Abstract

Visual search is a routine human behavior and canonical example of selectively sampling sensory information in service of attaining a goal. In an effort to extend insights from existing models of optimal visual search to naturalistic environments, we conducted a study of visual search in virtual reality. Participants viewed visual stimuli through a head-mounted display equipped with an eye tracker, while holding a handheld controller. Participants performed 300 trials during which they had 8 seconds to search for a target object in a cluttered room. Before each trial, we presented the participant with a blank environment containing only the target object, which rotated to ensure the participant could view it from multiple angles. After the target viewing period, we teleported participants to one of 6 possible rooms and at one of 5 possible viewpoints, and instantiated 10 different target/distractor sets for each room/viewpoint combination. Every participant experienced each of the 300 resulting scenes once, with order randomized across participants. We found that despite the large (>60) number of distractor objects and complexity of the environment, humans are remarkably efficient at finding the target, performing correctly on about 80% of the trials and typically finding the target within 3s. To better understand what features of the environment people are using to drive search, we annotated the sequence of gaze samples with semantic scene information. For each gaze sample, we extracted the label of which part of the scene is currently being foveated (e.g. “wall”, “mug”). When gaze was directed to a valid object (i.e. the target or a distractor), we used the 3D rendering of the object in the virtual environment to compute summary statistics about the shape and color of each object. We found evidence that people’s gaze is primarily directed to objects in the scene, and that the objects that people look at are more similar to the target along both shape and color. Furthermore, the similarity between sampled objects and target increased over time, suggesting that humans use similarity to target as a reward function which shapes their policy for which parts of the scene to sample during search. We discuss these results in the context of a formal model of feature selection as a meta-level Markov Decision Process, which we validate by predicting human fixation patterns in real time.