Using semantic content as cues for better scanpath prediction

Moran Cerf* E. Paxon Frady
Christof Koch
Computation and Neural Systems, California Institute of Technology, Pasadena, CA, 91125

Abstract

Under natural viewing conditions, human observers use shifts in gaze to allocate processing resources to subsets of the visual input. There are many computational models that try to predict these shifts in eye movement and attention. Although the important role of high level stimulus properties (e.g., semantic information) stands undisputed, most models are based solely on low-level image properties. We here demonstrate that a combined model of high-level object detection and low-level saliency significantly outperforms a low-level saliency model in predicting locations humans fixate on. The data is based on eye-movement recordings of humans observing photographs of natural scenes, which contained one of the following high-level stimuli: faces, text, scrambled text or cell phones. We show that observers - even when not instructed to look for anything particular, fixate on a face with a probability of over 80% within their first two fixations, on text and scrambled text with a probability of over 65.1% and 57.9% respectively, and on cell phones with probability of 8.3%. This suggests that content with meaningful semantic information is significantly more likely to be seen earlier. Adding regions of interest (ROI), which depict the locations of the high-level meaningful features, significantly improves the prediction of a saliency model for stimuli with high semantic importance, while it has little effect for an object with no semantic meaning.

CR Categories: I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Tracking; I.6.4 [Simulation and Modeling]: Model Validation and Analysis;

Keywords: Eye Tracking, Psychophysics, Natural Scenes

1 Introduction

Human visual attention serves to delegate the resources of the brain to quickly and efficiently process the vast amount of information that is available in the environment [James 1950]. Selective visual attention is beginning to be reasonably well understood and quantitative models have been derived to explain attentional and eye movement deployments in the visual scene [Itti and Koch 2001]. However, their predictive power has not reached its full potential [Peters et al. 2005; Oliva et al. 2003]. One of the dominant sensory-driven models of attention currently focuses on low-level attributes of the visual scene to evaluate the most salient areas. Features such as intensity, orientation, and color are commonly combined to produce maps through center-surround filtering at multi-scaled resolutions. These maps are normalized to create an overall saliency map, which predicts human fixations significantly above chance [Itti and Koch 2000]. Filling some of the gaps between the saliency map models current predictive power and the theoretical optimum is thought to be possible by incorporating higher-order statistics in saliency models [Einhäuser et al. 2006]. One way of doing this is by adding new feature channels that represent high-level attractive stimuli. For example, in recent work we have shown that faces, a high-level stimulus, attract gaze and can be used for better predictions of observer's scanpaths [Cerf et al. 2008]. This makes sense as faces were likely to have been an important feature to pick out in our evolutionary environment, and thus a process to allocate attention toward faces has reason to arise in the brain. This study looks at how a variety of high-level stimuli effects human attention in still images. Specifically, we are trying to address the question of whether any high-level object can be used for improved prediction, or only ones that have more meaningful properties. Although the claim that one is born with an innate ability to detect faces is still under debate [Hershler and Hochstein 2005; VanRullen 2006], we can surely claim that one is not born with a similar mechanism for text or cell phone detection. If we are good in detecting those in a natural scene it must be due to experience and adaptation with time. That said - it is obvious that reading text is very important for modern humans, and thus we can imagine text acting as an attractor similarly to faces.

In order to test the role of semantic content's effect on gaze, we had 36 subjects look at 4 different scenes, each containing a different high-level entity - faces, text, scrambled text and cell phones. We tested both the saliency of each of the high-level stimuli, and the predictive power of a saliency map model with and without the addition of a high-level object detector.

2 Methods

2.1 Experimental procedures

Thirty six subjects were divided to 4 groups, each viewing a set of images (1024×768 pixels). Group 1 viewed 150 images containing faces. Group 2 viewed 40 images of natural scenes containing text. Group 3 viewed the same 40 images as group 2 only that the text of the images was made of random letters drawn from an English based distribution of letters. The modified images were to look as if they had not been manipulated at all. The size, font, color, orientation and shape remained the same, only that the text in the scene was scrambled, such that it had no meaning. Group 4 viewed 37 images containing natural scenes with a cell phone located somewhere in the scene. The sizes of the cell phone, texts, and faces were chosen such that they cover no more than $5\% \pm 1\%$ (mean $\pm$ s.d.) of the entire image - between 1° to 5° of the visual field. All subjects were literate, had normal or corrected-to-normal vision, and were naïve to the purpose of the experiment.

The objects were chosen such that they vary from ones that carry highly semantic content (faces) to ones that carry lower semantic/emotional and personal content (cell phones). Although one can argue on the importance of faces in our life, it is clear that faces carry higher semantic content than cell phones. Images were presented to all groups in the same setup. Images were presented to subjects for 2 s, after which they were instructed to answer "How interesting was the image?" using a scale of 1-9 (9 being the most interesting). The task was chosen such that it would not bias subjects to look at anything in particular. No subject was shown more

*e-mail: moran@klab.caltech.edu

Figure 1: *Examples of images from the 4 categories: Faces, Text, Scrambled Text and cell phones, with scanpaths of one individual from each of the 4 groups superimposed. The triangle marks the first and the square the last fixation, the white line the scanpath, and the black circles the subsequent fixations. Faces and cell phone images were from the [Cerf et al. 2008] database. The trend of visiting the faces, text and even scrambled text first - typically within the 1^{st} or 2^{nd} fixation - is evident, while cell phones do not seem to draw eye gaze in the same manner. Note that images are in color, but are here printed in grayscale.*

than one category (cell phone, faces, text, or scrambled text) in order not to bias the viewings. The images were introduced as "regular images that one can expect to find in an everyday personal photo album". Scenes were indoors and outdoors still images (see examples in Fig. 1). Face images included faces in various skin colors, age groups, and positions. Faces had neutral expressions. No image had the face/text/cell phone at the center of the image as this was the starting fixation location in all trials. Subjects fixated on a cross in the center before each image onset. Eye-position data was acquired at 1000 Hz using an Eyelink1000 (SR Research, Osgoode, Canada) eye-tracking device. The images were presented on a CRT screen (120 Hz), using Matlab's psychophysics and eyelink toolbox extensions. Stimulus luminance was linear in pixel values. The distance between the screen and the subject was 80 cm, giving a total visual angle for each image of $28^\circ \times 21^\circ$. Subjects used a chin-rest to stabilize their head. Data was acquired from the right eye alone.

2.2 A model combining low-level saliency with high-level features

To determine whether high-level objects contribute more than their low-level attributes to power attention we tested how well the standard low-level feature driven saliency map does in comparison with the same model combined with an extra high-level entity detector. As a face-detector we used the widely used Viola and Jones algorithm for face recognition [Viola and Jones 2001]. For the other entities we manually defined minimal ROIs around the entity investigated. Each entity's saliency map was represented as a positive valued heat map over the image plane. The original saliency map is computed as an average of 3 channels: intensity, orientation and color [Itti et al. 1998]:

$$S = \frac{1}{3}(N(\overline{I}) + N(\overline{C}) + N(\overline{O})) \quad (1)$$

The modified map is made with the extra entity channel:

$$S = \frac{1}{4}(N(\overline{I}) + N(\overline{C}) + N(\overline{O}) + N(\overline{E})) \quad (2)$$

The combination was linear in nature with uniform weight distribution for maximum simplicity. Performance of the new saliency map was measured by the receiver operating characteristic (ROC). The hit rate was calculated by determining the locations where the saliency map is above threshold and there is a fixation present. Similarly, the false alarm rate was calculated by measuring the locations at which the saliency map is above threshold and there is no fixation present (for discussion see [Cerf et al. 2008]). The ROC curve was created by varying the threshold to cover all possible ranges of values the saliency map produces. The area under the curve (AUC) is a general measure of how well the saliency map predicts fixations. An AUC of 50% reflects chance and an AUC of 100% reflects perfect prediction.

3 Results

3.1 Psychophysical results

To evaluate the results of the 36 subjects' viewing of the images, we manually defined minimally sized rectangular ROIs around each target object in each image of the entire collection. We assessed how many of the first fixations went to a face/text/cell phone, how many of the second, third fixations and so forth. In group 1 a face was fixated on within the first two fixations in 89.3% of the trials (Fig. 2). Given that the face ROIs were chosen very conservatively (i.e. fixations just next to a face did not count as fixations on the face), this shows that faces, if present, are typically fixated on within the first two fixations ($327\ ms \pm 95\ ms$ on average). To verify that this is not due to chance we compared our results to an unbiased baseline. The baseline for a particular image is the fraction of all subjects' fixations from all other images that fall in the ROI of the particular image. The null hypothesis that we would see the same fraction of first fixations on a face at random is rejected at $p < 10^{-20}$ (t-test) (see Fig. 3 for illustration of the baseline calculation). Similar measures for the text and scrambled text show that text was fixated on with the first two fixations ($369\ ms \pm 158\ ms$) in 81.1% ($p < 10^{-15}$) and scrambled text ($415\ ms \pm 147\ ms$) in 51.1% ($p < 10^{-12}$).

Figure 2 shows that objects that carry higher semantic content draw more attention. Both text and faces - which are semantically important - are viewed significantly ($p < 10^{-5}$) earlier than cell phones, and present a similar "Poisson" pattern that decays with the estimated importance of the feature. One key reason to use only the first fixation as a measure for the text is the "reading" fixations. While for meaningful text subjects usually spend a few of their fixations actually reading the text, and for scrambled text they don't, in both cases they first gaze at the word, as if they are about to read it, and only then figure out if they are to follow it by actual reading or by saccading elsewhere.

3.2 Assessing the modified saliency model

In order to improve the predictive performance of our saliency algorithm, we studied the affects of semantic information on fixation allocation. We first tested the general importance of semantic con-

tent in a scene, and distinguished between different levels of importance. As control we used a cell phone which seems less important for our daily life than text and faces. Adding a fourth channel for each category to a standard saliency model and calculating a new saliency map we tried to see how well the new map does in predicting the locations of the subjects' fixations.

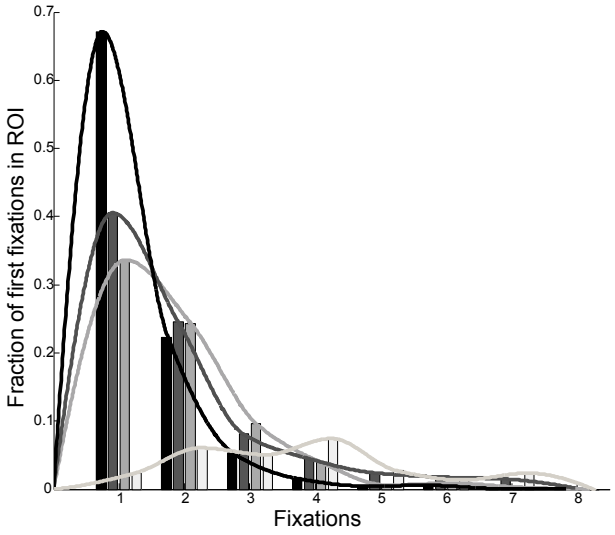

Figure 2: *Extent of fixation on regions-of-interest (ROIs). Bars depict percentage of trials, which reach the ROI the first time in the first, second, third, ... fixation. The solid curves depict the fitting of the data for each category. A Poisson-like fit indicates viewing of the ROI in early fixations. This is illustrated in the faces, text and scrambled text fits, whereas the cell phone fit indicates later fixations. This data shows that subjects tend to fixate on faces earlier than text, and even more so than scrambled text. All highly semantic cues are fixated on earlier than cell phones.*

The performance of the standard saliency map was on average 65.3% (AUC for the 150 face images). Adding the face channel increased the predictability to 79.3%. Predictions of fixation location for subjects viewing normal text improved from 71.4% on the standard saliency map, to 77.5%, and scrambled type improved from 72.9% to 77.1%. Both results are significant improvements ($p = 0.0022$, 0.0113 respectively). For the 40 text images, the text channel leads to improvements for every image. If successive fixations in the text ROI are removed - as they may constitute reading fixations, predictions for subjects viewing normal text still improves from 70.3% to 74.3% ($p = 0.0275$), and for scrambled text improves from 72.2% to 75.4% ($p = 0.0413$). Nevertheless, for the cell phone images, while the mean AUC for the standard saliency map prediction was 71.7%, it improved only to 72.1% for the new saliency map (not significant). The trend of significantly improving the predictability of a saliency map by adding a high-level semantic cue prevails only if the cue really seems to contain semantic content that is meaningful to us (be it innate or acquired). This trend is shown in Figure 4 where we compare the AUC for all images with the standard saliency map to the new saliency map and show that for faces and text there is an increase in predictability, while the cell phones show almost no improvement (all are on the diagonal).

In an attempt to better compare across groups we utilized our baseline measure as a normalization factor. This takes into account the varying size and locations of the ROI in all images (as these factors both influence how likely a certain region is to be fixated on by chance). By dividing the fraction of fixations in each entity's ROI by its baseline value, we form a normalized measurement which is

Figure 3: *Computation of the baseline. We consider all fixations, except the ones recorded for this image. Then, we compute the fraction of these which fall in the ROI for this image. Here, the ratio of white dots (i.e., those inside the ROI) to all dots. For this image, also shown in Fig. 1, out of 4497 fixations, 419 are in the ROI (9.32%).*

analogous to the number of times more likely the ROI is attended than chance. Our results show that fixations falling on faces produce a normalized score of 18.2 (times subjects are more likely to look at faces than baseline). Normal text produces a normalized score of 10.0, and scrambled text produces a value of 8.0. There is no significant difference between normal and scrambled text, but there is a significant difference between faces and text ($p < 10^{-5}$). Cell phones produce a relatively high score of 10, due to very low control values. That is, even a single fixation on a cell phone is very unlikely based on the baseline. Overall, faces, normal text and scrambled text are much more likely to be fixated on than chance in the first 2 fixations ($p < 10^{-10}$), whereas cell phones are typically visited only after the 3^{rd} fixation.

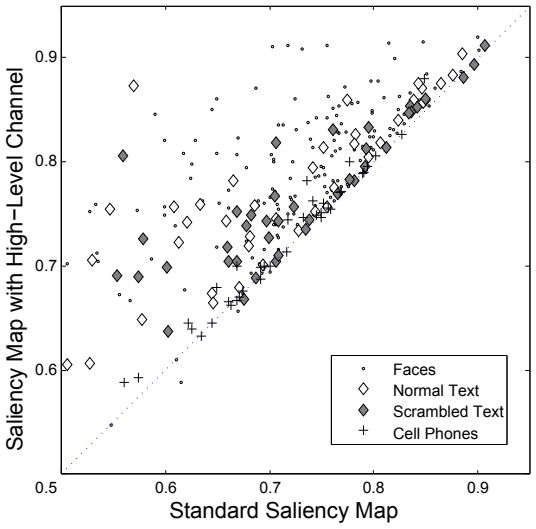

Figure 4: *Performance improvement for all images. Each symbol represents the model's performance predicting the subjects' fixations on a particular image, measured using an area under ROC curve (AUC) metric. Symbols above the diagonal indicate an improvement in the saliency map model when adding a high-level channel. Shapes and color-code in the scatter-plot indicate the different categories. Images with face/text/scrambled text channels are improved by the high-level channel. Images with faces are improved the most. Images with cell phone channel inclusion do not show an improvement in the predictive performance of the new model.*

4 Conclusion

The extent to which high-level cues such as faces are learned during early visual experience remains unclear [Johnson et al. 1991]. However, it is clear that faces are very important to us. More important than text, or scrambled text and surely more than cell phones. This is reflected in the results that show that adding high-level semantic cues to existing saliency models improves performance in predicting observers fixations. While the standard saliency model gives an average prediction of 69.8%, we here show that for images with semantic content in them, we can reach predictions levels of 77.9% on average. Saliency models with additional high-level channels can be beneficial not only for the improvement of fixations prediction which has applications in engineering and art, but can also serve as a measure (using the methods we demonstrate here for comparisons of chance viewing with observer viewing) to study the importance of high-level feature in a scene.

Acknowledgements

The authors wish to thank Kelsey Laird and Jonathan Harel for the valuable comments.

References

CERF, M., HAREL, J., EINHÄUSER, W., AND KOCH, C. 2008. Predicting human gaze using low-level saliency combined with face detection. In *Advances in Neural Information Processing Systems 20*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. MIT Press, Cambridge, MA.

EINHÄUSER, W., RUTISHAUSER, U., FRADY, E., NADLER, S., KÖNIG, P., AND KOCH, C. 2006. The relation of phase noise and luminance contrast to overt attention in complex visual stimuli. *Journal of Vision 6*, 11, 1148–1158.

HERSHLER, O., AND HOCHSTEIN, S. 2005. At first sight: a high-level pop out effect for faces. *Vision Res 45*, 13, 1707–24.

ITTI, L., AND KOCH, C. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research 40*, 10-12, 1489–1506.

ITTI, L., AND KOCH, C. 2001. Computational modeling of visual attention. *Nature Rev. Neurosci. 2*, 3, 194–203.

ITTI, L., KOCH, C., NIEBUR, E., ET AL. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 20*, 11, 1254–1259.

JAMES, W. 1950. *The Principles of Psychology*. Dover Publications.

JOHNSON, M., DZIURAWIEC, S., ELLIS, H., AND MORTON, J. 1991. Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition 40*, 1-2, 1–19.

OLIVA, A., TORRALBA, A., CASTELHANO, M., AND HENDERSON, J. 2003. Top-down control of visual attention in object detection. *Image Processing, 2003. Proceedings. 2003 International Conference on 1*.

PETERS, R., IYER, A., ITTI, L., AND KOCH, C. 2005. Components of bottom-up gaze allocation in natural images. *Vision Research 45*, 18, 2397–2416.

VANRULLEN, R. 2006. On second glance: Still no high-level pop-out effect for faces. *Vision Res 46*, 18, 3017–3027.

VIOLA, P., AND JONES, M. 2001. Rapid object detection using a boosted cascade of simple features. *Computer Vision and Pattern Recognition 1*, 511–518.