The First Subsection of the Second Section 18

B.2.2 The Second Subsection of the Second Section 18

B.2.2.1 The First Subsubsection of the Second Subsection of the Second Section 18

B.2.2.2 The Second Subsubsection of the Second Subsection of the Second Section 19

Appendix C. My Appendix #3 20

C.1 The First Section 20

C.2 The Second Section 20

Bibliography 21

List of Tables

List of Figures

1.1 The proposed interface can be used to teleoperate a simulation environment (left) and a real robot (right). 2

1.2 An overview of the hierarchical learning framework used in this work. Diagram credit to Mingyo. First, we get the desired hand trajectories and walking commands from the demonstrator. Second, we send these commands to the whole-body controller to produce joint-torque actions of the robot. Third, a behavior cloning policy is trained to imitate the human demonstrations. 3

Chapter 1

Introduction

Humanoid Robotics, Imitation Learning, and Virtual Reality (VR) are some of the most exciting and rapidly developing fields in technology today. It is perhaps no surprise that the marriage of these three technologies can have far-reaching impacts. Using VR, humans can control robots in an immersive simulation, or they can remotely teleoperate robots to perform dangerous tasks. The trajectories from the human demonstrations can then be used to train Imitation Learning policies, which can enable robots to perform the tasks autonomously. While deep Reinforcement Learning methods have been successful in teaching robot to learn from repeated interactions with an environment given a reward function, the sample inefficiency, need for online interactions, and reward engineering efforts make them difficult to implement in a humanoid platform. On the other hand, imitation learning skips over the requirement to create reward functions to shape a behavior and instead shows the robot directly what to do.

However, due to the complexity of humanoid robots, the lack of large-scale data for training, and the difficulty of creating an intuitive interface for humans to control robots, there has been no research on teaching humanoid

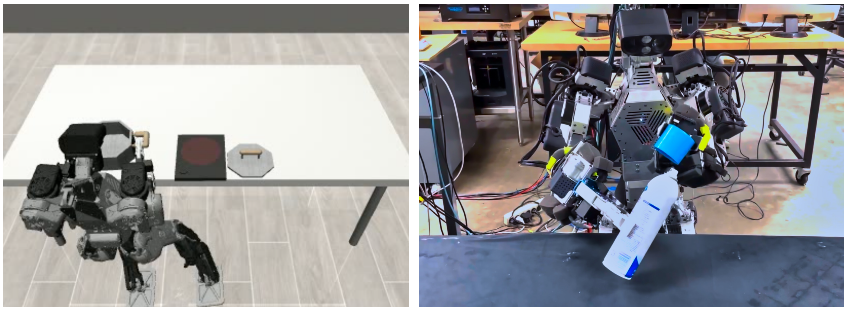

Figure 1.1: The proposed interface can be used to teleoperate a simulation environment (left) and a real robot (right).

robots to perform locomotion and bi-manipulation tasks, as far as I'm aware. There has been work to teach fixed-base robot arms through VR demonstrations [16], but humanoid robots are significantly more difficult to train due to the need to consider robot dynamics and ground reaction forces. Also, although there is a large body of work on teleoperating humanoids using VR and motion capture devices, I have not found any attempt to use the teleoperation data to train a neural network policy to control the humanoid autonomously. Furthermore, these "telepresence" systems are usually very complicated and require expensive hardware, which makes them unsuitable for scaling up data collection in a distributed manner.

In this thesis, I present a simple VR interface that uses the commonplace Oculus Quest 2 headset. I believe that compared to other humanoid teleoperation systems, the software architecture used in this project is uniquely positioned to be both adaptable to video games and accessible to the public. Although the interface is simple, it is nonetheless powerful enough to teach a humanoid robot to perform some simple tasks. In simulation, I hypothesize

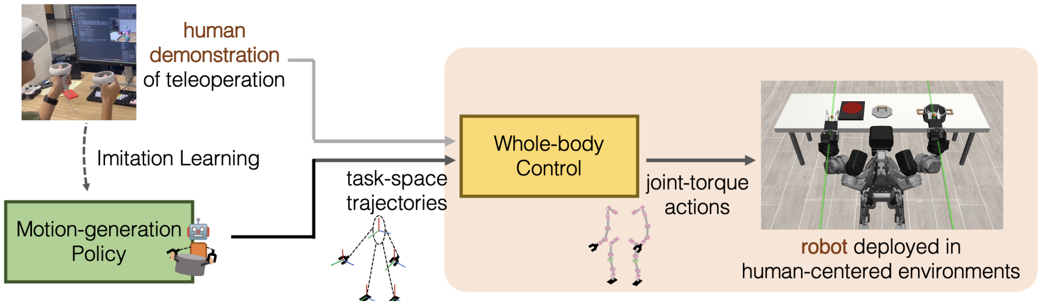

Figure 1.2: An overview of the hierarchical learning framework used in this work. Diagram credit to Mingyo. First, we get the desired hand trajectories and walking commands from the demonstrator. Second, we send these commands to the whole-body controller to produce joint-torque actions of the robot. Third, a behavior cloning policy is trained to imitate the human demonstrations.

that this setup can be incorporated in VR video games to massively scale up data collection. VR games contain rich interactions with the virtual world, have built-in rewards to label successes and failures, and integrate with powerful physical simulation engines. As a result, they are perfect for collecting human demonstrations to potentially teach humanoids to perform the same tasks in the real world. For example, Cook-Out is a VR game that requires players to cook in a kitchen, which is a valuable skill for humanoids to learn. In addition, on the real-robot side, we could potentially distribute the data collection by letting users with VR headsets control the robot remotely. This is similar to the RoboTurk crowd-sourcing platform for fixed-base arm robots [9]. I hope that by open-sourcing the teleoperation codebase (later this summer), research in humanoid robot learning can be made more accessible.

To show that we can use the collected demonstrations to train a humanoid to perform simple tasks, we present a hierarchical approach that learns a motion policy from teleoperation demonstrations with an underlying controller. The policy outputs the desired poses of the hands, while the whole-body controller takes care of balancing the robot and tracking the desired trajectory. An overview of this design is presented in Figure 1.2. First, we take advantage of the similar kinematic structures between the robot and demonstrators by collecting demonstrations using immersive VR. Second, we send the desired hand trajectories and walking commands to the whole-body controller to produce joint-torque actions of the robot. The controller prioritizes the robot's stability while tracking the hand and feet trajectories. The human demonstrator adapts to the tracking error by observing the effects of their actions in the VR headset. Third, a behavior cloning policy is trained to imitate the human demonstrations. The policy needs to learn the state-action distribution of whole-body control behaviors and adapt to it in a closed-loop manner similar to the human. During deployment, the generated trajectories are passed to whole-body control just like during demonstration.

In the below sections, I go into more details about the motivations for this project.

1.1 Humanoid Robots

Humanoid robots have gained a lot of attention in recent years. After Boston Dynamic's Atlas made headlines by jumping and dancing with human-like dexterity [5], Tesla also entered the market by developing the cheap and mass-producible humanoid named Optimus. If the cost of the robot could be kept below $20,000 like Elon Musk promised [14], we would be entering a world where general purpose humanoids could replace humans for unsafe and repetitive tasks. Many startups are also getting a lot of funding recently to pursue humanoid robots. Apptronik is an Austin-based startup that aims to create humanoids that can work alongside people. One of their prototypes, DRACO 3, is the platform used for this thesis.

This interest in humanoids is justified by their versatility and social capabilities. They can be used as personal assistants, as companions for the elderly, as workers in factories, and as first responders in disaster zones. The morphology of humanoids enables them to easily adapt to the human-centered world that we live in. Every tool, every environment, and every task in our society is designed for the human form. It wouldn't make sense to redesign power tools or get rid of stairs for the convenience of robots, so creating robots that can take advantage of the existing infrastructure made for humans is extremely valuable. Also, as humans, it's easier to provide demonstrations to a robot that has a similar form as us, rather than having to train our brain to adapt to the morphology of the robot.

1.2 VR Teleoperation Interface

In order for imitation learning to succeed on robots, high-quality demonstrations are essential. Using a VR interface is desirable since it closes the observation and embodiment gap. Instead of looking at the robot externally, the human will have the same perception as the robot. Instead of having to map the human body's joints to the robot's joints, the human is directly controlling the robot's joints. Because of the human brain's impressive capability to adapt, the demonstrator can quickly learn to treat the arms of the robot as their own arms, and perform tasks intuitively as they do in their own body. Indeed, there are various experiments that shows that humans can start to take ownership of their virtual body, even if the body is very different from their own, if appropriate multisensory correlations are provided [8] [13].

In addition, since humanoid robots have to maintain balance while following hand trajectories, the whole-body controller may fail to track the hand trajectory perfectly. So, in order to move the robot hand to a desired position, one needs to constantly observe the effects of their actions before deciding where to move next. In experiments, we notice that humans are great at adapting to the whole-body control's tracking error in this closed-loop manner. By using the VR interface, we are essentially borrowing the human brain's power to solve the embodiment mismatch issue and perform closed-loop actions.

1.3 Scaling up Demonstration Dataset for Humanoids

Recently, we have seen the successes of training deep learning algorithms on mind-blowingly huge datasets. For example, GPT-4 [10] is believed to be trained on most texts on the internet, which enables its scalable transformer architecture to produce human-like texts. The ability of the transformer to scale can also impact the robot learning field, since many transformer architectures for robotics has been proposed with impressive results [18] [7]. Notably, researchers at Google demonstrated that their transformer architecture is able to absorb a large dataset of multi-task demonstrations from different robots and even simulations, and the resulting model is able to generalize to unseen tasks, environments, and objects in a zero-shot manner [6]. Importantly, they demonstrated that by incorporating simulation data with the real-robot data, not only was the real-robot policy not degraded, but the generalization performance also improved significantly on objects only seen in simulation. This shows that even if the simulation isn't very realistic, the model can still absorb useful information that helps with generalization. The researchers stated that their dataset "consists of over 130k episodes, which contain over 700 tasks, and was collected with a fleet of 13 robots over 17 months." The scale of this dataset is very impressive, but the collection procedure seems very costly, and it is still small in scale compared to the billions of images used to train today's image generation models. In addition, the dataset isn't publicly available, and they are collected on single-arm wheeled robots instead of humanoids.

So, how can we scale up humanoid demonstration collection in a cheaper manner than investing manpower to collect demonstrations on the real robot? One proposal is to learn from the plethora of YouTube videos online. However, due to the difficulty of inferring 3D poses from video, the lack of proprioceptive view from the eyes of the human, and the ignorance of the person's internal state such as their joint positions, the wealth of information that online videos possess still remain out of reach for practical applications of robot learning.

Between learning from YouTube videos and collecting demonstrations on the real robot, taking advantage of the rich manipulation data from VR applications might just be practical enough to scale up humanoid data to satiate the appetite of today's deep learning methods. In VR, the human sees exactly what the robot sees, the hand poses are measured precisely by the headset and controllers, and the simulated robot provides full access to its internal states. As realistic simulation games such as Microsoft Flight Simulator rise up in popularity, we are presented with a valuable opportunity to collect high-quality human hand trajectories at scale for robot-learning research. This is not unreasonable since Meta is most likely already collecting information about the users' surroundings and actions for targeted ads [15].

Chapter 2

Background

Now that the motivation of the project has been discussed, I will introduce the formulations used in this work. Then, existing literature that relates to our goal will be surveyed.

2.1 Imitation Learning

2.1.1 RNN

2.2 Whole-body Control

2.3 Related Work

2.3.1 VR for Imitation Learning

The work that most closely relates to ours is [16], which also uses a consumer VR headset to teleoperate a robot. Their robot has a fixed base, so they can directly use their robot's built-in Jacobian-transpose based controller. Instead of streaming stereoscopic images to the headset, they use Unity to render the colorized point cloud from the robot's 3D camera to allow the user to look around freely in the VR world. They argue that since the robot's head has low degrees of freedom and moves slowly, controlling its movement using the headset's movement will cause motion sickness. Indeed, we are faced with the same problem. However, instead of using a point cloud that could look strange for the demonstrator, we simply disallowed movements of the head. This is sufficient for our tasks, and it does not seem to impact immersion. For learning, they also use a Behavior Cloning algorithm. However, they are only controlling one arm of the robot, whereas we are controlling both. They are also using the depth image as an input, but we only provide stereoscopic images. Finally, we are using an RNN to preserve information from previous time steps, while they only provide the 5 most recent end effector poses and no image history.

[2]

2.3.2 Humanoid Learning Locomotion

[]

https://ashish-kmr.github.io/a-rma/

https://humanoid-transformer.github.io

2.3.3 Humanoid Teleoperation

There is a recent surge in interest to remotely teleoperate robots. XPRIZE hosted the AVATAR competition, in which teams

2.3.4 Demonstration Interface

There are multiple ways to collect demonstrations for imitation learning in existing literature. First, a teleoperation input device with 6 degrees of freedom such as the SpaceMouse can be used to control the position and orientation of the robot's end effector [19]. However, it can be unintuitive for people to translate a 3D motion into the push and turn of a button, especially if there is a need to control 2 arms at once. In addition, since SpaceMouse controls the velocity instead of position, it requires training to perform actions involving precision. Second, humans can directly hold the robot to move it in a desired trajectory by applying force [1] [11]. This is called kinesthetic teaching, but it requires the human to come into the frame to control the robot, which becomes a problem when the policy is trained on vision data. To avoid this, we can also build a replica of the robot and move the replica manually, while the main robot follows its trajectory. For example, [17] used a low-cost replica of their bi-manipulation robot to collect demonstrations for fine manipulation tasks. To do so, a human demonstrator pushes the end-effectors of the replica robot to backdrive its joints, and the resulting joint positions are issued as commands for the actual robot to follow. While this method achieved impressive results for the fixed-base robot, they cannot handle the floating-base dynamics of humanoid platforms, which requires torque control to account for the dynamics of the robot.

Since we can manipulate a replica of the robot to create demonstrations, why can't we use our own body as this replica? After all, humanoids are design to mimic the morphology of humans.

Indeed, we can directly record the kinematics of human motions [3]. Using either a camera or a motion capture system, we can measure the angular displacement of the joints precisely, and then we can map the values to the robot's joints. However, robots can have different mass distributions and degrees of freedom than humans. So, the actions that humans perform may be impossible for robots or cause them to lose balance.

2.3.5 Videos

https://arxiv.org/abs/2104.07810 There are proposals to massively scale up human demonstration data using YouTube videos. For example, [4] shows that by watching YouTube videos of house tours, an off-policy Q-learning algorithm can learn the semantic cues in a human environment to improve navigation efficiency. To make it possible to learn dexterous manipulation skills from YouTube videos, [12] trained a neural network to retarget human finger poses from a video to a robotic hand.

Given the rapid development of humanoid hardware, cracking the code of human-like locomotion is the next big challenge. The most common approach is to use a combination of inverse kinematics and inverse dynamics to generate the joint torques that will move the robot to the desired pose. However, this approach is limited by the complexity of the inverse kinematics and inverse dynamics algorithms. Inverse kinematics is a non-convex optimization problem, and the inverse dynamics problem is also non-convex due to the non-linear constraints. This makes it difficult to find the optimal solution, and the solution is often suboptimal.

Human-to-Robot Imitation in the Wild

VideoDex: Learning Dexterity from Internet Videos

Learning to Manipulate Tools by Aligning Simulation to Video Demonstration

Zero-Shot Robot Manipulation from Passive Human Videos]Zero-Shot Robot Manipulation from Passive Human Videos

Learning from Observations Using a Single Video Demonstration and Human Feedback

Chapter 3

VR Interface

In this chapter, I'll go into the implementation details of the VR interface. The requirements for the interface are as follows:

1. It needs to report the 6-DOF poses of left and right hands as well as additional buttons for locomotion and gripper control.
2. Stereoscopic images need to be streamed and displayed on the VR headset to create depth perception for the wearer.
3. It needs to connect to both the simulation codebase written in Python and the real-robot controller written in C++.
4. The latency should be low and the computation speed should be fast so that the wearer can provide demonstrations with ease. The computation resource consumption should be low because the simulation and whole-body controller are resource-intensive.

Initially, I created a

Chapter 4

Future Works

Immersive demonstration with force feedback

https://arxiv.org/pdf/2301.09157.pdf

Appendices

Appendix A

Lerma's Appendix

The source LaTeX file for this document is no longer quoted in its entirety in the output document. A LaTeX file can include its own source by using the command `\verbatiminput{\jobname}`.

Appendix B

My Appendix #2

B.1 The First Section

This is the first section. This is the second appendix.

B.2 The Second Section

This is the second section of the second appendix.

B.2.1 The First Subsection of the Second Section

This is the first subsection of the second section of the second appendix.

B.2.2 The Second Subsection of the Second Section

This is the second subsection of the second section of the second appendix.

B.2.2.1 The First Subsubsection of the Second Subsection of the Second Section

This is the first subsubsection of the second subsection of the second section of the second appendix.

B.2.2.2 The Second Subsubsection of the Second Subsection of the Second Section

This is the second subsubsection of the second subsection of the second section of the second appendix.

Appendix C

My Appendix #3

C.1 The First Section

This is the first section. This is the third appendix.

C.2 The Second Section

This is the second section of the third appendix.

Bibliography

[1] Barış Akgün, Maya Cakmak, Karl Jiang, and Andrea Lockerd Thomaz. Keyframe-based learning from demonstration. *International Journal of Social Robotics*, 4:343–355, 2012.

[2] Sridhar Pandian Arunachalam, Irmak Güzey, Soumith Chintala, and Lerrel Pinto. Holo-dex: Teaching dexterity with immersive mixed reality, 2022.

[3] A. Billard and D. Grollman. Robot learning by demonstration. *Scholarpedia*, 8(12):3824, 2013. revision #138061.

[4] Matthew Chang, Arjun Gupta, and Saurabh Gupta. Semantic visual navigation by watching youtube videos, 2020.

[5] Boston Dynamics, YouTube. Do you love me? `https://www.youtube.com/watch?v=fn3KWM1kuAw`, 2020.

[6] Anthony Brohan et al. Rt-1: Robotics transformer for real-world control at scale, 2022.

[7] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts, 2022.

[8] Konstantina Kilteni, Antonella Maselli, Konrad P. Kording, and Mel Slater. Over my fake body: body ownership illusions for studying the multisensory basis of own-body perception. *Frontiers in Human Neuroscience*, 9, 2015.

[9] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, Silvio Savarese, and Li Fei-Fei. Roboturk: A crowdsourcing platform for robotic skill learning through imitation, 2018.

[10] OpenAI. Gpt-4 technical report, 2023.

[11] John Schulman, Jonathan Ho, Cameron Lee, and P. Abbeel. Learning from demonstrations through the use of non-rigid registration. In *International Symposium of Robotics Research*, 2013.

[12] Aravind Sivakumar, Kenneth Shaw, and Deepak Pathak. Robotic telekinesis: Learning a robotic hand imitator by watching humans on youtube, 2022.

[13] Mel Slater, Daniel Pérez Marcos, Henrik Ehrsson, and Maria Sanchez-Vives. Towards a digital body: the virtual arm illusion. *Frontiers in Human Neuroscience*, 2, 2008.

[14] Tesla, YouTube. Tesla AI day 2022. `https://www.youtube.com/watch?v=fn3KWM1kuAw`, 2022.

[15] XpertVR. Vr data collection: Traditional and contemporary data types. https://xpertvr.ca/vr-data-collection/, 2022.

[16] Tianhao Zhang, Zoe McCarthy, Owen Jow, Dennis Lee, Xi Chen, Ken Goldberg, and Pieter Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation, 2018.

[17] Tony Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023.

[18] Yifeng Zhu, Abhishek Joshi, Peter Stone, and Yuke Zhu. Viola: Imitation learning for vision-based manipulation with object proposal priors, 2023.

[19] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning, 2022.