Vision Research 107 (2015) 36–48

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Overt attention in natural scenes: Objects dominate features

Josef Stoll [a], Michael Thrun [a], Antje Nuthmann [b], Wolfgang Einhäuser [a,*]

[a] Neurophysics, Philipps-University Marburg, Germany
[b] School of Philosophy, Psychology and Language Sciences, Psychology Department, University of Edinburgh, UK

ARTICLE INFO

Article history:
Received 19 June 2014
Received in revised form 4 November 2014
Available online 3 December 2014

Keywords:
Attention
Fixation
Eye movements
Salience
Proto-objects
Natural scenes

ABSTRACT

Whether overt attention in natural scenes is guided by object content or by low-level stimulus features has become a matter of intense debate. Experimental evidence seemed to indicate that once object locations in a scene are known, salience models provide little extra explanatory power. This approach has recently been criticized for using inadequate models of early salience; and indeed, state-of-the-art salience models outperform trivial object-based models that assume a uniform distribution of fixations on objects. Here we propose to use object-based models that take a preferred viewing location (PVL) close to the centre of objects into account. In experiment 1, we demonstrate that, when including this comparably subtle modification, object-based models again are at par with state-of-the-art salience models in predicting fixations in natural scenes. One possible interpretation of these results is that objects rather than early salience dominate attentional guidance. In this view, early-salience models predict fixations through the correlation of their features with object locations. To test this hypothesis directly, in two additional experiments we reduced low-level salience in image areas of high object content. For these modified stimuli, the object-based model predicted fixations significantly better than early salience. This finding held in an object-naming task (experiment 2) and a free-viewing task (experiment 3). These results provide further evidence for object-based fixation selection – and by inference object-based attentional guidance – in natural scenes.

1. Introduction

Is attention guided by objects or by the features constituting them? For simple stimuli and covert shifts of attention, evidence for object-based attention arises mainly from the attentional costs associated with switching between objects as compared to shifting attention within an object (Egly, Driver, & Rafal, 1994; Moore, Yantis, & Vaughan, 1998). Such benefits extend to search in visual scenes with 3D objects (Enns & Rensink, 1991). For more natural situations, however, the question as to when a cluster of features constitutes an "object" does not necessarily have a unique answer (Scholl, 2001) and it may depend on the context and task. In the context of visual working memory, Rensink (2000) suggested that "proto-objects" form pre-attentively and gain their objecthood ("coherence") through attention. Extending the notion of objects to include such proto-objects, attention can be guided by "objects", even if more attentional demanding object processing has not yet been completed.

While for covert attention an object-based component to attention seems rather undisputed, for the case of overt attention, defined as fixation selection, in natural scenes two seemingly conflicting views have emerged, referred to as the "salience-view" and the "object-view". The "salience-view" states that fixated locations are selected directly based on a salience map (Itti & Koch, 2000; Itti, Koch, & Niebur, 1998; Koch & Ullman, 1985) that is computed from low-level feature contrasts. The term "salience" or "early salience" in this context is used in a restrictive sense to denote feature-based effects, and is thus not equivalent, but contained in "bottom-up", "stimulus-driven" or "physical" salience (Awh, Belopolsky, & Theeuwes, 2012). Put to the extreme, the salience-view assumes that these features drive attention irrespective of objecthood (Borji, Sihite, & Itti, 2013). The salience-view appears to be supported by the good prediction performance of salience-map models (Peters et al., 2005) and the fact that features included in the model (e.g., luminance contrasts) indeed correlate with fixation probability in natural scenes (Krieger et al., 2000; Reinagel & Zador, 1999). The "object-view", in turn, states that objects are the primary driver of fixations in natural scenes (Einhäuser, Spain, & Perona, 2008; Nuthmann & Henderson, 2010). As a corollary of this view, the manipulation of an object's features should leave

* Corresponding author at: Philipps-Universität Marburg, AG Neurophysik, Karl-von-Frisch-Str. 8a, 35032 Marburg, Germany. Fax: +49 6421 2824168.
E-mail address: wet@physik.uni-marburg.de (W. Einhäuser).

http://dx.doi.org/10.1016/j.visres.2014.11.006
0042-6989/

the pattern of preferably fixated locations unaffected, as long as the impression of objecthood is preserved.

The object-view is supported by two independent lines of evidence. One of them is based on the prediction of fixated locations within a scene, whereas the second one derives from distributional analyses of eye fixations within objects in a scene. With regard to the former, it is important to note that the robust correlation between fixations and low-level features, which seem to argue in favour of the salience-view, does not imply causality. Indeed, when lowering local contrast to an extent that the local change obtains an object-like quality, the reduced contrast attracts fixations rather than repelling them (Einhäuser & König, 2003), arguing against a causal role of contrast. Even though this specific result can be explained in terms of second-order features (texture contrasts, Parkhurst & Niebur, 2004), objects attract fixations and once object locations are known, early (low-level) salience provides little additional information about fixated locations (Einhäuser, Spain, & Perona, 2008). Together with the finding that object locations correlate with high values in salience maps (Elazary & Itti, 2008; Spain & Perona, 2011), it seems that salience does not drive fixations directly, but rather that salience models predict the locations of objects, which in turn attract fixations. This support for the object-view has, however, recently been challenged. In a careful analysis of earlier data, Borji, Sihite, and Itti (2013) showed that more recent models of early salience outperform the naïve object-based model of Einhäuser, Spain, and Perona (2008). This raises the question whether a slightly more realistic object-based model is again at par with early-salience models.

The second line of evidence for the "object-view" arises from the analysis of fixations relative to objects. Models of early salience typically predict that fixations target regions of high contrasts (luminance-contrasts, colour-contrasts, etc.), which occur on the edges of objects with high probability. Although the density of edges in a local surround indeed is a good low-level predictor of fixations (Mannan, Ruddock, & Wooding, 1996) and even explains away effects of contrast as such (Baddeley & Tatler, 2006; Nuthmann & Einhäuser, submitted for publication), fixations do *not* preferentially target object edges. Rather, fixations are biased towards the centre of objects (Foulsham & Kingstone, 2013; Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013). As a consequence of this bias, for edge-based early-salience models fixation prediction improves when maps are smoothed (Borji, Sihite, & Itti, 2013) and thus relatively more weight is put from the edges to the objects' centre (Einhäuser, 2013). Quantitatively, the distribution of fixations within an object is well-described by a 2-dimensional Gaussian distribution (Nuthmann & Henderson, 2010). The distribution has a mean close to the object centre, quantifying the so-called preferred viewing location (PVL), and a standard deviation of about a third of the respective object dimension (i.e., width or height). Since a PVL close to object centre in natural-scene viewing parallels a PVL close to word centre in reading (McConkie et al., 1998; Rayner, 1979), it seems likely that the PVL is a general consequence of eye-guidance optimizing fixation locations with respect to visual processing – at least when no action on the object is required: fixating the centre of an object (or word) maximizes the fraction of the object perceived with high visual acuity. A possible source for the variability in target position, as quantified by the variance or standard deviation of the PVL's Gaussian distribution, is noise in saccade programming (McConkie et al., 1998; Nuthmann & Henderson, 2010). Taken together, the existence of a pronounced PVL for objects in scenes suggests that fixation selection, and by inference attentional guidance, is object based.

Both lines of evidence for the object-view assume that object *locations* are known prior to deploying attention and selecting fixation locations. This does not require objects to be *recognized* prior to attentional deployment. Rather, a coarse parcellation of the scene into "proto-objects" could be computed pre-attentively (Rensink, 2000). If models of early salience in fact predict the location of objects or proto-objects, they could reach indistinguishable performance from object-based models, even if attention is entirely object based. The explanatory power of low-level feature models, like Itti, Koch, and Niebur (1998) salience, would then be explained by them incidentally modelling the location of objects or proto-objects. In turn, the existence of a PVL would be a critical test as to whether proto-objects as predicted by a model indeed constitute proto-objects that can guide attention in an object-based way. An early model that computed proto-objects in natural scenes explicitly in terms of salience (Walther & Koch, 2006) failed this test and showed no PVL for proto-objects, except for the trivial case in which proto-objects overlapped with real objects and the observed weak tendency for a central PVL for these proto-objects was driven by the real objects (Nuthmann & Henderson, 2010). In a more recent approach along these lines, Russell et al. (2014) developed a proto-object model that directly implements Gestalt principles and excels most existing models with respect to fixation prediction. Although a direct comparison of this model with real objects is still open, Russell et al.'s approach shows how object-based salience can act through proto-objects and can thus be computed bottom-up (and possibly pre-attentively) from scene properties.

In the present study, we test the object-view against the salience-view for overt attention in natural scenes. Two predictions follow from the object-view hypothesis.

(I) A model of fixation locations that has full knowledge of object locations in a scene and adequately models the distribution of fixations within objects ("PVL-model") does not leave any additional explanatory power for early salience. That is, salience-based models cannot outperform object-based models.

(II) Early-salience models that reach the level of object-based models do so, because they predict object (or proto-object) locations rather than guiding attention per se. Under the object-view hypothesis, any manipulation of low-level features that neither affects the perceived objecthood nor the location of the objects in the scene, will decrement the performance of the early-salience model more dramatically than that of the object-based model.

Here we test these predictions directly: using the object maps from Einhäuser, Spain, and Perona (2008) and a canonical PVL distribution from Nuthmann and Henderson (2010) we predict fixated locations for the images of the Einhäuser, Spain, and Perona (2008) stimulus set (S. Shore, uncommon places, Shore, Tillman, & Schmidt-Wulffen, 2004). In a first experiment, prediction (I) is tested on an independent dataset of fixations from 24 new observers who viewed the same Shore, Tillman, and Schmidt-Wulffen (2004) images. We compare an object-based model that incorporates the within-object PVL (PVL map) to the prediction of the Adaptive Whitening Salience Model (AWS, Garcia-Diaz et al., 2012a, 2012b), which is the best-performing model identified in the study by Borji, Sihite, and Itti (2013). In a second experiment, prediction (II) is tested by reducing saturation and contrast of the objects and testing how PVL map and AWS predict fixations of 8 new observers viewing these modified stimuli. In experiment 3, we repeat experiment 2 with a free-viewing task to rule out that object-based instructions biased the results in experiment 2.

2. Materials and methods

Stimuli for all experiments were based on 72 images from the Steven Shore "Uncommon places" collection (Shore, Tillman, & Schmidt-Wulffen, 2004; Fig. 1A), which constitute a subset of the 93 images used in Einhäuser, Spain, and Perona (2008) and correspond to the subset used in an earlier study ('t Hart et al., 2013). Our object-based modelling used the annotation data from the original study by Einhäuser, Spain, and Perona (2008), while all fixation data was obtained from an independent set of 40 new observers (24 in experiment 1, 8 in experiments 2 and 3, see Sections 2.2–2.4).

2.1. Models

All object-based models were computed based on the keywords provided by the 8 observers of Einhäuser, Spain, and Perona (2008) and the object outlines created in the context of this study. A list of all objects is available at http://www.staff.uni-marburg.de/~einhaeus/download/ObjectRecall.csv. From the outlines, bounding boxes were computed as the minimal rectangle that fully encompassed an object. In case an object had more than one disjoint part or more than one instantiation within the scene, separate bounding boxes were defined for each part and/or instantiation (Fig. 1B). Hereafter, both cases will be referred to as object "parts" for simplicity of notation. In total, the 72 images used for the present study contained 785 annotated objects consisting of a total of 2450 parts.

2.1.1. Original object map (OOM)

To test whether we could replicate the finding by Borji, Sihite, and Itti (2013) that AWS outperformed the trivial object map representation proposed in Einhäuser, Spain, and Perona (2008) on our new fixation dataset, we used the maps as defined there: the interior of each object named by any observer received a value of 1, the

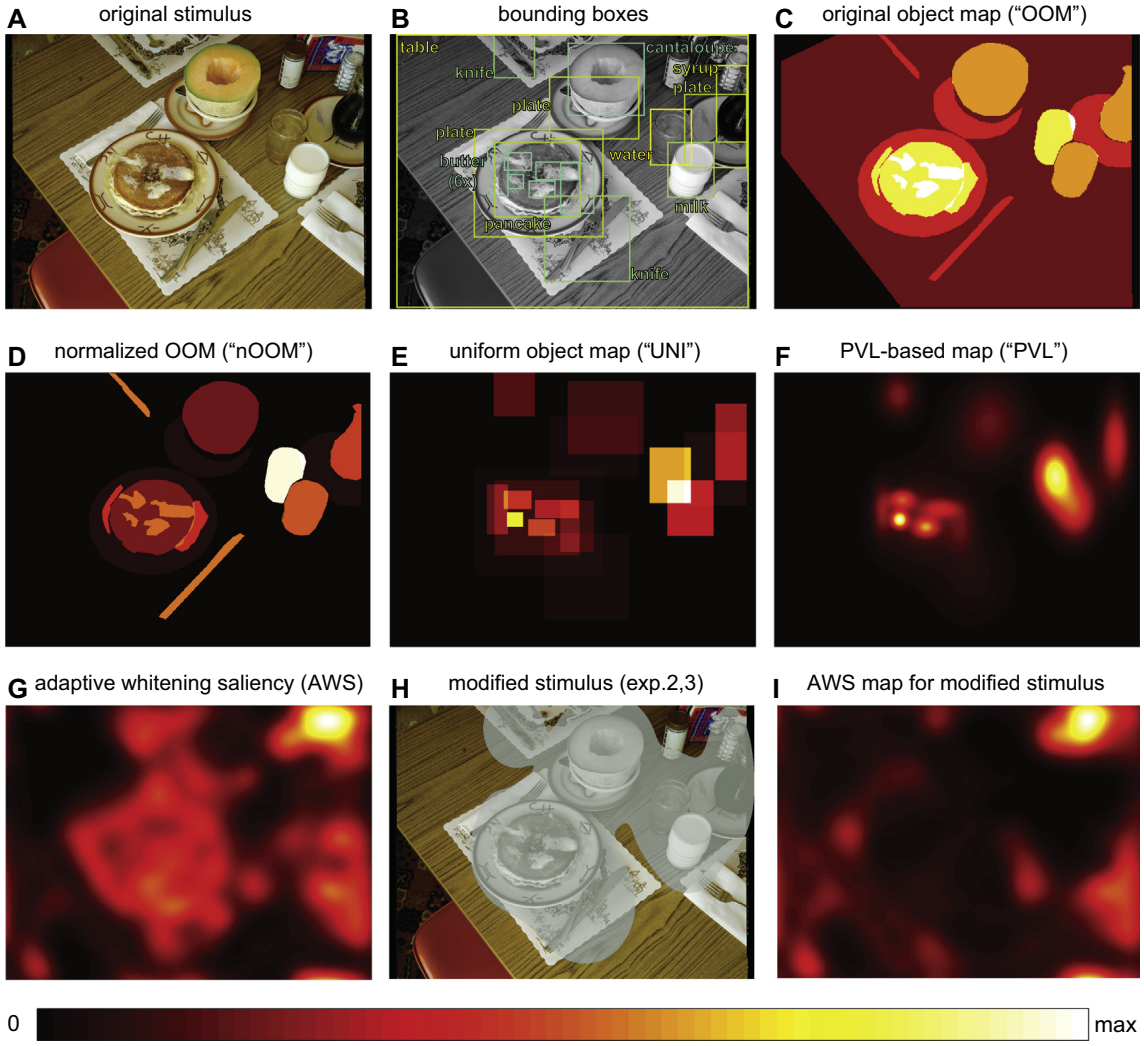

Fig. 1. Stimulus example and maps. (A) Example stimulus from the S. Shore set. (B) Bounding boxes of annotated objects, based on annotation data from Einhäuser, Spain, and Perona (2008). Same colour indicates same object. Note that some objects have multiple instances (knife, plate) or multiple disjoint parts (butter). (C) Object map as used in main analysis of Einhäuser, Spain, and Perona (2008) and Borji, Sihite, and Itti (2013). (D) Normalized version of the map in (C) ("nOOM"), in which each object is normalized to unit integral. (E) Object map based on bounding boxes from panel (B) with uniform sampling inside each object or object part. (F) Object map based on bounding boxes from panel (B) with sampling within each object or object part according to the Gaussian distribution of preferred viewing locations using the parameters from Nuthmann and Henderson (2010). (G) Salience map according to the AWS algorithm by Garcia-Diaz et al. (2012a, 2012b) for example image of panel (A). (H) Modified version of the stimulus of panel (A), as used in experiments 2 and 3. (I) AWS map for the modified stimulus. In panels (C) through (G) and (I), "hotter" colours indicate more weight, all maps are scaled to the same dynamic range for illustration. Note that maps in (D), (E), and (F) normalize each *object* to unit integral, hence large objects carry less weight per pixel, rendering the object "table" indistinguishable from background black at the colour-depth used for this illustration. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

exterior a value of 0 and the resulting object maps were added per image. That is, each pixel provides a count of how many objects cover this pixel (Fig. 1C). Here and for the following models, we ignored how many observers had named an object in the original study. This "unweighted" use of the maps had the rationale that the original data serves to provide a more or less exhaustive representation of objects in the scene, rather than providing their "representativeness" for the scene. Using "weighted" maps, however, yielded qualitatively similar results.

2.1.2. Normalized original object map (nOOM)

For comparison with the other models described in Sections 2.1.3 and 2.1.4, we normalized each object of the original object map to unit integral by dividing its contribution to the OOM by the number of pixels covered by it. This resulted in a normalized original object map (nOOM, Fig. 1D).

2.1.3. Uniform object map (UNI)

To provide a baseline for the PVL-based maps as described below (Section 2.1.4), we modelled the distribution of fixations within each bounding box to be uniform. To compute a uniform object map for each image, for each object we assigned each pixel within its bounding box the value of $1/A$, where A denotes the area in pixels (i.e., bounding box width w times its height h). If an object o consists of P_o parts, the contribution of each part was in addition multiplied by $1/P_o$. The maps obtained for each object were then added to obtain the map for the scene (Fig. 1E). By definition, the sum over all pixels of an object is 1, irrespective of the number of its parts; and each object makes the same contribution to the map, irrespective of its size or number of parts.

2.1.4. PVL-based object maps (PVL)

To model fixations within an object adequately, we started with the observation that fixations can be described by a 2-dimensional Gaussian distribution (Nuthmann & Henderson, 2010). We modelled the fixation distribution for each object as a Gaussian centred at the bounding box centre, with a vertical standard deviation of 34% of the bounding box height and a horizontal standard deviation of 29% of the bounding box width, using the numbers provided in Table 2 of (Nuthmann & Henderson, 2010). Following the procedure for the uniform maps, the Gaussians for each object were normalized to unit integral or – if there were P_o parts per object – the Gaussian for each part was normalized to integral $1/P_o$. Maps of the objects within each image were then added (Fig. 1F). As with the uniform maps, each object makes the same contribution to the map, irrespective of its size or number of parts.

2.1.5. Formal description of the models

Formally, we can write the description of Sections 2.1.1–2.1.4 as follows. For the OOM, we have

$$\mathrm{OOM}(x,y) = \sum_{o=1}^{N} S_o(x,y)$$

where S_o denotes the surface of the object o and N the number of objects in the image.

$$S_o(x,y) = \begin{cases} 1 & \text{if } (x,y) \text{ falls in the boundary of object } o \\ 0 & \text{otherwise} \end{cases}$$

For the nOOM, we have

$$\mathrm{nOOM}(x,y) = \sum_{o=1}^{N} \frac{S_o(x,y)}{A_o}$$

with A_o denoting the number of pixels in object o.

For the UNI maps, we first defined for each part p of object o

$$U_{o,p}(x,y) = \frac{1}{wh} B_{o,p}(x,y)$$

where $B_{o,p}$ denotes the bounding box of part p of object o, w is the bounding box width and h the bounding box height (indices of w and h have been omitted for simplicity, but w and h are to be understood to depend on o and p).

$$B_{o,p}(x,y) = \begin{cases} 1 & \text{if } (x,y) \text{ falls in the bounding box of part } p \text{ of object } o \\ 0 & \text{otherwise} \end{cases}$$

Then we summed over all objects and parts

$$\mathrm{UNI}(x,y) = \sum_{o=1}^{N} \frac{1}{P_o} \sum_{p=1}^{P_o} B_{o,p}(x,y)$$

Similarly, for the PVL maps, we first computed for each part p of object o

$$G_{o,p}(x,y) = \frac{\exp\left[-\frac{(x-x_0)^2}{2\sigma_x^2} - \frac{(y-y_0)^2}{2\sigma_y^2}\right]}{2\pi\sigma_x\sigma_y}$$

where (x_0, y_0) is the bounding box centre. Except for the analysis of Section 3.1.5, the standard deviations followed the Nuthmann and Henderson (2010) data: $\sigma_x = 0.29w$ and $\sigma_y = 0.34h$ with h and w denoting bounding box width and height for the respective object part.

Then we summed as above to obtain the PVL-based object map.

$$\mathrm{PVL}(x,y) = \sum_{o=1}^{N} \frac{1}{P_o} \sum_{p=1}^{P_o} G_{o,p}(x,y)$$

2.1.6. Adaptive whitening salience (AWS)

As an early-salience model we used the model that Borji, Sihite, and Itti (2013) identified to achieve best performance on our earlier data: adaptive whitening salience (AWS; Garcia-Diaz et al., 2012a, 2012b). We applied the matlab implementation as provided by the authors at http://persoal.citius.usc.es/xose.vidal/research/aws/AWSmodel.html using a scaling factor of 1.0 to the unmodified version of each image in its pixel intensity representation (Fig. 1G). Except for the scaling factor, which has a default value of 0.5 to reduce computation time for large images, default parameters as set in the authors' implementation were used. The effect of decreasing the scaling factor is explored in Appendix B.

2.1.7. Combined maps

To test whether adding an early-salience model to the object-based model improved fixation prediction, we combined normalized versions of the PVL and the AWS map. For each image, we computed a set of combined maps as

$$\mathrm{COM}_\alpha(x,y) = \alpha \frac{\mathrm{AWS}(x,y)}{\sum_{x,y}\mathrm{AWS}(x,y)} + (1-\alpha)\frac{\mathrm{PVL}(x,y)}{\sum_{x,y}PVL(x,y)}$$

In this equation α parameterizes the weight given to the early-salience model, with $\alpha = 0$ corresponding to the pure PVL map and $\alpha = 1$ to the pure AWS map.

For comparison, we also tested a multiplicative interaction between the maps

$$\left(\frac{\mathrm{AWS}(x,y)}{\sum_{x,y}\mathrm{AWS}(x,y)}\right)\left(\frac{\mathrm{PVL}(x,y)}{\sum_{x,y}\mathrm{PVL}(x,y)}\right)$$

2.2. Experiment 1

To test the models described herein, we recorded a new eye-tracking dataset using 24 observers (mean age: 24.6 years; 13

female, 11 male). Images were used in 3 different conditions, in their original colour (Fig. 1A) and in two colour-modified versions. Each observer viewed each of the 72 images once, 24 stimuli in each condition (24 unmodified, 24 with clockwise colour rotation and 24 with counter-clockwise colour rotation). Each condition of each image was in turn viewed by 8 observers. For the present study, only the unmodified images were analyzed; for completeness, the details of the colour modification and the main analysis for the colour modified stimuli are given in Appendix A. Stimuli were presented centrally at a resolution of 1024×768 pixels on a grey background ($18\ cd/m^2$) using a 19′ EIZO FlexScan F77S CRT monitor running at 1152×864 pixel resolution and 100 Hz refresh rate, which was located in 73 cm distance from the observer. Eye position was recorded at 1000 Hz with an Eyelink-1000 (SR Research, Ottawa, Canada) infrared eye-tracking device, and for fixation detection the Eyelink's built-in software with default settings (saccade thresholds of 35 deg/s for velocity and $9500\ deg/s^2$ for acceleration) was used. Observers started a trial by fixating centrally, before the image was presented for 3 s. The initial central (0th) fixation was excluded from analysis. After each presentation, observers were asked to rate the aesthetics of the preceding image and to provide five keywords describing the scene afterwards. Neither keywords nor ratings were analyzed for the present purposes. All participants gave written informed consent and all procedures were in accordance with the Declaration of Helsinki and approved by the local ethics committee (Ethikkommission FB04, Philipps-University Marburg).

2.3. Experiment 2 – modified stimuli, original task

In natural scenes, low-level salience and object presence tend to be correlated, which presents one possible explanation for good performance of salience models. To dissociate low-level salience from object presence, in experiment 2 we used a modified version of each stimulus. Specifically, we calculated the median value of the PVL map, and all pixels in the stimulus that exceeded this median (i.e., half of the image with largest PVL map values) were desaturated (transformed to greyscale) and halved in luminance contrast, while keeping the mean luminance unchanged (Fig. 1H). For these stimuli, any salience model that is based on low-level features such as colour-contrasts or luminance-contrast will therefore predict fixations in the unmodified (normal saturation, high luminance contrast) area or at the boundaries between saturated and unmodified regions (Fig. 1I). The PVL map remains unchanged by the experimental manipulation (all objects remain visible). Therefore, the salience model and the PVL-based object model now differ in their predictions with regard to fixation selection in scenes. Eight new observers participated in experiment 2 (mean age: 26.5, 4 male, 4 female). Other than using the modified stimuli, the experimental methods were identical to experiment 1.

2.4. Experiment 3 – modified stimuli, free viewing

Experiment 3 was identical to experiment 2 with the exception that observers were not asked to provide keywords after each stimulus, but the next trial started with a central fixation on a blank screen after each stimulus presentation. Observers received no specific instructions except that they were free to look wherever they liked as soon as the stimulus appeared. Eight new observers participated in experiment 3 (mean age 25.3; 6 female, 2 male).

2.5. Data analysis

To quantify how well a given map (AWS, OOM, nOOM, UNI, PVL) predicted fixation locations irrespective of spatial biases, we used a measure from signal-detection theory (SDT). For a given image *i*, we pooled the fixations of all observers and measured the values of the map at these locations. This defined the "positive set" for this image. We then pooled the fixations from all other images and measured the values at these locations for the map of image *i*. This defined the "negative" set for image *i*. This negative set includes all biases that are not specific for image *i*, and thus presents a conservative baseline. In some analyses, we restricted the dataset (e.g., to one colour condition, or to the nth fixation). In these cases, restrictions were applied to positive and negative set alike. To quantify how well the negative set could be discriminated from the positive set, we computed the receiver operating characteristic (ROC) and used the area under the ROC curve (AUC) as measure of prediction quality. Importantly, this measure is invariant under any strictly monotonic scaling of the maps, such that – except for combining maps – no map-wise normalization scheme needed to be employed for making the different maps comparable.

AUCs were obtained independently for each of the 72 images. Since AUCs were obtained by pooling over observers, all statistical analysis in the main text (ANOVAs, *t*-tests) was done "by-item". This by-item analysis allows for a robust computation of AUCs pooled across observers; however, for completeness we also report "by-subject" analyses (Appendix C).

In addition to parametric tests (ANOVAs, *t*-tests), for the main comparisons we also performed a sign test. The sign test is a non-parametric test that makes no assumptions on the distributions of AUCs across images. It tests whether the sign of an effect (i.e., is model "A" or model "B" better for a given image) is consistent across images, but ignores the size of the effect (i.e., by how much is model "A" better than model "B").

To analyze the effect of salience on object selection (Section 3.1.6), we used a generalized linear mixed model (GLMM). For the GLMMs we report *z*-values, that is, the ratio of regression coefficients to their standard errors ($z = b/SE$). Predictors were centred and scaled.

For the GLMM analysis we used the R system for statistical computing (version 3.1; R Core Team, 2014) with the glmer programme of the lme4 package (version 1.1–7; Bates et al., 2014), with the bobyqa optimizer. Data processing and all other analyses were performed using Matlab (Mathworks, Natick, MA, USA).

3. Results

3.1. Experiment 1

3.1.1. Object maps that consider the PVL are at par with the best early-salience model

Using all data from the fixation dataset, we test whether the prediction of fixated locations depended on the map used (AWS, OOM, nOOM, UNI, PVL). We find a significant effect of map type ($F(4,284) = 37.0$, $p < 0.001$, rm-ANOVA) on prediction. Post-hoc tests show that there are significant differences between all pairs of maps (all $ts(71) > 3.7$, all $ps < 0.001$) except between AWS and PVL ($t(71) = 1.29$, $p = 0.20$) and between AWS and UNI ($t(71) = 1.54$, $p = 0.13$). This confirms that AWS significantly outperforms naïve object-based maps. However, once the within-object PVL is taken into account, object-based maps are at par with state-of-the-art early salience maps; numerically, they even show a slightly better performance, though this is not statistically significant for the by-item analysis (Fig. 2A).

3.1.2. PVL is a necessary factor for the prediction of fixations

Already the UNI maps achieve better performance than the OOMs and reach indistinguishable performance from AWS. This

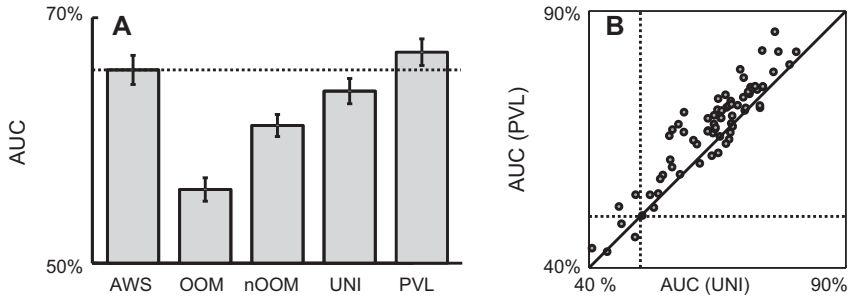

Fig. 2. PVL-based object map outperforms other maps. (A) Area under the ROC curve (AUC) as measure of fixation prediction by the five maps depicted in Fig. 1. Bars indicate mean AUC over images, errorbars s.e.m.s over the 72 images. (B) Image-wise comparison between AUCs for "PVL" and "UNI". The PVL-based map predicts fixations better than uniform sampling on the object in 59 images (points above diagonal), and worse only in 13 (points below diagonal).

raises the question whether the bulk of the benefit compared to the naïve object model arises from using bounding boxes rather than object outlines or from the normalization by object area. In other words, is there any true benefit of modelling the PVL within an object in detail? To address this question, we compare UNI maps to PVL maps image by image. We find that in 59/72 images, the maps that take the PVL into account outperform the UNI maps, with the reverse being true in 13/72 cases only (Fig. 2B). This fraction of images is significant ($p < 0.001$, sign-test). Similarly, the PVL outperforms the nOOM map (57/72, $p = 0.003$) and the OOM map (63/72, $p < 0.001$) for the vast majority of images. This shows that the benefit of considering the within-object PVL is robust across the vast majority of images.

3.1.3. Early salience provides little extra explanatory power, once object locations are known

Provided the location of objects are known, how much extra information does early salience add with regard to fixated locations? To address this question, we combine the PVL and AWS maps additively. We screen all possible relative weights ("α") of AWS relative to PVL in steps of 1%. When enforcing the same α for all images, as would be required for a model generalizing to unknown images, we find that even at the best combination ($\alpha = 52\%$, AWS adds only 2.2 percentage points to the PVL performance alone (69.4% as compared to 67.2%, Fig. 3). Even when allowing for adapting the weight for each image separately, the maximum AUC reaches 71.2%, such that the maximum possible gain by adding AWS to PVL is less than 4%. A multiplicative interaction (i.e., PVL "gating" AWS) is in the same range as the additive models (69.1% AUC).

3.1.4. Object salience and early salience are similar from the first free fixation on

In order to test whether the relative contributions of objects and early salience vary during prolonged viewing, we measured the fixation prediction of AWS and PVL separated by fixation number (Fig. 4). Even using a liberal criterion (uncorrected paired t-tests at each fixation number), we do not find any difference between PVL and AWS for any fixation number (all $ps > 0.17$; for fixation 0–9: all $ts(71) < 1.38$; for fixation 10 only 69 images contributed data: $t(68) = 0.82$). Neither do we find any clear main effect of fixation number (excluding fixation 0 and fixation 10) on the AUC for either PVL ($F(8,568) = 1.73$, $p = 0.09$) or AWS ($F(8,568) = 1.69$, $p = 0.10$). Thus, there is little evidence that fixation prediction by either early salience or objects changes over the course of a trial, and there is no evidence of any difference between PVL and AWS at any point in the trial.

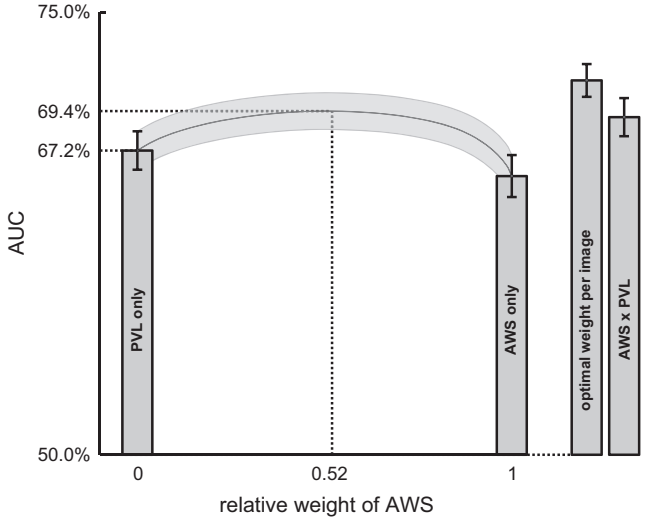

Fig. 3. Combination of PVL and AWS. AUC values for various combinations of PVL and AWS maps; *solid line*: AUCs for additively combined maps αAWS + $(1-\alpha)$PVL against weight α, mean over images (same weight for all images); *shaded area*: s.e.m. over images; *left bar*: PVL result ($\alpha = 0$; cf. Fig. 2A), *second bar from left*: AWS result ($\alpha = 1$, cf. Fig. 2A); *third bar from left*: AUC for choosing optimal weight per image, mean and s.e.m. over images; *right bar*: multiplicative interaction of AWS and PVL maps, mean and s.e.m. over images. Horizontal dashed lines indicate maximal gain in AUC for combining AWS and PVL compared to PVL alone: 2.2 percentage points.

3.1.5. Optimal PVL parameters generalize across datasets

For the analysis so far, we used the parameters of Nuthmann and Henderson (2010) to model the phenomenon of a PVL within objects. These were obtained on an entirely different stimulus set with observers performing distinct tasks (memory, search and preference judgements) and with a different setup. This raises the question, whether the average PVL generalizes across datasets. To test this, we modelled the PVL by 2-dimensional Gaussians with horizontal standard deviations ranging from 0.10 to 0.60 of bounding box width and vertical standard deviations with the same fraction of bounding box height, varied in 0.01 steps (i.e., we set $\sigma_x = \beta w$ and $\sigma_y = \beta h$, with h and w denoting bounding box height and width, and varied β). We find that prediction indeed reaches a maximum around $\beta = 0.31$ (Fig. 5A), in line with the values found in Nuthmann and Henderson (2010) and used throughout this paper. Interestingly, even the optimum value (67.15% AUC at an sd of 33% of bounding box dimensions) is very close to but slightly below the 67.19% found for the anisotropic Nuthmann & Henderson values. To test whether the result improves further for anisotropic (relative to the bounding box) PVL distributions, we vary σ_x and σ_y independently ($\sigma_x = \beta_x w$ and $\sigma_y = \beta_y h$) in 0.01 steps

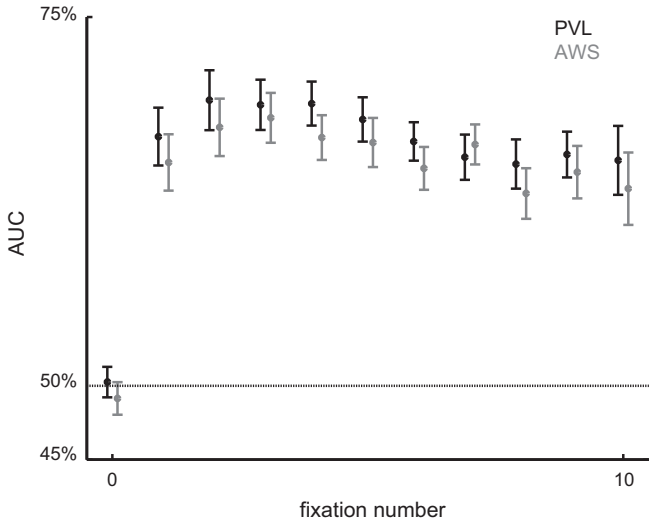

Fig. 4. Development of AUC over course of a trial. AUCs for AWS and PVL separated by fixation number, 0th fixation denotes the initial central fixation, which is not used for any other analysis.

around the Nuthmann and Henderson (2010) default value of $\beta_x = 0.29$ and $\beta_y = 0.34$. We find the AUC has its maximum off the diagonal (at $\beta_x = 0.29$, $\beta_y = 0.35$, Fig. 5B, circle) indicating an optimal PVL that is slightly anisotropic relative to the object's bounding box. This optimal AUC value is only 0.01% (percentage points) larger than the result for the original parameters (0.29, 0.34). Since these values were obtained on a different data set and experimental setup, it is tempting to speculate that the fraction of about 1/3 of bounding box dimensions, possible with a slight anisotropy, for the PVL might reflect a universal constant for object-viewing.

3.1.6. Prioritisation of objects by low-level salience

Is the prioritisation as to which object is selected, once the objects are given, biased by early salience? First, we test whether the probability that an object is fixated at all is related to its salience. To avoid obvious confounds with object area, we quantify an object's low-level salience by the maximum value of the normalized AWS map within the object surface ("peak AWS"). To keep all measures well defined, we restrict analysis to those 366 objects that consist of only one part. In addition to peak AWS, we consider two object properties, which could potentially confound peak AWS effects: object size (the number of pixels constituting the object), and object eccentricity (the distance of the object's centre of mass from the image's centre). For each observer, we allocate a "1" to a fixated object and a "0" to a non-fixated object, yielding a binary matrix with 366 × 8 (number of objects × number of observers, who viewed the respective image in unmodified colour) entries. We use a GLMM to determine the impact of the object properties on the thus defined fixation probability (cf., Nuthmann & Einhäuser, submitted for publication). The model includes the three object properties as fixed effects. With regard to the random effects structure, the model includes random intercepts for subjects and items as well as and by-item random slopes for all three fixed effects. We find a significant effect of peak AWS ($z = 4.55$, $p < 0.001$) above and beyond the effects of object size ($z = 5.09$, $p < 0.001$) and eccentricity ($z = -4.70$, $p < 0.001$). This indicates that among all objects, the objects with higher low-level salience are preferentially selected.

The analysis so far asks whether an object is fixated at all in the course of a 3 s presentation. For an infinite presentation duration, it seems likely that all objects would be eventually fixated; in turn, the salience of an object may be especially predictive for fixations early in the trial. To quantify this, we modify the analysis such that, rather than assigning a 1 to an object that is fixated at any time in the trial, we assign a 1 only to objects that are fixated at or before a given fixation number n. Computing the same GLMM for this definition and for each n, we find a significant prediction of fixation duration for each n (all $z > 3.3$, all $p < 0.001$). Z-values tend to increase with fixation number (i.e., with increasing n; Table 1). The effects of object size and eccentricity are also significant for all n, with the effect of eccentricity declining over fixation number (Table 1). These results offer a role for early salience that complements object-based fixation selection: attention is guided to objects, but among all the objects in a scene those with higher early salience may be preferentially selected.

3.2. Experiments 2 and 3 – modified natural stimuli

The PVL map performs as well as AWS, but it does not outperform the salience-based model. While this result already invali-

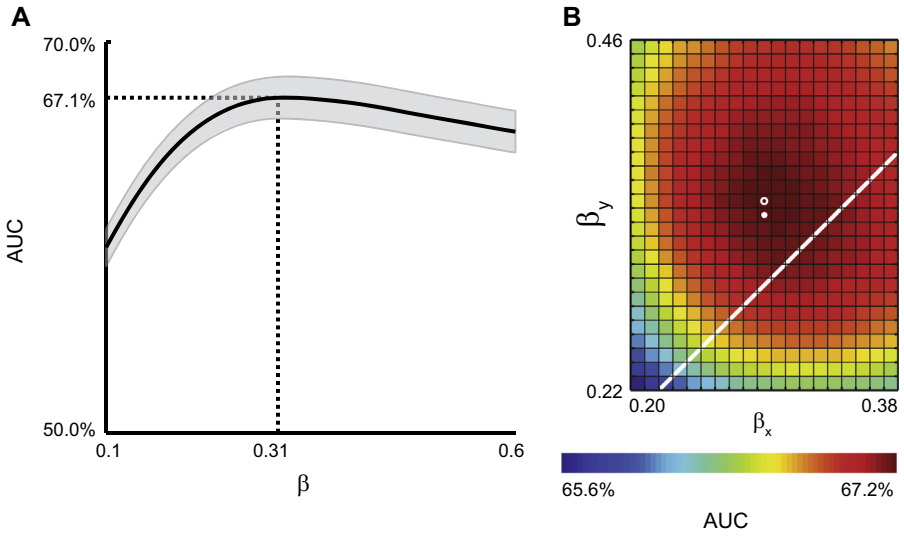

Fig. 5. Varying PVL-model. (A) AUC for varying size of the Gaussian that models the PVL within an object simultaneously for horizontal and vertical dimension, using the same ratio (β) relative to the respective bounding box dimension. Mean and s.e.m. over images; dashed lines indicate maximum. (B) Independent variation of horizontal and vertical standard deviation of the Gaussian that models the PVL, white line indicates diagonal (matching the values of panel A), white circle marks peak (0.29/0.38), white star the values by Nuthmann and Henderson (2010) used throughout the present paper (0.29/0.34).

Table 1
GLMM results: *z*-values for the fixed effects peak AWS, object size and object eccentricity on the probability of fixating labelled objects in scenes. Each column, labelled as fixation number *n*, reports data for a model that considers objects that are fixated at or before a given fixation number *n*.

Fixation number	1	2	3	4	5	6	7	8	9	10	Any
Peak AWS	3.551	3.348	4.206	4.345	4.596	4.558	4.206	4.348	4.552	4.682	4.554
Object size	3.915	4.284	4.482	4.462	4.362	4.741	5.215	5.272	5.214	5.079	5.088
Object eccentricity	−6.296	−7.279	−7.468	−6.223	−5.571	−5.389	−5.085	−5.267	−5.010	−4.850	−4.702

dates a key argument put forward against object-based fixation selection, namely that AWS was better than the naïve object-based approach (Borji, Sihite, & Itti, 2013), experiment 1 alone does not show a superiority of object-based models. In experiments 2 and 3, we dissociate the effect of objects from the effect of the features that constitute objects by manipulating low-level features at object locations.

3.2.1. Modification de-correlates AWS and PVL maps

The object-view states that early-salience models predict fixation selection through correlations of their features with object locations. To test this, we measure the correlation between AWS and PVL map values. To obtain sufficiently independent samples, we sample values from both maps on a central 11 × 8 grid of pixels 100 pixels apart (i.e., at (13,35), (13,135), ... (13,735), (113,35), ... (1013,735)) for each image. For the original images as used in experiment 1, AWS and PVL map values are indeed positively correlated ($r(6334) = 0.16$, $p < 0.001$).

The aim of the experimental manipulation in experiments 2 and 3 is to disentangle AWS from PVL predictions by reducing this correlation. We therefore generated new stimuli by halving contrast and removing saturation from the half of the image in which PVL was highest (Fig. 1H). This manipulation was effective in that the correlation between PVL and AWS for the modified stimuli is now negative ($r(6334) = -0.06$, $p < 0.001$) and – when individual images are considered – smaller than for the original image in 71/72 cases.

3.2.2. On modified images, PVL outperforms AWS

Using the fixation data of experiment 2 and computing AWS on the modified stimuli used, the PVL map now significantly outperforms AWS with respect to fixation prediction (AUC: 63.0 ± 1.3% vs. 56.6 ± 1.1% (mean ± s.e.m.); Fig. 6A, $t(71) = 3.41$, $p = 0.001$). On the level of individual images, the prediction of PVL is better for 51/72 images, a significant fraction ($p < 0.001$, sign-test). In the free-viewing task of experiment 3, the prediction by the PVL map remains virtually unchanged (AUC: 63.1 ± 1.3%) and is significantly better than the AWS performance (AUC: 58.1 ± 1.2%) both on average (Fig. 6B, $t(71) = 2.43$, $p = 0.02$) and for individual images (46/72, $p = 0.02$, sign-test). Both experiments show that when the prediction of an object-based model and a salience-based model are dissociated by experimentally manipulating the correlation between early salience and objecthood, object-based models outperform early salience. The result of experiment 3, in which observers had no specific instruction, furthermore rules out that the precedence of object-based fixation selection over low-level salience is a mere consequence of an object-related task.

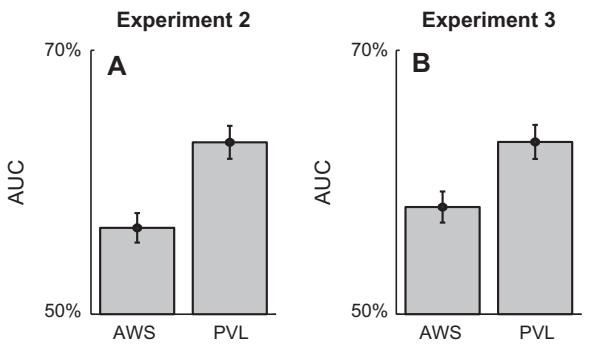

Fig. 6. Modified stimuli – PVL based map outperforms AWS. (A) Experiment 2 (object naming) and (B) experiment 3 (free viewing). Notation as in Fig. 2.

3.2.3. Dependence on fixation number

As for experiment 1, we analyzed the time course of PVL and AWS predictions. In experiment 2 (Fig. 7A), with the exception of the initial (0th) fixation, prediction is above chance for all fixations and both maps (all $ps < 0.007$, all $ts > 2.8$). Excluding the initial (0th) fixation and including all fixation numbers for which data from all images is available (1st through 9th), we find no effect of fixation number on AUC, neither for AWS ($F(8,568) = 0.53$, $p = 0.83$) nor for PVL ($F(8,568) = 1.37$, $p = 0.31$). In experiment 3, fixation durations were longer than in the other two experiments (270.6 ± 2.0 ms vs. 244.6 ± 1.8 ms and 243.3 ± 1.5 ms, excluding the initial fixation), such that from the 9th fixation on, data for some images are missing, and we only analyze fixations 1 through 8 further. For those fixations, AUCs are significantly different from chance for both maps (all $ps < 0.001$, all $ts > 4.0$). Again, we find no main effect of fixation number for AWS ($F(7,497) = 0.45$, $p = 0.87$). However, we find a main effect of fixation number for the PVL map ($F(7,497) = 4.79$, $p < 0.001$). Surprisingly, however, the prediction is best for the early fixations (Fig. 7B). The PVL model performs significantly better than AWS only for the 2nd and 3rd fixation ($t(71) = 3.30$, $p = 0.002$ in both cases), while for the other fixations performance is indistinguishable from AWS ($ts < 1.9$; $ps > 0.06$). Hence, especially early fixations, though not the first one, are guided rather by objects than by low-level salience if no object-related task is to be performed. At no time point during viewing a fixation is guided primarily by low-level salience.

3.2.4. AWS as object model

The object-view explains the performance of early-salience models by the correlation of their features to objects in natural scenes. Hence, if an experimental manipulation dissociates objects from their natural low-level features – like in our experiments 2 and 3 – the prediction performance of early-salience models should drop. Notably, we can derive an additional prediction from the object-view hypothesis: if the early-salience model is computed on the original (i.e., unmodified) stimulus, it predicts object locations. These object locations remain unaffected by the experimental manipulation. Consequently, the early-salience model computed on the unmodified image should still predict fixations on the modified image. We tested this hypothesis and found that AWS applied to the original image indeed predicts fixations on the modified image in experiment 2 better than AWS applied to the modified image itself (AUC: 66.0 ± 1.2% $t(71) = 7.41$, $p < 0.001$). The same holds for experiment 3 (AUC: 66.8 ± 1.2%; $t(71) = 6.15$; $p < 0.001$). This shows that the AWS model incidentally captures attention-guiding properties of natural scenes that still predict fixations when their correlation to the low-level features that are captured by low-level salience are removed.

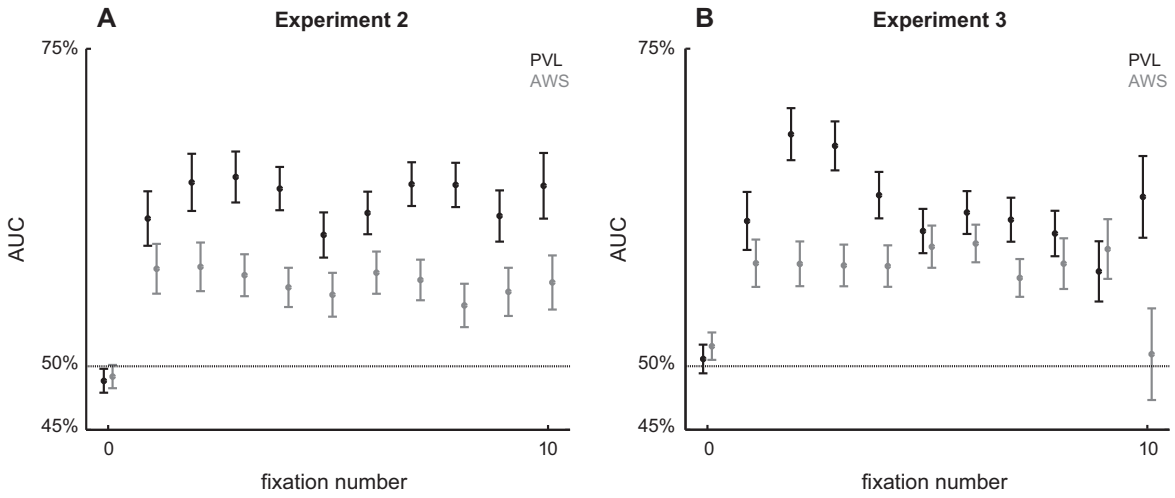

Fig. 7. Modified stimuli – development of AUC over course of a trial. (A) Experiment 2 and (B) experiment 3. Notation as in Fig. 4. Note that for the 10th fixation of panel (A), and for the 9th and 10th fixation in panel (B) not all images contributed data. Hence, statistical analysis was restricted to fixations 1 through 9 (experiment 2) and 1 through 8 (experiment 3).

4. Discussion

In this study we show that an object-based model that adequately models fixation distributions within objects (i.e., the preferred viewing location, PVL) performs at par with the best available model of early salience (AWS). The prediction by the object-based model is robust to small variations of the PVL's standard deviation and not substantially improved by any combination with the AWS model. Notably, when low-level features are manipulated while keeping objecthood intact, the object-based model outperforms the early-salience model. Together, these findings provide further support for the object-view of fixation selection: objects guide fixations and the prediction by early salience is mediated through its correlation with object locations.

If attention is indeed object-based, the question arises up to which level of detail object processing has to be performed prior to fixation selection and how such information can be extracted from the visual stimulus. The degree of object-knowledge required prior to attentional deployment has frequently been associated with "proto-objects" (Rensink, 2000). In the context of salience maps, Walther and Koch (2006) define proto-objects by extending the peaks of the Itti, Koch, and Niebur (1998) saliency map into locally similar regions. In this conception, proto-objects are a function of saliency (Russell et al., 2014). In principle, proto-objects can guide attention in two ways. First, proto-objects can be a proxy for real objects that is computed from stimulus properties. In this case, proto-objects, just like low-level salience, predict fixations through their correlation with object locations. Alternatively, proto-objects could constitute a "higher-level" feature that is causal in driving attention. Yu, Samaras, and Zelinsky (2014) provide indirect evidence for the latter view by showing that proto-objects are a proxy for clutter, and clutter is a possible higher-level feature for attention guidance (Nuthmann & Einhäuser, submitted for publication). In the present study, we show, however, that PVL-based maps outperform other object maps. For proto-objects that do not exhibit a PVL it seems therefore unlikely that they predict fixations better than real objects. An analysis testing proto-objects as defined by Walther and Koch (2006) showed that there was little evidence for a PVL for human fixations within these proto-objects (Nuthmann & Henderson, 2010). Importantly, there was no evidence for a PVL when only saliency proto-objects that did not spatially overlap with annotated real objects were analyzed. Therefore, proto-objects of that sort are not a suitable candidate for the unit of fixation selection in real-world scenes. In addition, AWS generates some notion of objecthood and can be used to extract proto-objects from a scene (Garcia-Diaz, 2011), presumably since the whitening aids figure–ground segmentation (see Russell et al., 2014, for a detailed discussion of this issue). Again, as shown by experiments 2 and 3, the features of AWS are dominated by object-based selection (PVL-based object maps), indicating that the implicit "proto-objects" of AWS do not match real objects with respect to fixation prediction. It is conceivable that the phenomenon of a PVL indeed constitutes an important property that distinguishes proto-objects from real objects, at least with respect to fixation selection. Consequently, the question whether proto-objects, whose computation is stimulus-driven, but not based exclusively on low-level features (Russell et al., 2014; Yu, Samaras, & Zelinsky, 2014), exhibit a PVL is an interesting question for future research.

Attention is likely to act in parallel with object processing rather than being a mere "pre-processing" step. There is a high structural similarity of salience-map models and hierarchical models of object recognition. Already the archetypes of such models, Koch and Ullman's (1985) salience map and Fukushima's (1980) Neocognitron, shared the notion of cascading linear filters and non-linear processing stages in analogy to simple and complex cells of primary visual cortex (Hubel & Wiesel, 1962). The computational implementation of the salience map (Itti, Koch, & Niebur, 1998) and the extension of the Neocognitron idea into a multi-stage hierarchical model (HMAX, Riesenhuber & Poggio, 1999) allowed both models to extend their realm to complex, natural scenes. Given the similarity between the salience map and HMAX, it is not surprising that more recent descendants of salience-map models, such as Itti and Baldi's (2005) "surprise", model human object recognition (Einhäuser et al., 2007) to a similar extent as HMAX itself (Serre, Oliva, & Poggio, 2007), and that in turn HMAX is a decisive ingredient in a state-of-the-art model of attentional guidance in categorical search tasks (Zelinsky et al., 2013). This modelling perspective – together with its roots in cortical physiology – argues that attentional selection and object recognition are not separated, sequential processes, but rather object processing and attention are tightly interwoven.

A challenge for both model testing and experimental research is that an object is not necessarily a static entity, but rather a perceptual and hierarchical construct that can change depending on the task and mindset of the observer. In the present study, we took a

Fig. A.1. Effects of colour modification in experiment 1. (A) Colour-modified (“rotation clockwise”) version of example stimulus in Fig. 1A. (B) Colour-modified (“rotation counter-clockwise”) version of stimulus in Fig. 1A. (C) AUCs for the models of Fig. 2 for colour-modified images (48 per observer). Notation as in Fig. 2. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

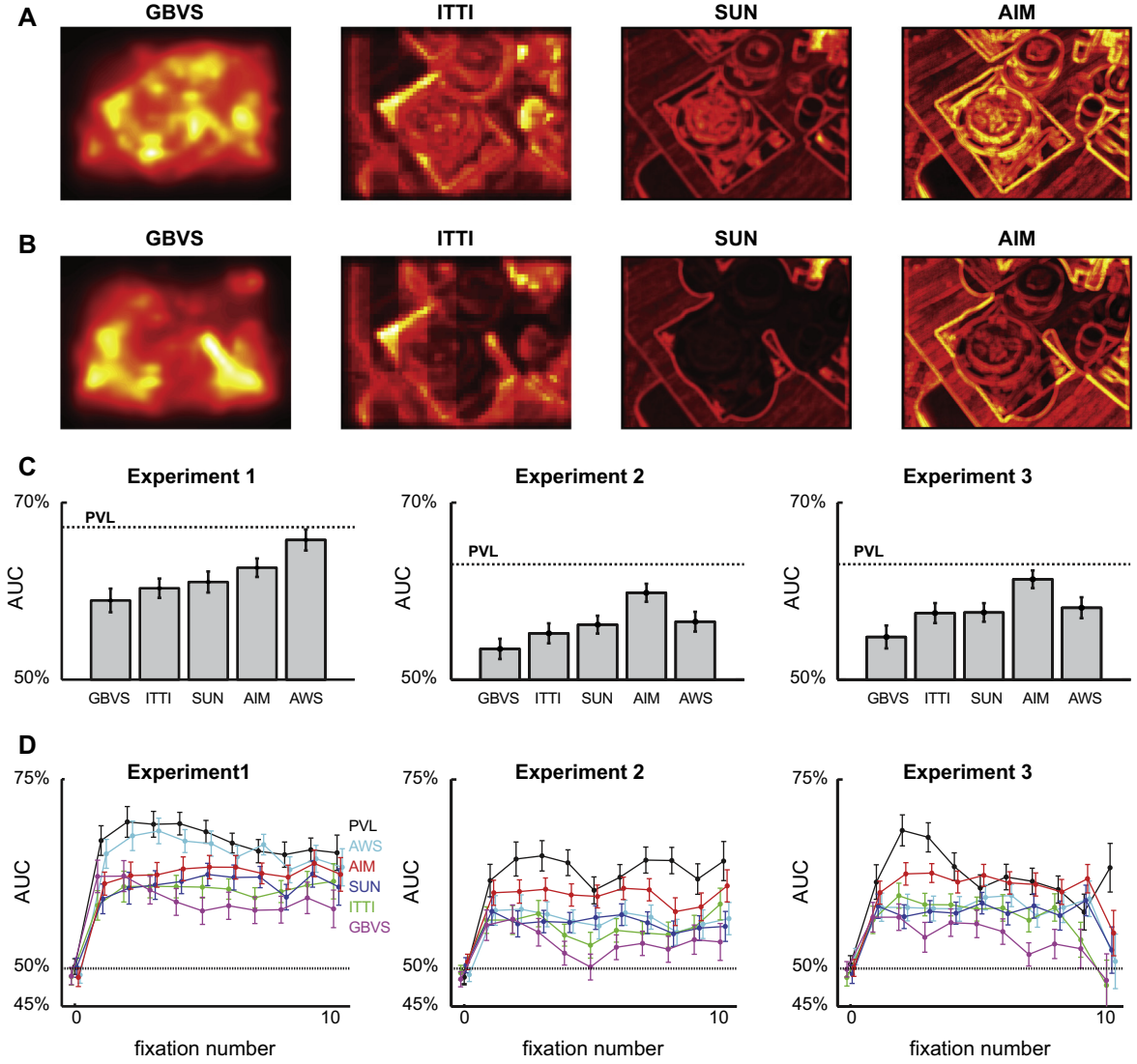

Fig. B.1. Other early-salience models. (A) Output of 4 different salience models (see text for details) on the example image of Fig. 1A. (B) Output of the models for the modified version of the stimulus used in experiments 2 and 3. Colourbar as in Fig. 1. (C) AUC for the 4 models, in comparison to PVL (dashed line) and AWS (right bar) for the 3 experiments. Notation as in Figs. 2 and 6. (D) AUC for the 4 models, AWS and PVL by fixation number. Notation as in Figs. 4 and 7. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

pragmatic approach, using all the keywords provided by at least one of the 8 observers in our original study (Einhäuser, Spain, & Perona, 2008). These ranged from large background objects (“sky”, “grass”, “road”, “table”) over mid-level objects (“car”, “house”, “woman”, “cantaloupe”) to objects that are part of other objects (“roof”, “window”, “purse”, “door”). Treating all of the objects equally, as done here, makes several simplifying assumptions. First, it assumes that the parameters of the PVL are indepen-

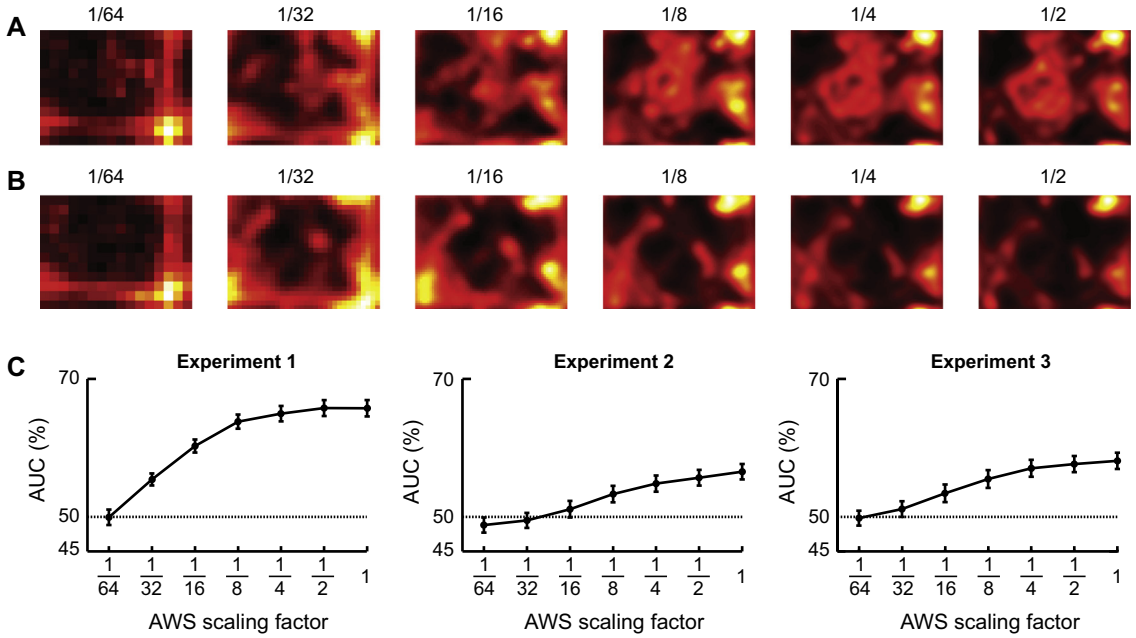

Fig. B.2. Effect of scaling factor in AWS model. (A) AWS map of example stimulus (Fig. 1A) at different scaling factors of the AWS model (given above each panel); scaling factor 1 (no scaling) is depicted in Fig. 1G. (B) AWS map of example modified stimulus (Fig. 1H). Scaling factors as in panel (A); scaling factor 1 is depicted in Fig. 1I. (C) AUC for scaling factors 1/64, 1/32, . . . , 1 for the three experiments. Rightmost datapoint in each panel (factor 1) corresponds to AWS data of the main text (Fig. 2A, Fig. 6A and Fig. 6B for exp. 1, 2 and 3, respectively).

dent of object size. Second, it puts more weight to objects that consist of multiple parts, provided that parts and object are named. Third, objects that are disjoint by occlusion are treated as separate objects. Forth, it does not respect any hierarchy of parts, objects or scene.

With regard to object size, Pajak and Nuthmann (2013) reported wider distributions of within-object fixation locations (i.e., larger variance) for smaller objects. Since – especially for very small and very large objects – the details of the presentation and measurement conditions may also have an effect on the exact distribution, we refrained from modelling this size dependence explicitly here. Since the PVL results are rather robust against the exact choice of the width of the Gaussian distribution, it is unlikely that the effects would be substantial, and – if anything – they should improve fixation prediction by the PVL maps further.

Putting more weight on objects with multiple named parts seems reasonable, at least as long as no clear hierarchy between parts and objects is established and both are likely to follow similar geometric rules to gain objecthood. By normalizing each object to unit integral, very large background objects in any case have a comparably small contribution, except in regions where no other (foreground) objects are present. In an extreme case, where the background object spans virtually the whole scene, the PVL for the background resolves to a model of the central fixation bias (Tatler, 2007), which in this view corresponds to a PVL at scene level. Indeed, the central bias is fit well by an anisotropic Gaussian for a variety of datasets (Clarke & Tatler, 2014). Note, however, that the present analysis is unaffected by generic biases through its choice of baseline.

Since disjoining objects by occlusions is rare in the present data set, this is more a technical issue than a conceptual one. Whether, from the perspective of fixation selection, occlusions are processed prior to attentional deployment (e.g., by means of estimating coarse scene layout prior to any object processing, cf. Hoiem, Efros, & Hebert, 2007; Schyns & Oliva, 2004) remains, however, an interesting question for further research in databases with substantial occurrences of such occlusions.

Finally, the issue concerning the relation between parts and objects has frequently been addressed in parallel in computational and human vision. Dating back to the works of Biederman (1987), human object recognition is thought to respect a hierarchy of parts. On the computational side, mid-level features seem ideal for object recognition (Ullman, Vidal-Naquet, & Sali, 2002), and many algorithms model objects as constellation (Weber, Welling, & Perona, 2000) or compositions (Ommer & Buhmann, 2010) of generic parts. The interplay between objects and parts is paralleled on the superordinate levels of scene and object: Humans can estimate scene layout extremely rapidly and prior to object content (Schyns & Oliva, 2004) and scene layout estimation aids subsequent computational object recognition (Hoiem, Efros, & Hebert, 2007). For human vision, this provides support for a "reverse hierarchy" (Hochstein & Ahissar, 2002) of coarse to fine processing after an initial quick feed-forward sweep (Bar, 2009). Transferring these results to the question of attentional guidance and fixation selection in natural scenes might provide grounds for some reconciliation between a pure "salience-view" and a pure "object-view". It is well conceivable that several scales and several categorical levels (scene, object, proto-objects, parts, features) contribute to attentional guidance. Indeed, recent evidence shows that the intended level of processing (superordinate, subordinate) biases fixation strategies (Malcolm, Nuthmann, & Schyns, 2014). The appropriate hierarchical level might then be dynamically adapted, and – for sufficiently realistic scenarios – be controlled by task demands and behavioural goals. The present data show, however, that for a default condition of comparatively natural viewing conditions, object-based attention supersedes early salience.

Acknowledgments

The research was supported by the German Research Foundation (DFG; grants: EI 852/1 and SFB/TRR 135). We thank the authors of the tested salience models for making their code publically available, D. Walper for support in running the experiments, and the Centre for Interdisciplinary Research Bielefeld (ZiF; FG

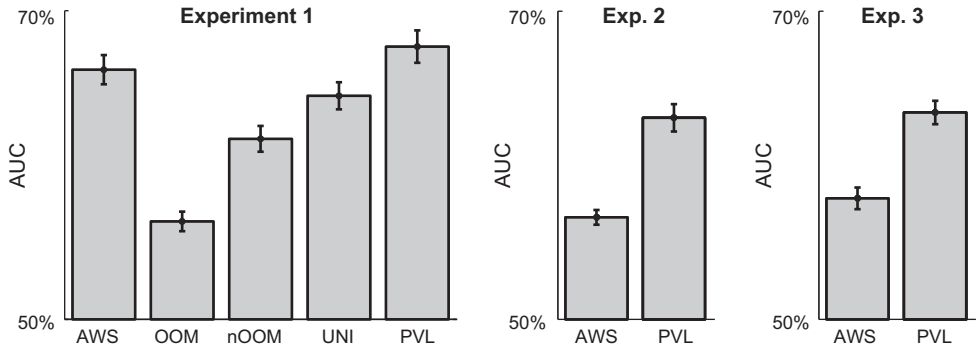

Fig. C.1. By-observer analysis. Comparison akin to Figs. 2 and 7 between object-based models and AWS, but first averaging across images and then analysing data by observer. Consequently, errorbars denote s.e.m. across observers ($N = 24$ for experiment 1, $N = 8$ for experiments 2 and 3).

"Priority") for providing the inspiring environment in which part of this study was conceived.

Appendix A. Colour conditions in experiment 1

Experiment 1 used stimuli in 3 different colour conditions: in their original colour (Fig. 1A) and in two colour-modified versions (Fig. A.1A and B): for the colour modification, images were transformed to DKL colour space (Derrington, Krauskopf, & Lennie, 1984) and each pixel was "rotated" by 90° (either clockwise or counter clockwise) around the luminance ($L + M$) axis. This manipulation (Frey, König, & Einhäuser, 2007) changes the hue of each pixel, but keeps saturation (or rather chroma) and luminance unchanged. Effectively, the manipulation swaps the $S - (L + M)$ axis (roughly: "blue–yellow") with the $L - M$ axis (roughly: "red–green") and therefore keeps the sum over these two "colour-contrasts", as used in most models of early salience, intact. That is, while the global appearance of the stimuli changes dramatically, for most commonly used salience models (including AWS) the effect of the modification is by definition negligible or absent. As the modification neither affected the AWS maps nor the PVL maps, we only used data from the unmodified stimuli for the present study. For completeness, we repeated the main analysis for the modified stimuli. As expected, the results are qualitatively very similar (Fig. A.1C) to the original colour data (Fig. 2A). This indicates that modifications to hue, at least if saturation and luminance are preserved, has little effect on the selection of fixated locations.

Appendix B. Other models and AWS parameter

For the main analysis, AWS has been chosen as reference, since Borji, Sihite, and Itti (2013) had identified it as best performing low-level salience model on the Einhäuser, Spain, and Perona (2008) data. Our data, especially the result that the AWS model applied to the original image predicts fixations better than the model applied to the actual modified stimulus (Section 3.2.4), however, casts doubt on the characterization of AWS as an "early" salience model. We therefore tested a series of other models (Fig. B.1A and B): Graph based visual saliency (GBVS; Harel, Koch, & Perona, 2007), saliency maps following the Itti, Koch, and Niebur (1998) model in the latest (as of September 2014) implementation available at http://ilab.usc.edu ("ITTI"), the "SUN" model (Zhang et al., 2008) and the "AIM" model (Bruce & Tsotsos, 2009). Since optimizing the model parameters for our dataset is not within the scope of the present paper, we used the default parameter settings as suggested by the respective authors throughout.

With the exception of the AWS model for experiment 1 ($t(71) = 1.29$, $p = 0.20$) and the AIM model for experiment 3 ($t(71) = 1.03$, $p = 0.31$), all models perform significantly worse than the PVL map (Fig. B.1C; all other $ts > 4.0$, $ps < 0.001$). However, even in experiment 3 and similar to AWS (Fig. 7), the AIM model still performs significantly worse than PVL for the 2nd and 3rd fixation (Fig. B.1D, right; $t(71) = 2.30$, $p = 0.02$ and $t(71) = 2.06$, $p = 0.04$, respectively). This indicates that our results are not specific to the AWS model, and further supports the view that early in the trial fixation selection is object-based even in the free-viewing task of experiment 3.

The implementation of the AWS model has one parameter, the factor by which the input image is scaled (Fig. B.2). For the unmodified images of experiment 1 (Fig. B.2A), a reduction by a factor of 0.5 does not change the prediction ($t(71) = 0.14$, $p = 0.89$), if anything, it yields a tiny improvement. With further reduction of the scaling factor, prediction performance monotonically decreases, but remains above chance ($ts(71) > 6.2$, $p < 0.001$) for all tested scales down to 1/32 (Fig. B.2C). At a factor of 1/64, prediction is indistinguishable from chance ($t(71) = 0.05$, $p = 0.96$). In experiments 2 and 3, a similar picture emerges: prediction performance decreases monotonically with decreasing scaling factor and becomes indistinguishable from chance at factors of 1/16 (exp. 2) or 1/32 (exp. 3, Fig. B.2C).

Appendix C. Alternative analyses by subject

For the main analysis, we first pool fixations within an image across all observers and then perform a "by-item" analysis, with the images being the items. Pooling over observers allows us to obtain a robust estimate of AUCs for each image. This is especially critical for those analyses that separate data by fixation number, as without pooling over observers the "positive set" for the AUC would contain only a single data point. For the main analysis, which aggregates over fixations, we alternatively could compute the AUC individually for each observer, then average over images and finally perform the statistical analysis over observers for these means. For completeness, we tested this "by-subject" analysis for the comparison between AWS and PVL for all three experiments as well as for all the object models for experiment 1 (Fig. C.1). The pattern of data looks similar to the main analysis (Figs. 2 and 7). For experiment 1, there is a significant effect of object model ($F(4,92) = 19.4$, $p < 0.001$, rmANOVA). Unlike in the main analysis, all pairwise comparisons, including the one between AWS and PVL, show significant differences (all $t(23) > 3.3$, all $ps < 0.003$): PVL performs better than any other model, followed by AWS, UNI, nOOM and OOM (Fig. C.1). Similarly, the difference between PVL and AWS for experiment 2 and 3 is significant (exp. 2: $t(7) = 8.96$, $p < 0.001$; exp. 3: $t(7) = 4.00$, $p = 0.005$): PVL outper-

forms AWS. This analysis not only supports our conclusions of object-based salience outperforming AWS, but also shows this effect already for experiment 1.

References

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences, 16*(8), 437–443.

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A Bayesian system identification analysis. *Vision Research, 46*(18), 2824–2833.

Bar, M. (2009). The proactive brain: Memory for predictions. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences, 364*(1521), 1235–1243. http://dx.doi.org/10.1098/rstb.2008.0310.

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. <http://CRAN.R-project.org/package=lme4>.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115–147.

Borji, A., Sihite, D. N., & Itti, L. (2013). Objects do not predict fixations better than early saliency: A re-analysis of Einhauser et al.'s data. *Journal of Vision, 13*(10), 18. http://dx.doi.org/10.1167/13.10.18.

Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*(3), 5. http://dx.doi.org/10.1167/9.3.5.

Clarke, A. D., & Tatler, B. W. (2014). Deriving an appropriate baseline for describing fixation behaviour. *Vision Research, 102*, 41–51.

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *Journal of Physiology, 357*, 241–265.

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General, 123*, 161–177.

Einhäuser, W. (2013). Objects and saliency: Reply to Borji et al.. *Journal of Vision, 13*(10), 20. http://dx.doi.org/10.1167/13.10.20.

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience, 17*(5), 1089–1097.

Einhäuser, W., Mundhenk, T. N., Baldi, P., Koch, C., & Itti, L. (2007). A bottom-up model of spatial attention predicts human error patterns in rapid scene recognition. *Journal of Vision, 7*(10), 6.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14), 18. http://dx.doi.org/10.1167/8.14.18.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*(3), 3. http://dx.doi.org/10.1167/8.3.3.

Enns, J. T., & Rensink, R. A. (1991). Preattentive recovery of three-dimensional orientation from line drawings. *Psychological Review, 98*(3), 335–351.

Foulsham, T., & Kingstone, A. (2013). Optimal and preferred eye landing positions in objects and scenes. *Quarterly Journal of Experimental Psychology, 66*(9), 1707–1728.

Frey, H. P., König, P., & Einhäuser, W. (2007). The role of first- and second-order stimulus features for human overt attention. *Perception & Psychophysics, 69*(2), 153–161.

Fukushima, K. (1980). Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*(4), 193–202.

Garcia-Diaz, A. (2011). *Modeling early visual coding and saliency through adaptive whitening: Plausibility, assessment and applications*. Ph.D. thesis, Higher Technical Engineering School, University of Santiago de Compostela.

Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012a). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing, 30*(1), 51–64.

Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R., & Pardo, X. M. (2012b). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision, 12*(6), 17.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. *Advances in Neural Information Processing Systems (NIPS*2006), 19*, 545–552.

't Hart, B. M., Schmidt, H. C., Roth, C., & Einhäuser, W. (2013). Fixations on objects in natural scenes: Dissociating importance from salience. *Frontiers in Psychology, 4*, 455. http://dx.doi.org/10.3389/fpsyg.2013.00455.

Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron, 36*(5), 791–804.

Hoiem, D., Efros, A. A., & Hebert, M. (2007). Recovering surface layout from an image. *International Journal of Computer Vision, 75*(1).

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology, 160*, 106–154.

Itti, L., & Baldi, P. (2005). A principled approach to detecting surprising events in video. In *Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR)* (Vol. 1, pp. 631–637).

Itti, L., & 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., 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. http://dx.doi.org/10.1109/34.730558.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4*(4), 219–227.

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision, 13*(2–3), 201–214.

Malcolm, G. L., Nuthmann, A., & Schyns, P. G. (2014). Beyond gist: Strategic and incremental information accumulation for scene categorization. *Psychological Science, 25*(5), 1087–1097.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision, 10*(3), 165–188.

McConkie, G. W., Kerr, P. W., Reddix, M. D., & Zola, D. (1998). Eye movement control during reading: I. The location of initial eye fixations on words. *Vision Research, 28*(10), 1107–1118.

Moore, C. M., Yantis, S., & Vaughan, B. (1998). Object-based visual selection: Evidence from perceptual completion. *Psychological Science, 9*, 104–110.

Nuthmann, A., & Einhäuser, W. (submitted for publication). A new approach to modeling the influence of image features on fixation selection in scenes.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*(8), 20. http://dx.doi.org/10.1167/10.8.20.

Ommer, B., & Buhmann, J. M. (2010). Learning the compositional nature of visual categories for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*, 501–515.

Pajak, M., & Nuthmann, A. (2013). Object-based saccadic selection during scene perception: Evidence from viewing position effects. *Journal of Vision, 13*(5), 2. http://dx.doi.org/10.1167/13.5.2.

Parkhurst, D. J., & Niebur, E. (2004). Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience, 19*(3), 783–789.

Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision Research, 45*(18), 2397–2416.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.

Rayner, K. (1979). Eye guidance in reading: Fixation locations within words. *Perception, 8*, 21–30.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network, 10*(4), 341–350.

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition, 7*, 17–42.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2*(11), 1019–1025.

Russell, A. F., Mihalaş, S., von der Heydt, R., Niebur, E., & Etienne-Cummings, R. (2014). A model of proto-object based saliency. *Vision Research, 94*, 1–15. http://dx.doi.org/10.1016/j.visres.2013.10.005.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80*(1–2), 1–46.

Schyns, P. G., & Oliva, A. (2004). From blobs to boundary edges: Evidence for time and spatial-scale-dependent scene recognition. *Psychological Science, 5*(4), 195–200.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences of the United States of America, 104*(15), 6424–6429.

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

Spain, M., & Perona, P. (2011). Measuring and predicting object importance. *International Journal of Computer Vision, 91*(1), 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*(14), 4.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience, 5*(7), 682–687.

Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*(9), 1395–1407.

Weber, M., Welling, M., & Perona, P. (2000). Unsupervised learning of models for recognition. In *Proc. 6th European conf. computer vision (ECCV)*.

Yu, C.-P., Samaras, D., & Zelinsky, G. J. (2014). Modeling visual clutter perception using proto-object segmentation. *Journal of Vision, 14*(7), 4.

Zelinsky, G. J., Peng, Y., Berg, A. C., & Samaras, D. (2013). Modeling guidance and recognition in categorical search: Bridging human and computer object detection. *Journal of Vision, 13*(3), 30. http://dx.doi.org/10.1167/13.3.30.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7), 32. http://dx.doi.org/10.1167/8.7.32.