International Journal of Multimedia Data Engineering and Management
Volume 9 • Issue 2 • April-June 2018

A Randomized Framework for Estimating Image Saliency Through Sparse Signal Reconstruction

Kui Fu, State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China

Jia Li, State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, International Research Institute for Multidisciplinary Science, Beihang University, Beijing, China

ABSTRACT

This article proposes a randomized framework that estimates image saliency through sparse signal reconstruction. The authors simulate the measuring process of ground-truth saliency and assume that an image is free-viewed by several subjects. In the free-viewing process, each subject attends to a limited number of regions randomly selected, and a mental map of the image is reconstructed by using the subject-specific prior knowledge. By assuming that a region is difficult to be reconstructed will become conspicuous, the authors represent the prior knowledge of a subject by a dictionary of sparse bases pre-trained on random images and estimate the conspicuity score of a region according to the activation costs of sparse bases as well as the sparse reconstruction error. Finally, the saliency map of an image is generated by summing up all conspicuity maps obtained. Experimental results show proposed approach achieves impressive performance in comparisons with 16 state-of-the-art approaches.

KEYWORDS

Activation Cost, Randomized Framework, Reconstruction Error, Sparse Coding, Visual Saliency

INTRODUCTION

In an image, there always exist certain visual stimuli that demonstrate impressive capabilities in attracting human attention. As a consequence, it is important to detect such salient visual content before conducting complex segmentation and recognition tasks. In this manner, limited computational resources can be assigned to salient visual content with high priority so that images can be efficiently processed as the human-being does.

Inspired by various psychological and neurobiological theories (e.g., Guided Search Model (Wolfe et al., 1989) and Feature Integration Theory (Treisman et al., 1980)), numerous saliency models have been proposed in the past two decades. In these models, a common solution is to divide an image into non-overlapping rectangular patches (i.e., macro-blocks) at a single scale or multiple scales. After that, the saliency of a patch is measured by its rarity in the fixed local and/or global contexts. For example, such visual rarity can be computed as local contrast (Itti et al., 1998), surprise (Itti & Baldi, 2005), coding length increment (Hou & Zhang, 2009) as well as the global viewing time (Harel et al., 2007), entropy rate (Wang et al., 2010) and co-occurrence frequency (Lu et al., 2013). In particular, some approaches (e.g., (Hou & Zhang, 2007; Fang et al., 2012; Li et al., 2013; Li et al.,

DOI: 10.4018/IJMDEM.2018040101

2015)) first transform image into the frequency domain and then measure patch rarity via spectrum analysis. Moreover, since rarity can be simultaneously measured from multiple feature channels, some researchers proposed to derive patch saliency by combining various kinds of rarities with a heuristic framework (Borji & Itti, 2012) or a "feature-saliency" mapping function learned from user data (Judd et al., 2009; Borji, 2012; Zhao & Koch, 2012).

Generally speaking, these saliency models can achieve impressive performance in many cases. However, they still have two drawbacks. First, these models often embed an image patch into fixed local and/or global contexts, while each patch actually appears along with different neighbors when an image is free-viewed by different subjects (i.e., flexible contexts). Second, many models propose to directly measure the saliency values of small patches with fixed sizes (e.g., 8×8 blocks), while human attention in the free-viewing process is often attached to regions with changing sizes (i.e., flexible regions). Actually, it is believed that the performance of saliency estimation can be greatly improved if the elementary saliency unit and its context are both flexibly selected.

Before addressing these two issues, the authors first turn to a fundamental problem: how to measure the "ground-truth" saliency of an image? In most eye-tracking experiments, the ground-truth saliency map of an image is considered to be the fixation density map formed by collecting fixations of multiple subjects during the free-viewing processes. As shown in Figure 1, a number of subjects are requested to free-view the same image for several seconds (usually 3-4 seconds) and their fixations are recorded with a high -speed eye tracking apparatus (e.g., 30Hz (Ramanathan et al., 2010), 60Hz (Li et al., 2013), or even 240Hz (Itti, 2008)). Eventually, a location that captures more fixations (i.e., high fixation density) becomes more salient. By inspecting these procedures, the authors simplify the measurement of ground-truth saliency into four main steps, including: 1) each image gets free-viewed by multiple subjects, 2) each subject (with different prior knowledge) attends to different regions, 3) each region receives a number of fixations that reflects its conspicuity score, and 4) image saliency is measured by the fixation density of multiple subjects (i.e., the conspicuity maps generated by multiple subjects).

Figure 1. In an eye-tracking experiment, different subjects may free-view different regions in the same image. As a result, their fixations are recorded so as to form a fixation density map that reflects the ground-truth saliency