Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser et al.'s data

Ali Borji
Department of Computer Science, University of Southern California, Los Angeles, CA, USA

Dicky N. Sihite
Department of Computer Science, University of Southern California, Los Angeles, CA, USA

Laurent Itti
Neuroscience Program, Department of Computer Science, University of Southern California, Los Angeles, CA, USA

Einhäuser, Spain, and Perona (2008) explored an alternative hypothesis to saliency maps (i.e., spatial image outliers) and claimed that "objects predict fixations better than early saliency." To test their hypothesis, they measured eye movements of human observers while they inspected 93 photographs of common natural scenes (Uncommon Places dataset by Shore, Tillman, & Schmidt-Wulen 2004; Supplement Figure S4). Subjects were asked to observe an image and, immediately afterwards, to name objects they saw (remembered). Einhäuser et al. showed that a map made of manually drawn object regions, each object weighted by its recall frequency, predicts fixations in individual images better than early saliency. Due to important implications of this hypothesis, we investigate it further. The core of our analysis is explained here. Please refer to the Supplement for details.

Introduction

We compare the hypothesis that objects predict fixations better than early saliency (hereafter called object map) against 11 saliency models. We employ three types of saliency model maps: Two of them are different versions of the Itti, Koch, and Niebur (1998) model, ITTI98 and ITTI, which correspond to different normalization schemes. In ITTI98, each feature map's contribution to the saliency map is weighted by the squared difference between the globally most active location and the average activity of all other local maxima in the feature map (Itti et al., 1998). This gives rise to smooth saliency maps, which tend to correlate better with noisy human eye-movement data. In the ITTI model, the spatial competition for saliency is much stronger which gives rise to much sparser saliency maps (Itti & Koch, 2000). As we will see, these sparser maps tend to score lower than smoother maps when compared to noisy human eye-movement data, as human fixations that occur away from the few saliency peaks in this model's maps strongly penalize the model's score (i.e., this model yields more misses than a model with smooth maps). We also use the exact saliency maps of Einhäuser, Spain, and Perona (2008) (denoted here as ITTI* and which appear to be thresholded versions of the ITTI98 maps), to make our results directly comparable.

We perform three analyses. Our first analysis regards handling center bias (cb). Instead of the classic area under curve (AUC) score (employed by Einhäuser et al., 2008), we use the shuffled AUC (sAUC) score as it discounts center bias (spatial priors) in fixation data (Tatler, Baddeley, & Gilchrist, 2005). Briefly, in both scores, human fixations are considered as the positive set, but while in the AUC score some points from the image are sampled uniformly random as the negative set, in the sAUC score negative points are sampled from fixations of observers over other images. This allows sAUC to discount systematic spatial biases in human gaze patterns (e.g., center bias). sAUC score varies between 0.5 (chance level) and 1.0 (perfect agreement between model and gaze data). Figure 1A shows sAUC values for all models. There is no significant difference between the object-map model and ITTI* in their ability to predict human gaze (t test, $p = 0.234$, $\alpha = 0.05/n = 0.0045$ with Bonferroni

Citation: Borji, A., Shite, D. N., & Itti, L. (2013). Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser et al.'s data. *Journal of Vision*, *13*(10):18, 1–4, http://www.journalofvision.org/content/13/10/18, doi:10.1167/13.10.18.

doi: 10.1167/13.10.18 Received February 14, 2013; published August 29, 2013 ISSN 1534-7362 © 2013 ARVO

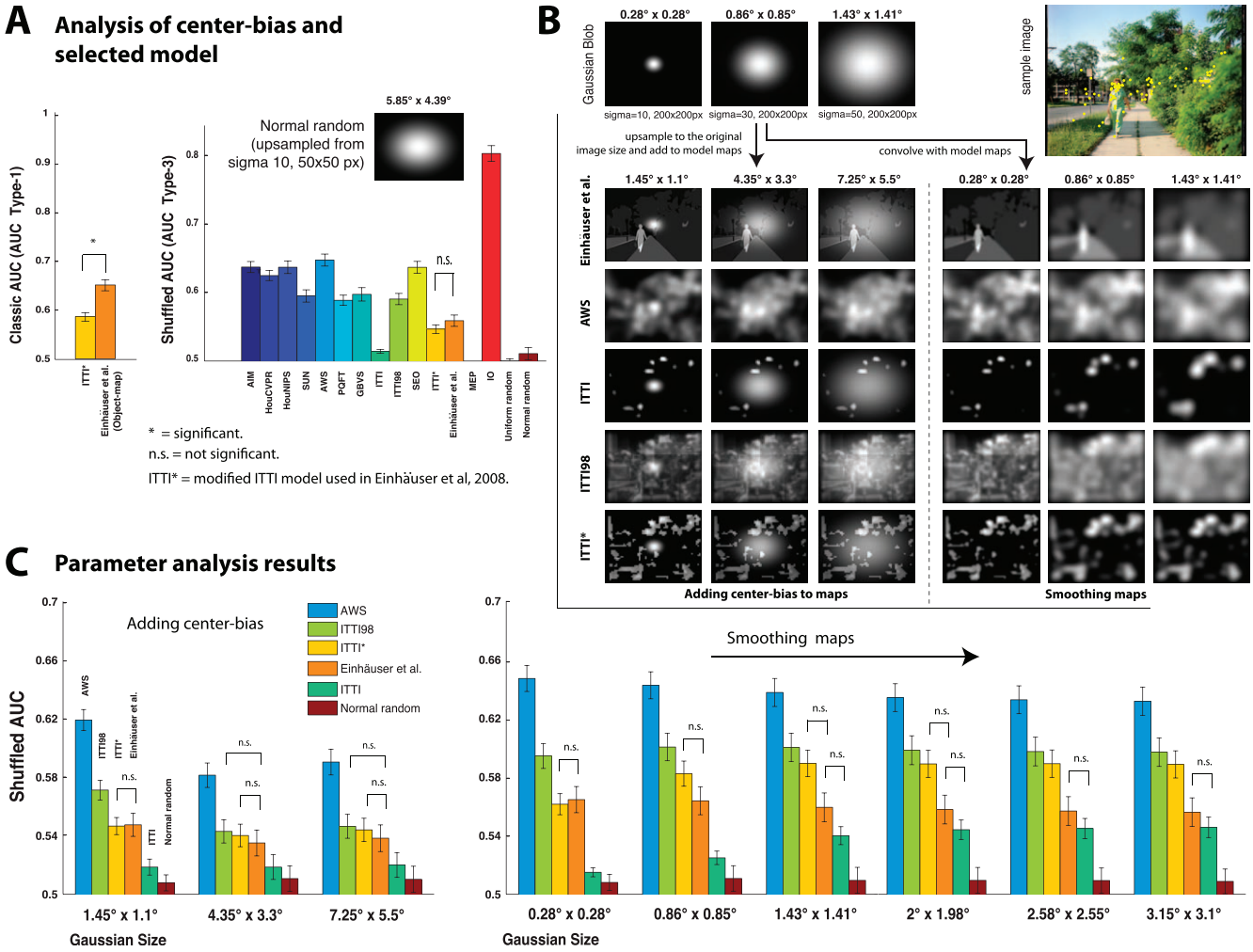

Figure 1. (A) Left: The object-map model outperforms the ITTI* significantly above chance when using the classic AUC score (AUC Type-1; Supplement) aligned with the original analysis of Einhäuser et al. (2008) (*t* test, $p < 0.05$; See also Figure S2A). Right: Using the shuffled AUC score (sAUC; AUC Type-3; Supplement), which discounts center bias in eye data, every difference between the object map and other models is statistically significant except the ITTI* model. This result shows the importance of appropriate tackling of center bias in fixation data (our first analysis; no significant difference between the object-map model and the ITTI* model). With newer saliency models or even with the original ITTI98 model (with a different normalization scheme than the newer ITTI model resulting in smoother maps), image-based outliers predict fixations significantly better. AWS model scores the best. MEP and random models score lowest about 0.5. This supports our second analysis about choosing the right model for data analysis (i.e., dependency of conclusions on the used model). To build a Gaussian of sigma size $0.28^\circ \times 0.28^\circ$ on this dataset, we used this Matlab script: myGauss = fspecial('gaussian',50,10) where 50 and 10 are image size and standard deviation of the Gaussian, respectively. To build the normal random model, we used the Gaussian shown in (A) and upsampled it to 1024×768 pixels (original image size presented at $29^\circ \times 22^\circ$ in Einhäuser et al.'s, 2008, study, thus ~ 35 pixel/°), resulting in $5.85^\circ \times 4.39^\circ$. (B) Gaussian blobs (kernels) of three different sizes were added to model prediction maps (shown here for five models). Gaussian kernels are built with size *x* and image size 200 using this Matlab script: myGauss = fspecial('gaussian',200,*x*) where $x \in \{10, 30, 50\}$ which leads to these sizes in degrees: $1.45^\circ \times 1.1^\circ$, $4.35^\circ \times 3.3^\circ$, and $7.25^\circ \times 5.5^\circ$ for the used dataset. Each prediction map of a model was smoothed by convolving with a Gaussian filter (for the shown image). Gaussian sizes for smoothing are: $0.28^\circ \times 0.28^\circ$, $0.86^\circ \times 0.86^\circ$, and $1.43^\circ \times 1.41^\circ$. (C) Prediction accuracies of models using the sAUC score: (left) center bias added and (right) smoothed saliency maps. Significance values are according to Bonferroni-corrected *t* test ($\alpha = 0.05/5 = 0.01$). By adding center bias, the object-map model is significantly above the ITTI model but not the ITTI98 model. Adding center bias does not dramatically change prediction power of models. With smoothing, the object-map model is significantly below the ITTI98 model. Smoothing more rises the accuracy of the ITTI model to the point that there is no longer a significant difference between this model and the object-map model. There is also no significant difference between the object-map model and ITTI* model with small amount of smoothing while with large amount of smoothing the ITTI* model outperforms the object-map model significantly. This supports our third analysis on parameterization. Error bars indicate standard error of the mean (*SEM*): $\sigma / m^{0.5}$, where σ is the standard deviation and $m = 93$ is the number of images. Please refer to Borji, Sihite, and Itti (2013) for details on fixation prediction models used here. See main text and Supplement for more details.

correction to the number of compared non-trivial models, thus $n = 11$ comparisons; Supplement Table 1). Therefore, properly accounting center bias, which is strong in these data and likely due to photographer bias (photographs were shot by humans who tend to place the most interesting objects at the center; another reason might be the viewing strategy), already negates the object-based hypothesis. ITTI98 scores better than ITTI on this particular comparison to human gaze data, thanks to the smoother maps of the ITTI98 model (See Figure S2). A simple sum of the object map and the ITTI* model (i.e., adding maps) yields sAUC of 0.576 ± 0.069 which is significantly above the ITTI* alone (t test; $p = 2.23 \times 10^{-6} < 0.05$) as well as the object-map alone (t test; $p = 0.0048 < 0.05$). Thus, object information helps fixation prediction (i.e., provides an independent source of information than what is conveyed by the ITTI* model), but alone does not perform significantly above saliency. See Supplement for reasons why center bias was not correctly addressed in Einhäuser et al. (2008).

Our second analysis regards using different saliency models as representatives of the saliency hypothesis. Please see Borji and Itti (2013) for a review of saliency models. Every difference between the object-map model and other models is significant (except the ITTI* model) using the sAUC score, with the object-map model being significantly above the very sparse ITTI model but significantly below the smoother ITTI98 and all other tested models. Performances of these four models: adaptive whitening saliency (AWS) (Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2012), attention based on information maximization (AIM) (Bruce & Tsotsos, 2009), ITTI98, and MEP (mean eye position), in order, are: 0.647 ± 0.084, 0.637 ± 0.077, 0.590 ± 0.079, and 0.491 ± 0.059. The human interobserver model (IO), a smoothed map built from fixations of other subjects over the same image, achieves the highest score of 0.803 ± 0.111 (mean $\pm$ standard deviation). The normal random model, a central Gaussian with the sigma of $5.85° \times 4.39°$ (Figure 1A; inset), is discounted using the sAUC, scoring near 0.5 (and likewise for the uniform random model). Note, in particular, how MEP scores at the chance level.

Our third analysis regards model parameters and their influence on the accuracy of the object-map hypothesis. We take a closer look at model differences by systematically investigating two parameters that impact scores: (a) center bias in a model (as opposed to in human data) by explicitly adding center preference to a model prediction map using three central Gaussian kernels with increasing sigmas $1.45° \times 1.1°$, $4.35° \times 3.3°$, and $7.25° \times 5.5°$. This analysis was prompted by several models which add a central Gaussian to maps and which tend to correlate strongly with human fixations, and it is complementary to our first analysis of center bias in scoring metrics above. (b) Smoothing by convolving the prediction map of a model with six variable-size Gaussian kernels ($0.28° \times 0.28°$, $0.86° \times 0.85°$, $1.43° \times 1.41°$, $2° \times 1.98°$, $2.58° \times 2.55°$, and $3.15° \times 3.1°$). We chose six models for this analysis: AWS, ITTI, ITTI98, normal random, ITTI*, and the object map. Figure 1B shows a sample image and its corresponding prediction maps with added center bias and with smoothing (left and right columns, respectively).

By adding center bias (Figure 1C; left panel), there is no significant difference between the object-map model and the ITTI* model (similar to Figure 1A; Bonferroni-corrected t test, $\alpha = 0.05/5 = 0.01$). The ITTI98 model is significantly above the object-map model only using the first Gaussian kernel. The object-map model is significantly above the ITTI model in all cases ($p = 3.007 \times 10^{-5}$, $p = 0.00016$, $p = 9.793 \times 10^{-5}$; Bonferroni-corrected t test). The accuracy of the normal random model does not increase with further adding center-bias and is not significantly better than chance. The AWS model is significantly above the object-map model using all three Gaussian kernels.

With smoothing (Figure 1C; right panel), we observed an interesting pattern. With small amounts of smoothing (first two Gaussian kernels), the object-map model is significantly better than the ITTI model. This difference is not statistically significant using the third Gaussian kernel ($p = 0.0850$). To further investigate this, we smoothed saliency maps more with larger Gaussian sizes (fourth, fifth, and sixth Gaussians). Accuracies of these two models (object map and ITTI) become closer to each other and there is no significant difference between them anymore. Interestingly, with mild amounts of smoothing, there is no significant difference between the object-map and ITTI* models but with further smoothing, the ITTI* model outperforms the object map significantly. The AWS and ITTI98 models score significantly higher than the object-map model using sAUC with all Gaussian kernels.

In summary, by introducing perturbations in Einhäuser et al.'s (2008) original analysis in three directions, (a) evaluation score and how it may be affected by center bias; (b) selected model; and (c) smoothness of saliency maps and object maps; Figure S3, we find that the conclusion of Einhäuser et al. is fragile: It is negated in a vast majority of the perturbation cases we examined—and especially in the case that best captures the state of the art (sAUC score, AWS model, any added central Gaussian or smoothing). Thus, contrary to Einhäuser et al.'s claim, we conclude that objects do not predict fixations better than early saliency (although objects score above chance, suggesting that they still play a role in guiding attention). Our results support that early image-based

representations based on spatial outliers guide attention more strongly than object representations in free viewing of pictures of natural scenes.

Acknowledgments

This work was supported by the National Science Foundation (grant number CMMI-1235539), the Army Research Office (W911NF-11-1-0046 and W911NF-12-1-0433), and U.S. Army (W81XWH-10-2-0076). The authors affirm that the views expressed herein are solely their own and do not represent the views of the United States government or any agency thereof. We would like to thank John Shen for his comments on the manuscript.

Commercial relationships: none.
Corresponding author: Ali Borji.
Email: borji@usc.edu.
Address: Department of Computer Science, University of Southern California, Los Angeles, CA.

References

Borji, A., & Itti, L. (2013). State-of-the-art in Visual Attention Modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207.

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, *22*(1), 55–69.

Bruce, N., & Tsotsos, J. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*(3):5, 1–24, http://www.journalofvision.org/content/9/3/5, doi:10.1167/9.3.5. [PubMed] [Article]

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14):18, 1–26, http://www.journalofvision.org/content/8/14/18, doi:10.1167/8.14.18. [PubMed] [Article]

Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing, 30,* 51–64.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20,* 1254–1259.

Shore, S., Tillman, L., & Schmidt-Wulen, S. (2004). *Stephen shore: Uncommon places: The Complete works*. New York: Aperture.

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45,* 643–659.