SALICON: Saliency in Context

Ming Jiang†, Shengsheng Huang†, Juanyong Duan†, Qi Zhao*
Department of Electrical and Computer Engineering, National University of Singapore.

Saliency in Context (SALICON) is an ongoing effort that aims at understanding and predicting visual attention. This paper presents a new psychophysical paradigm to collect large-scale human attentional data during natural explorations on images. With this paradigm, we build the SALICON dataset with 10,000 natural images, by crowdsourcing the data collection with Amazon Mechanic Turk (AMT). The SALICON dataset is by far the largest in both scale and context variability. The human viewing data during the assumption-free exploration also provides insights to other vision tasks and complement them to better understand and describe image contents (see Figure 1).

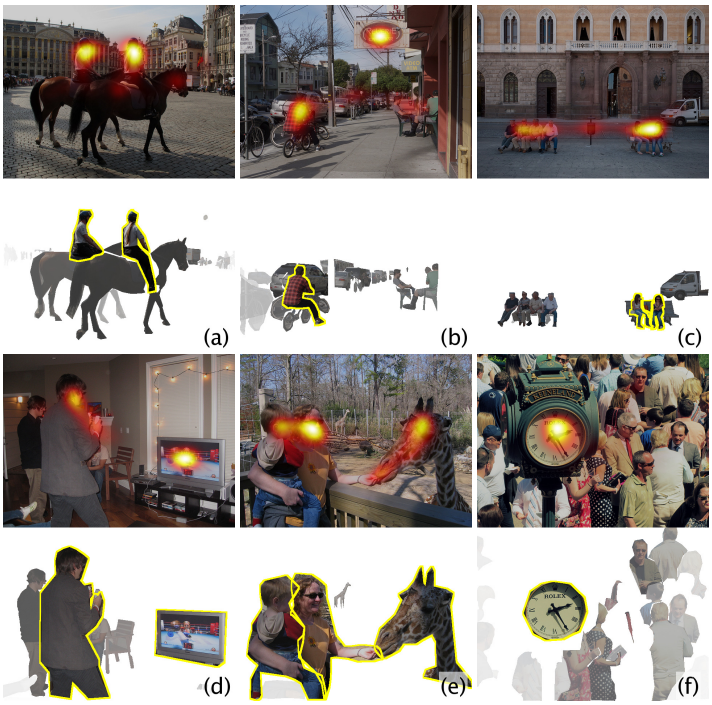

Figure 1: We propose a new method to collect large-scale attentional data (SALICON, 1st row) for in visual understanding. With the annotated object segments, our attentional data naturally highlights key components in an image (ranked object segments in the 2nd row, with key objects outlined in yellow) to (a) rank object categories, (b) suggest new categories important to characterize a scene (text in this example), (c-e) convey social cues, and (f) direct to places designed for attention in advertisement.

Human visual system shows a well-defined contrast sensitivity by retinal eccentricity relationship. Specifically, contrast sensitivity to higher spatial frequencies drops off as a function of retinal eccentricity. To simulate the free-viewing patterns of human visual attention with mouse tracking, we generated a resolution map to simulate the sensitivity drop-off in peripheral vision. It was defined as a function $R:\Theta\rightarrow[0,1]$, where Θ is the set of viewing angles θ with respect to the retinal eccentricity, and $[0,1]$ represents the set of relative spatial frequency. The resolution map approximates a normal adult's vision with the exclusion of the blind spot. A higher $R(\theta)$ indicates a higher resolution at the visual eccentricity θ. Specifically, the resolution map is formulated as

$$R(x,y)=\frac{\alpha}{\alpha+\theta(x,y)}, \quad (1)$$

where $\alpha=2.5^\circ$ is the half-height angle, meaning that when $\theta(x,y)=\alpha$ the image became only half the resolution of the pixel in the center of attention ($\theta(x,y)=0$). An example of the produced multi-resolutional images is shown in Figure 2.

Figure 2: An example of the mouse-contingent stimuli. The red circles indicate the movement of mouse cursor from one object to another.

We deployed the experiment on the Amazon Mechanical Turk (AMT) using 10,000 images from the MS COCO dataset [1] and 700 images from the OSIE dataset [2]. The OSIE eye-tracking data were compared as a baseline to evaluate the mouse-tracking performance. For verification, we also recruited 16 subjects to view all the 700 OSIE images in the lab environment with the proposed paradigm. The data similarity was measured with shuffled AUC (sAUC) [3]. We also included the state-of-the-art saliency algorithms in the comparison. As shown in Figure 3a, the lab and AMT mouse models scored closely in sAUC (0.86). The mouse-tracking performances were much closer to the human performance in eye tracking (0.89) than the computational models. The high mouse-eye agreement was observed in most images. Figure 3b presents the images with high and low sAUC scores in mouse tracking (with AMT). With the achieved similarity between the two modalities, we further exploited the mouse tracking as a benchmark to evaluate computational saliency algorithms, and the model rankings were also found consistent across mouse-tracking and eye-tracking datasets.

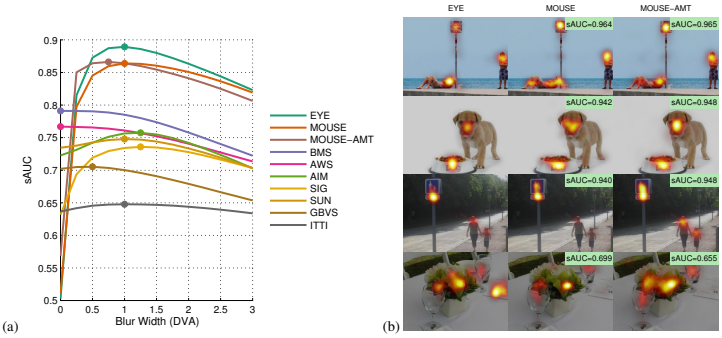

Figure 3: (a) Eye fixation prediction performance with mouse tracking and the highly referred/state-of-the-art computational saliency models. (b) Image examples with high and low eye-mouse similarities evaluated with sAUC. Eye fixation maps and mouse maps are overlaid.

[1] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014.

[2] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *J. Vis.*, 14(1):28.1–20, 2014.

[3] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. SUN: A bayesian framework for saliency using natural statistics. *J. Vis.*, 8(7):32.1–20, 2008.

†The three authors contribute equally to this work.
*Corresponding author. eleqiz@nus.edu.sg
This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.