Predicting Perceptions: Proceedings of the 3rd International Conference on Appearance, pp. 122-124, Edinburgh, UK, ISBN 978-1-4716-6869-2, April 2012.

Towards Object-Based Saliency

M. Dziemianko
Institute for Language, Cognition and Computation
University of Edinburgh
Edinburgh, UK
m.dziemianko@sms.ed.ac.uk

A. D. F. Clarke
Institute for Language, Cognition and Computation
University of Edinburgh
Edinburgh, UK
a.clarke@ed.ac.uk

F. Keller
Institute for Language, Cognition and Computation
University of Edinburgh
Edinburgh, UK
keller@ed.ac.uk

ABSTRACT
Eye movements during scene comprehension can be seen as a series of decisions where and when to look. There has been substantial work towards determining likely fixation locations and the best known methods involve the computation of saliency maps which assign "interestingness" values to each coordinate of analyzed image. However these methods are usually purely bottom-up and only consider low-level visual features. In this paper we present preliminary work towards creating a computational framework based on the alternative cognitive relevance hypothesis.

Keywords
Saliency, Cognitive Relevance hypothesis, Eye tracking

1. INTRODUCTION
Sequences of fixations are important indicators of the processing performed by attentional systems and a number of models have been proposed to predict eye-movements during scene comprehension. They can be broadly divided into two categories. The first one consists of bottom-up models exploiting low-level visual features to predict areas likely to be fixated. A number of studies have shown that certain features and their statistical unexpectedness attract human attention [1]. The best-known example is Itti and Koch's model [6] which builds saliency maps based on color, orientation, and scale filters inspired by neurobiological results. The second group of models assume the existence of top- down supervision of attention which contributes to the selection of fixation targets. A number of models have been proposed to capture context effects on visual attention; a prominent example is the Contextual Guidance Model [9], which combines bottom-up saliency with a prior encoding global scene information.

On the other end of continuum there is the cognitive relevance hypothesis which holds that fixations are directed according to the requirements of the current task [4]. Although the attentional processing and fixation locations are generated from visual input they are assumed not to be ranked on basis of saliency, but rather based on their relevance to the current task. There is considerable experimental evidence supporting this hypothesis [5, 7].

Predicting Perceptions: The 3rd International Conference on Appearance, 17-19 April, 2012, Edinburgh, UK. Conference Proceedings Publication ISBN: 978-1-4716-6869-2, Pages: 122-124

These two views – visual salience and cognitive relevance – differ in the representation over which attentional selection is made. Saliency requires low-level image based representation, while cognitive relevance framework needs higher-level object based representation. To the best of our best knowledge, there is no computational model based on the cognitive relevance hypothesis. The closest work is perhaps that of Spain and Perona [8], who developed a model for object importance (defined as the probability of an object in a scene being named) which includes several features derived from saliency maps. Related work [2] shows that the location of objects in a scene is a better predictor of fixations than low- level (pixel based) saliency. However, it seems clear that some objects will naturally attract more fixations than others. The aim of this paper is to investigate the feasibility of a computational model of attentional selection based on the cognitive relevance hypothesis – addressing the question why some objects are fixated more than others.

2. MODEL
Our model is based on a simplified Factored Shapes and Appearances (FSA) representation [3]. The central assumption of the representation is that the pixels corresponding to each object have been generated by W fixed Gaussians in a fea- ture space (we found Lab to be the most effective in our initial experiment).

In first phase the means μ and covariances Σ of these Gaussians are extracted by fitting a Gaussian Mixture Model (GMM) with W components over all pixels in the image. At this stage object boundaries and locations are ignored. In subsequent step, pixels are clustered into W clusters ac- cording to the associated GMM components by selecting component $w^$ that maximizes probability of a pixel being drawn from the Gaussian distribution with mean μ_w and covariance Σ_w :

$$\hat{w} = argmax_w \frac{1}{(2\pi)^{k/2}|\Sigma_w|^{1/2}} e^{-\frac{1}{2}(x-\mu_w)^T\Sigma^{-1}(x-\mu_w)} \quad (1)$$

where x is feature vector representing a pixel, while k dimensionality of this vector. The value of W was chosen to be 15 (following [3]).

The final step of the first phase consists of computing global histograms H of the pixel assignments $w^$. Each histogram is then normalized, dividing each bucket count by total number of pixels, so that it represents proportions of pixels belonging to each cluster rather than absolute counts. The whole process is shown in Figure 1.

Table 1: Results of a preliminary evaluation. Our model, object based saliency, and the mean saliency value of [9] clearly outperform the baseline (cross- trial reference). The average area occupied by the top five objects is indicated next to each model.

Model	Hit rate	Average area
Object-based saliency	14.52%	6.84%
Mean [9] saliency	15.78%	5.13%
Cross-trial reference	6.98%	6.84%

The saliency map is created in the second phase. At this stage the model assumes that the image is fully annotated (i.e., boundaries for each object within the scene are provided). For each of the objects o_i an additional histogram h_i is computed considering only the pixels and their assignments w^ within the boundaries of the object. The histogram h_i is also normalized by the total number of pixels within the object. Histograms computed this way are distributions over the different pixel types present in the scene.

In the following step an interestingness value Ii is assigned to each object o_i. In preliminary experiments, we assigned the Kullback-Leibler (KL) divergence between local (object) pixel distribution hi and the global distribution H:

$$I_i = D_{KL}(h_i||H) = \sum_{w=1}^{W} h_i(w) \log \frac{h_i(w)}{H(w)} \qquad (2)$$

The KL divergence measures the expected number of extra bits required to encode samples from h_i when using a code based on H; intuitively, it represents how different the object is from its surroundings (and thus interesting), with a larger value meaning more interesting. The saliency map is constructed by filling the area corresponding to each object with the interestingness value Ii assigned to it.

3. EVALUATION

Our initial experiment was conducted using eye-tracking data. Our dataset consisted of 100 fully annotated photo realistic scenes. The eye-movements of 17 participants were recorded using an Eyelink II eye-tracker in an object naming task. Each participant was presented with 72 different scenes from the dataset and an additional 64 fillers, totalling 1088 trials usable for evaluation. In our experiment we considered only first five fixations (except for the initial fixation on the center cross), resulting in 5937 fixations.

For evaluation, we selected the five objects with the highest interestingness value in each scene. The results calculated per subject were analysed with an ANOVA. Table 1 presents the average fraction of fixations falling onto the selected regions calculated on per-participant basis.

For comparison we constructed "object-based" saliency maps using the formulation of salience proposed by [9]. We combined the pixel saliency values by calculating the mean, max or median value over the object area. The mean performed best, with median slightly worse, and max unusable due to large areas being assigned the maximum possible value. As a baseline we cross-applied selected areas to other trials with an average result of just under 7% over 100 rounds.

The model performs clearly above chance with over 14% of fixations falling onto selected regions. This is significantly better than the baseline ($F(1, 16) = 281.30$, $p < 0.001$), and not significantly different from the performance of transformed pixel-based saliency ($F(1, 16) = 2.21$, $p = 0.147$).

The analysis of the results obtained with the two methods shows that, despite the similar performance, the set of predicted fixation locations is different for the two models. Only about 35.72% of the fixations found by object-based saliency were also found with transformed pixel saliency. It is important to note that the object-based saliency measure presented in this paper only uses color information, while traditional salience measures make extensive use of contrast and orientation features.

4. CONCLUSION AND FURTHER WORK

In this paper we presented an initial attempts at building a computational model of cognitive relevance. The results of a preliminary evaluation are encouraging, although they do not yet match the performance of a bottom-up approach.

Future work will focuses on completing the model by introducing a more appropriate pixel vector representation and interestingness metric. We will also work towards resolving problems arising from the relations between the objects in a scene such as inclusion, occlusion, crowding, and from the compound structure of the objects themselves. Note that we do not discount bottom-up approaches, aim towards a combined model that takes into account both object-level features and low-level visual cues. Finally we want to incorporate semantic and contextual information and explore the interaction between scene gist and object importance in the proposed framework.

5. REFERENCES

[1] N. Bruce and J. Tsotsos. Saliency based on information maximization. In Advances in Neural Information Processing Systems 18, pages 155–162. Cambridge, MA: MIT Press, 2006.
[2] Wolfgang Einhauser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. Journal of Vision, 8(14):1–26, 11 2008.
[3] S.A. Eslami and C. Williams. Factored shapes and appearances for parts-based object understanding. In Proceedings of the British Machine Vision Conference, pages 18.1–18.12. BMVA Press, 2011.
[4] J. Henderson, J. Brockmole, and M. Castelhano. Visual saliency does not account for eye-movements during visual search in real-world scenes. Eye movements research: insights into mind and brain, 2007.
[5] J.M. Henderson, G.L. Malcolm, and C. Schandl. Searching in dark: Cognitive relevance drives attention in real-world scenes. Psycholomic Bulletin & Review, 16:850–856, 2009.
[6] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11):1254–1259, 1998.
[7] A. Nuthmann and J.M. Henderson. Object-based attentional selection in scene viewing. Journal of Vision, 10(8), 2010.
[8] Merrielle Spain and Pietro Perona. Measuring and predicting object importance. International Journal of Computer Vision, 91:59–76, 2011.
[9] A. Torralba, A. Oliva, M.S. Castelhano, and J.M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. Psychological Review, 113:766–786, 2006.

Predicting Perceptions: Proceedings of the 3rd International Conference on Appearance, pp. 122-124, Edinburgh, UK, ISBN 978-1-4716-6869-2, April 2012.

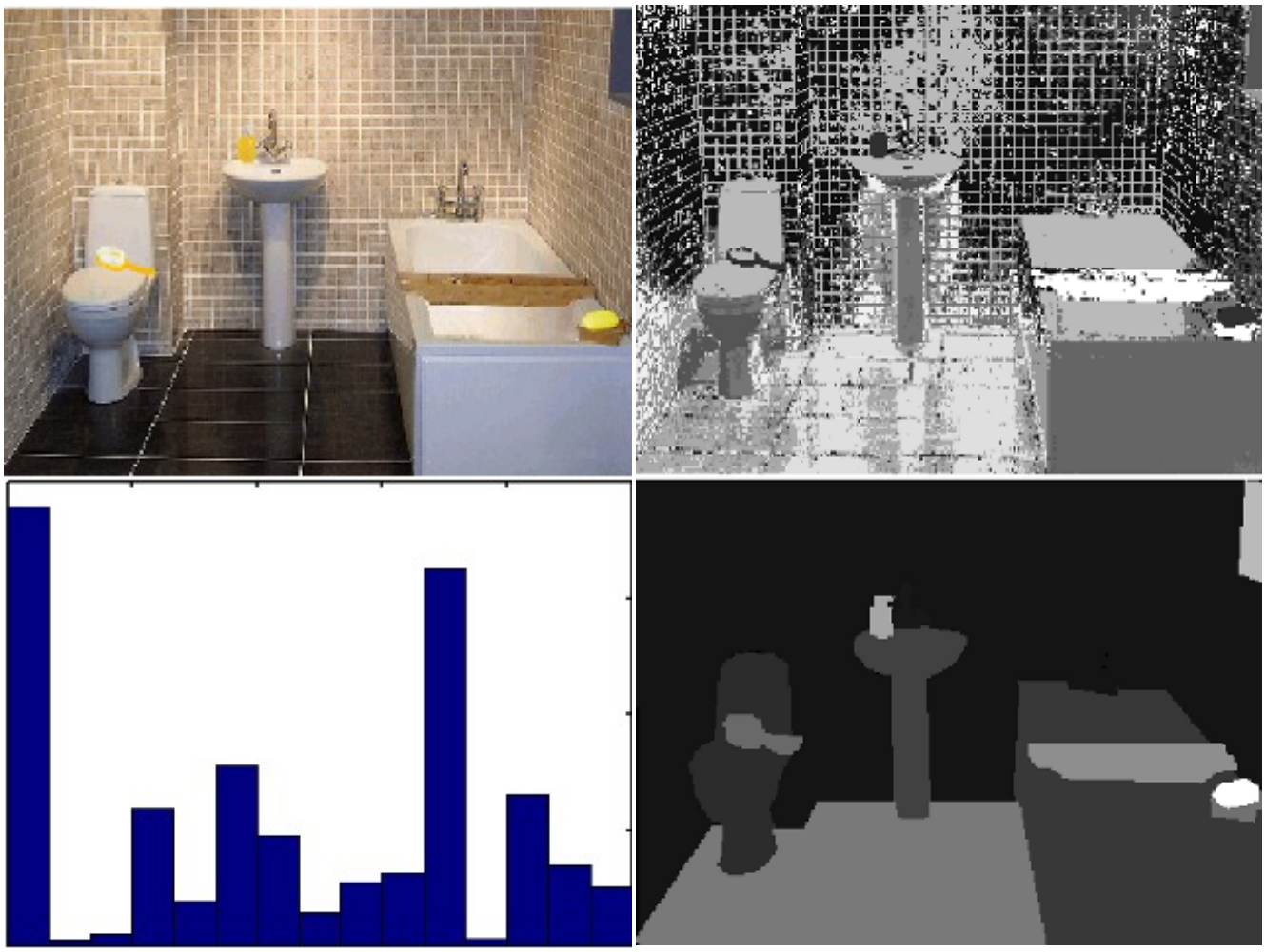

Figure 1: Calculation of the global histogram H : from left to right: original image, clustering of pixels to different Gaussian components, histogram of the assignments, and objects interestingness map