Crowd-sourced Data Collection Pipeline for Human Attention Tracking in Virtual Environment

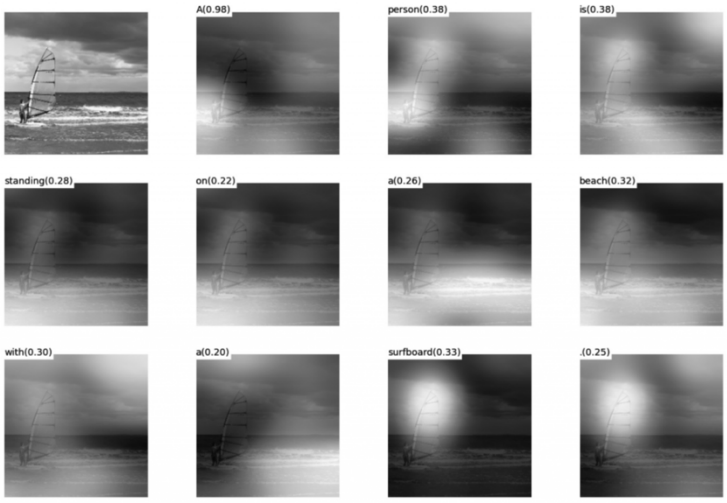
EE 267 Project Proposal

Zhiyang He

In recent years' research in computer vision and robotics, attention-based image understanding has become a highly active field. As described by Ilya Sutskever, the research director of OpenAI in an interview last year, "Attention Mechanisms are one of the most exiciting advancements". Compared to traditional vision methods that take entire images as input, attention based learning offers several advantages:

1. Better adaptation to human visual system: human eyes do not accept every pixel simultaneously as equally accurate. Our foveated vision (implemented in HW3) chooses a sub-area to focus, while our brain chooses the next area to look at. Convolutional Neural Networks oversimplify this feedback-control mechanism as input-output system. Attention-based methods fill in this gap, and allow us to study neural networks that better aligns with human beings.
2. Lower computational cost: CNN runs slow on large images because the computation on 2D image is linearly correlated with image resolution, but Human beings are able to reason about large pictures based on small image areas. Attention-based learning allows us to aggregate local information and improve the speed of visual algorithms.

Works on Attention learning have combined latest deep learning architecture (https://arxiv.org/pdf/1312.6110.pdf) with former findings (http://papers.nips.cc/paper/4089-learning-to-combine-foveal-glimpses-with-a-third-order-boltzmann-machine, Learning where to Attend with Deep Architectures for Image Tracking). In Show, Attend and Tell, researchers managed to reconstruct what the model is looking at, by visualizing the weights.



(b) A person is standing on a beach with a surfboard.

These projects have been focused on standard 2D image, and few works have combined attention visual learning with Virtual Reality. However we know that human vision is most well adapted to stereo scenes, and in order to really understand the difference and similarity of our system and human vision, we need to extend vision research to 3D virtual discipline.

This is confirmed by recent work by Stanford researchers collected attention information for human in VR environment using eye-tracking. They created a visual saliency map for human to understand scenes, and argued that human attention in VR scenes bears differences from 2D image attention. However, due to the costly set-up using DK2 head-mounted display and pupil-lab stereoscopic eye tracker, the dataset collected in the experiment is not enough for conducting training using neural network.

The goal of this project is to propose a solution that can bridge the gap between the need for visual data and the difficulty in collecting. I will build an end-to-end data collection system, where I use human head orientation in virtual scene as an approximation for eye direction in order to collect human visual attention under specific VR tasks like scene understanding. (research has confirmed that human tend to rotate head with >7 degree view of focus). To scale up data collection, I rely on low-end Google Cardboard and Amazon AMT to deploy the application. Given the low cost and growing

Project Specification

What to Collect

1. Task 1: free exploration in scenes. Here we will follow the procedures specified in Sitzmann et al. (Saliency in VR: How do people explore virtual environments?), where we place users in randomized scenes, each at a randomized starting point. Then we will make user choose when to stop, and write a sentence describing the scene they have seen. The entire head movement of the user while exploring the scene will be recorded.
2. Task 2: semantic exploration. The goal of this task is to study how users extract useful semantic information. We will place users in indoors scenes scanned by

How to Collect

1. On AMT interface, we will provide user with a web link to open up in their phone.
2. For the ease of development within course project period, I implement the system using Three.js on mobile browser, incorporating code from course homework project.

Visualization

1. A second part of the system is visualizing collected human attention data stream, in the manner similar to the above Show, Attend and Tell project. We will render a sliding window in dark background, indicating which area is seen by user in real time. This provides us a window to compare what machine considers to be useful, with visual cues that human beings rely on understanding scenes.

Cost of Collecting

1. I will conduct cost estimation of data collection on the proposed system, and compare it with data collection based on eye-tracking. Using the comparison, I will discuss the viability and potential challenges of using the system.

Future Work

1. Future work include deploying the platform on AMT, collecting data and maintaining the system
2. After data collection is finished, we can do multiple studies with the data. This includes: calculating human VR saliency map and comparing it with machine predictions, inputting the visualized attention image stream into LSTM to do human gaze prediction.