Murat Ambarkutuk {murata@vt.edu}, 03/01/2016

1 Learning to Predict Where Humans Look

In their paper, the authors propose a method to create saliency maps for given images in order to predict where humans would look in the images. The contributions of the paper are an experimental dataset and a method using different features to model visual fixation by viewers of images. The authors noted that the proposed algorithm performs better than state-of-the-art methods, while demystifying the visual perception process.

2 Approach

The proposed approach takes advantage of the dataset provided with it. The dataset was obtained from 15 testee's gaze information, in particular the parts of the images they look at as they view images for three seconds (per image). Between two images, the authors reported that a gray scene was displayed for one second to (roughly speaking) reset the point upon which the testee's eyes are fixated. The testees were also told that it was a form of memory test in order to insentivize attention. Personally, if I was told the same thing, I would find three seconds too short a time window for a memory test. I will discuss more about the dataset and the data acquisition process in the following section.

To train a linear SVM, three levels of features were extracted. Intensity, color contrast, local energy and similar information constituted low-level features. A horizon detector is used to capture the place where the objects rest (constituting a mid-level features). Finally, a Viola-Jones face detector and DPM are used to capture higher level features. Given the features and the samples acquired from the saliency maps, a linear SVM is trained to predict where a human would look.

3 Impressions

Given the authors' superior results compared to previous approaches, the paper bridges the gap in our understanding in visual perception and cognition. It also makes sense to to use low-, mid-, and high-level features to model the approach. The data acquisition process is described in detail, which I, personally, find very enlightning. As a graduate student starting my career in academia, the data acquisition process has given me intuitions on how I should conduct experiments and report results. However, using a heavily biased dataset which can be accurately modeled with a Gaussian function centered at the image center does not seem reasonable for the given task. It is expected to see objects of interest at the center area of images and the authors have no control on that. However, as is mentioned in the paper, the testees can be seated at an angle to the screen. This, I suspect, would reduce the bias. Another point I would like to discuss is that the proposed method takes advantage of fixation points to obtain samples. However, I believe that this is not an accurate representation of the way humans look at their environments by scanning the points. Instead, the way we percieve the environment is to understand the whole scene by looking at the patches. For example, as we read the lines of books and papers, we tend to see at word-level, or in some cases full sentences. To the best of my understanding, this paper uses features capturing mostly local information. Thus, it would be a better approach if they would account for that. One thing that can be done to model this is to use soft voting as they capture the data set.