Saliency, attention, and visual search: An information theoretic approach

Neil D. B. Bruce

Department of Computer Science & Engineering and Centre for Vision Research, York University, Toronto, ON, Canada

John K. Tsotsos

Department of Computer Science & Engineering and Centre for Vision Research, York University, Toronto, ON, Canada

A proposal for saliency computation within the visual cortex is put forth based on the premise that localized saliency computation serves to maximize information sampled from one's environment. The model is built entirely on computational constraints but nevertheless results in an architecture with cells and connectivity reminiscent of that appearing in the visual cortex. It is demonstrated that a variety of visual search behaviors appear as emergent properties of the model and therefore basic principles of coding and information transmission. Experimental results demonstrate greater efficacy in predicting fixation patterns across two different data sets as compared with competing models.

Keywords: saliency, visual attention, visual search, eye movements, information theory, efficient coding, pop-out, search asymmetry, independent components, redundancy, statistical inference

Citation: Bruce, N. D. B., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision, 9*(3):5, 1–24, http://journalofvision.org/9/3/5/, doi:10.1167/9.3.5.

Introduction

Humans perform visual search tasks constantly from finding a set of keys to looking for a friend in a crowded place. However, despite the importance of this task and its ubiquity in our everyday lives, the current understanding of the neural underpinnings of this behavior falls short of forming a consensus opinion. The steep drop-off in visual acuity from the fovea to the periphery necessitates an efficient system for directing the eyes onto those areas of the scene that are relevant to satisfying the goals of an observer. Moreover, a related and important task is the direction of the focus of attention cortically; that is, the cortical mechanisms underlying the direction of focused processing onto task relevant visual input.

Over the last several decades, a great deal of research effort has been directed toward further understanding the mechanisms that underlie visual sampling, either through observing fixational eye movements, or in considering the control of focal cortical processing. Consideration of fixational eye movements necessarily involves two distinct components, one being the top-down task-dependent influence on these behaviors, and the second characterized by bottom-up stimulus-driven factors governed by the specific nature of the visual stimulus.

The importance of the former of these categories is well documented and perhaps most prominently demonstrated by Yarbus (1967). In the experiments of Yarbus, observers were asked a variety of different questions about a specific scene while having their eye movements tracked. The resulting data demonstrates wildly different patterns of eye movements depending on the question posed. More recent efforts have continued in the same vein (Hayhoe & Ballard, 2005; Hayhoe, Shrivastava, Mruczek, & Pelz, 2003; Land, Mennie, & Rusted, 1999), observing eye movements in a variety of real-world settings and further demonstrating the role of task in the direction of visual and presumably cortical sampling.

Certain visual events such as a bright flash of light, a vividly colored sign, or sudden movement will almost certainly result in an observer's gaze being redirected, independent of any task-related factors. These behaviors reflect the bottom-up stimulus-driven component of visual sampling behavior. Even in the absence of such remarkable visual patterns, the specific nature of the visual stimulus at hand no doubt factors appreciably into the visual sampling that ensues. A number of studies have attempted to expound on this area by observing correlation between fixations made by human observers and basic features such as edges or local contrast (Parkhurst, Law, & Niebur, 2002; Tatler, Baddeley, & Gilchrist, 2005). The general finding of such studies is that there is no simple single basic feature that adequately characterizes what comprises salient content across all images. An additional limitation of such an approach is that any result of such a study says little about the underlying neural basis for such computation or the corresponding neural implementation.

An additional domain in which saliency is considered is in the context of attention models that posit the existence

of what has been called a saliency map. The introduction of saliency maps came conceptually with Treisman and Gelade's (1980) Feature Integration Theory in the form of what they describe as a *master map of locations*. The basic structure of the model is that various basic features are extracted from the scene. Subsequently the distinct feature representations are merged into a single topographical representation of saliency. In later work this representation has been deemed a saliency map and includes with it a selection process that in vague terms selects the largest peak in this representation, and the *spotlight* of attention moves to the location of this peak (Koch & Ullman, 1985). In this context, the combined pooling of the basic feature maps is referred to as the saliency map. Saliency in this context then refers to the output of an operation that combines some basic set of features into a solitary representation.

Although models based on a saliency map have had some success in predicting fixation patterns and visual search behavior, there exists one significant methodological shortcoming of the definition of saliency captured by these saliency map models. The definition of saliency is emergent from a definition of local feature contrast that is loosely based on observations concerning interaction among cells locally within primate visual cortex. Although the models succeed at simulating some salience-related behaviors, they offer little in explaining *why* the operations involved in the model have the structure that is observed and, specifically, what the overall architecture translates into with respect to its relationship to the incoming stimulus in a principled quantitative manner. As such, little is offered in terms of an explanation for design principles behind observed behavior and the structure of the system.

In this paper, we consider the role that the properties of visual stimuli play in sampling from the stimulus-driven perspective. The ambition of this work lies in explaining *why* certain components implicated in visual saliency computation behave as they do and also presents a novel model for visual saliency computation built on a first principles information theoretic formulation dubbed Attention based on Information Maximization (AIM). This comprises a principled explanation for behavioral manifestations of AIM and contributions of this paper include:

1. A computational framework for visual saliency built on first principles. Although AIM is built entirely on computational constraints, the resulting model structure exhibits considerable agreement with the organization of the human visual system.
2. A definition of visual saliency in which there is an implicit definition of context. That is, the proposed definition of visual salience is not based solely on the response of cells within a local region but on the relationship between the response of cells within a local region and cells in the surrounding region. This includes a discussion of the role that context plays in the behavior of related models.
3. Consideration of the impact of principles underlying neural coding on the determination of visual saliency and visual search behavior. This includes a demonstration that a variety of visual search behaviors may be seen as emergent properties of principles underlying neural coding combined with information seeking as a visual sampling strategy.
4. A demonstration that the resulting definition of visual saliency exhibits greater agreement with fixational eye movement data than existing efforts.

As a whole, we establish that an information maximization strategy for saliency-related neural gain control is consistent with the computation observed in the visual cortex. These results are discussed in terms of implications with respect to how attentional selection in general is achieved within the visual cortex.

Information maximization and visual sampling

The central core of the model is built on computational constraints derived from efficient coding and information theory. The intuition behind the role of these elements in saliency computation may be introduced by considering an example from an early influential paper by Attneave (1954) that considers aspects of information theory as they pertain to visual processing. Within this work, Attneave provides the following description and associated figure (labeled Figure 1):

> "Consider the very simple situation presented in Figure 1. With a modicum of effort, the reader may be able to see this as an ink bottle on the corner of a desk. Let us suppose that the background is a

uniformly white wall, that the desk is a uniform brown, and that the bottle is completely black. The visual stimulation from these objects is highly redundant in the sense that portions of the field are highly predictable from other portions. In order to demonstrate this fact and its perceptual significance, we may employ a variant of the "guessing game" technique with which Shannon has studied the redundancy of printed English. We may divide the picture into arbitrarily small elements, which we "transmit" to a subject (S) in a cumulative sequence, having him guess at the color of each successive element until he is correct…. If the picture is divided into 50 rows and 80 columns, as indicated, our S will guess at each of 4,000 cells as many times as necessary to determine which of the three colors it has. If his error score is significantly less than chance [2/3 × 4,000 + 1/2(2/3 × 4,000) = 4,000], it is evident that the picture is to some degree redundant. Actually, he may be expected to guess his way through Figure 1 with only 15 or 20 errors."

The intent of Attneave's example is to demonstrate that there exists significant redundancy in natural visual stimuli and that human subjects appear to have some degree of an internal model of this redundancy. A second observation that is not made in the original description, but that is fundamental to the subject matter of this paper, is that one might also suggest that the areas of the scene where subjects make the greatest number of errors on average in guessing are those that contain content of interest. This is equivalent to the Shannon (1948) self-information associated with each pixel location in this context.

One may also imagine a hypothetical generalization of the game described by Attneave in which a human participant is required to describe the contents of a region of a scene containing arbitrary structure, lightness, contrast, and colors. Although it is not practical to carry out an experiment of this type, most would agree that there exists some general intuition concerning what a certain portion of a scene is expected to contain on the basis of its context. Consider for example Figure 2: Under the blacked out regions (left) labeled A, B, and C, one would likely claim to have some general intuition concerning the contents of each hidden region on the basis of the surround and the contents of the scene. It is also evident from the frame on the right that the image content hidden within regions A and B come very close to this intuition whereas the content that lies beneath region C is very far from our expectation and would almost certainly require the greatest number of guesses within the hypothetical guessing game. This region is also the most informative in a Shannon sense on this basis.

Recently, a variety of proposals based on information theory concerning attention and fixation behavior have emerged. An earlier proposal of ours presents in preliminary form a demonstration of correlation between an information-based definition of visual saliency and fixation behavior in human observers (Bruce, 2005; Bruce & Tsotsos, 2006). In this paper, a more developed presentation of these ideas is put forth along with additional supporting evidence. In the later part of this section, this proposal is contrasted against other existing proposals.

The following provides an overview of AIM and each of its constituent components. For additional details pertaining to the specifics of the implementation and a more mathematical treatment, the reader may also refer to Appendix A. The premise of our proposal is that the saliency of visual content may be equated to a measure of the information present locally within a scene as defined by its surround, or more specifically, how unexpected the content in a local patch is based on its surround. This quantity corresponds to the surprisal, or the expected number of guesses required in the general version of the

game described by Attneave. The machinery involved in this computation is depicted in Figure 3. Boxed regions (surrounded by a dotted line) contain the inputs to and outputs of the various operations involved in the computation, and these computational operations are depicted by the rounded rectangles. A description of each of these components follows.

Independent feature extraction

For each coordinate location i, j in the scene, the response of various learned filters with properties reminiscent of V1 cortical cells are computed. This stage may be thought of as measuring the response of various cortical cells coding for content at each individual spatial location and corresponds roughly to Gabor-like cells that respond to oriented structure within a specific spatial frequency band and color opponent cells. This yields a set of coefficients for each local neighborhood of the scene $C_{i,j}$ that may be assumed mutually independent. Operations involved in this stage are depicted in Figure 3b and a description of these operations as well as discussion of the assumption of mutual independence follows the overview description of the model. More specific details may also be found in Appendix A.

Density estimation

The content of a local neighborhood $C_{i,j,k}$ (i, j corresponding to position of the local neighborhood) of the image is characterized by several coefficients a_k corresponding to the various basis filters that code for that location. Let us consider one of these coefficients that, choosing an arbitrary example, might correspond to the presence of edge content corresponding to a specific orientation and spatial frequency at that location. In a larger region $S_{i,j,k}$ surrounding the location in question, one also has for each spatial location in the surround, a single coefficient corresponding to this same filter type. Considering all spatial locations in the surround the coefficients corresponding to the filter in question form a distribution (based on a non-parametric or histogram density estimate) that may be used to predict the likelihood of the response of the coefficient in question for $C_{i,j,k}$. For computational parsimony, the definition of surround in the simulations shown is such that each pixel in the image contributes equally to the density estimate and is performed based on a 1000 bin histogram density estimate with the number of bins chosen to be in a range where the likelihood estimate is insensitive to a change in the number of bins. That said, the proposal is amenable to computation based on a local surround and results concerning the quantitative evaluation are included based on such a definition. It is worth noting that in the presence of the sort of parallel hardware with which the brain is equipped, the computation of a likelihood estimate based on the local surround is highly efficient. For more discussion related to this issue, the readers may refer to the section on related literature and Appendix A.

Joint likelihood and self-information

A density estimate for any single coefficient based on coefficients corresponding to the same filter type from the surround affords a likelihood estimate corresponding to the single filter type in question. Based on the independence assumption emergent from a sparse representation, an overall likelihood for all coefficients corresponding to a single location is given by the product of the likelihoods associated with each individual filter type. That is, the likelihood of responses corresponding to the entire cortical column corresponding to a specific spatial location is given by the product of the likelihoods associated with the individual filters. The Shannon Self-Information of this overall likelihood $p(x)$ is then given by $-\log(p(x))$. Note that this is equivalent to the sum of the self-information of the individual cell responses. The resulting information map depicts the saliency attributed to each spatial location based on the Shannon information associated with the joint likelihood of all filters in the cortical column. An additional point of interest is that the depiction in what appears as a saliency map can then be thought of as the average Shannon self-information of cells across a cortical column corresponding to content appearing at each spatial location. It should be noted, however, that the saliency-related computation takes place at the level of a single cell, which is an important consideration in addressing different architectures concerning how attentional selection is achieved; this is an issue that is considered in the Discussion section.

The independent feature extraction stage involves some specific details pertaining to the computational motivation for the proposal of AIM as well as its relationship to properties of cortical cells. Operations involved in independent feature extraction are depicted in Figure 3b and are given as follows.

ICA

A large number of local patches were randomly sampled from a set of 3600 natural images. Images were drawn from the Corel stock photo database and consist of a variety of photographs of outdoor natural scenes captured in a variety of countries. In total, 360,000 patches of 31 × 31 × 3 (and 21 × 21 × 3 for a few results noted in the text) width × height × RGB pixels form the training set for ICA based on the random selection of 100 patches from each image. Independent Component Analysis (Cardoso, 1999; Lee, Girolami, & Sejnowski, 1999) is applied to the data in order to learn a

sparse spatiochromatic basis. The impetus for this stage of the computation lies in the computational complexity associated with density estimation. Taken at face value, characterizing the likelihood of a local region of the visual field on the basis of its surround seems a difficult problem computationally. To consider the metaphor of an image, even for a region as small as a 5 × 5 pixel region of RGB values, one is faced with a probability density estimate on a 75-dimensional space. The amount of data required for a likelihood estimate in such a space appears infeasible in its raw form. A solution to this dilemma comes in the coding observed in the visual cortex wherein as suggested by Attneave and others (Atick, 1992; Attneave, 1954; Barlow, 1961) an efficient representation of natural image statistics is established. There is considerable evidence in favor of sparse coding strategies in the cortex, with sparse coding an important and ubiquitous property of cortical coding (Foldiak & Young, 1995) and having the implication that the response of cells coding for content at a particular spatial location are relatively independent. This is a consideration that is critical as it implies that computation of the self-information of a local patch may be reduced from a high-dimensional patch-wise estimate of self-information to a low-dimensional feature-wise estimate (Bruce, 2004). This reduces the problem of a density estimate in 75 dimensions as given in the example, to 75 one-dimensional density estimates. An independent representation is therefore central to realizing the required density estimate and additionally has implications from the perspective of observed behavior, which are discussed in more detail later in this paper. It is also encouraging from the perspective of biological plausibility that a representation of this form exhibits considerable correspondence with cortical computation and the properties of cortical cells: A number of influential studies have demonstrated that learning a sparse encoding of natural image statistics may yield a sparse V1-like set of cells (Bell & Sejnowski, 1997; Olshausen & Field, 1996), including those that code for spatiotemporal content (van Hateren & Ruderman, 1998) and color opponency (Wachtler, Lee, & Sejnowski, 2001). The output of this operation is a set of basis functions with properties akin to those appearing in early visual areas including responding to oriented structure at various spatial frequencies and red–green/blue–yellow color opponency. This provides a transformation between raw pixel elements and cortical cell responses for which the responses of cells may be assumed independent and in a representation having a close correspondence with early visual cortex.

Matrix pseudoinverse and matrix multiplication

ICA assumes that a local image patch is comprised of a linear combination of the basis filters. The pseudoinverse of the mixing matrix provides the unmixing matrix, which may be used to separate the pixel content within any local region into independent components. More specifically, for each local neighborhood of the image C_k, multiplication of the local pixel matrix with the unmixing matrix produces a set of coefficients that corresponds to the relative contribution of the various basis functions in representing the local neighborhood. These coefficients may be thought of as the responses of V1-like cells across a cortical column, corresponding to the location in question.

There have been other relevant proposals centered around the role of information or likelihood estimation in determining the deployment of fixations. Najemnik and Geisler (2005) consider fixation behavior predicted by a Bayesian ideal observer with the focus on predicting sequences of fixations. They demonstrate that human observers appear to compute an accurate posterior probability map in the search for a target within 1/*f* noise and that inhibition of return proceeds according to a very coarse representation of past fixations. An important element of this work lies in showing that target search appears to operate according to maximizing the information about the location of the target in its choice of fixations. Another effort that leans more toward a stimulus-driven approach in the sense that there is no specific target is that of Renninger, Verghese, and Coughlan (2007). The task involved determining whether the silhouette of a particular shape matched with a subsequently presented silhouette. Eye movements were tracked during the presentation to observe the strategy underlying the selection of fixations. Renninger et al. demonstrate that the selection of fixation points proceeds according to a strategy of minimizing local uncertainty, which equates to a strategy of maximizing information assuming information equates to local entropy. This will typically correspond to regions of the shape silhouette, which contain several edgelets of various orientations. In agreement with the work of Najemnik and Geisler, it was found that there is little benefit to the optimal integration of information across successive fixations. Mechanisms for gain control at the level of a single neuron have been observed, which have been shown to correspond to a strategy based on information maximization (Brenner, Bialek, & de Ruyter van Steveninck, 2000). Although the proposal put forth in this paper is distinct from a description that involves sequences of fixations, the search for a specific target, or specific task conditions, it is nevertheless encouraging that there do appear to be mechanisms at play in visual search that serve to maximize some measure of information in sampling, and it is also the case that the findings of these studies may be viewed as complementary to our proposal rather than conflicting.

One critique that might be levied against an entropy-based definition of the sort described in Renninger et al. (2007) is that a definition based on minimizing local

uncertainty or entropy is inherently local. The difference between the measure of information proposed in AIM and one based on local entropy is that a definition based on local entropy amounts roughly to a measure of local activity. In the context of AIM, entropy is defined on the surround of the local region under consideration with information equated to the self-information of local content in the context of the surround. As such, this information-based definition has an inherent contrast component to it, rather than relying on local activity only.

One can imagine a stimulus such as a region containing lines of homogeneous orientation in a sea of randomly oriented lines, or a shape silhouette of the sort employed in the study of Renninger et al. (2007) that has significant edge content on the entire boundary with the exception of a single smooth region. In such cases, one may expect fixations to be directed to the relatively more homogeneous regions of the scene or figure. Therefore, a local definition of what is salient that fails to include the context in which a localized stimulus is presented will fail to adequately characterize some aspects of behavior, since local activity seems less of a determining factor in what may draw an observers' gaze in these cases. It is also worth stating that such a measure based on self-information is likely to exhibit some degree of correlation with an entropy-based definition for certain stimulus choices since the nature of natural stimuli is biased in favor of having relatively few cells active at a time by construction (Foldiak & Young, 1995).

Fixational eye movement data

A means of considering model plausibility is in considering correlation between the proposed definition of visual saliency and eye movements made by human observers in a task designed to minimize the role of top-down factors. We therefore have considered the extent to which the behavior of AIM agrees with two sets of eye tracking data, one that includes a variety of colored natural images, and the other that comprises a wide variety of different videos that include natural spatiotemporal content, television advertisements, and video games, which was the basis for evaluation for the Surprise model of Itti and Baldi (2006). The resulting model performance is compared against the saliency models of Itti, Koch, and Niebur (1998) and Itti and Baldi (2006) where appropriate.

Single image eye tracking evaluation

Methods

The data that form the basis for performance evaluation are derived from eye tracking experiments performed while participants observed 120 color images. Images were presented in random order for 4 seconds each with a gray mask between each pair of images appearing for 2 seconds. Participants were positioned 0.75 m from a 21-inch CRT monitor and given no particular instructions except to observe the images. Images consist of a variety of indoor and outdoor scenes, some with very salient items, others with no particular regions of interest. The eye tracking apparatus consisted of an ERICA workstation including a Hitachi CCD camera with an IR emitting LED at the center of the camera lens. The infrared light was reflected off two mirrors into the eye facilitating segmentation of the pupil. Proprietary software from ABB corporate research was used to analyze the data. The parameters of the setup are intended to quantify salience in a general sense based on stimuli that one might expect to encounter in a typical urban environment. Data were collected from 20 different participants for the full set of 120 images.

The issue of comparing between the output of a particular algorithm and the eye tracking data is non-trivial. Previous efforts have selected a number of fixation points based on the saliency map and compared these with the experimental fixation points derived from a small number of subjects and images (7 subjects and 15 images in a recent effort (Privitera & Stark, 2000)). There are a variety of methodological issues associated with such a representation, the most important such consideration being that the representation of perceptual importance is typically based on a saliency map. Observing the output of an algorithm that selects fixation points based on the underlying saliency map obscures observation of the degree to which the saliency maps predict important and unimportant content and, in particular, ignores confidence away from highly salient regions. Secondly, it is not clear how many fixation points should be selected. Choosing this value based on the experimental data will bias output based on information pertaining to the content of the image and may produce artificially good results in that it constrains assessment of fixations to a number of locations that may be correlated with the number of salient regions in the image, reducing the importance of model predictions away from these regions.

The preceding discussion is intended to motivate the fact that selecting discrete fixation coordinates based on the saliency map for comparison may not present the most appropriate representation to use for performance evaluation. In this effort, we consider two different measures of performance. Qualitative comparison is based on the representation proposed in Koesling, Carbone, and Ritter (2002). In this representation, a fixation density map is produced for each image based on all fixation points and subjects. This is given by a 2D Gaussian kernel density estimate wherein the standard deviation σ is chosen to approximate the drop-off in visual acuity moving peripherally from the center of the fovea based on the viewing distance of participants in the experiment. The density map then comprises a measure of the extent to which each

pixel of the image is sampled on average by a human observer based on observed fixations. This affords a representation for which similarity to a saliency map may be considered at a glance. Quantitative performance evaluation is achieved according to the procedure of Tatler et al. (2005). The saliency maps produced by each algorithm are treated as binary classifiers for fixation versus non-fixation points. The choice of several different thresholds for the saliency maps treated as binary classifiers in predicting fixated versus not fixated pixel locations allows an ROC curve to be produced for each algorithm. An overall quantitative performance score is then given by the area under the ROC curve. For a further explanation of this method, refer to Tatler et al. (2005).

Results

Figure 4 demonstrates a qualitative comparison of the output of AIM with fixation density maps, as compared with the output of the saliency maps produced by the Itti et al. algorithm. The frames from left to right are given as follows: The image under consideration, the output of AIM as applied to the image appearing in the leftmost frame. The output of the Itti et al. algorithm as described in Itti et al. (1998). The human fixation density map depicting the extent to which each pixel location is sampled on average by a human observer accounting for foveation. The original image modulated by the output of AIM showing the localization of saliency-related activation. As can be observed there is considerable qualitative similarity between AIM output and the human density maps. Figure 5 demonstrates ROC curves associated with AIM and the Itti et al. algorithm when treated as classifiers for fixation points, along with 99% confidence intervals. The area under the curves is 0.781 ± 0.0087 and 0.729 ± 0.0082 for AIM and the Itti et al. algorithms, respectively. These results are determined using identical computation to that appearing in Bruce and Tsotsos (2006) with each pixel location contributing equally to the density estimate for computational parsimony. We have also performed analysis based on a local surround with the drop-off corresponding to the approximate drop-off observed in surround suppression moving from the center of a target cell in V1 as shown in Figure 7 of Petrov and McKee (2006) as fit to a 2D Gaussian kernel. Due to the additional computation required for such a simulation, analysis was performed for a 21×21 window size, which yields an ROC score of 0.762 ± 0.0085 based on the local surround. As a basis for comparison, the ROC score for a 21×21 window size in which each pixel location contributes equally to the estimate yields an ROC score of 0.768 ± 0.0086. This suggests that one might also expect an ROC score in the vicinity of 0.78 for a simulation based on a local surround definition based on a 31×31 window size. Note also that a more careful analysis of the extent and shape of the local surround may yield further gains with respect to ROC score as we have considered only a single sensible choice for the local simulation. This, however, does not impact upon the conclusions of this paper, or the demonstration that self-information based on a local surround yields performance that exceeds that of previous efforts (Itti et al., 1998). As a whole the results are encouraging in the validation of saliency as defined within AIM as a quantity that correlates with the direction of human eye movement patterns. For further implementation details pertaining to these results, the reader is urged to consult Appendix A.

Video-based eye tracking evaluation

Methods

The data employed for evaluation of fixations in video examples are identical to that employed in Itti and Baldi (2006). The following gives an overview of the details concerning the data set. For a detailed description of the methods involved please refer to the aforementioned work. Eye tracking data were collected from eight subjects aged 23–32 with normal or corrected-to-normal vision. A total of 50 video clips consisting of a variety of categories of scenes including indoor scenes, outdoor scenes, television clips, and video games were employed for the experiment. Video clips were displayed at a resolution of 640×480 and consist of over 25 minutes of playtime at approximately 60 Hz. Total analysis is based on 12,211 saccades.

The generality of AIM means that it may be applied to any arbitrary set of neurons provided they form a sparse representation. A natural extension to considering neurons that code for combinations of angular and radial frequency and color opponency is that of considering spatiotemporal patterns. To carry out this evaluation, we learned a spatiotemporal basis using the Lee et al. (1999) extended Infomax algorithm. The data set employed was the van Hateren data employed to learn the basis presented in van Hateren and Ruderman (1998) and consists of grayscale image sequences at 50 frames per second of natural data. The basis was learned by randomly sampling spatiotemporal volumes of $11 \times 11 \times 6$ frames from throughout the

Figure 4. A qualitative comparison of the output of AIM with the experimental eye tracking data for a variety of images. Also depicted is the output of the Itti et al. algorithm for comparison: From left to right: Original unprocessed image. Output of AIM, *hotter* areas correspond to more salient regions. Saliency as computed by the Itti et al. (1998) algorithm. Eye tracking density maps from experimental data averaged across 20 subjects depicts the extent to which each pixel location was sampled on average by human observers. The original image modulated by the output of AIM, demonstrating the localization of patterns and affording a sense of cortical modulation associated with various stimulus patterns.

video taking every second frame so that the basis corresponds to data sampled at 25 frames per second. The result of the ICA algorithm is a set of cells selective for oriented content at different spatial frequencies and for different velocities of motion not unlike those appearing in V1 as reported in van Hateren and Ruderman (1998). For the sake of comparison, we considered the same set of videos used in Itti and Baldi (2006) and employed the same strategy for performance evaluation with respect to predicting fixations. In short, this process involves examining the saliency of randomly selected locations relative to saliency values at fixations. The KL-divergence between these two distributions is then used to rank algorithms. A detailed description of this procedure may be found in Itti and Baldi (2006).

Results

A sample of frames from a variety of qualitatively different videos is shown in Figure 6 (left of each pair) along with the associated saliency (right of each pair). It is interesting to note that the output agrees qualitatively with our intuition of what is salient across a wide range of types of spatiotemporal data, including situations with low contrast structure, crowding, and with an inherent tradeoff between stationary and moving structure. The resulting distributions from consideration of the eye tracking data described, subject to AIM, are shown in Figure 7. Results are compared with those that appear in Itti and Baldi (2006) and include three static metrics of salience corresponding to local intensity variance, orientation contrast, and Entropy as well as a characterization of motion energy (for details, see Itti & Baldi, 2006) and, in addition, the output of the saliency model of Itti et al.

(1998) and the Surprise model of Itti and Baldi (2006). The KL-divergence associated with this evaluation is 0.328. This is a 36% improvement over the Surprise model of Itti and Baldi with a KL score of 0.241 and a 60% improvement over the saliency model of Itti et al. (1998) with a KL score of 0.205. Importantly, this apparently strong performance comes from the same biologically plausible setup that yielded favorable performance for spatiochromatic data, without modification or any additional assumptions required for consideration of spatiotemporal neurons. This evaluation supports the claim of generality of information as a strategy in saliency computation and additionally offers a means of characterizing spatiotemporal saliency. Additionally, no prior model of scene content or memory is involved as in Itti and Baldi (2006), but rather the prediction is based on the current state of neurons that code for spatiotemporal content. Overall, the results provide further support of the generality of AIM in predicting fixation and visual-search-related behaviors and demonstrates the efficacy of the proposal in predicting fixation patterns on a qualitatively different data set than that comprised of still images.

Visual search behavior

The study of visual search has been influential in shaping the current understanding of computation related to attention and the determination of visual saliency. Owing to the large body of psychophysics work within this area, in addition to some of the peculiarities that are observed within the visual search paradigm, it is natural to consider how model predictions measure up against the wealth of psychophysics results in this area. It is with this in mind that we revisit a variety of classic results derived from the psychophysics literature revealing that AIM exhibits considerable explanatory power and offers some new insight on certain problem domains.

It is important to state that the visual search literature largely focuses on situations in which there is a specific target definition (e.g., find the red horizontal bar). Within the context of this paper we consider only the bottom-up determination of saliency and there is no specific notion of task or target. A more general account of visual search behavior requires treatment of at least these two

considerations in combination. As the focus of this paper is on a detailed proposal for computation related to bottom-up salience, we only aim to consider visual search insofar as saliency contributes to the efficiency with which search is performed. While more general conclusions might be drawn in including the impact of task specific bias on the response of cells involved based on, for example, a simple multiplicative gain applied to cells that accounts for their relevance to a specific task, this is a complex mechanism in itself and is outside of the scope of the study at hand. That said, the bottom-up salience of items within a display is nevertheless an important contribution to any visual search task and thus the results considered are relevant in the context of the more general body of visual search results in which there is a specific definition associated with the visual search task to be performed. It is also worth noting that the conclusions put forth in this paper when combined with accounts of visual search for which the focus is on the role of bias in the guidance toward target items (e.g., Wolfe, Cave, & Franzel, 1989) have as a result a treatment of visual search that is applicable to the more general body of visual search results for which the impact of task is a central component. That said, it is important to bear in mind that the results presented focus on but one element of visual search, and caution should be exercised in considering how conclusions drawn from the examples that follow relate to visual search performance in general.

Generally, models of attention assume that the focus of attention is directed according to a competitive Winner-take-all process acting on some neural representation in the cortex. An important element of this representation is the saliency of a target item relative to the saliency of the distractors since this is the determinant of search efficiency according to various selection mechanisms (Desimone & Duncan, 1995; Itti et al., 1998; Koch & Ullman, 1985; Tsotsos et al., 1995). It is assumed then throughout the discussion that search efficiency is a function of the ratio of target to distractor saliency in line with other similar efforts (Li, 2002). This assumption allows the consideration of saliency to be disentangled from the mechanisms that underlie attentional gating, which remains a contentious issue.

Serial versus parallel search

An observation that has been influential in earlier models of attention is that certain stimuli seem to be found effortlessly from within a display, while others require considerable effort to be spotted seemingly requiring elements of the display to be visited in turn. Consider for example Figure 8. In the top left, the singleton item distinguished by its orientation is found with little effort seemingly drawing attention automatically. This phenomenon is sometimes referred to as "pop-out." The same may be said of the singleton defined by color in the top middle frame; however, the singleton in the top right frame requires examining the elements of the frame in turn to locate the target. These observations form

the motivation for Treisman's Feature Integration Theory (Treisman & Gelade, 1980), a seminal work in attention modeling based on the observation that some targets are found effortlessly and seemingly in parallel while others seem to require a serial search of target items with the search time increasing as a linear function of the number of distracting elements. In particular, the distinction between these two cases is when a target item is defined by a conjunction of features rather than a single feature. On the bottom row of Figure 8 is the output of the AIM algorithm with the saliency scale shown on the left-hand side. *Warmer* colors are more salient, and this scale is used in all examples scaled between the maximum and minimum saliency values across all examples within an experiment. As can be seen in Figure 8 the target relative to distractor saliency is very high for the first two cases, but the target saliency is indistinguishable from that of the distractors in the third case, suggesting no guidance toward the target item and hence requiring a visit of items in serial order. Thus, the distinction between a serial and parallel search is an emergent property of assuming a sparse representation and saliency based on information maximization. Since the learned feature dimensions are mutually independent, the likelihood is computed independently for uncorrelated feature domains implying unlikely stimuli for singletons based on a single feature dimension but equal likelihood in the case of a target defined by a conjunction. This behavior seen through the eyes of AIM is then a property of a system that seeks to model redundancy in natural visual content and overcome the computational complexity of probability density estimation in doing so. An additional example of a conjunction search is featured in Figure 9: The small, rotated, and red 5's are easily spotted, but finding the 2 requires further effort. It is worth noting that this account of visual search has been revised to some extent with more recent experiments demonstrating an entire continuum of search slopes ranging from very inefficient to very efficient (Wolfe, 1998). This is a consideration that is also supported by AIM as more complex stimuli that give rise to a distributed representation may yield very different ratios of target versus distractor saliency.

Target–distractor similarity

An additional area of psychophysics work that has been very influential is that of observing the effects of target–distractor similarity on difficulty in search tasks. Generally, as a target item becomes more similar in its properties to the distracting items, the search becomes more difficult (Duncan & Humphreys, 1989; Pashler, 1987). An example of this modeled on the experiment of Duncan and Humphreys (1989) and based on the example shown in Wolfe and Horowitz (2004) is shown in Figure 10 (top). Moving from the top left to top right frame, a shift of the target away from the distractors in color space occurs. The resulting saliency appears below each example and the ratio of distractor to target saliency is 0.767, 0.637, 0.432, and 0.425, respectively. This ratio is given by the saliency score at the center of the singleton and by the mean of the saliency score at the center of distractors, respectively. There is one important element appearing in this example that perfectly matches the data of Duncan and Humphreys: In the two rightmost stimulus examples, the distractor to target saliency ratio remains the same. This implies that beyond a certain distance for a particular feature dimension, a further shift along this feature dimension makes no difference in search efficiency. This is exactly the effect reported in Duncan and Humphreys (1989). In AIM, the effect emerges due to a single neuron type responding to both target and distractor items. Once the target–distractor distance increases to the extent that there are no cells that respond strongly to both the target and distractor items, a further shift in feature space has no effect on task difficulty. Hence the specific

effect observed in Duncan and Humphreys (1989) also appears as an emergent property of modeling redundancy and with saliency equated to information. Interestingly, the resulting data are almost identical to the experimental results despite the simplifying assumptions in learning the V1-like neural representation.

Distractor heterogeneity

A question that follows naturally from consideration of the role of target–distractor similarity is that of whether distractor–distractor similarity has any effect on search performance. The most telling effect in this domain is that increasing the heterogeneity of the distractors yields a more difficult search (Duncan & Humphreys, 1989; Nagy & Thomas, 2003; Rosenholtz, Nagy, & Bell, 2004). Consider for example Figure 11. In the top left case, the item 15 degrees from horizontal appears to pop-out. This effect is diminished in the top middle frame and severely diminished in the top right frame. The saliency attributed to each of these cases appears below each stimulus example. The finding that an increase of distractor heterogeneity results in a more difficult search is consistent with the behavior of AIM. Distributing the distractors over several different cell types rather than a single type of neuron means that the distractors are considered less probable and hence more informative thus decreasing the ratio of target to distractor saliency. There is also a secondary effect in the example given of target–distractor similarity since broad tuning means that cells tuned to a particular orientation may respond weakly to a distractor type other than that for which they are tuned, or the target. This serves to highlight the importance of the specifics of a neural code in the determination of visual saliency and also offers insight on why the determination of efficiency in visual search tasks may be difficult to predict. It is worth noting that this basic effect captures behaviors that models based on signal detection theory (Verghese, 2001) fail to. For example, a horizontally oriented bar among distractors at 30 degrees is much more salient than a horizontal bar among distractors 1/3 oriented at 30 degrees, 1/3 at 50 degrees, and 1/3 at 70 degrees as observed in Rosenholtz (2001a). This is an important peculiarity of visual search that is inherent in an information seeking model but absent from many competing models of saliency computation.

Search asymmetries

A stimulus domain that has generated a great deal of interest involves so-called search asymmetries, due to their potential to reveal peculiarities in behavior that may further our understanding of visual search. One asymmetry that has received considerable attention is an asymmetry attributed to presence vs. absence of a feature as in Figure 12 (Treisman & Gormican, 1988). In this example, a search for a dash among plus signs is much more difficult than the converse. In examining the associated saliency maps as computed by AIM, it is evident that this behavior is also inherent in the information-based

definition. Note that this is simply a specific case of a more general phenomenon and the same might be observed of a Q among O's or any instance where a singleton is defined by a feature missing as opposed to its presence. This phenomenon can be explained by the fact that in the feature present case, the feature that distinguishes the target is judged to be improbable and hence informative. In the case of the feature absent, there is nothing about the location that distinguishes it from background content in the context of the missing feature since the background regions also elicit a zero response to the "missing" feature. Rosenholtz (2001b) reveals an additional class of asymmetries, which she points out are examples of poor experimental design as opposed to true asymmetries. An example of such a stimulus appears in Figure 13 (top). Rosenholtz points out that the asymmetry appearing in Figure 13, which corresponds to the task of finding a red dot among pink being easier than the converse (top left and top second from left), may be attributed to the role of the background content (Rosenholtz et al., 2004); a change in background color (top right and top second from right) causes a reversal in this asymmetry. From the resultant saliency maps, it is evident that AIM output also agrees with this consideration (Figure 13, bottom). Reducing the contrast between the background and the target/distractors would also be expected to give rise to a more pronounced asymmetry as the response of a cell to target/distractors and background become less separable. This is indeed the behavior reported in Rosenholtz et al. (2004).

An important point to note is the fact that viewed in the context of AIM, the color background asymmetry arises from the same cause as the feature presence–absence asymmetry, both a result of the role of the background in determining feature likelihood. In each case, it is the role of the background content in determining the likelihood associated with any particular firing rate. In the colored background examples, the background causes greater suppression of the target or distractors depending on its color. One example Rosenholtz (1999) describes as an asymmetry in experimental design is that of a moving target among stationary distractors versus a stationary target among moving distractors, suggesting that the design be rectified by ensuring the motion of the distractors is coherent. Under these conditions, the stationary search becomes more efficient but still remains significantly less efficient than the moving target case. This is an aspect of search performance that is captured by the behavior of AIM: If there exists units that elicit a response to non-target background locations and also to the stationary target, this may have an effect of suppressing target saliency that will be absent in the moving target case. Encouraging is the fact that this effect emerges due to the role of background content in the determination of saliency consistent with the model of Rosenholtz (2001b).

Related proposals

There are a few related proposals that include similar definitions of the saliency of visual content. We have deferred discussion of these models to this point in the text so that specific reference to some of the results appearing in this paper may be made. A central element of the proposal put forth in this paper, as mentioned briefly, is that under the assumption of a sparse representation, the

likelihood of local content as characterized by a sparse ensemble of cells may be reduced to a computationally feasible problem of many likelihood estimates of one dimension. This was a point that was the focus of Bruce (2004), which presented this point along with the suggestion that this computation might be performed for any arbitrary definition of context. In Bruce (2004) qualitative results of this measure were presented for a definition of context based on the entirety of a single natural image under consideration, or for a definition based on *ecological statistics* in which a large set of natural images forms the likelihood estimate. Zhang, Tong, Marks, Shan, and Cottrell (2008) have presented analysis of the relationship between this latter definition and locations fixated by human observers. The results they present show comparable performance to results for which the estimation is based on the image in question or based on a local surround region. However, such a definition precludes the possibility of context specific determination of saliency and thus will not produce any of the behaviors associated with the various psychophysics paradigms we have considered. There are a few behaviors that Zhang et al. describe, which they suggest a context specific model of saliency fails to capture, such as the asymmetric performance observed in a visual search for a bar oriented 5 degrees from vertical among many vertical bars versus a bar oriented vertically among many bars oriented 5 degrees from vertical, with the suggestion that a likelihood based on natural image statistics is necessary to account for this effect. There is however a significant oversight associated with this statement. An encoding based on ICA is optimal with respect to the statistics of the natural environment and therefore there is some representation of natural image statistics in general inherent in the context specific model in the definition of the receptive fields themselves. Therefore one also observes this specific asymmetry in the context of our model as the units that respond most strongly to content oriented 5 degrees from vertical also respond to a vertically oriented edge, but the converse is not the case (or the response is weaker) as the nature of the coding dictates a smaller orientation bandwidth for vertical edges. Combined with a suppressive surround, this results in the observed performance asymmetry. The same may be said of novelty of stimuli (e.g., Shen & Reingold, 2001; Wang, Cavanagh, & Green, 1994) assuming familiarity with a specific character set may have as a consequence a more efficient neural representation. It is also interesting to note that as receptive field properties (and image statistics) vary with position in the visual field, that behavior in tasks for which performance is anisotropic with respect to the location in the visual field might also be explained by AIM. This however is an issue that is difficult from an implementation perspective, requiring different cell types for different locations in the visual field and an explicit model of dependencies between different cell types. There are a few additional points of interest that appear in the work of Zhang et al., which are discussed at the end of this section. A definition that is closer to the former definition appearing in Bruce (2004) in which the likelihood estimate is based on the content of the entirety of a single image under consideration appears in Torralba, Oliva, Castelhano, and Henderson (2006). In Torralba et al. (2006), the focus is on object recognition and how context may guide fixations in the search for a specific object. They propose the following definition: $P(O = 1, X|L, G)$, where $O = 1$ indicates that the object O in question is present, X is the location within the scene, and L and G are the local and global features, respectively. Via Bayes rule and excluding certain terms that appear in reformulating this definition, one arrives at an expression for saliency $S(x) = \frac{1}{P(L|G)}\, p(X|O = 1, G)$. While the focus of Torralba et al. (2006) is on how context informs the saliency within the context of an object recognition task given by the location likelihood conditioned on the global statistics for instances in which the object appears, the formulation also results in a term that is the inverse function of the likelihood of some set of local features conditioned on the global features. In the model of Torralba et al. (2006), they propose that image structure is captured on the basis of global image features. These global features consist of a coarse spatial quantization of the image and the features themselves pool content across many feature channels for each spatial location. Given this formulation, evaluation of $P(L|G)$ directly is infeasible. For this reason, an estimate of $P(L|G)$ is computed on the basis of the joint likelihood of a vector of local features based on a model of the distribution of said features over the entire scene. The likelihood $P(L|G)$ is fit to a multivariate power exponential distribution with assumptions on the form of the distribution allowing an estimate of the joint likelihood of a local feature vector. Aside from the most obvious difference between the proposal put forth in this paper and that appearing in Torralba et al. (2006); that being computation of saliency based on local context as mediated by surround suppression, versus global receptive fields), there are a few comments that may be made in regards to the relationship to the proposal put forth in this paper. A significant point that may be made is that in considering the joint likelihood of a set of features, one once again fails to predict a variety of the behaviors observed in the psychophysics examples. For example the independence assumption is central to some of the psychophysics behaviors discussed, such as the distinction between a pop-out and conjunction search, or the feature presence/absence asymmetry. Secondly, it is unclear how the computational machinery proposed to achieve this estimate will scale with the number of features considered and it is likely that a local surround contains an insufficient number of samples for the required covariance matrix estimate. Therefore, it is once again the case that this proposal does not correspond to behavior observed psychophysically and also seems to prohibit computation of a measure of information in which the context is local. It should be noted that this

quantity is not the main focus of the proposal of Torralba et al. (2006) but does serve as a useful point of contrast with the proposal at hand and highlights the importance of sparsity for likelihood estimation involved in a case where the data contributing to such an estimate is limited. It is interesting to note that the circuitry required to implement AIM is consistent with the behavior of local surround suppression with the implication that surround suppression may subserve saliency computation in line with recent suggestions (Petrov & McKee, 2006). There are in fact several considerations pertaining to the form of a local surround-based density estimate that mirror the findings of Petrov and McKee. Specifically, suppression in the surround comes from features matching the effective stimulus for the cell under consideration, is spatially isotropic, is a function of relative contrast, is prominent in the periphery and absent in the fovea, and the spatial extent of surround suppression does not scale with spatial frequency. It is also interesting to note that suppression of this type is observed for virtually all types of features (Shen, Xu, & Li, 2007). It is important to note that AIM is the sole proposal that is consistent with the entire range of psychophysical results considered and has a strong neural correlate in its relationship to behavior observed in the recent surround suppression literature.

An additional consideration that is also noted in the study of Zhang et al. (2008) and also in Le Meur, Le Callet, Barba, and Thoreau (2006) is that the nature of eye tracking data sets is such that there is a considerable central bias. This effect is sufficiently strong that a central Gaussian appears to better predict the locus of fixation points than any model based on the features themselves. This is almost certainly as Zhang et al. suggest, due to the bias that emerges from images consisting of composed photographs in which the photographer centers items of interest and possibly the fact that images are presented on a framed display. In light of this, a model for which the basic machinery results in higher scores away from the borders of the image will have a score that is artificially inflated. This is an important consideration with respect to the quantitative evaluation presented. As the border effects are especially strong in the implementation of Itti et al. (1998), relative to those presented in this study (as shown by Zhang et al.), one might expect an even larger difference in the performance metric depicted in Figure 5 if edge effects were accounted for. This consideration however does not impact on the conclusions of this study and for a detailed treatment of this issue, readers may wish to consult Zhang et al. (2008).

Discussion

It is interesting to consider how the content discussed in the previous sections fits in with the "big picture" as far as attention modeling is concerned. There are a variety of different schools of thought on the computational structure underlying attentional selection in primates ranging from those that posit the existence of a "saliency map" (Itti et al., 1998; Koch & Ullman, 1985; Li, 2002), in the cortex to those that claim a distributed representation over which winner-take-all behavior or competitive interaction facilitates attentional selection (Desimone & Duncan, 1995; Tsotsos et al., 1995). Thus far we have depicted saliency in a manner more consistent with the former of these categories demonstrating the total information at any spatial location as the sum of information attributed to all cells that code for content at that location. The correspondence between the proposal and models based on a saliency map can then be thought of as considering the average salience of cells across a cortical column corresponding to a particular location. What is perhaps more interesting is the relationship between the proposal and distributed models of attention. It is evident that as the observation likelihood is computed at the level of a single cell, it is possible that this signal is used to control its gain at the single cell level in accord with neurophysiological observations. It is evident that the proposal put forth is amenable to a saliency map style representation, but it is our opinion that recent results are more consistent with a distributed selection strategy in which gating is achieved via localized hierarchical winner-take-all competition and saliency-related computation achieved via local modulation based on information.

In this vein, the following discussion considers evidence in favor of a distributed representation for attentional selection as put forth in Tsotsos et al. (1995) and the relationship of such a representation to the proposal put forth by AIM.

Visual processing appears to constitute a dichotomy of rapid general perception on one hand versus slower detailed processing on the other as evidenced by such paradigms as change blindness (Rensink, O'Regan, & Clark, 1997). Many studies demonstrate that certain quantities are readily available from a scene at a glance such as Evans and Treisman (2005) and Huang, Pashler, and Treisman (2007) while other judgments require considerably more effort. This is evidently a product of a visual hierarchy in which receptive fields cover vast portions of the visual field and representations code for more abstract and invariant quantities within higher visual areas. Attentional selection within the model of Tsotsos et al. (1995) proceeds according to this assumption with attentional selection implemented via a hierarchy of winner-take-all processes that gradually recover specific information about an attended stimulus including the specific conjunction of features present and, in particular, the precise location of a target item. In line with this sort of architecture, recent studies have shown that a variety of judgements can be made on a visual stimulus with a time course shorter than that required for localization of a target item (Evans & Treisman, 2005; Huang et al., 2007).

It should be noted that within the traditional saliency map paradigm, there is nothing inherent in the structure of the model that is consistent with this consideration as spatial selection forms the basis for determining the locus of attention. Furthermore, the *forest before trees* priority in visual perception appears to be general to virtually any category of stimulus including the perception of words preceding that of letters (Johnston & McClelland, 1974) and scene categories more readily perceived than objects (Biederman, Rabinowitz, Glass, & Stacy, 1974) in addition to a more general global precedence effect as demonstrated by Navon (1977). As a whole, the behavioral studies that observe early access to general abstract quantities prior to more specific simple properties such as location seem to support an attentional architecture that consists of a hierarchical selection mechanism with higher visual areas orchestrating the overall selection process. Further evidence of this arrives in the form of studies that observe pop-out of high-level features such as depth from shading (Ramachandran, 1988), facial expressions (Ohman, Flykt, & Esteves, 2001), 3D features (Enns & Rensink, 1990), perceptual groups (Bravo & Blake, 1990), surface planes (He & Nakayama, 1992), and parts and wholes (Wolfe, Friedman-Hill, & Bilsky, 1994). As mentioned, the important property that many of these features may share is an efficient cortical representation. Furthermore, pop-out of simple features may be observed for features that occupy regions far greater than the receptive field size of cells in early visual areas. It is unclear then how a pooled representation in the form of a saliency map mediating spatial selection can explain these behaviors unless one assumes that it comprises a pooled representation of activity from virtually every visual area. The only requirement on the neurons involved is sparsity and it may be assumed that such computation may act throughout the visual cortex with localized saliency computation observed at every layer of the visual hierarchy in line with more general models of visual attention (Desimone & Duncan, 1995; Tsotsos et al., 1995). There also exists considerable neurophysiological support in favor of this type of selection architecture. In particular the response of cells among early visual areas appears to be affected by attention at a relatively late time course relative to higher visual areas (Martínez et al., 1999; Nobre et al., 1997; Roelfsema, Lamme, & Spekreijse, 1998) and furthermore the early involvement of higher visual areas in attention-related processing is consistent with accounts of object-based attention (Tipper & Behrmann, 1996; Somers, Dale, Seiffert, & Tootell, 1999).

In a recent influential result, it was shown that focused attention gives rise to an inhibitory region surrounding the focus of attention (Hopf et al., 2006). This result is a prediction of a hierarchical selection architecture (Tsotsos et al., 1995) along with the ability to attend to arbitrarily sized and shaped spatial regions (Müller & Hübner, 2002); these considerations elude explanation within the traditional saliency map paradigm in its current form and are more consistent with a distributed hierarchical selection strategy (Kastner & Pinsk, 2004). It is also important to note that even for *basic* features, an important consideration is the scale at which analysis is performed with regard to conclusions that emerge from the proposal. It is evident that varying the extent of surround suppression might have some effect on the pop-out observed in a case such as that appearing in Figure 9. Under the assumption of a hierarchical representation in which features are represented at each layer with increasing extent of receptive field size and surround, one has a definition that is less sensitive to scale (for example, in embedding AIM within the hierarchical selective attention architecture of Tsotsos et al., 1995). It is also worth noting that in order to explain a result such as that of Enns and Rensink (1990) whereby targets defined by unique 3D structure pop-out, or that of Ramachandran (1988) whereby shape defined by shading results in pop-out, the global definition proposed by Torralba et al. (2006) would require a summary representation of more complex types of features such as 3D structures or shape from shading based on global receptive fields. These considerations raise questions for any definition of saliency in which the determination is based on global scene statistics. The case is even stronger if one considers pop-out effects associated with faces although there remains some contention that demonstrations of pop-out effects associated with faces are a result of confounds associated with simpler features (Hershler & Hochstein, 2005, 2006). As a whole, a hierarchical representation of salience based on a local judgment of information is the only account that appears to be consistent with the entire range of effects described.

The preceding discussion serves to establish the generality of the proposal put forth by AIM. The portion of saliency computation that is of interest is the normalization or local gain control observed as a product of the context of a stimulus. This is an aspect of computation that is only a minor consideration within other models and accounted for based on a crude or general mechanism within a normalization operation with only loose ties to visual circuitry (Itti et al., 1998).

In conclusion, we have put forth a proposal for saliency computation within the visual cortex that is broadly compatible with more general models concerning how attention is achieved. In particular, the proposal serves to provide the *missing link* in observing pop-out behaviors that appear within models that posit a distributed strategy for attentional selection; a subset of attention models for which favorable evidence is mounting. The proposal is shown to agree with a broad range of psychophysical results and allows the additional possibility of simulating apparent high-level pop-out behaviors. Finally, the model demonstrates considerable efficacy in explaining fixation data for two qualitatively different data sets demonstrating the plausibility of a sampling strategy based on information seeking as put forth in this paper.