Thesis Proposal

Eliot Laidlaw

February 2022

1 Introduction

Controlling robots to complete tasks that require dexterous manipulation or acute awareness of their surroundings proves very difficult. Adding a human to the loop can make precise control easier, but the interface by which the human controls the robot is critical to success. Using virtual reality (VR) to teleoperate the robot should theoretically offer the user the most effective visual information by placing them virtually into the robot’s environment. In actuality, it is not trivial to reconstruct the robot’s environment in VR, based on the following necessities for a well-functioning system:

1. The user must have up-to-date, high-fidelity visual information that allows for dexterous manipulation with the robot’s end effectors.
2. The user must have enough visual information to give them the situational awareness required to move the robot around its environment.
3. The system must present data in a way that does not induce discomfort or nausea for the user.

Improving the technologies required for building such a VR teleoperation system will be the focus of my thesis. I plan to explore methods of quickly reconstructing the robot’s 3D environment using sensor data that is streamed in real time.

2 Background

Two inherent problems in creating a convincing VR reconstruction of the robot’s environment are that 1) we’re using primarily 2D sensor data to reconstruct a 3D environment, and 2) when the user moves or rotates their head, we have to render the scene from a viewpoint for which we don’t explicitly have sensor data. Much progress has been made on these problems in the overlapping fields of scene reconstruction and novel view synthesis (NVS).

Much of the recent development in scene reconstruction and NVS has followed the introduction of Neural Radiance Fields (NeRFs) by Mildenhall et al.
in 2020. NeRF reconstructs scenes with impressive, near-photorealistic quality. The downside however, is that optimizing a NeRF for a single scene takes several hours or days to complete, and even rendering a single frame post-optimization takes on the order of several seconds to minutes. Since NeRF was introduced, many advances have been made on both optimization and rendering speed. While these methods still are not able to optimize scenes and render novel views at the framerates required for interactive VR, it may be possible to adapt them for VR applications.

3 Research Direction

My goal is to adapt a NeRF-based method to work for VR teleoperation. The recent work by Müller et al, Instant Neural Graphics Primitives, improves training time for NeRF from several hours to a few seconds for decent quality reconstructions. I hope to adapt this method for VR applications by 1) selectively re-optimizing only the parts of the scene that change from frame to frame, 2) using scene flow to transform already-optimized parts of the scene to their new positions when they move, and 3) using an intermediate 3D representation to speed up rendering to the framerates required for VR.

Current NeRF based methods that reconstruct scenes over time require the whole video sequence at training time. By having all of the data a priori these methods can learn to separate static and dynamic content in the scene, often using a separate network to reconstruct each part. This allows several frames of video to be reconstructed without excessive training parameters, because much of the scene never changes.

This optimization is harder to leverage in an online method, because we cannot determine which parts of the scene are static and which parts are dynamic from the first frame of video when the application starts. When we receive a new frame of video, therefore, we must do as little re-optimization as possible to adjust the scene for what has changed. This will likely involve separating static and dynamic content “on the fly.” If we can segment each image into sections that have changed and not changed from one frame to the next, we only need to devote re-training time to the sections of the scene that change. We also may be able to use inferred 3D scene flow to transform implicit content to its new position before actually re-optimizing the scene, saving valuable compute power.

It is the hope that these improvements will allow for real-time VR reconstruction of scenes.