Hand Tracking in VR

Group: Royce Cheng-Yue and Hershed Tilak

Technical Approach

We are planning to approach this problem in three steps:

1) Finger tip and joint detection using RGB-D camera
2) Rendering 2D points into corresponding 3D points in the VR 3D world
3) Use rendered hands to interact with the VR environment in a game

Finger Tip and Joint Detection

There are many hand datasets but most of them are not egocentric and do not have joint annotations. We plan to use General-HANDS [1], which contains 23,640 frames containing hands and their corresponding joint annotations. We plan to train a baseline random forest regressor and eventually a CNN, such as Faster R-CNN [2] or YOLO [3] (for their latencies), to determine the relevant bounding boxes for the joints. We could also look into using LSTMs or a Kalman filter to incorporate temporal information.

At test time, the input to the model will be images of RGB-D values, streamed from an RGB-D camera (such as the Kinect), and the output will be the fingertip and finger joint bounding boxes.

Hand Rendering

Once we are able to detect the relevant fingertips, we would need to reconstruct the corresponding hand in the virtual world. Since we know the depth of each pixel in the 2D image, we could extrapolate the point into 3D space. Once we construct the joints in 3D space as a skeleton for the human hand, we could render material on top of this skeleton in Unity.

VR Interactions

After we render hands in the virtual world, we would need to figure out a way to facilitate interactions. This would involve understanding the depth of the hands in the VR world and detecting actions on a desired object. Some possible interactions could involve grasping objects or drawing an image. This could be achieved using object collisions and the physics engine in Unity.

In the end, we plan to make an interactive demo of the hand tracking feature in the form of a game, such as a VR Fruit Ninja.

Potential Issues

The following are some potential issues with this project.

1) Occlusions of the fingertips and finger joints. If some of these joints are occluded, we would not be able to render the joints that were occluded. For now, we assume all of the joints are visible. If not, we would need to either estimate the positions of the joints using temporal data (maybe via an LSTM).
2) Camera rotations and translations. If the HMD rotates and translates but the hand does not move, we would need to make sure that the rendered hand remains still. This could be difficult due to the noisy data from depth cameras.
3) We are currently waiting for a response to access the General-HANDS dataset (annotations are retrieved if access is approved via email). If we are unable to retrieve this dataset, we must look into other egocentric hand datasets. A possible but time consuming approach would be to collect our own dataset using colored gloves.

References

[1] Wetzler, A., Slossberg, R. & Kimmel, R. Rule Of Thumb: Deep derotation for improved fingertip detection.

[2] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497, 2015.
[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. arXiv preprint arXiv: 1506.02640, 2015.