Feasibility of Computational Estimation of Task-Oriented Visual Attention

Yorie Nakahira\textsuperscript{1} \hspace{1cm} Minoru Nakayama\textsuperscript{1,2}

\textsuperscript{1} Tokyo Institute of Technology, Email: nakahira.y.ab@m.titech.ac.jp
\textsuperscript{2} Tokyo Institute of Technology, Email: nakayama@cradle.titech.ac.jp

Abstract

Eye movements are elicited by a viewer’s task-oriented contextual situation (top-down) and the visual environment (bottom-up). The former and its computability were investigated by measuring a subjects’ eye movements while identical graphs were viewed under different mental frameworks. The prerequisite conditions for predicting task-oriented attention from the visual environment were determined.

Keywords: eye movement prediction; top-down attention; saliency; fixation; gaze duration

1 Introduction

Usability assessment of what we view by means of visual comprehensiveness requires the repeated labor-intensive measurement and interpretation of data (Robert et al. 2003). Alternatively, a computational model can be used to generate attention information from the visual environment. However, eye movement predictions need to integrate bottom-up image-based saliency prompts and top-down task-oriented contextual situations (Itti & Koch 2001). Current research has difficulty with the modeling of the latter (Benjamin, 2009) when quantitative models such as Itti’s saliency map (1998) are used.

To get around this problem, this study intends to determine if it is possible to predict top-down attention using only the visual environment, by exploring the following: (1) which characteristics of images, or the meanings the images possess, and (2) what contextual situation, if given to viewers, enable the modeling of task-oriented attention using features of images.

2 Method

2.1 Eye Movement Recordings

Seven subjects (4 male, 3 female) viewed eight bar-graphs (Figure 1) shown in two different constructs. In the first trial, subjects were told they would be shown rectangular objects, without being conscious that they were bar-graphs. In fact, eight bar-graphs without any axes were shown. In the second trial, subjects were assigned the task of understanding the quantitative data represented, using graphs. The same eight bar-graphs were shown but with axes. Calibration preceded both trials. Each image appeared randomly, one at a time for three seconds, followed by a one second interval during which subjects fixated on a central cross. Each subject repeated this experiment for 3 times. An eye tracker (nac: EMR-(NL, 640x480 pixel resolution, 60Hz) recorded subjects left eye positions.

2.2 Definitions of Metrics and Fixation Maps

In this paper, fixation is defined as a stable eye position with a velocity below the threshold of 20 degrees per second (Robert et al. 2003), gaze duration is defined as the sum of each fixation duration at a particular region of interest, and scan paths are the spatial arrangements of a sequence of fixations. Scan paths, the total number of fixations on each bar, and gaze duration on each bar were generated for later analysis.

The fixation location and duration were used to generate two kinds of fixation maps (David 2002), which were three dimensional matrices: the x-y axes were the size of the image, with the third dimension being an indication of the number of fixations or duration of gaze on each area. The Fixation Number Maps (Figure 3) were calculated by dropping an identical Gaussian kernel of the same height at the location for each fixation. The Gaze Duration Maps were calculated using a Gaussian kernel of a height proportional to the fixation duration. The half-height width for the Gaussian kernel is determined according to the area over which a fixation can be said to exist. The map is then normalized so that the final value of the highest peak equals one. The initial fixation locations, which are specified by the central cross at the moment when the images appear, are ignored during the calculation of this map.

3 Results

3.1 Modeling Attention Level

As previous research has suggested, Itti’s saliency model was not adaptable to both trials, because it fails to predict either the point on the graph which the subjects’ eye moved to initially, or the scan paths. Nevertheless, one notable thing is that when images are viewed sufficiently often (21 times per image), the average amount of attention paid to each region of interest as a whole (in this case, each bar) does have a correlational relationship with one of the saliency features: intensity. This implies that the saliency of intensity may be used to indicate the level of attention given to regions of interest, and to compute the probable amount of visual attention.
The intensity map (Figure 1) of Itti’s saliency model may be used as a measure of conspicuousness, to predict the distribution of fixations (define this as ‘Attention Level’). For each trial, in every bar in every image, the sum of the intensity corresponding to the position of each of the bars was calculated, then normalized so that the sum of the intensity of all bars in one graph equaled one. Similarly, on the fixation maps, the sum of the values of all pixels viewed which corresponded to the positions of each bar (Attention Level) was calculated for both the Fixation Number and Gaze Duration Maps, as a measure of the amount of attention given to each bar, followed by a normalization process so that the sum of the Attention Level for each bar on one bar-graph equaled one.

Table 1 shows the correlational coefficient between the Attention Level in terms of fixation number or gaze duration and intensity for both trials. Figure 2 illustrates the relationship between the proportional fixation number and the relative intensity; each dot represents a particular bar in a bar-graph with attention level on the horizontal axis and intensity level on the vertical axis. Both the fixation number and intensity are normalized so that the sum of all bars in the bar-graph equals one.

Table 1: Correlation between Attention Level and Intensity

<table>
<thead>
<tr>
<th></th>
<th>trial1</th>
<th>trial2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixation Number</td>
<td>0.90</td>
<td>0.65</td>
</tr>
<tr>
<td>Gaze Duration</td>
<td>0.49</td>
<td>0.69</td>
</tr>
</tbody>
</table>

3.2 Adaptability of the Prediction Model

Contrary to the expectation that saliency features are insufficient when predicting top-down attention from the second trial, Table 1 indicates that the correlational relationship for the first trial is no better than for the second trial. To illustrate this tendency, Figure 3 plots the Gaze Duration Map for the image in Figure 1 of the first trial (right) and the second trial (left). The most significant difference in eye movement between the two trials is that subjects in the first trial focus more on a couple of positions corresponding to the largest points of intensity on the map. However, subjects in the second trial look at each bar thoroughly, according to the relative level of intensity, which contributes to a better adaptability of the bars which have a relatively low level of intensity.

Since the concept of a saliency map is conventionally used for the first fixation prediction, it works well when the intensity is high, but not surprisingly, loses its accuracy when the intensity is low. Yet, the intensity feature is a better predictor in situations where subjects intentionally pay attention to every bar, in order to grasp the meaning of each graph as a whole, or in other words, to understand the quantity of each bar in relation to the others. This is noteworthy, as it implies that the difference between the two trials—the additional task imposed in the second trial—actually helps, rather than working against the accuracy of the model.

In conclusion, it can be inferred that task-oriented Attention Level (not the first fixation) can be predicted by measuring the intensity features from the visual environment when the following two conditions are satisfied: (1) there is an underlying top-down force that directs gaze toward each region of interest, according to the level of perceived importance implicit in its saliency of intensity; and (2) the visual environment (the graphs used in this study) possesses an implicit meaning which can be calculated from the intensity of the image.

4 Summary

The study confirmed that top-down attention is capable of being predicted merely by processing the images in the visual environment. A future direction of study would be to investigate the differences in prediction accuracy between each image, to determine which may or may not permit accurate predictions to be calculated.

![Figure 1: Example of a displayed Image and its Intensity Map](image1)

![Figure 2: Attention Levels vs conspicuousness](image2)

![Figure 3: Fixation Maps of Figure 1. The map for the first trial is on right, second trial on left.](image3)

References


Itti, L., Jacob & Keith, S. Karn(2003), Eye tracking in human-computer interaction and usability research: ready to deliver to promises, in The Mind’s Eye, pp574-605.