RIA DOSHI
riadoshi@gmail.com | Monta Vista High School

Enhancing Multisensory Feedback to Improve Virtual Reality Experiences

I have been working in Stanford's Computer Science Department in the IPRL lab under the guidance of Professor Jeannette Bohg since the summer of 2019. Virtual reality and video processing currently lack rich and advanced auditory and visual signals humans expect in real-life interactions. This prevents users from becoming truly immersed in their online experience. The goal of this research is to assemble a multi-sensory dataset to produce sounds dynamically reflecting ways we interact with objects (touch, tap, bump, etc), pair this dataset with a neural-network model trained on optical flow to detect nuanced hand movements, and ultimately improve the virtual reality experience. This research has the potential to help increase student engagement in distance-learning during global health crises such as COVID-19.

The datasets leveraged were based on ASMR (autonomous sensory meridian response) videos, in which creators tap, touch, and interact with a variety of objects to create a satisfying experience for the listener. In many cases, these are high quality videos recorded with objects with a microphone, making it an ideal way to extract vision and sound. I applied artificial intelligence algorithms and techniques on this data set to: a) Detect Hands in Videos using MediaPipe b) Extract Coordinates of Hand Landmarks c) Identify Occluded Hand Frames d) Investigate Several Optical Flow Techniques for Hand-Tracking e) Run MaskRCNN on an Annotated ASMR Dataset.

However, when the hands of the person interacting with the object are occluded (either by the microphone or by other obstruction), extracting visual image is not feasible. This is because while MediaPipe can detect hands, it does not offer the functionality to extract the landmarks (x,y,z coordinates). So, I integrated MediaPipe with a third-party library to capture the coordinates and then coded an algorithm to filter out the occluded frames if >75% of the z coordinates switched signs from one frame to another for a specific object. Additionally, Professor Michael Black's optical flow techniques were successfully used and tested on the ASMR dataset to augment the visual representation of hand motion.

In addition to visual representations, the MaskRCNN model was successfully trained to enhance detection of objects used in ASMR data sets. Ultimately, these visual representations of objects result in improved virtual reality experiences by enhancing the detection of higher quality sounds produced during day-to-day object interaction.