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.