Object-based Saliency as a Predictor of Attention in Visual Tasks

Michal Dziemianko (m.dziemianko@sms.ed.ac.uk)
Alasdair Clarke (a.clarke@ed.ac.uk)
Frank Keller (keller@inf.ed.ac.uk)
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, UK

Abstract

The top-down guidance of visual attention is an important factor allowing humans to effectively process incoming visual information. Our understanding of the processes governing attention is not complete, with growing evidence for attention selection based on *cognitive relevance*. In this paper, we investigate whether models for *salient object detection* from computer vision can be used to predict attentional shifts in visual tasks. Our results show that the object-based interpretation of saliency provided by these models is a substantially better predictor of fixation locations than traditional pixel-based saliency.

Keywords: eye-tacking; saliency; visual attention.

Introduction

Virtually every human activity occurs within a visual context and many tasks require visual attention in order of be successfully accomplished (Land & Hayhoe, 2001). When processing a visual scene, humans have to localize objects, identify them, and establish their spatial relations. The eye-movements involved in this process provide important information about the cognitive processes that unfold during scene comprehension.

A number of models have been proposed to predict eye-movements during scene processing and they can be broadly divided into two categories. The first category consists of bottom-up models that exploit 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 (e.g., Bruce & Tsotsos, 2006). Moreover, low-level features are believed to contribute to the selection of fixated areas, especially when the visual input does not provide useful high-level information (Peters et al., 2005). These experimental results are captured by models that detect salient areas in the visual input and use them to predict attention. The best-known example is the model of Itti et al. (1998), which builds a pixel-based saliency map using color, orientation, and scale filters inspired by neurobiological results.

The second group of models assumes that top-down supervision of attention contributes to the selection of fixation targets (e.g., Torralba et al., 2006). Various types of such supervision have been observed experimentally. Humans show the ability to learn general statistics of the appearance, position, size, spatial arrangement of objects, and their relationships (e.g., Zelinsky, 2008). They also exploit visual memory during scene comprehension tasks (e.g., Shore & Klein, 2000). Moreover, studies such as those of Chun & Jiang (1998) show that participants benefit from learning spatial arrangement of the objects in consecutive searches. Theoretically, such results can be accommodated by the Cognitive Relevance Framework (Henderson et al., 2009), which assumes that attention is allocated to locations that are cognitively relevant for the task performed.

Cognitive relevance predicts that objects should have a privileged status in visual processing, which is in line with experimental evidence suggesting that the allocation of attention is object-based rather than pixel-based. For example, Henderson et al. (2007) argue that saliency does not account for fixated areas in visual search, while Nuthmann & Henderson (2010) show that the preferred fixation point or *landing position* is the center of an object: fixations are distributed normally around an object's center of mass, where the spread might be explained by oculomotor errors. Consistent with this, Einhauser et al. (2008) show that the position of objects is a better predictor of fixations than early saliency in tasks such as artistic evaluation, analysis of content, and search.

An alternative view on saliency comes from the computer vision literature, which deals with task of salient object detection: the objects that are perceived by humans as visually most interesting have to be separated from the background. Typically this involves image segmentation and the calculation of visual features in order to select pixels belonging to salient objects. In this context, saliency is a feature of an object, rather than an early pixel-based attractor of attention.

In this paper, we investigate the extent to which methods proposed for salient object detection can be applied to the prediction of fixations. We are not concerned with the prediction of salient image patches, but rather with the selection of objects that are likely to be fixated. This approach allows us to develop computational models of attentional selection based on cognitive relevance defined over objects (Henderson et al., 2007, 2009). We compare the performance of this approach to traditional models which predict fixation locations using pixel-based saliency maps.

Background

As discussed above, there is experimental evidence for the object-based allocation of attention. Additionally, some objects seem to inherently attract more attention than others, a fact that has been conceptualized using *proto-objects*: pre-recognition entities that draw attention (Rensink, 2000). Proto-objects have been incorporated into saliency-based models (Walther & Koch, 2006) and have also been applied

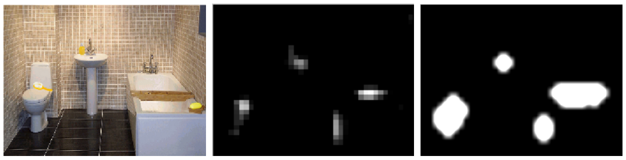

Figure 1: Example of proto-objects extracted from an image using the model of Walther & Koch (2006). From left to right: original image, saliency map computed according to Itti et al. (1998), proto-object mask. The salient patches, and hence the proto-objects, do not necessarily correspond to the real objects in the scene.

in robotics to create attentional systems for virtual and physical agents (see e.g., Yu et al., 2010). These models perform image segmentation to identify proto-objects: the image is divided into a collection of regions that correspond to areas enclosed by constant, high saliency values. Figure 1 shows an example of such proto-objects extracted from an image using the model of Walther & Koch (2006).

While Walther and Koch's model is conceptually interesting, its cognitive status is questionable, as there is evidence that it does not predict fixation locations well (Nuthmann & Henderson, 2010). Alternative models of attention selection based on objects rather than proto-objects have been proposed in computer vision. For example, the work of Liu et al. (2011) focuses on detecting objects annotated by people as salient. These models use machine learning techniques to compute which arrangements of visual features such as center-surround histograms, orientation, scale are perceived as salient. However, in a computer vision context, attentional selection is regarded merely as an engineering task: the aim is to identify areas matching pre-annotated training data, rather than to gain a greater understanding of human behavior.

Models

We implemented and evaluated three models for salient object detection. Throughout our work we assume that the images are fully annotated with object boundaries, therefore the problem of segmentation and separation of objects from the background does not need to be solved within the models. This assumption makes it possible to evaluate object-based saliency models separately from image segmentation algorithms, which can vary widely in their performance.

A. Conversion of Standard Saliency

Standard, pixel-based saliency is the baseline against which we evaluate object-based models. The baseline model we use is Torralba et al.'s (2006), which approximates saliency as the probability of the local images feature L in a given location based on the global distribution of these features:

$$p(L) \quad \propto \quad e^{-\frac{1}{2}[(L-\mu)^T\Sigma^{-1}(L-\mu)]} \tag{1}$$

Here, μ is the mean vector and Σ the covariance matrix of the Gaussian distribution of local features estimated over the currently processed image. The local features are computed as a set of Steerable pyramid responses computed over three color channels for six orientations and four scales, totaling 72 values at each position.

Based Torralba et al.'s model, we can define a group of models which convert pixel-based saliency values to object-based salience scores. Such a conversion can be performed by computing functions such as the *maximum*, *mean*, *median*, or *mode* of the pixels that make up an object. Examples for the use of this method exist in the literature (e.g., Spain & Perona, 2011), with maximum and mean being common. These models will be referred to as *converted* in this paper.

B. Liu et al. Features

Liu et al. (2011) describes a system for salient object detection based on conditional random fields, which simultaneously segments pixels into areas corresponding to objects and computes the pixel's salience. The model is based on three feature channels – contrast, center-surround histograms and spatial color – which are described below. The salience of a pixel is defined to be the a weighted sum of these three feature maps, while the salience of an object is defined as the sum over all pixels within the object's boundary. The full specification of our implementation of Liu's model can be found in Dziemianko (2013). Examples of the feature channels are given in Figure 3.

Multiscale Contrast Contrast is one of the most commonly used features in saliency models and is implemented over a multiscale Gaussian pyramid. In each layer of the pyramid, the contrast at pixel (x,y) is defined to be the mean squared difference of the intensity of pixel at (x,y) and its adjacent neighbors. The multiscale contrast for $I(x,y)$ is then taken to be the sum over the layers of the corresponding pyramid. This has the effect of approximating human receptive field by highlighting high-contrast boundaries while omitting homogeneous regions within objects.

Center–Surround Histograms One of the weaknesses of previous measures of visual salience is that, due to their reliance on high-contrast center-surround features, they tend to emphasis the boundaries of objects while giving very low scores to pixels within an object's boundary (see Figure 2). To tackle this issue, Liu et al. (2011) propose to use region-based features in addition to the center-surrounds described above. These are computed by considering the histogram of colors within an object's bounding box, and comparing it with a surrounding region of equal area (see Figure 2). The χ^2 metric is used to measure the distance between histograms and the full details on how these regions are constructed can be found in Liu et al. (2011).

Color Spatial Distribution The last feature used by Liu et al. (2011) is the spatial color distribution, motivated by the observation that salient objects are less likely to contain colors that are distributed widely throughout the image. A simple method for quantifying this is to compute the spatial vari-

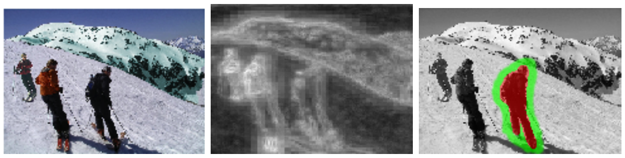

Figure 2: An example of high saliency values being assigned to object boundaries due to its reliance on high-contrast features. From left to right: original image, traditional saliency, an object (red) and its surrounding area (green).

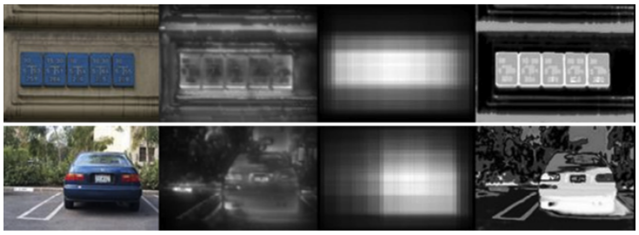

Figure 3: Examples of features from Liu et al. (2011). From left to right: original image, multiscale contrast, center-surround histogram, color spatial distribution (image from Liu et al. (2011) with modifications).

ance of color. This involves representing the distribution of colors contained in the image by a Gaussian mixture model. We then carry out soft assignment: for each of these Gaussians, c, we then calculate $p(c|I(x,y))$, the probability of assigning pixel $I(x,y)$ to the Gaussian $\mathcal{N}(\mu_c,\Sigma_c)$. Using these we can calculate the weighted mean and variance for each color component along the horizontal axis:

$$M_h(c) = \frac{1}{|X|_c}\sum_x p(c|I(x,y))\cdot x \quad (2)$$

$$V_h(c) = \frac{1}{|X|_c}\sum_x p(c|I(x,y))\cdot |x-M_h(c)|^2 \quad (3)$$

where $|X|_c = \sum_x p(c|I_x)$. The vertical spatial variance, V_v, is computed in the same way, and $V(c)$, the spatial variance of each color component, is then simply defined as:

$$V(c) = V_h(c)+V_v(c) \quad (4)$$

Finally, the feature function $f_s(x,y)$ is defined as:

$$f_s(x,y) \propto \sum_c p(c|I(x,y))\cdot(1-V(c)) \quad (5)$$

C. Color-component Histograms

In addition to the models described above, we have implemented our own model based on a simplified *factored shapes and appearances* representation (Eslami & Williams, 2011). This model shares some characteristics with the spatial color distribution described above, as it assumes that the pixels corresponding to each object have been generated by a number of Gaussians in a feature space (we found *Lab*-space to be the most effective). However, it performs a comparison of histograms of color cluster assignments within the object and its surrounding area.

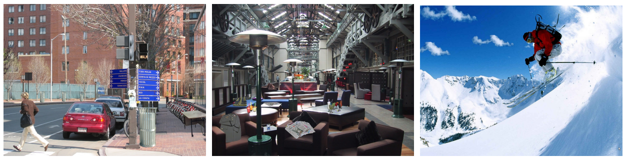

Figure 4: Examples of scenes used in the visual counting experiment. Targets on the images on the left and in center are *man*, while for the image on the right it is *goggle*.

In the 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. Similar to Eslami & Williams (2011), we use $W = 15$ Gaussians. At this stage object boundaries and locations are ignored. In the subsequent step, pixels are clustered into W clusters according to the associated GMM components by selecting a component, $\hat{w}$, that maximizes the probability of a pixel being drawn from the Gaussian distribution. The final step of the first phase consists of computing global histograms H of the pixel assignments $\hat{w}$ representing the proportion of pixels belonging to each cluster.

The saliency scores are computed 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 object in the scene, we calculate the histogram of pixel assignments over the pixels within the object's boundary. We then define an *interestingness* value for each object as the Kullback-Leibler (KL) divergence between the local (object) pixel distribution and the global distribution H. Intuitively, I represents how different the object is from its surroundings and thus interesting.

Evaluation

Method

We evaluate the performance of the models discussed on eye-tracking data collected in a visual counting and an object naming task. In the visual counting task, 25 participants were asked to count the number of occurrences of a cued target object, which was either animate (e.g., man) or inanimate (e.g., goggle). The data set consisted of 72 fully object-annotated photo-realistic scenes (both indoor and outdoor), with total of 1809 polygons with mean of 25.12±11 and a median of 25 polygons per image, containing zero to three instances of the target object. The data was collected using an Eyelink II head-mounted eye-tracker with a sampling rate of 500 Hz. The images were displayed with a resolution of 1024 × 768 pixels, subtending a visual field of approximately 34 × 30 degrees. The data set consists of 54,029 fixations. Figure 4 presents examples of scenes used in the experiment.

The object naming dataset (Clarke et al., under revision) contains data collected during an object naming experiment.

Figure 5: Examples of stimuli used in the object naming experiment. Typical responses are: *cars, crossing, person* for the left, *bench, man* for the center, and *barbecue, charcoal, chimney* for the right image.

The stimuli consists 132 fully object-annotated images with a total of 2,858 polygons with mean of 14.2 ± 5 and a median of 26 polygons per image. The images were presented to 24 participants after the task was explained using written instructions. Before each trial, participants were asked to fixate a central cross. The image was then displayed for 5000 ms, followed by a beep, after which the participants named objects present in the scene. The image was displayed until the participant finished the trial. Image presentation and apparatus were the same as in the visual counting data set. A total of 2,904 usable trials were collected, resulting in 88,371 fixations. Examples of images used as stimuli are shown in Figure 5.

Analysis

As well as the models described above, we test two baselines that do not use saliency in any form. The first one weights objects by their Euclidean distance from the center of the image, normalized by object area. This approach is inspired by experimental evidence of center bias in scene viewing (e.g. Tatler, 2007), and will be referred to as *center bias*.

Secondly, based on the findings of Nuthmann & Henderson (2010), we also include a baseline that predicts fixations by selecting object centers. In this case, a map is built as a sum of Gaussians centered on the bounding boxes of the object in the image. The covariances of the Gaussians are dependent on object's size, with a factor fitted using 10-fold cross-validation to avoid overfitting the datasets. This baseline is referred to as *object overlay*.

In the Results and Discussion section below, we show how the different models perform by using receiver operating characteristic (ROC) plots, which indicate the sensitivity (i.e., true positive rate vs. false positive rate) of a classifier as its discrimination threshold varies. Moreover, in order to statistically compare model performance, we calculate the area under the ROC curve (AUC) of each participant. The AUC measures the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.[1] We submit the AUC means to an ANOVA analysis to compare the performance of the different models pairwise, e.g., *saliency* against *converted (mean)*. For standard pixel-based saliency, the ROC curve is constructed by thresholding the saliency values to select the desired proportion of pixels. The ROC plots for object-based models can not be constructed this method as it would not ensure that entire objects are selected. Instead, an increasing number of objects with the highest saliency values is iteratively selected, and their total area is plotted in the ROC curve. The ROC curves constructed this way are incomplete, representing only selection of up to about 50% of the image area. Constructing ROC plot for larger selections would result in significant discontinuities due to the fact of all small objects being already selected and essentially only large objects corresponding to surfaces such as floor, sky, or wall being left.

Model	Obj. counting	Obj. naming
Saliency	61.66	55.87
Object overlay	63.60	59.78
Center bias	68.02	69.17
Converted (max)	55.27	64.66
Converted (mean)	70.44	68.65
Liu et al. 2011 features	66.67	67.42
Color-component hist.	66.73	67.40

Table 1: Estimated percentage areas under the ROC curves presented in Figure 6.

Results and Discussion

The results are presented in Figure 6. The ROC curves show that selection based on object overlay is better than saliency for thresholds smaller than 40%. Object-based saliency models in turn outperform object overlay. Center bias turns out to be a very competitive baseline, which is only matched by converted (mean).

An analysis of the areas under the ROCs, summarized in Table 1, confirm these observations. The ANOVAs reveal that for both datasets, object position overlay is significantly better than saliency with $F(1,24) = 9.27, p < 0.005$ for object counting, and $F(1,23) = 9,84, p < 0.005$ for object naming.

The calculation of area under ROC curve for object-based models is not trivial due to the discontinuity of the plot. We estimated the AUC by interpolating the missing values.[2] The analysis of the interpolated curves shows that for both datasets, object-based selection is superior to traditional saliency, and to object overlay. These differences are statistically significant, for example converted (mean) is better than saliency with $F(1,24) = 165.60, p < 0.001$ for counting and $F(1,23) = 279.30, p < 0.001$ for naming; for color histogram the values are $F(1,24) = 34.67, p < 0.001$ and $F(1,24) = 227.40, p < 0.001$ respectively.

The pattern for Converted (max) is more complicated. On the naming data, it is significantly better than saliency ($F(1,24) = 132.10, p < 0.001$), but not as good as any of the other methods. On the counting data, it is significantly weaker than standard saliency ($F(1,23) = 245.70, p <$

[1] The AUC is equivalent to a Wilcoxon test of ranks.

[2] The discontinuities were interpolated by plotting linear segments between end points of the ROC curve.

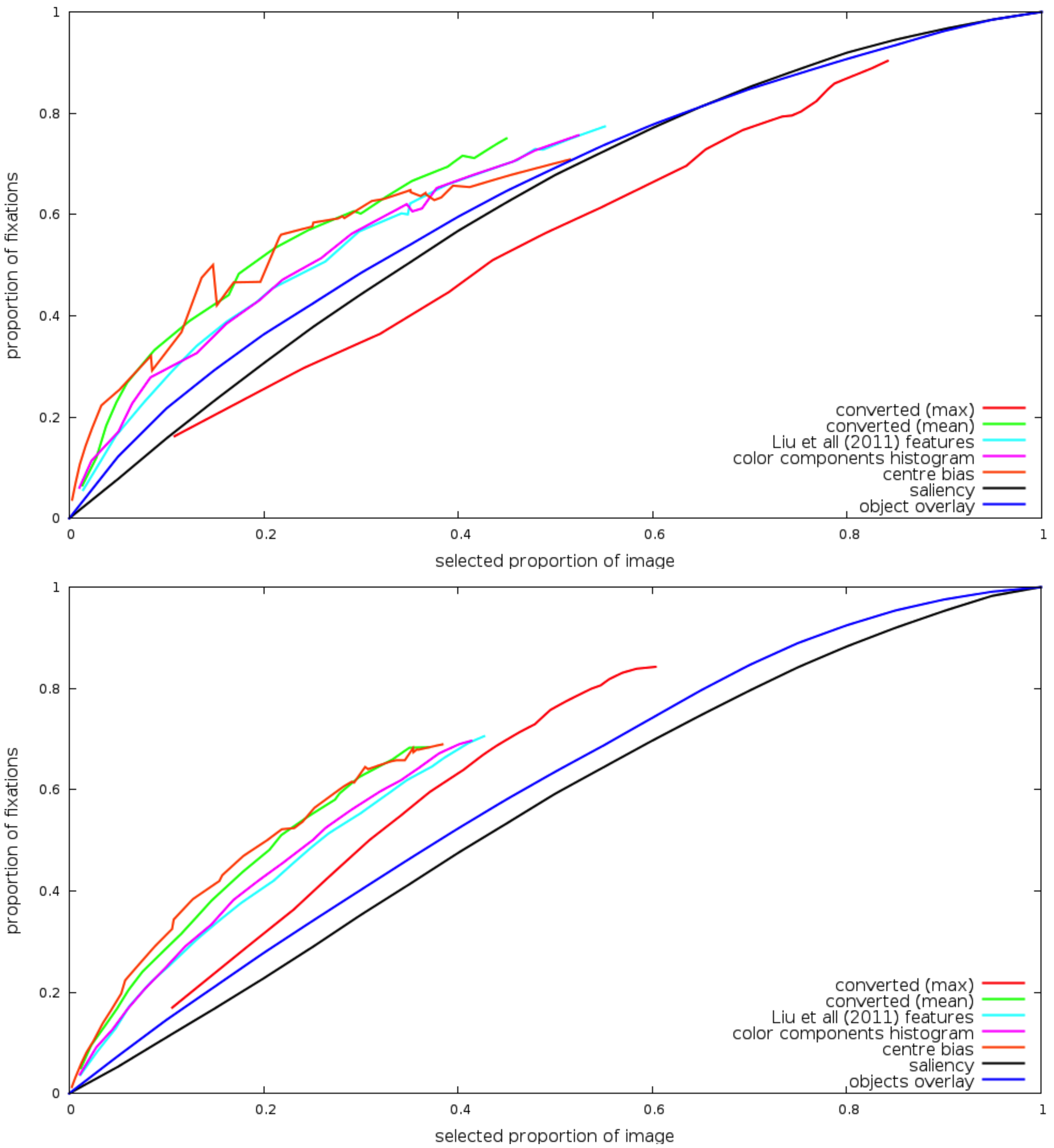

Figure 6: Performance of object-based selection of fixation locations on the Visual Count (top) and Object Naming (bottom) datasets. Note that traditional saliency and object-based models cannot be compared directly due to differences in the selection method, see text for details.

0.001), operating around chance level. This can be explained by the fact that saliency is sensitive to high contrast edges, usually corresponding to object boundaries. As such, the highest saliency values corresponding to the object might not fall within the object, but rather belong to its neighbors.

A surprising results is that object-based selection does not outperform selection based on center bias. However, closer investigation of the object rankings based on center bias and

Converted (mean) reveals that the average correlation coefficient between the respective rankings is only 0.50 for the naming and 0.43 for the counting data. This indicates that different sets of objects are selected by the two model for a given threshold, accounting for different subsets of fixations. A combined model would be a promising next step.

Conclusion

In this paper, we discussed the issue of objectness and its relation to the allocation of visual attention. We demonstrated that it is possible to develop object-based version of saliency. Object-based saliency is not calculated as a value for each of the image pixels (or coordinates), but rather over an area within the boundaries of an object. In this approach, saliency is treated as a feature of an object, similar to other features such as position. This approach is compatible with theories assuming an object-based allocation of attention, such as the Cognitive Relevance Framework (Henderson et al., 2009).

The evaluation we presented used an object counting and an object naming data set. In spite of both of these tasks being object-centric by definition, we believe that our results generalize to other experimental tasks. Such tasks are often either object-centric as well (e.g., visual search), or evidence exists that attentional access is object-based even if the task defined in terms of objects (e.g., in aesthetic judgment or interestingness judgment, see Nuthmann & Henderson 2010; Einhauser et al. 2008). Indeed it was shown that visual attention is object-based during everyday interaction with the surrounding world (Land et al., 1999). Finally, it has been suggested that *free viewing* does not mean that viewers look at images without any task constraints, but rather with constraints to which experimenters do not have access (see Tatler et al., 2011, for further discussion).

Even though the intuition that salience is a property of objects has been utilized before, we are not aware of any extensive experimental study aiming to investigate whether object-based saliency and techniques used to detect salient objects in computer vision can reliably predict human fixations. We showed that the prediction of fixations based on objects and their visual features is not only possible, but superior to standard saliency. However, using the maximum value of saliency within an object was not confirmed as a reliable predictor of whether object is going to be fixated, which is a important result considering the popularity of this feature in previous modeling studies.

Acknowledgments

The support of the European Research Council under award number 203427 “Synchronous Linguistic and Visual Processing” is gratefully acknowledged.

References

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In *Advances in Neural Information Processing Systems 18*, (pp. 155–162). Cambridge, MA: MIT Press.

Chun, M., & Jiang, Y. (1998). Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, *36*, 28–71.

Clarke, A., Coco, M., & Keller, F. (under revision). The impact of attentional, linguistic and visual features during object naming. In G. Zielinsky, T. Berg, & M. Pomplun (eds.), *Frontiers in Perception Science: Research Topic on Scene Understanding: Behavioural and computational perspectives*.

Dziemianko, M. (2013). *Modelling Eye Movements and Visual Attention in Synchronous Visual and Linguistic Processing*. Ph.D. thesis, University of Edinburgh.

Einhauser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*, 1–26.

Eslami, S., & Williams, C. (2011). Factored shapes and appearances for parts-based object understanding. In *Proceedings of the British Machine Vision Conference*, (pp. 18.1–18.12). BMVA Press.

Henderson, J., Brockmole, J., & Castelhano, M. (2007). Visual saliency does not account for eye-movements during visual search in real-world scenes. *Eye movements research: insights into mind and brain*.

Henderson, J., Malcolm, G., & Schandl, C. (2009). Searching in the dark: cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*, 850–856.

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.

Land, M., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, *41*, 3559–3565.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in control of activities of daily living. *Perception*, *28*.

Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., & Shum, H. (2011). Learning to detect salient objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*, 353–367.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *20*, 1–19.

Peters, R., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research*, *45*, 2397–2416.

Rensink, R. (2000). Seeing, sensing, and scrutinizing. *Vision Research*, *10-12*, 1469–1487.

Shore, D., & Klein, R. (2000). On the manifestations of memory in visual search. *Spatial Vision*, *14*, 59–75.

Spain, M., & Perona, P. (2011). Measuring and predicting object importance. *International Journal of Computer Vision*, *91*, 59–76.

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*, 4.

Tatler, B. W., Hayhoe, M., Land, M., & Ballard, D. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, *113*, 766–786.

Walther, D., & Koch, C. (2006). Modelling attention to salient proto-objects. *Neural Networks*, *19*, 1395–1407.

Yu, Y., Mann, G., & Gosine, R. (2010). An object-based visual attention model for robotic applications. *IEEE Transactions on Systems, Man, and Cybernetics: Cybernetics*, *5*, 1398–1412.

Zelinsky, G. (2008). A theory of eye movements during target acquisition. *Psychological Review*, *115*, 419–433.