Vision Research 94 (2014) 1–15

Contents lists available at ScienceDirect

Vision Research

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

A model of proto-object based saliency

Alexander F. Russell [a], Stefan Mihalaş [b,c], Rudiger von der Heydt [b,c], Ernst Niebur [b,c], Ralph Etienne-Cummings [a,*]

[a] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, United States
[b] Department of Neuroscience, Johns Hopkins University, Baltimore, MD 21218, United States
[c] Zanvyl-Krieger Mind Brain Institute, Johns Hopkins University, Baltimore, MD 21218, United States

ARTICLE INFO

Article history:
Received 22 March 2013
Received in revised form 6 August 2013
Available online 31 October 2013

Keywords:
Attention
Saliency
Gestalt
Proto-object

ABSTRACT

Organisms use the process of selective attention to optimally allocate their computational resources to the instantaneously most relevant subsets of a visual scene, ensuring that they can parse the scene in real time. Many models of bottom-up attentional selection assume that elementary image features, like intensity, color and orientation, attract attention. Gestalt psychologists, however, argue that humans perceive whole objects before they analyze individual features. This is supported by recent psychophysical studies that show that objects predict eye-fixations better than features. In this report we present a neurally inspired algorithm of object based, bottom-up attention. The model rivals the performance of state of the art non-biologically plausible feature based algorithms (and outperforms biologically plausible feature based algorithms) in its ability to predict perceptual saliency (eye fixations and subjective interest points) in natural scenes. The model achieves this by computing saliency as a function of proto-objects that establish the perceptual organization of the scene. All computational mechanisms of the algorithm have direct neural correlates, and our results provide evidence for the interface theory of attention.

1. Introduction

The brain receives an overwhelming amount of sensory information from the retina – estimated at up to 100 Mbps per optic nerve (Koch et al., 2004; Strong et al., 1998). Parallel processing of the entire visual field in real time is likely impossible for even the most sophisticated brains due to the high computational complexity of the task (Broadbent, 1958; Tsotsos, 1991). Yet, organisms can efficiently process this information to parse complex scenes in real time. This ability relies on selective attention which provides a mechanism through which the brain filters sensory information to select only a small subset of it for further processing. This allows the visual field to be subdivided into sub-units which are then processed sequentially in a series of computationally efficient tasks (Itti & Koch, 2001a), as opposed to processing the whole scene simultaneously. Two different mechanisms work together to implement this sensory bottleneck. The first, top down attention, is controlled by the organism itself and biases attention based on the organism's internal state and goals. The second mechanism, bottom up attention, is based on different parts of a visual scene having different instantaneous saliency values. It is thus a result of the fact that some stimuli are intrinsically conspicuous and therefore attract attention.[1]

Most theories and computational models of attention surmise that it is a feature driven process (Itti, Koch, & Niebur, 1998; Koch & Ullman, 1985; Treisman & Gelade, 1980; Walther et al., 2002). However, there is a growing body of evidence, both psychophysical (Cave & Bichot, 1999; Duncan, 1984; Egly, Driver, & Rafal, 1994; Einhauser, Spain, & Perona, 2008; He & Nakayama, 1995; Ho & Yeh, 2009; Kimchi, Yeshurun, & Cohen-Savransky, 2007; Matsukura & Vecera, 2006; Scholl, 2001) and neurophysiological (Ito & Gilbert, 1999; Qiu, Sugihara, & von der Heydt, 2007; Roelfsema, Lamme, & Spekreijse, 1998; Wannig, Stanisor, & Roelfsema, 2011), which shows that attention does not only depend on image features but also on the structural organization of the scene into perceptual objects.

[1] The dichotomy between top-down and bottom-up attention has recently been challenged by Awh, Belopolsky, and Theeuwes (2012). These authors argue that the traditional definition of top-down attention, which includes all signals internal to the organism, conflates the effects of current selection goals with selection history. Although selection history can influence an organism's goals through associative learning and memory, it also includes effects which countermand goal driven selection. Awh, Belopolsky, and Theeuwes (2012) thus propose that the notion of top-down attention should be changed so that current goals and selection history are distinct categories. Attention would thus consist of three processes: current goals, selection history and physical salience (bottom up attention). Our work is only concerned with the last of these factors.

* Corresponding author.
E-mail address: retienne@jhu.edu (R. Etienne-Cummings).

0042-6989/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.visres.2013.10.005

In the Kimchi, Yeshurun, and Cohen-Savransky (2007) experiment a display of 9 red and green, 'L'-shaped elements was used to show that objects can automatically attract attention in a stimulus driven fashion. Subjects were tasked with identifying the color of a target element in the display. In a subset of the trials the elements were arranged, using Gestalt factors, to form an object (see Fig. 1) which was task irrelevant (the task being to report the color of a tagged L shape). Reaction times were fastest when the target formed part of the object, slowest when the target was outside of the object and intermediate when there was no object present. These results suggest that attention is pre-allocated to the location of the object giving rise to a benefit when the target forms part of the object and a cost when the target is outside of the object. Consequently, a model of salience should identify the object as the most salient location in the visual field. However, as shown in Fig. 1, feature based algorithms such as those by Itti, Koch, and Niebur (1998), Harel, Koch, and Perona (2007), Garcia-Diaz, Fdez-Vidal, et al. (2012), Garcia-Diaz, Leborn, et al. (2012)and Hou and Zhang (2008) are unable to do this. Instead image features ('L'-shapes) are recognized as the most salient regions for both the no-object and object cases.

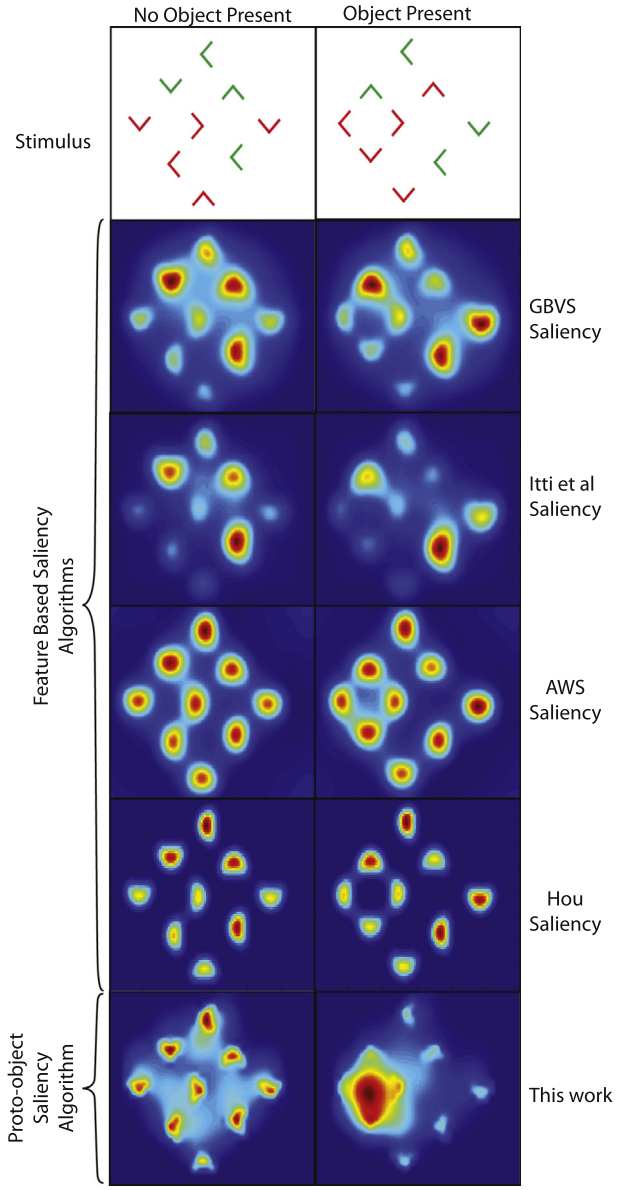

Fig. 1. Top row: Stimuli used by Kimchi, Yeshurun, and Cohen-Savransky (2007). 'L'-shaped elements were arranged to form a no object (left) or object (right) condition. It was found that, in the object present case, attention is automatically drawn to the location of the object. Second row: results of the Graph Based Visual Saliency (GBVS) algorithm (Harel, Koch, & Perona, 2007) in predicting the locations of highest saliency. Third row: results of the Itti, Koch, and Niebur (1998) algorithm in predicting the locations of highest salience. Fourth row: results of the Adaptive Whitening Saliency (AWS) model (Garcia-Diaz, Fdez-Vidal, et al., 2012; Garcia-Diaz, Leborn, et al., 2012). Fifth row: results of the Hou algorithm by Hou and Zhang (2008). All feature based algorithms fail to identify the object in the second display. Bottom row: results of the proto-object saliency algorithm described in this work. The algorithm uses Gestalt cues to perform perceptual scene organization. The formation of the object is clearly identified by the algorithm. In all figures red is the highest salience and blue the lowest. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

In the work that follows we present a biologically plausible model of object based visual salience. The model utilizes the concept of border ownership cells, which have been found in monkey visual cortex (Zhou, Friedman, & von der Heydt, 2000), to provide components of the perceptual organization of a scene. Hypothetical grouping cells use Gestalt principles (Koffka, 1935; Kanizsa, 1979) to integrate the global contour information of figures into tentative proto-objects. Local image saliency is then computed as a function of grouping cell activity. As shown in Fig. 1, these mechanisms allow the model to correctly assign the location of highest salience to the object in the stimuli used by Kimchi, Yeshurun, and Cohen-Savransky (2007). If no object is present, individual elements are awarded the highest saliency. In the remainder of this paper the model is used to investigate whether visual saliency is better explained through image features or through (proto)-objects, and whether the bottom up bias in subjective interest (Masciocchi et al., 2009; Elazary & Itti, 2008) is object or feature based. Our results strongly support the "interface theory" of attention (Qiu, Sugihara, & von der Heydt, 2007) which states that figure-ground mechanisms provide structure for selective attention. This work has important benefits in not only understanding the visual processes of the brain but also in designing the next generation of machine vision search and object recognition algorithms.

2. Related work

Early theories of visual attention were built on the Feature Integration Theory (FIT) proposed by Treisman and Gelade (1980). FIT is a two stage hypothesis designed to explain the differences between feature and conjunction search. It proposes that feature search, where objects are defined by a unique feature, occurs rapidly and in parallel across the visual field. Conjunction search, where an object is defined by a combination of non-unique features, occurs serially and requires attention. In the first, pre-attentive, stage of FIT the features which constitute an object are computed rapidly and in parallel in different feature maps. This allows for the rapid identification of a target defined by a unique feature. However, if an object is defined by a combination of non-unique features then attention is needed to bind the features into a single object and the search must be performed serially. The situation is more complex than described in this early view (see, e.g. Wolfe, 2000) but it suffices as a starting point to understand the basic principles.

How independent, parallel feature maps give rise to the deployment of bottom up attention can be explained by a saliency (Koch & Ullman, 1985) or master (Treisman, 1988) map which guides visual search towards conspicuous targets (Wolfe, 1994, 2007). These maps represent visual saliency by integrating the conspicuity of individual features into a single, scalar-valued 2D retinotopic map. The activity of the map provides a ranking of the salient locations in the visual field with the most active region describing the next location to be attended. Several structures in the pulvinar (Robinson & Petersen, 1992), posterior parietal cortex (Bisley &

Goldberg, 2003; Constantinidis & Steinmetz, 2005; Kusunoki, Gottlieb, & Goldberg, 2000), superior colliculus (Basso & Wurtz, 1998; McPeek & Keller, 2002; Posner & Petersen, 1990; White & Munoz, 2010), the frontal eye fields (Thompson & Bichot, 2005; Zenon et al., 2010), or visual cortex (Koene & Zhaoping, 2007; Mazer & Gallant, 2003; Zhaoping, 2008) have been suggested as physiological substrates of a saliency map.

Numerous computational models (Itti & Koch, 2001b; Itti, Koch, & Niebur, 1998; Milanese, Gil, & Pun, 1995; Niebur & Koch, 1996; Walther et al., 2002) were developed to explain the neural mechanisms responsible for computing a saliency map. The Itti, Koch, and Niebur (1998) model, based off the conceptual framework proposed by Koch and Ullman (1985), is arguably the most influential of all saliency models. The model works as follows: an input image is decomposed into various feature channels (color, intensity and orientation in this model, plus temporal change in the closely related Niebur & Koch (1996) model; other channels can be added easily). Within each channel, a center surround mechanism and normalization operator work together to award unique, conspicuous features high activity and common features low activity. The results of each channel are then normalized to remove modality specific activation differences. In the last stage, the results of each channel are linearly summed to form the saliency map. This model, which uses biologically plausible computation mechanisms, is able to reproduce human search performance for images featuring pop out (Itti, Koch, & Niebur, 1998) and it predicts human eye fixations significantly better than chance (Parkhurst, Law, & Niebur, 2002). More recent saliency algorithms have improved on the normalization method (Itti & Koch, 2001a; Parkhurst, 2002), changed the way in which features are processed to compute saliency (Garcia-Diaz, Fdez-Vidal, et al., 2012; Garcia-Diaz, Leborn, et al., 2012; Harel, Koch, & Perona, 2007; Hou & Zhang, 2008), incorporated learning (Zhang et al., 2008) or added additional feature channels; however all of these models incorporate the ideas of feature contrast and feature uniqueness, as introduced by Koch and Ullman (1985), to compute saliency.

In contrast to FIT and feature based attention, Gestalt psychologists argue that the whole of an object is perceived before its individual features are registered. This is achieved by grouping features into perceptual objects using principles like proximity, similarity, closure, good continuation, common fate and good form (Koffka, 1935; Kanizsa, 1979). This view is backed by an increasing amount of evidence, both psychophysical (Cave & Bichot, 1999; Duncan, 1984; Egly, Driver, & Rafal, 1994; Einhauser, Spain, & Perona, 2008; He & Nakayama, 1995; Ho & Yeh, 2009; Matsukura & Vecera, 2006; Scholl, 2001) and neurophysiological (Ito & Gilbert, 1999; Qiu, Sugihara, & von der Heydt, 2007; Roelfsema, Lamme, & Spekreijse, 1998; Wannig, Stanisor, & Roelfsema, 2011). These results show that attention does not only depend on image features but also on the structural organization of the scene into perceptual objects.

One theory of object based attention is the integrated competition hypothesis (Duncan, 1984) which states that attention is allocated through objects in the visual field competing for limited resources across all sensorimotor systems. When an object in one system gains dominance, its processing is supported across all systems while the representation of other objects is suppressed. Sun and Fisher (2003) and Sun et al. (2008) utilized this theory in their design of an object based saliency map which could reproduce human fixation behavior for a number of artificial scenes. Although it is based on a biologically motivated theory, their model does not use biologically plausible computational mechanisms; instead, machine vision techniques are used. Consequently the model does not provide insight into the biological mechanisms which can account for object based attention.

An alternative hypothesis for object based attention is coherence theory (Rensink, 2000). It uses the notion of proto-objects, which are pre-attentive structures with limited spatial and temporal coherence. They are rapidly computed in parallel across the visual field and updated whenever the retina receives a new stimulus. Focused attention acts to stabilize a small number of proto-objects generating the percept of an object with both spatial and temporal coherence. Because of temporal coherence, any changes to the retina at the object's location are treated as changes to the existing object and not the appearance of a new one. During this stage the object is said to be in a coherence field. Once attention is released from the object, it dissolves back into its dynamic proto-object representation. In coherence theory, proto-objects serve the dual purpose of being the "highest-level output of low-level vision as well as the lowest-level operand on which high-level processes (such as attention) can act" (Rensink, 2000). Consequently proto-objects must not only provide a representation of the visual saliency of a scene but also a mechanism through which top down attention can act.

Walther and Koch (2006) used the concept of proto-objects to develop a model of object based attention. Their model uses the Itti, Koch, and Niebur (1998) feature based saliency algorithm to compute the most salient location in the visual field. The shape of the proto-object at that location is then calculated by the spreading of activation in a 4-connected neighborhood of above threshold activity in the map with the highest saliency contribution at that location. They demonstrated that proto-object based saliency could improve the performance of the biologically motivated HMAX (Riesenhuber & Poggio, 1999) image recognition algorithm. However, there are two drawbacks associated with this model. First, the model does not extend the Itti, Koch, and Niebur (1998) algorithm to account for how the arrangement of features into potential objects can affect the saliency of the visual scene. As a result, this model cannot explain the results obtained by Kimchi, Yeshurun, and Cohen-Savransky (2007) (see Fig. 1). Second, although the computational mechanisms in the model can theoretically be found in the brain, it is unclear if the spreading of brightness or color signals, as used in the algorithm to extract proto-objects, actually occurs in the visual cortex (Rossi & Paradiso, 1999; Roe, Lu, & Hung, 2005; von der Heydt, Friedman, & Zhou, 2003).

In the following, we present a neurally plausible proto-object based model of visual attention. Perceptual organization of the scene through figure-ground segregation is achieved through border ownership assignment – the one sided assignment of a border to a region perceived as figure. Neurons coding border ownership have been discovered in early visual cortex, predominantly area V2 (Zhou, Friedman, & von der Heydt, 2000), see Fig. 2. Li (1998) suggests that border-ownership signals originate from the lateral propagation of edge signals through primary visual cortex but such a mechanism seems unlikely because the signals would have to travel along slow intra-areal connections. This is not compatible with the observed time course of border-ownership responses which appear as early as 20 ms after the edge signals of a visual stimulus arise. An alternative hypothesis, supported by recent neurophysiological evidence (Zhang & von der Heydt, 2010), is that "grouping ($\mathcal{G}$) cells" communicate with border-ownership ($\mathcal{B}$) neurons via (fast) white matter projections (Craft et al., 2007). The grouping neurons integrate object features into tentative proto-objects without needing to recognize the object (Craft et al., 2007; Mihalaş et al., 2011). The high conduction velocity of the myelinated fibers connecting the border ownership neurons and the grouping cells accounts for the fast development of border ownership responses. The proto-object saliency model draws inspiration from the neuronal model of Craft et al. (2007) which uses a recurrently connected network

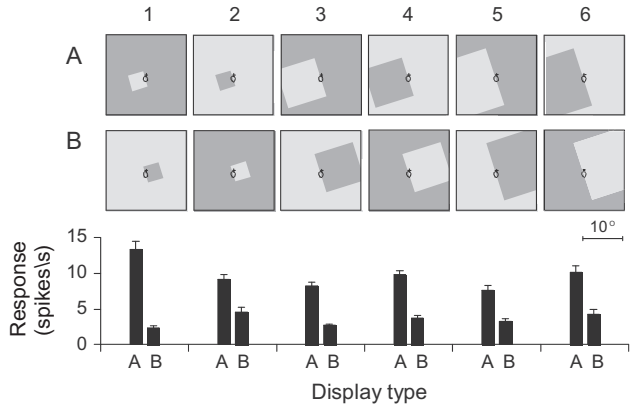

Fig. 2. Response of a border ownership cell in monkey V2 to stimuli of varying sizes. Border ownership cells only respond when their receptive field falls over a contrast edge and their response is modulated by which side of their receptive field the figure appears on. Rows A and B show the stimuli which are, for a given trial, identical within the receptive field (black ellipse) of the border ownership selective cell. Bar graphs below Row B show the mean firing rate of the cell to the stimuli. For all sizes and both contrast polarities, the cell's preferential response occurred when the square was located to the left of the receptive field. Reproduced with permission from Zhou, Friedman, and von der Heydt (2000).

of $\mathcal{B}$ and $\mathcal{G}$ cells to model border ownership assignment for a number of synthetic images.

3. Model

The core of our model is a grouping mechanism which estimates the location and spatial scale of proto-objects within the input image. This mechanism, described in Section 3.1, provides saliency information through the perceptual organization of a scene into figure and ground. In Section 3.2 the grouping mechanism is extended to operate across multiple feature channels and to incorporate competition between proto-objects of similar size and feature composition.

Objects in a visual scene can occlude each other partially. Our convention is that in the occlusion zone, we always refer to the object that is closest to the observer as the "figure," and that behind it as the "background." To achieve scale invariance, the algorithm successively down samples the input image, $\beta(x,y)$, in steps of $\sqrt{2}$ to form an image pyramid spanning 5 octaves. The kth level of the pyramid is denoted using the superscript k. Unless explicitly stated any operation applied to the pyramid is applied independently to each level. Each layer of the network represents neural activity which propagates from one layer to the next in a feed forward fashion – the model does not have any recurrent connections. This was done to ensure the computational efficiency of the model. However, if computation time is not an issue then recurrent connections can be added to ensure more accurate border-ownership and grouping assignment (see Craft et al. (2007) and Mihalaş et al. (2011) for examples of such circuits). Receptive fields of neurons are described by correlation kernels and correlation is used to calculate the neural response to an input, see below for details. The model was implemented using MATLAB (Mathworks, Natick, MA, USA).

3.1. A feed forward model of grouping

This model (shown in Fig. 3) is responsible for estimating the location and spatial scale of proto-objects within the input image. The first stage of processing extracts object edges using 2D Gabor filters (Kulikowski, Marcelja, & Bishop, 1982), which approximate the receptive fields of simple cells in the primary visual cortex (Jones & Palmer, 1987; Marcelja, 1980). Both even, $g_{e,\theta}(x,y)$, and odd, $g_{o,\theta}(x,y)$, filter kernels are used, where:

$$\begin{aligned} g_{e,\theta}(x,y) &= e^{-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}}\cos(\omega x') \\ g_{o,\theta}(x,y) &= e^{-\frac{x'^2+\gamma^2y'^2}{2\sigma^2}}\sin(\omega x') \end{aligned} \tag{1}$$

where $\theta \in \{0, \pi/4, \pi/2, 3\pi/4\}$ radians, γ is the spatial aspect ratio, σ is the standard deviation of the Gaussian envelope, ω is the spatial frequency of the filter, and x' and y' are coordinates in the rotated reference frame defined by

$$\begin{aligned} x' &= x\cos(\theta) + y\sin(\theta) \\ y' &= -x\sin(\theta) + y\cos(\theta) \end{aligned} \tag{2}$$

Simple cell responses are computed according to

$$\begin{aligned} \mathcal{S}^k_{e,\theta}(x,y) &= \beta^k(x,y) * g_{e,\theta}(x,y) \\ \mathcal{S}^k_{o,\theta}(x,y) &= \beta^k(x,y) * g_{o,\theta}(x,y) \end{aligned} \tag{3}$$

where $\mathcal{S}^k_{e,\theta}(x,y)$ and $\mathcal{S}^k_{o,\theta}(x,y)$ are the even and odd edge pyramids at angle θ and $*$ is the correlation operator defined as

$$f(x,y) * g(x,y) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} f(m,n)g(x+m,y+n) \tag{4}$$

Using an energy representation (Adelson & Bergen, 1985; Morrone & Burr, 1988), contrast invariant complex cell responses are calculated from a simple cell response pair as

$$\mathcal{C}^k_\theta(x,y) = \sqrt{\mathcal{S}^k_{e,\theta}(x,y)^2 + \mathcal{S}^k_{o,\theta}(x,y)^2} \tag{5}$$

where $\mathcal{C}^k_\theta(x,y)$ is the complex cell's response at angle θ.

To infer whether the edges in $\mathcal{C}^k_\theta(x,y)$ belong to figure or ground, knowledge of objects in the scene is required. This context information is retrieved from a center surround mechanism, as commonly implemented in the retina, lateral geniculate nucleus and cortex (Reid, 2008). Center-surround receptive fields have a central region (the center) which is surrounded by an antagonistic region (the surround) which inhibits the center. Our model uses center-surround mechanisms of both polarities, with ON-center receptive fields identifying light objects on dark backgrounds and OFF-center surround operators detecting dark objects on light backgrounds. This is implemented as

$$\begin{aligned} \mathcal{CS}^k_D(x,y) &= \lfloor \beta^k(x,y) * cs_{off}(x,y) \rfloor \\ \mathcal{CS}^k_L(x,y) &= \lfloor \beta^k(x,y) * cs_{on}(x,y) \rfloor \end{aligned} \tag{6}$$

where $\lfloor\cdot\rfloor$ is a half-wave rectification, and $\mathcal{CS}_D$ and $\mathcal{CS}_L$ are the dark and light object pyramids. cs_{off} and cs_{on} are the OFF-center and ON-center center surround mechanisms generated using a difference of Gaussians as follows:

$$\begin{aligned} cs_{on}(x,y) &= \frac{1}{2\pi\sigma_i^2}e^{-\frac{x^2+y^2}{2\sigma_i^2}} - \frac{1}{2\pi\sigma_o^2}e^{-\frac{x^2+y^2}{2\sigma_o^2}} \\ cs_{off}(x,y) &= -\frac{1}{2\pi\sigma_i^2}e^{-\frac{x^2+y^2}{2\sigma_i^2}} + \frac{1}{2\pi\sigma_o^2}e^{-\frac{x^2+y^2}{2\sigma_o^2}} \end{aligned} \tag{7}$$

In these equations, σ_i is the standard deviation of the center (inner) Gaussian and σ_o is the standard deviation of the surround (outer) Gaussian.

Next, for a given angle θ, antagonistic pairs of border ownership responses, $\mathcal{B}_\theta$ and $\mathcal{B}_{\theta+\pi}$, are created by modulating $\mathcal{C}$ cell responses with the activity from the $\mathcal{CS}$ pyramids. $\mathcal{B}$ cells can code borders for light objects on dark backgrounds or dark objects on light backgrounds. This is achieved by computing $\mathcal{B}$ cell activity independently for each contrast case and then summing to give a

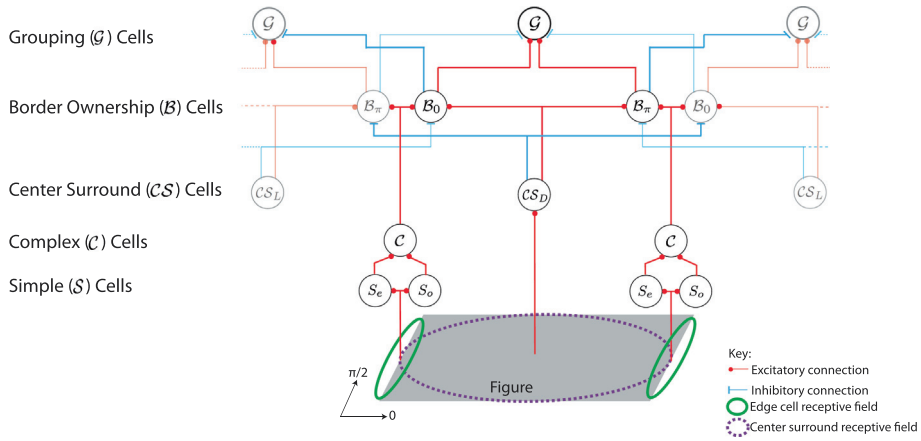

Fig. 3. Grouping network architecture. Input to the system is the gray parallelogram at the bottom of the figure. For simplicity only cells and connections at a single scale are shown. High (low) contrast connections and cells indicate high (low) activation levels. Red (blue) lines represent excitatory (inhibitory) connections. Green ellipses indicate simple ($\mathcal{S}$) cell receptive fields. Simple cells are activated by figure edges and the outputs of the even (S_e) and odd (S_o) cells are combined to form complex ($\mathcal{C}$) cells. The complex cells directly excite the Border Ownership ($\mathcal{B}$) neurons. Note that, in this example, $\mathcal{B}_0$ neurons represent right borders and $\mathcal{B}_\pi$ neurons represent left borders; in general, border-ownership selective neurons always come in two populations whose preferred side of figure differ by 180 degrees. Also note that the mechanism rendering $\mathcal{B}$, neurons insensitive to contrast are not shown in this figure (see Eq. (12) and preceding equations). An estimate of objects in the scene is extracted using $\mathcal{CS}$ neurons which have large center surround receptive fields. $\mathcal{CS}_L$ neurons have on-center receptive fields and extract bright objects from dark backgrounds. $\mathcal{CS}_D$ neurons have off-center receptive fields and extract dark objects from light backgrounds (the subscripts stand for "light" and "dark."). Thus, $\mathcal{CS}_D$ and $\mathcal{CS}_L$ neurons have identical receptive fields but of opposite polarity, at each location. For clarity only the receptive field of the most responsive $\mathcal{CS}$ neuron is shown at a given location (purple dashed ellipse). The $\mathcal{B}$ cell responses to $\mathcal{C}$ cells are modulated by the $\mathcal{CS}$ cells. Excitation from the $\mathcal{CS}$ cell coding for figure on the $\mathcal{B}$ cell's preferred side enhances $\mathcal{B}$ cell activity whilst inhibition from the $\mathcal{CS}$ cell coding for an object on the non preferential side suppresses $\mathcal{B}$ cell activity. This is the start of global scene context integration as objects with well defined contours will strongly bias the $\mathcal{B}$ cells to code for their borders whilst they inhibit $\mathcal{B}$ cells coding for ground. $\mathcal{B}$ cell activity is then integrated in an annular manner to give the grouping activity represented by the $\mathcal{G}$ cells. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

border ownership response independent of figure-ground contrast polarity. This mimics the behavior of the border ownership cell shown in Fig. 2.

$\mathcal{B}_{\theta,L}$, the border ownership activity for a light object on a dark background is given by

$$\mathcal{B}^k_{\theta,L}(x,y) = \left[\mathcal{C}^k_\theta(x,y) \times \left(1 + \sum_{j\geq k}\frac{1}{2^j} v_{\theta+\pi}(x,y) * \mathcal{CS}^j_L(x,y) - w_{opp}\sum_{j\geq k}\frac{1}{2^j} v_\theta(x,y) * \mathcal{CS}^j_D(x,y)\right)\right] \tag{8}$$

and $\mathcal{B}_{\theta,\mathcal{D}}$, the border ownership activity for a dark object on a light background is given by

$$\mathcal{B}^k_{\theta,D}(x,y) = \left[\mathcal{C}^k_\theta(x,y) \times \left(1 + \sum_{j\geq k}\frac{1}{2^j} v_{\theta+\pi}(x,y) * \mathcal{CS}^j_D(x,y) - w_{opp}\sum_{j\geq k}\frac{1}{2^j} v_\theta(x,y) * \mathcal{CS}^j_L(x,y)\right)\right] \tag{9}$$

where v_θ is the kernel responsible for mapping object activity in the $\mathcal{CS}$ pyramids back to the objects edges (see Fig. 4 for details), w_{opp} is the synaptic weight of the inhibitory signal from the opposite polarity $\mathcal{CS}$ pyramid and the term 2^{-j} normalizes the v_θ operator such that the influence across spatial scales is constant. v_θ is generated using the von Mises distribution as follows:

$$v_\theta(x,y) = -\frac{\exp\left[\left(\sqrt{x^2+y^2} - R_0\right)\sin\left(\tan^{-1}\left(\frac{y}{x}\right) - \theta\right)\right]}{2\pi I_0\left(\sqrt{x^2+y^2} - R_0\right)} \tag{10}$$

R_0 is the zero crossing radius of the center surround masks, and θ is the desired angle of the mask. The factor $\frac{\pi}{2}$ rotates the mask to ensure it is correctly aligned with the edge cells. I_0 is the modified Bessel function of the first kind. v_θ is then normalized according to

$$v_\theta(x,y) = \frac{v_\theta(x,y)}{\max(v_\theta(x,y))} \tag{11}$$

The border ownership responses coding for light and dark objects are then combined to give the contrast polarity invariant response

$$\mathcal{B}^k_\theta(x,y) = \mathcal{B}^k_{\theta,L}(x,y) + \mathcal{B}^k_{\theta,D}(x,y) \tag{12}$$

The sign of the difference $\mathcal{B}_\theta(x,y) - \mathcal{B}_{\theta+\pi}(x,y)$ determines the direction of border ownership at pixel (x,y) and orientation θ. Its magnitude gives a confidence measure for the strength of ownership which is also used for determining the local border orientation, see Eqs. (13) and (14).

In the above, the activity of a $\mathcal{B}$ cell is facilitated by $\mathcal{CS}$ activity on its preferred side and suppressed by $\mathcal{CS}$ activity on its non-preferred side. This is motivated by neurophysiological results which show that when an edge is placed in the classical receptive field of a border ownership neuron, image fragments placed within the cell's extra-classical field can cause enhancement of the cell's activity if the image fragment is placed on its preferred side, and suppression if it is placed on the non-preferred side (Zhang & von der Heydt, 2010). Furthermore, modulating the $\mathcal{B}$ cell responses with the $\mathcal{CS}$ activity summated across spatial scales ensures that the $\mathcal{B}$ cell response is invariant to spatial scale — a behavior exhibited by the $\mathcal{B}$ cells' biological counterpart, see Fig. 2 (Zhou, Friedman, & von der Heydt, 2000). Furthermore, by biasing the $\mathcal{B}$ cell activity by the $\mathcal{CS}$ activity at lower spatial scales, the model is made robust in its border ownership assignment when small concavities occur in larger convex objects.

At each pixel multiple border ownership cells exist coding for each direction of ownership at multiple orientations. However, a pixel can only belong to a single border. The winning border owenership response ($\hat{\mathcal{B}}$) is selected, from the pool of all border ownership responses, according to

$$\hat{\mathcal{B}}^k(x,y) = \mathcal{B}^k_{\hat{\theta}}(x,y) \tag{13}$$

where

$$\hat{\theta} = \arg\max_\theta\left(\mathcal{B}^k_\theta(x,y) - \mathcal{B}^k_{\theta+\pi}(x,y)\right) \tag{14}$$

In words, the orientation of $\hat{\mathcal{B}}$ is assigned according to the pair of $\mathcal{B}$ cell responses with the highest difference, and the direction of

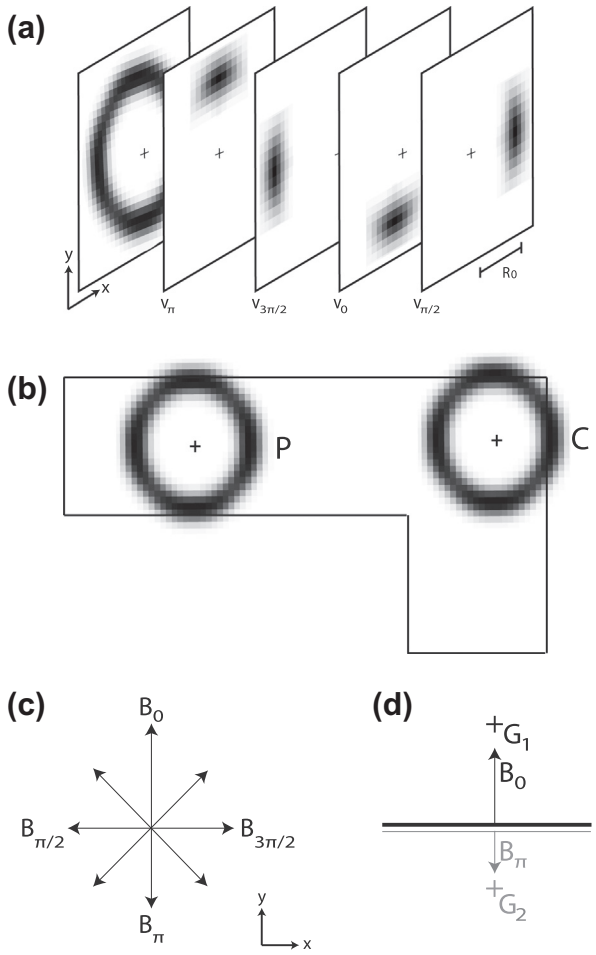

Fig. 4. (a) The annular receptive field of the grouping cells (adapted from Craft et al. (2007)). In the model this is realized by using eight individual kernels (v_θ in the text) whose combined activity produces the desired annular receptive field. Each kernel is generated using Eq. (10) and the kernels v_0, $v_{\pi/2}$, v_π and $v_{3\pi/2}$ are shown in the figure. Identical kernels are also used as the connection pattern to map the activity of the $\mathcal{CS}$ cells to the $\mathcal{B}$ cells during border ownership assignment. See Eq. (8). (b) The annular receptive field of the grouping cells bias the $\mathcal{G}$ cells to have a preference for continuity (C) and proximity (P) (adapted from Craft et al. (2007)). (c) Conventions for the display of $\mathcal{B}$ cell activity at a given pixel. The length of the arrows indicates the magnitude of each cell's activity. (d) Border ownership is assigned to a given pixel by selecting, from the pool of all potential $\mathcal{B}$ cells (shown in (c)), the pair of cells with the greatest activity difference. Within that pair, the cell with the greater activity will own the border – in the case shown, $\mathcal{B}_0$. The winning border will then excite its corresponding grouping cell, $\mathcal{G}_1$. This is done by mapping the activity of $\mathcal{B}_0$ to $\mathcal{G}_1$ using v_0. $\mathcal{G}_1$ also receives a small inhibitory signal from $\mathcal{B}_\pi$ (not shown here). $\mathcal{G}_2$ denotes the $\mathcal{G}$ cell which corresponds to $\mathcal{B}_\pi$. Black (gray) lines indicate high (low) activity.

ownership is assigned to the $\mathcal{B}$ cell in that pair with the greater response.

The final stage of the algorithm calculates grouping ($\mathcal{G}$) cell responses by integrating the winning $\hat{\mathcal{B}}$ cell activity in an annular fashion, see Fig. 4. This biases $\mathcal{G}$ cells to show preference for objects whose borders exhibit the Gestalt principles of continuity and proximity (Fig. 4b). $\mathcal{G}$ cell activity is defined according to

$$\mathcal{G}^k(x,y) = \sum_\theta \left\lfloor \delta\left(\mathcal{B}^k_\theta(x,y), \hat{\mathcal{B}}^k\right) \times \left[\mathcal{B}^k_\theta(x,y) - w_b \times \mathcal{B}^k_{\theta+\pi}(x,y)\right] * v_\theta(x,y) \right\rfloor \tag{15}$$

where $\delta\left(\mathcal{B}^k_\theta(x,y), \hat{\mathcal{B}}^k\right) = 1$ if $\mathcal{B}^k_\theta(x,y) = \hat{\mathcal{B}}^k$ and zero otherwise. w_b is the synaptic weight of the inhibitory signal from the $\mathcal{B}$ cell coding for the opposite (non-preferred) direction of ownership.

The $\mathcal{G}$ pyramid from Eq. (15) is the output of the grouping algorithm.

3.2. Proto-object based saliency

In this section the basic model of Section 3.1 is extended to account for multiple feature channels and to incorporate competition between proto-objects of similar size and feature composition. Note that saliency is obtained based on the primitives generated for the proto-object computation and that the basic mechanisms are shared.

The extended model, shown in Fig. 5, accepts an input image which is decomposed into 9 feature channels: 1 intensity channel, 4 color-opponency channels and 4 orientation channels. A normalization operator added to the grouping mechanism allows for competition between proto-objects of similar size and feature composition. The effect of this operator is that the grouping activity of maps with few proto-objects is promoted and the grouping activity of maps with multiple proto-objects is suppressed. This operator, $\mathcal{N}_1(\cdot)$, which is very similar to that used by Itti, Koch, and Niebur (1998), works as follows: The two center surround pyramids, $\mathcal{CS}_D$ and $\mathcal{CS}_L$, are simultaneously normalized so that all values in the maps are in the range $[0,\ldots,M]$. If, before normalization, the maximum of $\mathcal{CS}_D$ was twice the maximum of $\mathcal{CS}_L$, then after the normalization the maximum of $\mathcal{CS}_D$ will be M and the maximum of $\mathcal{CS}_L$ will be $M/2$. Next, the average of all local maxima, $\bar{m}$, is computed across both maps. In the final stage of normalization each center surround map is multiplied by $(M-\bar{m})^2$. $\mathcal{N}_1(\cdot)$ is simultaneously applied to $\mathcal{CS}_D$ and $\mathcal{CS}_L$ to preserve the local ordering of activity in the maps. The effects of the normalization propagate forward through the grouping mechanism – maps with high $\mathcal{CS}$ activity, will have high border ownership activity which results in high grouping activity. Conversely, maps with low $\mathcal{CS}$ activity will have weak grouping activity.

The intensity channel, $\mathcal{I}$, is generated according to

$$\mathcal{I} = \frac{r+g+b}{3} \tag{16}$$

where r, g and b are the red, green and blue channels of the RGB input image (Itti, Koch, & Niebur, 1998).

The four color opponency channels – red–green ($\mathcal{RG}$), green–red ($\mathcal{GR}$), blue–yellow ($\mathcal{BY}$) and yellow–blue ($\mathcal{YB}$) are generated by decoupling hue from intensity through normalizing each of the r, g, b color channels by intensity. However, because hue variations are not perceivable at very low luminance the normalization is only applied to pixels whose intensity value is greater than 10% of the global intensity maximum of the image. Pixels which do not meet this requirement are set to zero. This ensures that hue variations at very low luminance do not contribute towards object saliency. The normalized r,g,b values are then used to create four broadly tuned color channels, red ($\mathcal{R}$), green ($\mathcal{G}$), blue ($\mathcal{B}$) and yellow ($\mathcal{Y}$) (Itti, Koch, & Niebur, 1998), according to

$$\begin{aligned}
\mathcal{R} &= \left\lfloor r - \frac{g+b}{2} \right\rfloor \\
\mathcal{G} &= \left\lfloor g - \frac{r+b}{2} \right\rfloor \\
\mathcal{B} &= \left\lfloor b - \frac{r+g}{2} \right\rfloor \\
\mathcal{Y} &= \left\lfloor \frac{r+g}{2} - \frac{|r-g|}{2} - b \right\rfloor
\end{aligned} \tag{17}$$

and the opponency signals $\mathcal{RG}$, $\mathcal{GR}$, $\mathcal{BY}$ and $\mathcal{YB}$ are then created as follows:

$$\begin{aligned}
\mathcal{RG} &= \lfloor \mathcal{R} - \mathcal{G} \rfloor \\
\mathcal{GR} &= \lfloor \mathcal{G} - \mathcal{R} \rfloor \\
\mathcal{BY} &= \lfloor \mathcal{B} - \mathcal{Y} \rfloor \\
\mathcal{YB} &= \lfloor \mathcal{Y} - \mathcal{B} \rfloor
\end{aligned} \tag{18}$$

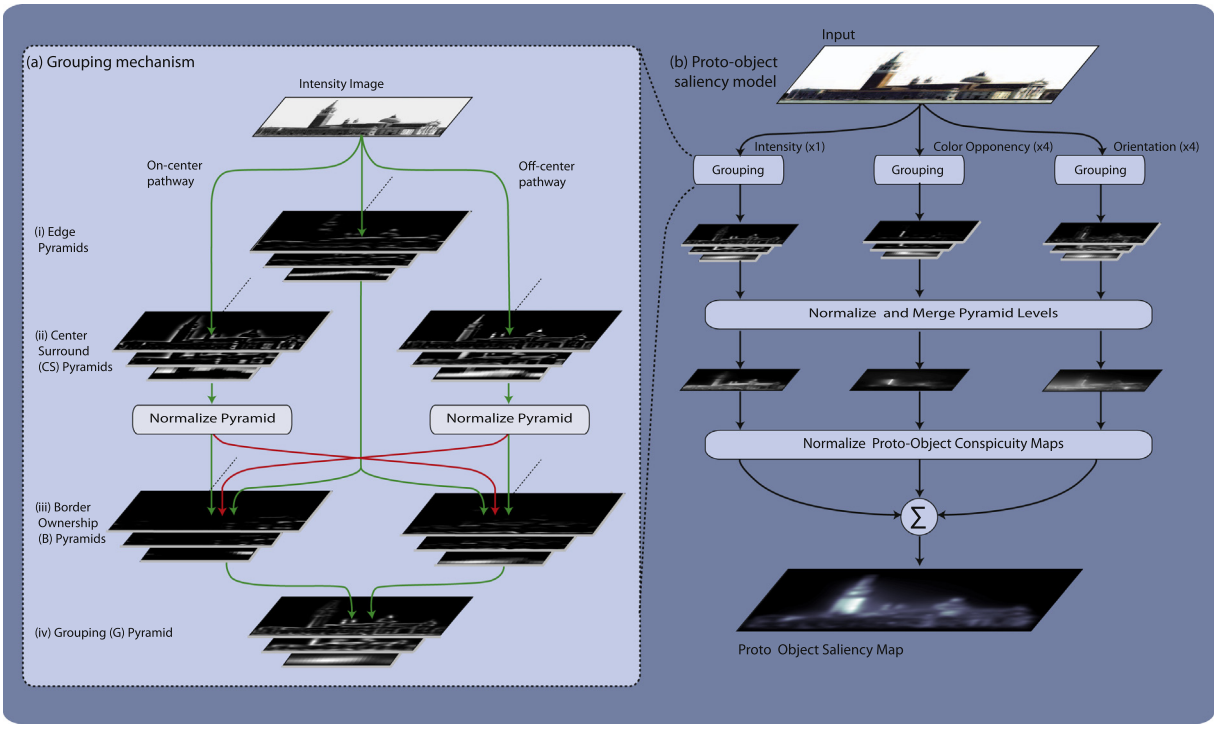

Fig. 5. (a) Overview of the feed forward grouping mechanism acting on the Intensity channel. Image pyramids are used to provide scale invariance. The first stage of processing extracts object edges for angles ranging between 0 and $3\pi/4$ radians in $\pi/4$ increments. Only the extracted edges orientated at 0 radians are shown. Next, bright and dark objects are extracted from the intensity input image using ON-center and OFF-center center surround mechanisms. The center-surround ($\mathcal{CS}$) pyramids are then normalized before being combined with the edge information to assign border ownership to object edges. Border ownership activity is then integrated in an annular fashion to generate a grouping pyramid. (b) The proto-object saliency algorithm. An input image is separated into Intensity, Color Opponency and Orientation channels. The activity of each channel is then passed to the grouping mechanism. The grouping mechanism for both the intensity and color-opponency channels is identical to that shown in (a). However, to ensure that the orientation channels respond only to proto-objects at a given angle, the grouping mechanism for the orientation channels differs slightly. In these channels the ON-center center-surround mechanism has been replaced by an even Gabor filter (with a positive center lobe) and the OFF-center center-surround mechanism has been replaced by an even Gabor filter (with a negative center lobe). These filters are tuned to the orientation of the channel and the width of their center lobes match the zero-crossing diameter of the center-surround mechanisms used in the intensity and color-opponency channels. Like the center-surround mechanisms they replace, the output of the Gabor filters provides an estimation of the location of light objects on dark backgrounds and of dark objects on light backgrounds; however, their response is also modulated by the orientation of proto-objects in the visual scene. The outputs of the Gabor filters are then normalized and the algorithm continues as in (a). The resulting grouping pyramids are then normalized and collapsed to form channel specific conspicuity maps. The conspicuity maps are then normalized, enhancing the activity of channels with unique proto-objects and suppressing the activity of channels with multiple proto-objects. The normalized conspicuity maps are then linearly combined to form the proto-object saliency map.

The four orientation channels, $\mathcal{O}_\alpha$ where $\alpha \in \{0, \pi/4, \pi/2, 3\pi/4\}$ radians, are created using $\mathcal{I}$ as the input to the grouping algorithm. Orientation selectivity is obtained by replacing the center-surround mechanisms in the grouping algorithm with even Gabor filters orientated at[2] α. Specifically,

$$\begin{aligned} cs_{on}(x,y) &= \exp\left[-\frac{x'^2+\gamma_1^2 y'^2}{2\sigma_1^2}\right]\cos(\omega_1 x') \\ cs_{off}(x,y) &= -\exp\left[-\frac{x'^2+\gamma_1^2 y'^2}{2\sigma_1^2}\right]\cos(\omega_1 x') \end{aligned} \tag{19}$$

where x' and y' are the rotated coordinate system defined according to

$$\begin{aligned} x' &= x\cos(\alpha) + y\sin(\alpha) \\ y' &= -x\sin(\alpha) + y\cos(\alpha) \end{aligned} \tag{20}$$

The spatial frequency, ω_1, of the Gabor filters is set so that the width of the central lobe of the filters matches the zero crossing diameter of the original center surround mechanisms. The result of this is that $\mathcal{CS}_D$ still codes for dark objects on light backgrounds and $\mathcal{CS}_L$ still codes for light objects on dark backgrounds; however the activity in these maps is modulated by the orientation of the proto-objects.

Each of the above feature channels is processed independently by the grouping mechanism to form feature specific grouping pyramids, $\mathcal{G}_i$ where i is the channel type. These grouping pyramids are then collapsed to form proto-object conspicuity maps – $\bar{\mathcal{I}}$ for intensity, $\bar{\mathcal{C}}$ for color-opponency and $\bar{\mathcal{O}}$ for orientation. This is achieved through a second normalization, $\mathcal{N}_2(\cdot)$, and a cross scale addition $\oplus$ of the pyramid levels. $\mathcal{N}_2(\cdot)$ is identical to $\mathcal{N}_1(\cdot)$ except that it operates on a single map. $\oplus$ is achieved by scaling each map to scale $k = 8$ and then performing a pixel-wise addition.

For the intensity channel, $\bar{\mathcal{I}}$ is generated according to

$$\bar{\mathcal{I}} = \oplus_{k=1}^{k=10}\mathcal{N}_2\left(\mathcal{G}_I^k\right) \tag{21}$$

$\bar{\mathcal{C}}$ is generated according to

$$\bar{\mathcal{C}} = \oplus_{k=1}^{k=10}\left(\mathcal{N}_2\left(\mathcal{G}_{RG}^k\right) + \mathcal{N}_2\left(\mathcal{G}_{GR}^k\right) + \mathcal{N}_2\left(\mathcal{G}_{BY}^k\right) + \mathcal{N}_2\left(\mathcal{G}_{YB}^k\right)\right) \tag{22}$$

and the orientation conspicuity map is generated according to

$$\bar{\mathcal{O}} = \sum_{\alpha\in\{0^\circ,45^\circ,90^\circ,180^\circ\}} \mathcal{N}_2\left(\oplus_{k=1}^{k=10}\mathcal{N}_2(\mathcal{O}_\alpha)\right) \tag{23}$$

The conspicuity maps are then normalized and linearly combined to form the proto-object saliency map $\mathcal{S}$:

[2] This is only done in the orientation channels. The intensity and color opponency channels use the regular center surround operators as described in Section 3.1.

Table 1
Model parameters.

Parameter	Value
γ	0.5000
σ	2.2400
ω	1.5700
σ_i	0.9000
σ_o	2.7000
w_b	1.0000
R_0	2.0000
w_{opp}	1.0000
σ_l	3.2000
γ_l	0.8000
ω_l	0.7854
M	10.0000

$$\mathcal{S} = \frac{1}{3}\left(\mathcal{N}_2(\bar{\mathcal{I}}) + \mathcal{N}_2(\bar{\mathcal{C}}) + \mathcal{N}_2(\bar{\mathcal{O}})\right) \tag{24}$$

The parameters used in the proto-object saliency algorithm are shown in Table 1.

4. Results

Saliency models are often designed to predict either eye-fixations or salient objects. Models do not generalize well across these categories (Borji, Sihite, & Itti, 2013) and so comparisons of model performance should be conducted between models of the same class. Models designed to predict salient objects work by first detecting the most salient location in the visual scene and then segmenting the entire extent of the object corresponding to that location (Borji, Sihite, & Itti, 2013). This differs from our model of proto-object saliency which calculates saliency as a function of proto-objects (as opposed to image features). The resulting saliency map can then be used to predict eye-fixations. Consequently our model will be tested against four algorithms which are also designed to predict eye-fixations. The first algorithm we test against is the much used Itti, Koch, and Niebur (1998) algorithm. Although more recent models score better in their ability to predict eye-fixations (Borji, Sihite, & Itti, 2012, 2013), it is biologically plausible and this model shares many computational mechanisms (the input color scheme, normalization operator and method to combine component feature maps into the final saliency map) with ours. It is thus useful to compare our model with the Itti, Koch, and Niebur (1998) algorithm since any improvements of the performance of our model can be attributed to the figure-ground organization in the proto-object saliency map. The primary goal of our work is to evaluate the influence of the representation of proto-object on saliency computations, rather than to design a model that is optimized for best eye-prediction performance, without regard to biological relevance. Nevertheless, it is interesting to compare the eye position prediction performance of our model with the best performance of saliency models that have been optimized for this purpose. We therefore compare our model with three other methods, the Graph Based Visual Saliency (GBVS) model (Harel, Koch, & Perona, 2007), the Adaptive Whitening Saliency (AWS) model (Garcia-Diaz, Fdez-Vidal, et al., 2012; Garcia-Diaz, Leborn, et al., 2012) and the Hou model by Hou and Zhang (2008). These algorithms are less biologically plausible than the Itti, Koch, and Niebur (1998) algorithm and ours; however they offer state of the art performance for feature based saliency algorithms (Borji, Sihite, & Itti, 2013). The AWS and Hou models were ranked first and second in their ability to predict eye-fixations across a variety of image categories (Borji, Sihite, & Itti, 2012).

Fig. 6(a) shows the response of these algorithms to a vertical bar. It can be seen that the highest activation in the proto-object saliency map corresponds to the center of the bar, while in the four feature based algorithms, the peaks of saliency tend to concentrate at the edges of objects. Fig. 6(b) shows a simple application of the proto-object saliency map, showing that it can detect pop-out stimuli, as can all the other algorithms (data not shown).

To quantify the performance of the algorithms, in their ability to predict perceptual saliency, two experiments were performed. The first measures the ability of the algorithms to predict attention by using eye fixations as an overt measure of where subjects are directing their covert attention, This idea, first used by Parkhurst, Law, and Niebur (2002) to quantify the performance of the Itti, Koch, and Niebur (1998) algorithm, is based on the premotor theory of attention (Rizzolatti et al., 1987) which posits that the same neural circuits drive both attention and eye fixations. Numerous psychological (Hafed & Clark, 2002; Hoffman & Subramaniam, 1995; Kowler et al., 1995; Sheliga, Riggio, & Rizzolatti, 1994, 1995), physiological (Kustov & Robinson, 1996; Moore & Fallah, 2001, 2004; Moore, Armstrong, & Fallah, 2003) and brain imaging (Beauchamp et al., 2001; Corbetta et al., 1998; Nobre et al., 2000) studies provide strong evidence for this link. Recent experiments (Elazary & Itti, 2008; Masciocchi et al., 2009) have found that subjective interest points are also biased by bottom up factors. Experiment 2 investigates whether or not this bias is better explained by proto-objects or features.

Both experiments used the image database of Masciocchi et al. (2009) which consists of 100 images, with 25 images in each of four categories (buildings, home interiors, fractals and landscapes). For each image both fixation and interest data are available. Eye fixation data was collected from 21 subjects during a free viewing task. The images were presented to the subjects in a random order and all trials began with the subject fixating at a cross located at the center of the screen. Each image subtended approximately $30.4^\circ \times 24.2^\circ$ of visual angle. Each image was displayed for 5 s and subjects' eye movements were recorded using an eye tracker. Eye movements that travelled less than 1° in 25 ms, and were longer than 100 ms were counted as a single fixation. On average participants made 12.89 ± 3.11 fixations per trial (i.e., within 5 s). Interest point data was collected from 802 subjects in an online experiment. Each subject was given a set of 15 images randomly selected from the data set and the subjects were told to click on the 5 most interesting points in the image. The experiment was self-paced. For full details of how fixation and interest data was collected see Masciocchi et al. (2009). Examples of the images and their corresponding saliency maps are shown in Fig. 7.

To measure the saliency map's performance, two popular metrics are used, the area under the Receiver Operating Characteristic curve (Green & Swets, 1966) and the Kullback Leibler divergence (Itti & Baldi, 2005, 2006). In their original forms these metrics are extremely sensitive to edge effects caused by the way in which image edges are handled during the filtering stages of the algorithms (Zhang et al., 2008). These edge effects inadvertently introduce different amounts of center bias to each algorithm. This gives the algorithms varying (false) abilities to explain the center bias[3] found in human fixation and interest data due to the effects of priming, the fact that the center of the screen is the optimal viewing point (Tatler, 2007; Vitu et al., 2004) and that photographers tend to put the subject in the center of the image (Parkhurst & Niebur, 2003; Reinagel & Zador, 1999; Schumann et al., 2008; Tatler, 2007; Tatler, Baddeley, & Gilchrist, 2005; Tseng et al., 2009). To provide a fair

[3] The Itti, Koch, and Niebur (1998), AWS, Hou and proto-object saliency algorithms were not designed to account for center bias. However, by more heavily connecting graph nodes at the center of the saliency map, the GBVS algorithm explicitly incorporates a center bias. To ensure a fair evaluation of the algorithms, the GBVS algorithm was configured to have uniform connections across the saliency map, eliminating the center bias.

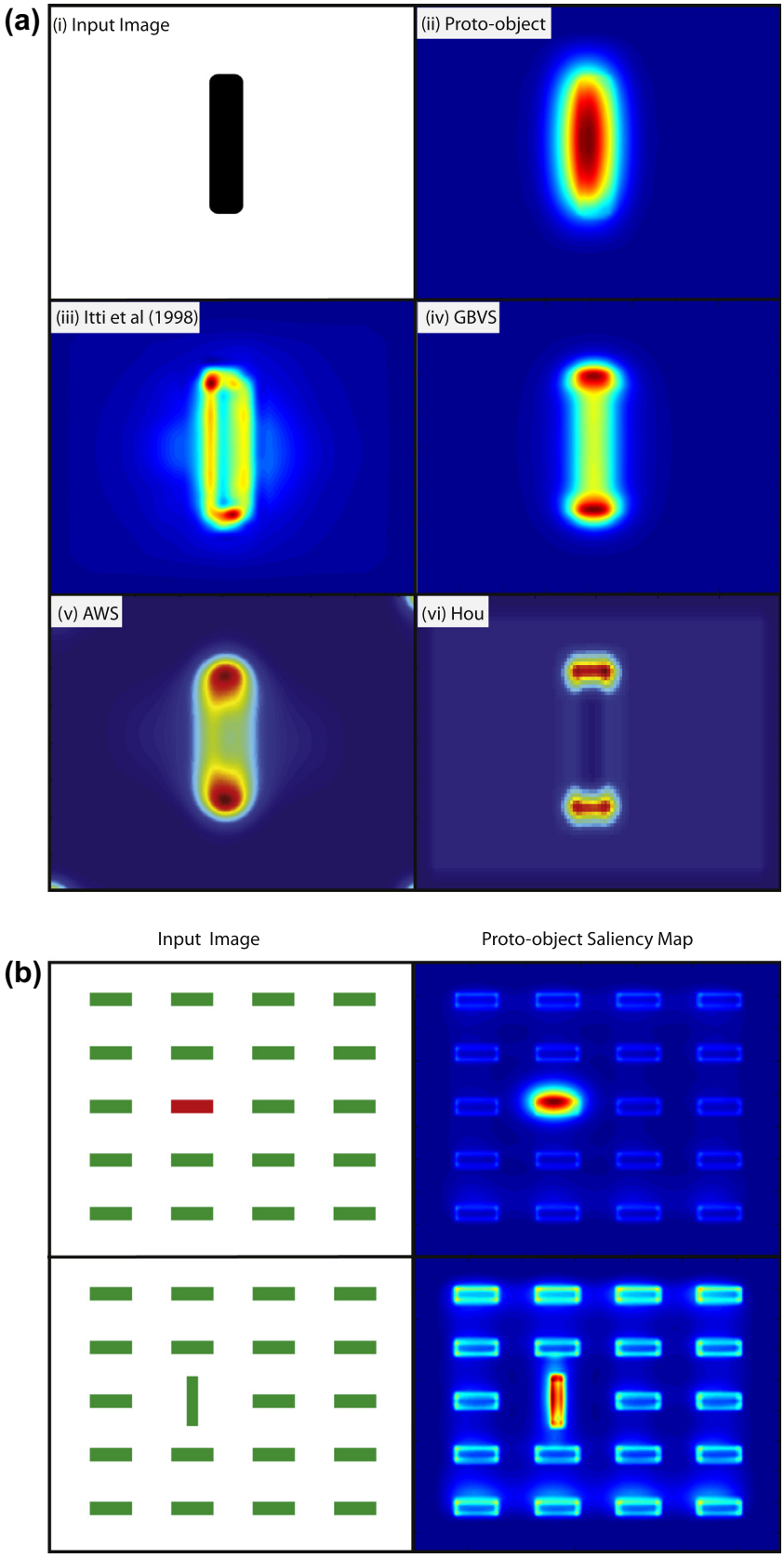

Fig. 6. (a-i) Image of a vertical bar. (a-ii) Proto-object saliency map of the bar. (a-iii) Itti, Koch, and Niebur (1998) saliency map of the bar. (a-iv) GBVS saliency map of the bar. (a-v) AWS saliency map of the bar. (a-vi) Hou saliency map of the bar. Red indicates high activity and blue the lowest activity. (b) Proto-object saliency maps for two feature search tasks. In both cases the item described by the unique feature is awarded the highest saliency. Red indicates high activity and blue the lowest activity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

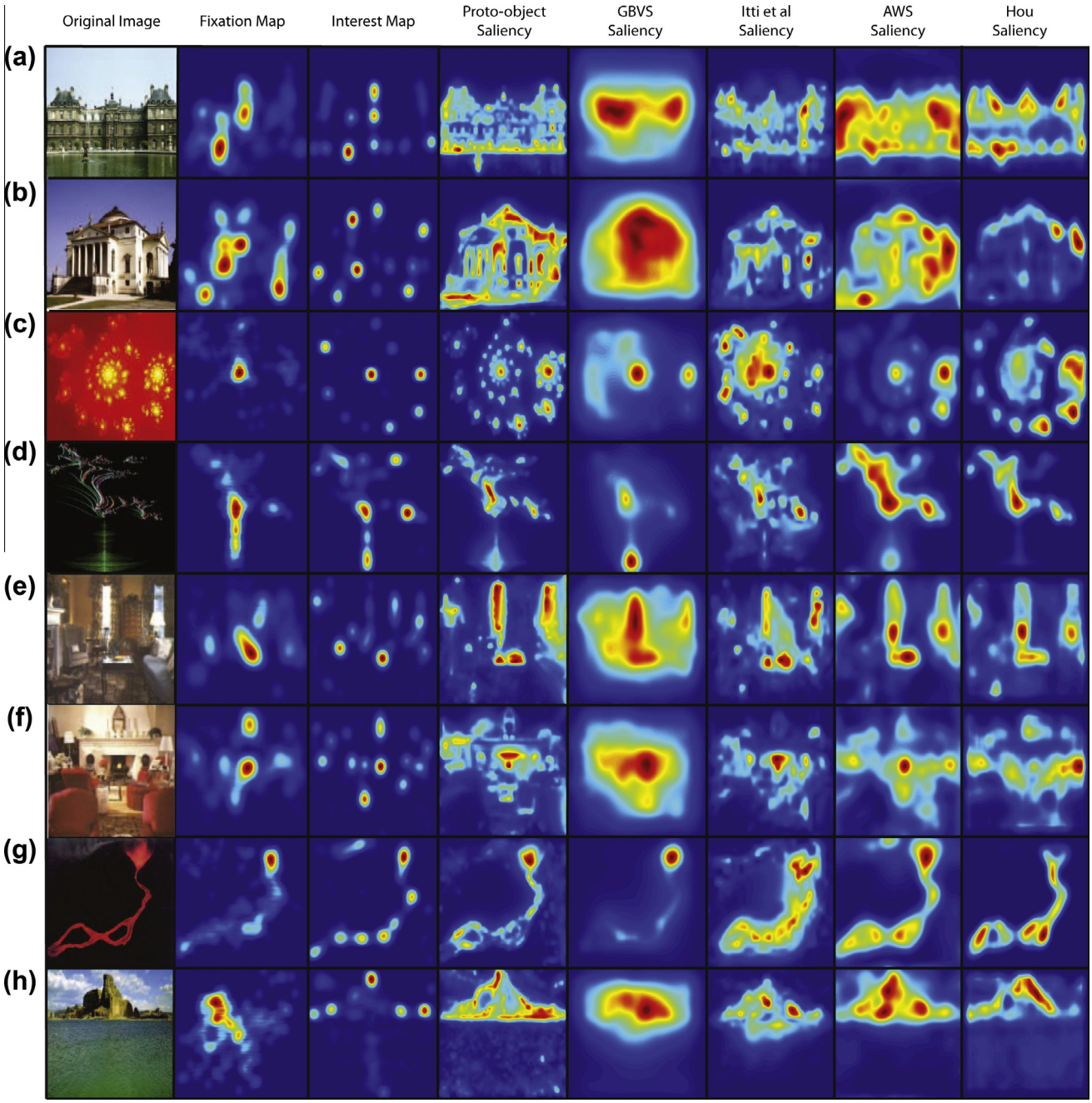

Fig. 7. Examples of images and their associated eye-fixation maps, interest maps and saliency maps calculated using the three algorithms discussed in the text. Two images are shown for each image category: buildings (a and b), fractals (c and d), home interiors (e and f) and landscapes (g and h). The eye fixation (interest) maps were generated by combining the fixation (interest) points across all images and then convolving the combined points with a 2D Gaussian with a standard deviation of 27 pixels, the standard deviation of the eye tracker error.

comparison the metrics are modified to only use saliency values sampled at human fixation (interest) points. This diminishes the influence of edge effects as human eye fixations and interest points are less likely to be near image edges. Furthermore, by only using salient values sampled at fixation (interest) locations the center bias effects all aspects of the metrics equally.

The modified Receiver Operating Characteristic (ROC) examines the sensitivity (as a function of true and false positives) of the saliency map to predict fixation (interest) points. To calculate an ROC curve the saliency map is treated as a binary classifier, where pixels below a threshold are classified as not fixated and pixels above a threshold are classified as fixated. By varying the threshold, and using eye fixations as ground truth, an ROC curve can be drawn. The area under the curve provides a metric as to how well the algorithm performed, with an area of 1 meaning perfect prediction and an area of 0.5 meaning chance; values below 0.5 correspond to anti-correlation. Following Tatler, Baddeley, and Gilchrist (2005) the eye fixation (interest) points for the image under consideration are used to sample the saliency map when computing true positives and the set of all fixation (interest) points pooled across all other images (drawn with no repeats) is used to sample the saliency map when computing false positives. The ROC scores are then normalized by the ROC score describing the ability of human fixation (interest) points to predict other human fixation (interest) points. To do this, the test subjects were randomly partitioned into two equally sized groups. As in Masciocchi et al. (2009), the

fixation points from one group were then convolved with a 2D Gaussian (standard deviation 27 pixels, equal to the standard deviation of the eye tracker errors) to generate a fixation map. The average ROC score of the fixation map's ability to predict the remaining fixation points was then calculated. This process was repeated 10 times and the average was used to perform the normalization. The modified ROC metric is the most reliable metric to use when quantifying a saliency map's ability to predict eye-fixations (Borji, Sihite, & Itti, 2012).

For the modified Kullback Leibler divergence (KLD) metric, the method described in Zhang et al. (2008) is used. For a given image, the KL divergence is computed between the histogram of saliency sampled at fixation (interest) points for that image, against a histogram of saliency sampled at the same fixation (interest) points but from the saliency map of a different image. This is repeated 99 times, once for each alternative saliency map in the data set, and the average result is the KLD score for that image. If the saliency map performs better than chance, then the histogram of saliency values computed at fixated (interesting) locations should have higher values than the histogram of saliency values computed at the same locations but sampled from a different saliency map. This will cause a high KL divergence between the two distributions. If the saliency map performs at chance then the KL divergence between the two histograms is low. The average KLD score, across all images, is then normalized by the KLD score describing the ability of human fixation (interest) points to predict other human fixation (interest) points. This was calculated in a similar way to the average ROC score described above.

4.1. Experiment 1: Predicting fixation points

Table 2 shows the ability of the five algorithms to predict human fixation points as computed by the ROC and KLD metrics and Table 3 shows the significance, in difference, of these results. Significance was calculated using a paired t-test between the proto-object model's score and the score of the competing algorithm. The results show that the AWS algorithm is significantly better ($p < 10^{-5}$) than the proto-object saliency algorithm as judged by the ROC metric. There is no significance difference between the results of the proto-object and AWS algorithms as judged by the KLD metric. The results also show that there is no significant difference between the proto-object and Hou algorithms for both metrics and that the proto-object based saliency algorithm is significantly better at predicting fixation points than the Itti, Koch, and Niebur (1998) and GBVS algorithms ($p < 10^{-2}$ for the difference between the KLD score for the proto-object algorithm and the GBVS algorithm, $p < 10^{-7}$ for all other cases).

4.2. Experiment 2: Predicting interest points

Table 2 shows the ability of the five algorithms to predict human interest points as computed by the ROC and KLD metrics and Table 3 shows the significance, in difference, of these results. The results show that the AWS algorithm is significantly better ($p < 10^{-5}$) than the proto-object saliency algorithm as judged by the ROC metric. There is no significance difference between the results of the proto-object and AWS algorithms as judged by the KLD metric. The results also show that there is no significant difference between the proto-object and Hou algorithms as computed by the ROC metric, however the proto-object algorithm is significantly better ($p < 10^{-2}$) at predicting human interest points as judged by the KLD metric. The proto-object saliency map scores significantly higher ($p < 10^{-7}$) in its ability to predict subjective interest points than both the Itti, Koch, and Niebur (1998) and GBVS algorithms.

Table 2
Average ability of the saliency maps to predict human eye fixations and subjective interest points across all images.

Algorithm	Fixation points		Interest points	
	Area ROC	KLD	Area ROC	KLD
Proto-object saliency	0.9208	1.3048	0.7874	0.4851
Feature saliency (Itti et al.)	0.8707	0.7197	0.7473	0.2696
Feature saliency (GBVS)	0.8668	1.1056	0.7111	0.3387
Feature saliency (AWS)	0.9483	1.2239	0.8100	0.4678
Feature saliency (Hou)	0.9213	1.1796	0.7828	0.4143

Table 3
Significance of the ability of proto-object saliency to predict eye fixations and subjective interest points compared to the ability of the feature based methods.

Algorithm	Metric	Fixation points		Interest points	
		Significant	p-Value	Significant	p-Value
Feature saliency (Itti et al.)	ROC	✓	$<10^{-9}$	✓	$<10^{-8}$
	KLD	✓	$<10^{-9}$	✓	$<10^{-12}$
Feature saliency (GBVS)	ROC	✓	$<10^{-7}$	✓	$<10^{-16}$
	KLD	✓	$<10^{-2}$	✓	$<10^{-7}$
Feature saliency (AWS)	ROC	✓	$<10^{-5}$	✓	$<10^{-5}$
	KLD	✗	–	✗	–
Feature saliency (Hou)	ROC	✗	–	✗	–
	KLD	✗	–	✓	$<10^{-2}$

5. Discussion

5.1. Proto-object saliency

In agreement with the work by Borji, Sihite, and Itti (2012) our results show that the AWS and Hou algorithms are the top performing feature based saliency models. However, contrary to the results of Borji, Sihite, and Itti (2012), we find that the performance of the Itti, Koch, and Niebur (1998) and GBVS algorithms is approximately equal. The results also show that AWS algorithm outperforms the proto-object algorithm (significantly according to the ROC metric, not significantly according to the KLD metric), but the proto-object algorithm is equal to (if not better than) the Hou algorithm and that it significantly outperfoms both the GBVS and Itti, Koch, and Niebur (1998) algorithms. When contrasting the results of the algorithms it should be noted that the approach taken in the design of the Itti, Koch, and Niebur (1998) and proto-object models was to build models in as biologically plausible a fashion as possible. This differs from the approach taken in the design of the AWS, Hou and GBVS algorithms which, although motivated by biology, take a higher level approach and do not attempt to model specific visual functions.

The AWS algorithm performs so well for two main reasons. Firstly, the Hou, GBVS, Itti and proto-object algorithms all use a limited, fixed feature space. As a result, each algorithm's performance depends on how well the features used in the algorithms match those found in the images being tested. In contrast, the feature space of the AWS algorithm is adapted to the statistical structure of each image. As a result the AWS algorithm uses the optimal features for each image being tested, as opposed to the other algorithms which use a fixed, non-optimal feature space for all images in the data set.

A second factor which can explain the AWS algorithm's performance is that, although not explicitly designed into the AWS algorithm, the AWS algorithm includes basic aspects of object based attention. Figs. 2 and 3 of Garcia-Diaz, Leborn, et al. (2012) show that the AWS algorithm exhibits early stages of figure-ground segregation (Garcia-Diaz, Fdez-Vidal, et al., 2012; Garcia-Diaz, Leborn,

et al., 2012). These results can be attributed to the whitening used by the AWS algorithm. Whitening is one property of center-surround mechanisms (Doi & Lewicki, 2005; Graham, Chandler, & Field, 2004). Thus, the whitening stages of the AWS algorithm are similar to filtering using center-surround mechanisms and the output of the whitening stage of the AWS algorithm approaches that of the CS_D and CS_L center-surround cells in the proto-object algorithm. If the normalization used in the AWS algorithm correctly enhances the figure and suppresses the ground in this output then the output saliency map from the AWS algorithm will include a basic representation of proto-objects. However, because of the adaptive nature of the AWS algorithm, these proto-objects may be at more optimal scales than the fixed proto-object sizes used in our algorithm. Furthermore, unlike the proto-object algorithm the AWS algorithm does not have any competition between the figure-ground responses created during whitening. Thus the saliency map from the AWS algorithm will be blurrier than that of the proto-object algorithm. This can be seen in Fig. 7. As mentioned above, blurrier maps tend to outperform sharper saliency maps (Borji, Sihite, & Itti, 2013).

It should be noted that although the AWS algorithm contains basic notions of figure-ground organization we still classify it as a feature based algorithm. The figure-ground representation in the AWS algorithm is tightly coupled to the spatial location of features in the image. The AWS algorithm does not include any mechanisms, such as the grouping cells in the proto-object algorithm, to process the arrangement of features in a scene into tentative proto-objects. Consequently, in scenarios where features and object are decoupled, such as in the Kimchi, Yeshurun, and Cohen-Savransky (2007) experiment shown in Fig. 1, the AWS algorithm is unable to award the highest saliency to the object in the scene. Our proto-object algorithm is the only model able to explain the Kimchi, Yeshurun, and Cohen-Savransky (2007) results.

The proto-object based model shares many computational mechanisms with the Itti, Koch, and Niebur (1998) algorithm. These results show that by incorporating perceptual scene organization into the algorithm it is possible to match the performance of sophisticated, non-biologically plausible models such as the Hou model which uses predictive coding techniques to calculate saliency. It should also be noted that in the proto-object model the spread of activation in the saliency map is localized to the figures in the images while the background receives very low saliency values. In contrast, in the feature based algorithms the activation of saliency is not localized to the figures. This is especially evident in Fig. 7(c) and (d). Blurrier saliency maps tend to outperform sharper saliency maps (Borji, Sihite, & Itti, 2013).

The figure ground organization in the proto-object saliency algorithm is a result of the $\mathcal{G}$ cells which integrate object features into proto-objects using large annular receptive fields (see Fig. 4). These receptive fields bias the grouping cell activity, and consequently salient locations, to fall on the centroids and the medial axis of the proto-objects (Ardila et al., 2012). An example of this is shown in Fig. 6(a-ii) where the highest activation in the saliency map corresponds to the center of the bar. Note that, in the feature based algorithms, the peaks of saliency tend to concentrate at the edges of objects. This is shown in Fig. 6(a-iii) through (a-vi). The results of the proto-object saliency map are confirmed by results obtained by Einhauser, Spain, and Perona (2008) and Nuthmann and Henderson (2010) who show that object centers are a better predictor of human fixations than object features.

In the Einhauser, Spain, and Perona (2008) study, subjects were presented with 99 images. The test subjects were asked to perform artistic evaluation, analysis of content and search on the data set whilst their eye fixations were recorded. Immediately after an image was presented, the subjects were asked to list objects which they saw. A hand-segmented object based saliency map was then created where the saliency of objects within the map was proportional to the recall frequency of the objects across all test subjects. From this, Einhauser, Spain, and Perona (2008) concluded that saliency only has an indirect effect on attention by acting through recognized objects. Consequently, they suggest that saliency is not just a result of preprocessing steps to higher vision but instead incorporates cognitive functions such as object recognition.

In a complementary experiment, Nuthmann and Henderson (2010) presented participants with 135 color images of real-world scenes. The pictures were divided into blocks of 45 images and while viewing each block, the subject was asked to either take a memory test, perform a search task, or evaluate aesthetic preferences. Eye fixations were recorded and it was found that the Preferred Viewing Location (PVL) of an object is always towards its center. In a second experiment the PVL of "saliency proto-objects", generated using the Walther and Koch (2006) algorithm, was investigated. A PVL existed for proto-objects which overlapped real objects; however, no PVL was found for saliency proto-objects which did not overlap real objects. Thus, when the influence of real objects is removed from saliency proto-objects, little evidence for a PVL remains. Consequently, Nuthmann and Henderson (2010) argue that saliency proto-objects are not selected for fixation and attention. Instead they hypothesize that a scene is parsed into constituent objects which are prioritized according to cognitive relevance.

At first glance it may appear as if the findings of Nuthmann and Henderson (2010) are contradictory to the work presented in this paper; however a distinction must be made between the Walther and Koch (2006) proto-objects and the proto-objects of this work. Both are close to the notion of proto-objects as defined by Rensink (2000), but their implementations and interpretations are fundamentally different. The Walther and Koch (2006) model uses the Itti, Koch, and Niebur (1998) feature based saliency algorithm to compute the most salient location in the visual field. The shape of the proto-object at that location is then calculated by a spreading of activation around the most conspicuous feature at that location. The proto-objects are purely a function of individual features—there is no notion of object in their proto-objects. In contrast, the proto-objects in this paper are represented by grouping cells whose activity is dependent on not only the individual features of an object but also on Gestalt principles of perceptual organization. The grouping cells provide a handle for selective attention by not only providing the spatial location of potential objects within a scene but also by acting as pointers to the features which constitute an object, akin to how a symbolic pointer in a computer program can point to a structure composed of many individual elements. When cast in the saliency framework of Section 3, the normalized grouping cell activity provides a measure of how unique an object is and consequently a measure of its saliency. This is in line with recent neurophysiological results which show that border ownership is computed independently of (top down) attention (Qiu, Sugihara, & von der Heydt, 2007). This suggests that saliency is a function of proto-objects and not the other way around, as implicitly assumed in the Walther and Koch (2006) implementation. In line with Rensink (2000)'s proto-object definition, the grouping cells are the highest output level of low level visual processes and provide a purely feed forward measure of object based attention. However, again following Rensink, grouping cells can also act as the lowest level of top down processes. In fact, by using a similar network of border ownership and grouping cells, Mihalaş et al. (2011) have shown that grouping cells can explain many psychophysical results of top down attention (this is discussed in more detail below). Using this distinction, the experiments of Nuthmann and Henderson (2010) do not exclude proto-objects

as a mechanism through which object based attention can be explained. Instead, their results bolster the growing literature (Cave & Bichot, 1999; Duncan, 1984; Egly, Driver, & Rafal, 1994; He & Nakayama, 1995; Ho & Yeh, 2009; Ito & Gilbert, 1999; Matsukura & Vecera, 2006; Qiu, Sugihara, & von der Heydt, 2007; Roelfsema, Lamme, & Spekreijse, 1998; Scholl, 2001; Wannig, Stanisor, & Roelfsema, 2011) that supports the idea that attention is object and not feature based.

Both Einhauser, Spain, and Perona (2008) and Nuthmann and Henderson (2010) posit that their results can only be explained through higher order neural mechanisms, such as object recognition, which are used to guide object based attention. While there is no denying that attention has a strong top down component, our the results provide an alternative explanation, namely that object based saliency, acting through proto-objects, can also direct attention. A recent study by Monosov, Sheinberg, and Thompson (2010) shows that, in a visual search task, attention is applied before object recognition is performed. This agrees with a proto-object based theory of attention as neurophysiological results show that border ownership signals emerge 50–70 ms after stimulus presentation (Qiu, Sugihara, & von der Heydt, 2007; Zhou, Friedman, & von der Heydt, 2000) while object recognition is a relatively slower process occurring 120–150 ms after stimulus presentation (Johnson & Olshausen, 2003).

5.2. Object based bias for interest

Experiments by Masciocchi et al. (2009) and Elazary and Itti (2008) show that subjective interest points are biased by a bottom up component. Using the same image database as that used in this paper (see Section 4 for details), Masciocchi et al. (2009) investigated the correlation between subjective interest points, eye fixations and salient locations generated using the Itti, Koch, and Niebur (1998) saliency algorithm. Interestingly, Masciocchi et al. (2009) found that participants agreed in which things were "interesting" even though the interest point selection process was entirely subjective and performed independently by all participants. Furthermore, positive correlations were found between interest points and both eye-fixations and salient locations. This suggests that the selection of interesting locations is in part driven by bottom up factors and that interest points can serve as an indicator of bottom up attention (Masciocchi et al., 2009). In the Elazary and Itti (2008) experiment, participants were not explicitly asked to label interesting locations, instead interest was defined as the property that makes participants label one object over another in the *LabelMe* database (available at http://labelme.csail.mit.edu). The Itti, Koch, and Niebur (1998) saliency algorithm was then applied to the database and its ability to predict interesting objects was analyzed. Elazary and Itti (2008) found that saliency was a significant predictor of which objects participants chose to label. In 76% of the images, the saliency algorithm finds at least one labeled object within the three most salient locations of the image. This indicates that interesting objects within a scene are not only dependent on higher cognitive processes but also on low level visual properties (Elazary & Itti, 2008).

Both Masciocchi et al. (2009) and Elazary and Itti (2008) surmise that the segmentation of a scene into objects is an important factor in interest point selection. Our experiment 2 investigates this by testing whether or not the bottom up bias of selective interest is better explained through proto-objects or through features. The results, see Section 4, show that proto-object based saliency matches or outperforms feature based saliency algorithms (except for the AWS algorithm) in its ability to predict interest points. This indicates that the bottom up component of interest is not only dependent on saliency but also on the perceptual organization of a scene into tentative objects.

5.3. The interface theory of attention

The work presented in this paper uses a feed forward network of $\mathcal{B}$ and $\mathcal{G}$ cells to compute figure ground organization and saliency. In a complementary study, Mihalaş et al. (2011) demonstrated that a recurrent model, using a similar network of $\mathcal{B}$ and $\mathcal{G}$ cells, could perform top-down object based attention. When a broadly tuned, spatial top-down signal was applied to the grouping neurons representing a given proto-object, attention backpropagated through the network enhancing the local features ($\mathcal{B}$ cells) of the object — top down attention auto-localized and auto-zoomed to fit the contours of the proto-object (Mihalaş et al., 2011). Using this network Mihalaş et al. (2011) were able to reproduce the psychophysical phenomena described by Egly, Driver, and Rafal (1994) and Kimchi, Yeshurun, and Cohen-Savransky (2007).

Together these two studies provide support for the "interface" theory of attention (Craft et al., 2007; Qiu, Sugihara, & von der Heydt, 2007; Zhou, Friedman, & von der Heydt, 2000) where the neuronal network that creates figure ground-organization provides an interface for bottom-up and top-down selection processes. This is a natural fit with coherence theory (Rensink, 2000), where the $\mathcal{G}$ cells provide a handle or interface to the proto-objects for both bottom up and top down processing. In the interface theory the magnitude of attentional enhancement is not dependent on where in the cortical hierarchy the attentional processing is performed but rather on how involved the local circuits are in processing contextual scene information (Kastner & McMains, 2007).

5.4. Grouping cell receptive fields and local features

An important aspect of the grouping algorithm are the large annular receptive fields of the $\mathcal{G}$ cells which bias grouping activity for figures exhibiting the Gestalt principles of continuity and proximity. Although, there is no direct electrophysiological evidence yet that shows that cells with such receptive fields exist, psychophysical evidence points to special integration mechanisms for concentric circular patterns (Sigman et al., 2001; Wilson, Wilkinson, & Asaad, 1997).

Furthermore, there is neurophysiological evidence for neurons selective for concentric gratings (Gallant et al. (1996)). As an alternative, the $\mathcal{G}$ cells do not need complete annular receptive fields (Craft et al., 2007); instead their responses could be computed through intermediate cells tuned to curved contour segments or combinations of such segments (as was performed in our computations). Cells exhibiting such properties have been shown to exist in extrastriate cortex (Brincat & Connor, 2004; Pasupathy & Connor, 2001; Yau et al., 2013).

In this work grouping is assigned through the Gestalt principles of proximity and continuity. Although excluded from this iteration of the model, other Gestalt principles, such as symmetry are also important. Indeed, a saliency map which uses local mirror symmetry has been found to be a strong predictor of eye fixations (Kootstra, de Boer, & Schomaker, 2011). Furthermore, in this model, border ownership assignment is purely a result of (estimated) $\mathcal{G}$ cell activity and local orientation features. Additional features such as T-junctions (Craft et al., 2007) and stereoscopic cues (Qiu & von der Heydt, 2005) are also important for correct border ownership assignment. For a more accurate description of perceptual scene organization, future iterations of the model should include such mechanisms.

6. Conclusion

A biologically plausible model of proto-object saliency has been presented. The model is constructed out of basic computation

mechanisms with known biological correlates, yet is able to match the performance of state of the art, non bio-inspired algorithms. The performance of the algorithm strengthens the growing body of evidence which suggests attention is object based. In addition, the model was used to investigate whether the bottom up bias in subjective interest is object or feature based. To the authors' knowledge this is the first experiment of this kind. The results support the idea that the bias is object based. Lastly, the model supports the "interface" hypothesis which states that attention is a result of how involved the local circuitry of perceptual organization is in processing the visual scene.

Acknowledgments

We acknowledge support by the Office of Naval Research under MURI Grant N000141010278 and by NIH under R01EY016281-02.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America – A, 2*, 284–299.

Ardila, D., Mihalas, S., von der Heydt, R., & Niebur, E. (2012). Medial axis generation in a model of perceptual organization. In *46th Annual conference on information sciences and systems* (pp. 1–4). IEEE Press.

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

Basso, M. A., & Wurtz, R. H. (1998). Modulation of neuronal activity in superior colliculus by changes in target probability. *The Journal of Neuroscience, 18*, 7519–7534.

Beauchamp, M. S., Petit, L., Ellmore, T. M., Ingeholm, J., & Haxby, J. V. (2001). A parametric fMRI study of overt and covert shifts of visuospatial attention. *Neuroimage, 14*(2), 310–321.

Bisley, J. W., & Goldberg, M. E. (2003). Neuronal activity in the lateral intraparietal area and spatial attention. *Science, 299*, 81–86.

Borji, A., Sihite, D. N., & Itti, L. (2012). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*.

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *In Proc European conference on computer vision (ECCV), 22*(1), 55–69. http://dx.doi.org/10.1109/TIP.2012.2210727, Epub 2012 Jul 30.

Brincat, S. L., & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience, 7*, 880–886.

Bisley, J. W., & Goldberg, M. E. (2003). Neuronal activity in the lateral intraparietal area and spatial attention. *Science, 299*, 81–86.

Borji, A., Sihite, D. N., & Itti, L. (2012). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*.

Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *In Proc European conference on computer vision (ECCV), 22*(1), 55–69. http://dx.doi.org/10.1109/TIP.2012.2210727.

Brincat, S. L., & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience, 7*, 880–886.

Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon.

Cave, K. R., & Bichot, N. P. (1999). Visuospatial attention: Beyond a spotlight model. *Psychonomic Bulletin & Review, 6*, 204–223.

Constantinidis, C., & Steinmetz, M. A. (2005). Posterior parietal cortex automatically encodes the location of salient stimuli. *The Journal of Neuroscience, 25*, 233–238.

Corbetta, M., Akbudak, E., Conturo, T. E., Snyder, A. Z., Ollinger, J. M., Drury, H. A., et al. (1998). A common network of functional areas for attention and eye movements. *Neuron, 21*, 761–773.

Craft, E., Schütze, H., Niebur, E., & von der Heydt, R. (2007). A neural model of figure-ground organization. *Journal of Neurophysiology, 97*(6), 4310–4326.

Doi, E., & Lewicki, M. S. (2005). Relations between the statistical regularities of natural images and the response properties of the early visual system. In *Japanese cognitive science society*, Kyoto University, July 2005 (pp. 1–8).

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 113*, 501–517.

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.

Einhauser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14), 1–26.

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision, 8*, 1–15.

Gallant, J. L., Connor, C. E., Rakshit, S., Lewis, J. W., & Van Essen, D. C. (1996). Neural responses to polar, hyperbolic, and Cartesian gratings in area v4 of the macaque monkey. *Journal of Neurophysiology, 76*(4), 2718–2739.

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

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

Graham, D. J., Chandler, D. M., & Field, D. J. (2004). Decorrelation and response equalization with center-surround receptive fields. *Journal of Vision, 4*(8), 276.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: John Wiley.

Hafed, Z. M., & Clark, J. J. (2002). Microsaccades as an overt measure of covert attention shifts. *Vision Research, 42*(22), 2533–2545. cited By (since 1996) 72.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *NIPS* (pp. 545–552). Cambridge, MA: MIT Press.

He, Z. J., & Nakayama, K. (1995). Visual attention to surfaces in three-dimensional space. *Proceedings of the National Academy of Sciences of the United States of America, 92*, 11155–11159.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics, 57*(6), 787–795.

Hou, X., & Zhang, L. (2008). Dynamic visual attention: Searching for coding length increments, In *NIPS* (pp. 681–688).

Ho, M. C., & Yeh, S. L. (2009). Effects of instantaneous object input and past experience on object-based attention. *Acta Psychologica (Amst), 132*, 31–39.

Ito, M., & Gilbert, C. D. (1999). Attention modulates contextual influences in the primary visual cortex of alert monkeys. *Neuron, 22*, 593–604.

Itti, L., & Baldi, P. (2005). A principled approach to detecting surprising events in video. In *IEEE proc. CVPR*, Washington, DC, September 2005 (Vol. 1, pp. 631–637).

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. *NIPS* (Vol. 18, pp. 1–8). Cambridge, MA: MIT Press.

Itti, L., & Koch, C. (2001a). Computational modeling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194–203.

Itti, L., & Koch, C. (2001b). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging, 10*, 161–169.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE PAMI, 20*(11), 1254–1259.

Johnson, J. S., & Olshausen, B. A. (2003). Timecourse of neural signatures of object recognition. *Journal of Vision, 3*(7), 499–512.

Jones, J. P., & Palmer, A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology, 58*(6), 1233–1258.

Kanizsa, G. (1979). *Organization in vision: Essays on Gestalt perception*. New York, NY: Praeger Publishers.

Kastner, S., & McMains, S. A. (2007). Out of the spotlight: Face to face with attention. *Nature Neuroscience, 10*, 1344–1345.

Kimchi, R., Yeshurun, Y., & Cohen-Savransky, A. (2007). Automatic, stimulus-driven attentional capture by objecthood. *Psychonomic Bulletin & Review, 14*, 166–172.

Koch, K., McLean, J., Berry, M., Sterling, P., Balasubramanian, V., & Freed, M. A. (2004). Efficiency of information transmission by retinal ganglion cells. *Current Biology, 14*(17), 1523–1530.

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

Koene, A. R., & Zhaoping, L. (2007). Feature-specific interactions in salience from combined feature contrasts: Evidence for a bottom-up saliency map in v1. *Journal of Vision, 7*(7), 6.

Koffka, K. (1935). *Principles of Gestalt psychology*. NY: Hartcourt.

Kootstra, G., de Boer, B., & Schomaker, L. (2011). Predicting eye fixations on complex visual stimuli using local symmetry. *Cognitive Computation, 3*(1), 223–240. ISSN 1866-9956.

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research, 35*(13), 1897–1916.

Kulikowski, J., Marcelja, S., & Bishop, P. (1982). Theory of spatial position and spatial frequency relations in the receptive fields of simple cells in the visual cortex. *Biological Cybernetics, 43*, 187–198.

Kustov, A. A., & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. *Nature, 384*(6604), 74–77.

Kusunoki, M., Gottlieb, J., & Goldberg, M. E. (2000). The lateral intraparietal area as a salience map: The representation of abrupt onset, stimulus motion, and task relevance. *Vision Research, 40*, 1459–1468.

Li, Z. (1998). Primary cortical dynamics for visual grouping. In K. M. Wong, I. King, & D. Y. Yeung (Eds.), *Theoretical aspects of neural computation* (pp. 154–164). Springer-Verlag.

Marcelja, S. (1980). Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America, 70*, 1297–1300.

Masciocchi, C. M., Mihalas, S., Parkhurst, D., & Niebur, E. (2009). Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision, 9*(11), 25.1–22.

Matsukura, M., & Vecera, S. P. (2006). The return of object-based attention: Selection of multiple-region objects. *Perception & Psychophysics, 68*, 1163–1175.

Mazer, J. A., & Gallant, J. L. (2003). Goal-related activity in v4 during free viewing visual search. evidence for a ventral stream visual salience map. *Neuron, 40*(6), 1241–1250.

McPeek, R. M., & Keller, E. L. (2002). Saccade target selection in the superior colliculus during a visual search task. *Journal of Neurophysiology, 88*, 2019–2034.

Mihalaş, S., Dong, Y., von der Heydt, R., & Niebur, E. (2011). Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention

to objects. *Proceedings of the National Academy of Sciences of the United States of America, 108*(18), 7583–7588.
Milanese, R., Gil, S., & Pun, T. (1995). Attentive mechanisms for dynamic and static scene analysis. *Neural Computation, 34*, 2428–2434.
Monosov, I. E., Sheinberg, D. L., & Thompson, K. G. (2010). Paired neuron recordings in the prefrontal and inferotemporal cortices reveal that spatial selection precedes object identification during visual search. *Proceedings of the National Academy of Sciences of the United States of America, 107*(29), 13105–13110.
Moore, T., Armstrong, K. M., & Fallah, M. (2003). Visuomotor origins of covert spatial attention. *Neuron, 40*(4), 671–683.
Moore, T., & Fallah, M. (2001). Control of eye movements and spatial attention. *Proceedings of the National Academy of Sciences of the United States of America, 98*(3), 1273–1276.
Moore, T., & Fallah, M. (2004). Microstimulation of the frontal eye field and its effects on covert spatial attention. *Journal of Neurophysiology, 91*(1), 152–162.
Morrone, M. C., & Burr, D. C. (1988). Feature detection in human vision: A phase-dependent energy model. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 235*, 221–245.
Niebur, E., & Koch, C. (1996). Control of selective visual attention: Modeling the "where" pathway. In D. S Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.). *Advances in neural information processing systems* (Vol. 8, pp. 802–808). Cambridge, MA: MIT Press.
Nobre, A. C., Gitelman, D. R., Dias, E. C., & Mesulam, M. M. (2000). Covert visual spatial orienting and saccades: Overlapping neural systems. *NeuroImage, 11*(3), 210–216.
Nuthmann, A., & Henderson, J. (2010). Object-based attentional selection in scene viewing. *Journal of Vision, 10*(8), 1–19.
Parkhurst, D. (2002). Selective attention in natural vision: Using computational models to quantify stimulus-drive attentional allocation. PhD Thesis, The Johns Hopkins University, Baltimore, MD.
Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research, 42*(1), 107–123.
Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 16*, 125–154.
Pasupathy, A., & Connor, C. E. (2001). Shape representation in area v4: Position-specific tuning for boundary conformation. *Journal of Neurophysiology, 83*, 2505–2519.
Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience, 13*, 25–42.
Qiu, F. T., Sugihara, T., & von der Heydt, R. (2007). Figure-ground mechanisms provide structure for selective attention. *Nature Neuroscience, 10*, 1492–1499.
Qiu, F. T., & von der Heydt, R. (2005). Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. *Neuron, 47*, 155–166.
Reid, R. C. (2008). The visual system. In M. Zigmond, F. Bloom, S. Landis, J. Roberts, & L. Squire (Eds.), *Fundamental neuroscience*. San Diego: Academic Press.
Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network, 10*, 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*, 1019–1025.
Rizzolatti, G., Riggio, L., Dascola, I., & Umilt, C. (1987). Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychologia (1, Part 1), 25*, 31–40.
Robinson, D. L., & Petersen, S. E. (1992). The pulvinar and visual salience. *Trends in Neurosciences, 15*(4), 127–132. ISSN 0166-2236.
Roelfsema, P. R., Lamme, V. A., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature, 395*, 376–381.
Roe, A., Lu, H., & Hung, C. (2005). Cortical processing of a brightness illusion. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 3869–3874.
Rossi, A., & Paradiso, M. (1999). Neural correlates of perceived brightness in the retina, lateral geniculate nucleus, and striate cortex. *The Journal of Neuroscience, 19*, 6145–6156.
Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition, 80*, 1–46.
Schumann, F., Einhauser-Treyer, W., Vockeroth, J., Bartl, K., Schneider, E., & Konig, P. (2008). Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of Vision, 8*, 1–17.
Sheliga, B. M., Riggio, L., & Rizzolatti, G. (1994). Orienting of attention and eye movements. *Experimental Brain Research, 98*(3), 507–522.
Sheliga, B. M., Riggio, L., & Rizzolatti, G. (1995). Spatial attention and eye movements. *Experimental Brain Research, 105*(2), 261–275.
Sigman, M., Cecchi, G. A., Gilbert, C. D., & Magnasco, M. O. (2001). On a common circle: Natural scenes and gestalt rules. *Proceedings of the National Academy of Sciences of the United States of America, 98*(4), 1935–1940. http://dx.doi.org/10.1073/pnas.031571498.
Strong, S. P., Koberle, R., de Ruyter van Steveninck, R. R., & Bialek, W. (1998). Entropy and information in neural spike trains. *Physical Review Letters, 80*, 197–200.
Sun, Y., & Fisher, R. (2003). Object based visual attention for computer vision. *Artificial Intelligence, 146*, 77–123.
Sun, Y., Fisher, R., Wang, F., & Gomes, H. (2008). A computer vision model for visual-object-based attention and eye movements. *Computer Vision and Image Understanding, 112*(2), 126–142.
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*, 1–17.
Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research, 45*, 643–659.
Thompson, K. G., & Bichot, N. P. (2005). A visual salience map in the primate frontal eye field. *Progress in Brain Research, 147*, 251–262.
Treisman, A. (1988). Features and objects: The fourteenth Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology: Section A, 40*(2), 201–237.
Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97–136.
Tseng, P., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision, 9*(7). http://dx.doi.org/10.1167/9.7.4.
Tsotsos, J. (1991). Is complexity theory appropriate for analysing biological systems? *Behavioral and Brain Sciences, 14*(4), 770–773.
Vitu, F., Kapoula, Z., Lancelin, D., & Lavigne, F. (2004). Eye movements in reading isolated words: Evidence for strong biases towards the center of the screen. *Vision Research, 44*, 321–338.
von der Heydt, R., Friedman, H. S., & Zhou, H. (2003). Filling-in: From perceptual completion to cortical reorganization. Oxford University Press, pp. 106–127 (Chapter Searching for the neural mechanism of color filling-in).
Walther, D., Itti, L., Riesenhuber, M., Poggio, T., & Koch, C. (2002). Attentional selection for object recognition – A gentle way. *Lecture Notes in Computer Science* (Vol. 2525, pp. 472–479). Springer.
Walther, D., & Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks, 19*(9), 1395–1407.
Wannig, A., Stanisor, L., & Roelfsema, P. R. (2011). Automatic spread of attentional response modulation along Gestalt criteria in primary visual cortex. *Nature Neuroscience, 14*, 1243–1244.
White, B., & Munoz, D. (2010). Independent influence of luminance and color on saccade initiation during target selection in the superior colliculus. *Journal of Vision, 10*(7), 1320.
Wilson, H., Wilkinson, F., & Asaad, W. (1997). Concentric orientation summation in human form vision. *Vision Research, 37*(17), 2325–2330.
Wolfe, J. M. (1994). Guided search 2.0 – A revised model of visual search. *Psychonomics Bulletin & Review, 1*(2), 202–238.
Wolfe, J. M. (2000). Visual attention. *Seeing, 2*, 335–386.
Wolfe, J. M. (2007). Guided search 4.0. *Integrated models of cognitive systems*, pp. 99–119.
Yau, J., Pasupathy, A., Brincat, S., & Connor, C. (2013). Curvature processing dynamics in macaque area V4. *Cerebral Cortex, 23*, 198–209.
Zenon, A., Filali, N., Duhamel, J. R., & Olivier, E. (2010). Salience representation in the parietal and frontal cortex. *Journal of Cognitive Neuroscience, 22*, 918–930.
Zhang, L., Tong, M., Marks, T., Shan, H., & Cottrell, G. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7), 32.1–20.
Zhang, N. R., & von der Heydt, R. (2010). Analysis of the context integration mechanisms underlying figure-ground organization in the visual cortex. *The Journal of Neuroscience, 30*(19), 6482–6496.
Zhaoping, L. (2008). Attention capture by eye of origin singletons even without awareness—A hallmark of a bottom-up saliency map in the primary visual cortex. *Journal of Vision, 8*(5).
Zhou, H., Friedman, H., & von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience, 20*, 6591–6611.