An Information Theoretic Model of Saliency and Visual Search

Neil D.B. Bruce and John K. Tsotsos

Department of Computer Science and Engineering and
Centre for Vision Research
York University, Toronto, ON, Canada
{neil,tsotsos}@cse.yorku.ca
http://www.cse.yorku.ca/~neil

Abstract. In this paper, a proposal which quantifies visual saliency based on an information theoretic definition is evaluated with respect to visual psychophysics paradigms. Analysis reveals that the proposal explains a broad range of results from classic visual search tasks, including many for which only specialized models have had success. As a whole, the results provide strong behavioral support for a model of visual saliency based on information, supplementing earlier work revealing the efficacy of the approach in predicting primate fixation data.

Keywords: Attention, Visual Search, Saliency, Information Theory, Fixation, Entropy.

1 Introduction

Visual search is an important task in everyday functioning, but a consensus on the precise details of the system underlying visual search in primates has yet to be reached. Consideration of specific stimulus sets in a lab setting has allowed observation of some of the peculiarities of visual search in primates revealing surprising efficiency for some visual search tasks and surprising inefficiency for others. Despite the considerable interest and effort placed on the problem, and the growing body of data on visual search, explanation for various effects exists in many instances within only specialized models. One might view the ultimate aim of modeling in visual search to be a single model with the minimum set of requirements that captures all observed visual search behavior and additionally is based on some basic well defined principle. It is our view that our proposal Attention based on Information Maximization (AIM) satisfies the last of these requirements, and the intention of the remainder of the discussion is to address the extent to which the first of these requirements is satisfied. In the sections that follow, it is established that the model exhibits considerable agreement with a broad range of psychophysical observations lending credibility to the proposal that attentional selection is driven by information.

In [1] we described a first principles definition for visual saliency built on the premise that saliency may be equated to the amount of information carried

L. Paletta and E. Rome (Eds.): WAPCV 2007, LNAI 4840, pp. 171–183, 2007.

by a neuron or neuronal ensemble. It was demonstrated that such an approach reveals surprising efficacy in predicting human fixation patterns and additionally carries certain properties that make the proposal plausible from a biological perspective. An additional and perhaps more favorable test for a model that claims to represent the process underlying the determination of visual saliency in the primate brain, is the extent to which the model agrees with behavioral observations, and in particular, those behaviors that on first inspection may seem counterintuitive. It is with this in mind that we revisit the proposal that visual saliency is driven fundamentally by *information*, with consideration to a variety of classic psychophysics results. In this paper, we extend the results put forth in [1] to consideration of various classic psychophysics paradigms and examine the relation of qualitative behavioral trends to model behavior. It is shown that the model at hand exhibits broad compatibility with a wide range of effects observed in visual search psychophysics.

2 Saliency Based on Information Maximization

The following describes briefly the procedure for computing the information associated with a given neuron response or ensemble of neurons. For a more detailed description, including details pertaining to neural implementation, the reader should refer to [1]. Prior efforts at characterizing the information content of a spatial location in the visual field appeal to measures of the entropy of features locally. Some shortcomings of such a measure are highlighted in [1], but in short, local activity does not always equate to informative content (consider a blank space on an otherwise highly textured wallpaper). In the context of AIM, the information content of a neuron is given by $-log(p(x))$ where x is the firing rate of the neuron in question and $p(x)$ the observation likelihood associated with the firing rate x. The likelihood of the response a neuron elicits is predicted by the response of neurons in its support region. In the work presented here, we have assumed a support region consisting of the entire image for ease of computation, but it is likely that in a biological system the support region will have some locality with the contribution of neighbouring units to the estimate of $p(x)$ proportional to their proximity to the unit exhibiting the firing rate x. This discussion is made more concrete in considering a schematic of the model as shown in figure 1. A likelihood estimate based on a local window of image pixels appears to be an intractable problem requiring estimate of a probability density function on a high-dimensional space (e.g. 75 dimensions for a 5x5 RGB patch). The reason this estimate is possible is that the content of the image is not random but rather is highly structured. The visual system exploits this property by transforming local retinal responses into a space in which correlation between different types of cell responses is minimized [2,3]. We have simulated such a transformation by learning a basis for spatiochromatic 11x11 RGB patches based on the JADE ICA algorithm [4]. This is depicted in the top left of figure 1. This allows the projection of any local neighborhood into a space in which feature dimensions may be assumed mutually independent. The

likelihood of a given cell response can then be characterized by observing the distribution of responses of cells of that type in the surround allowing a likelihood estimate of the response of the cell in question which is readily converted to a measure of information via an inverse logarithm. The likelihood estimate in the implementation shown is performed as follows: For each image and a specific feature type, a histogram based on 100 bins is produced based on the response of all units of the type in question across the entire image. The likelihood of any individual response may then be computed on the basis of a lookup on the histogram. It is worth noting, that the property of considering only those units of the same type in the surround emerges from the nature of the learned basis for representing visual content. By construction, dependence across different feature types is minimized allowing a tractable multidimensional density estimate based on many 1-D histograms. In practice, there does exist residual correlation between similar features at a given location and a complete model might take this into account. In this implementation, across feature interactions have been ignored in the interest of computational parsimony. The information attributed to any given location can then be computed as a sum of the information attributed to all features for a given location. It is interesting to note the relation of this notion of saliency to an alternative recent approach by Itti and Baldi [5]. In the work of Itti and Baldi, saliency is defined as content that is *surprising* on the basis of an information theoretic measure based on the KL-divergence between prior and posterior models of visual content. The proposal based on information maximization is also a measure of surprise corresponding to the likelihood of observing a particular neuronal response based on the response of nearby neurons that characterize the surround in space-time. One might argue that this is a simpler more intuitive definition of surprise that may be evaluated on the current state of the neurons involved and with no memory requirements. The relation of this notion of surprise to neuroanatomy is also perhaps more explicit in the case of information maximization as detailed in the discussion section of the paper.

3 Attention and Visual Search

To consider whether the proposal put forth in [1] extends to basic results pertaining to attention, and is not merely correlated with some quantity that drives saccades, predictions of an information theoretic formulation are considered in the context of classic psychophysics results. It is shown in addition to predicting a wide range of attention related results, that the analysis sheds light on some visual search effects offering a different perspective on their interpretation and cause.

Despite the considerable effort that has been devoted to understanding visual search behavior, a consensus on the exact nature of mechanisms underlying selective attention has yet to be reached. The following section demonstrates that an explanation based on information seeking, while parsimonious, is able to account for a substantial proportion of basic results drawn from the psychophysical

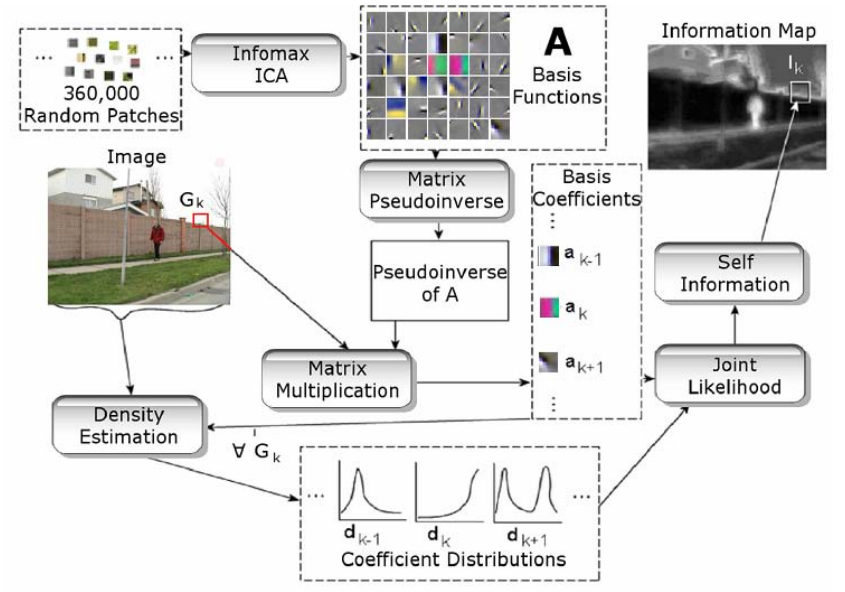

Fig. 1. A schematic of the computation underlying the determination of local saliency. Local content is projected into a basis optimized for mutual independence of coefficients. The likelihood of a response can then be computed within each feature domain by observing the distribution of responses of the same type of cell based on its support region.

literature on visual search including domains for which only specialized models have had success.

The time taken to find a specified target from among an array of elements is often assumed to reflect some measure of the saliency of this target relative to the saliency of competing distractors. In this work, we assume the amount of information determines relative saliency. Often attention models also prescribe a particular mechanism by which saliency translates into a shift in the focus of attention. The common element of such mechanisms, is that typically the *window* of attention gradually shifts from more salient to less salient targets. Search efficiency in this effort is thus equated with the saliency of the target relative to the saliency of distractors in line with other similar work (e.g. [6]).

3.1 Serial Versus Parallel Search

Curious is the observation that when searching for a variety of targets among distractors, some targets appear to "pop-out" while others require considerable effort to be found. This is exemplified in figures 2 and 3. In figure 2 the elements that are distinguished by a single feature (color or orientation) immediately pop-out. On the other hand, the singleton stimulus defined by a conjunction of features in figure 2 (top right) requires closer consideration of the stimulus elements to be spotted. In the case of figure 3 the smaller, red, and rotated 5's are

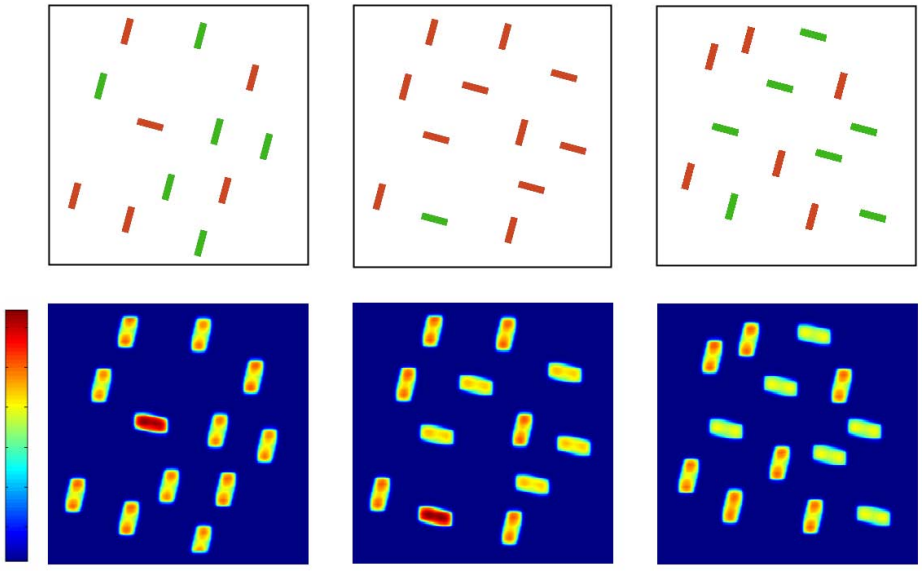

Fig. 2. Stimulus patterns for singletons defined by orientation, color and a conjunction of both (top, left to right) and their associated saliency maps (bottom, left to right)

found immediately, while finding the red 2 requires further effort. These sort of observations form the basis for Treisman's Feature Integration Theory (FIT), an influential contribution to our current understanding of visual search [7]. Treisman proposed that visual search consists of a two stage process. In the first stage, various basic features are measured in parallel across the entire visual field such as color, orientation and spatial frequency. If the first stage does not signal the presence of a target, a second stage occurs which considers single, or clusters of stimuli in turn. When target and distractor saliency are characterized in terms of information, the apparent distinction between parallel and serial search tasks is inherent in the difference between target and distractor saliency. The critical consideration is that within a sparse representation, the constituent features are assumed to be mutually independent. This implies that targets defined by a single feature are highly salient relative to the distractors, while those defined by a conjunction of features are indistinguishable from the distractor elements on the basis of saliency alone. Figure 4 shows a probability density representation of the response of a small number of hypothetical cell responses (idealized examples for the purpose of exposition) to the stimuli appearing in figure 2. For the case shown in figure 2 (top left), a large number of units respond to the stimuli oriented 15 degrees from vertical, and only a small number to the bar 15 degrees from horizontal. On the basis of this, the likelihood of the response associated with the singleton is lower and thus it is more informative. Since an approximately equal number of units respond to both green and red stimuli, this stimulus dimension dictates that all of the stimuli are equally informative. The

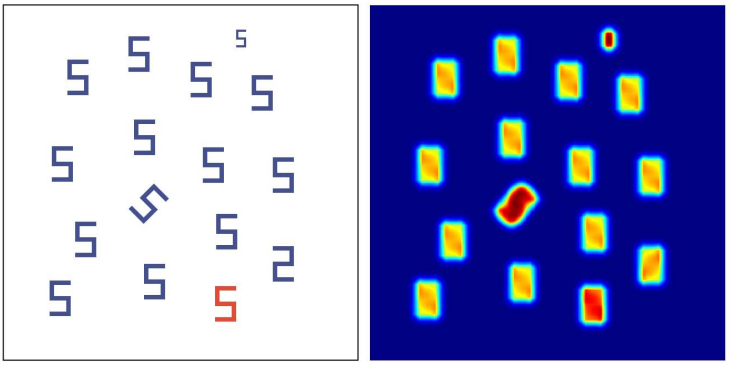

Fig. 3. An additional example of a conjunction search

situation for the stimulus shown in figure 2 (top middle) is analogous except that color is the discriminating dimension and orientation dictates all stimuli are equally salient. In the case of figure 2 (top right), there is a singleton element, but the number of units responding to all four cell types is approximately equal and as such, a serial search of the elements is required. The output of the implementation of AIM applied to the examples shown in figures 2 and 3 is shown below each of the stimulus examples in figure 2 and on the right in figure 3 revealing agreement between model output and the expected response to the stimuli in question. A scale of relative saliency is displayed (bottom left) based on maximum and minimum values for saliency equated across all conditions and is used in the remainder of the figures depicting the relative saliency equated across trials within each experiment.

The large body of visual search psychophysics that has ensued following Treisman's original proposal has revealed that behavior in search tasks is somewhat more involved than the dichotomy in search performance put forth by FIT. More specifically, it has been demonstrated that an entire continuum of search slopes may be observed ranging from very shallow to very steep in the number of display elements [8]. In the example of the conjunction search we have shown, we considered only a single unit for each of the two orientations present, and only a single unit for each color present. The assumption in this case is reasonable based on what is known about cell properties in V1 and is useful for the sake of demonstration. However, there are many types of stimuli that may require a representation in V1 by a large number of different cell types. Such types will not yield examples that are so clear cut. That being said, one important consideration that may be stated is that one would expect a continuum of saliency measures for such stimuli. That is, the saliency of targets relative to distractors depends on a complex distributed representation based on a large ensemble of many different types of cells. Without specific knowledge of the neural encoding on which attentive processes are operating, it may be difficult to form an

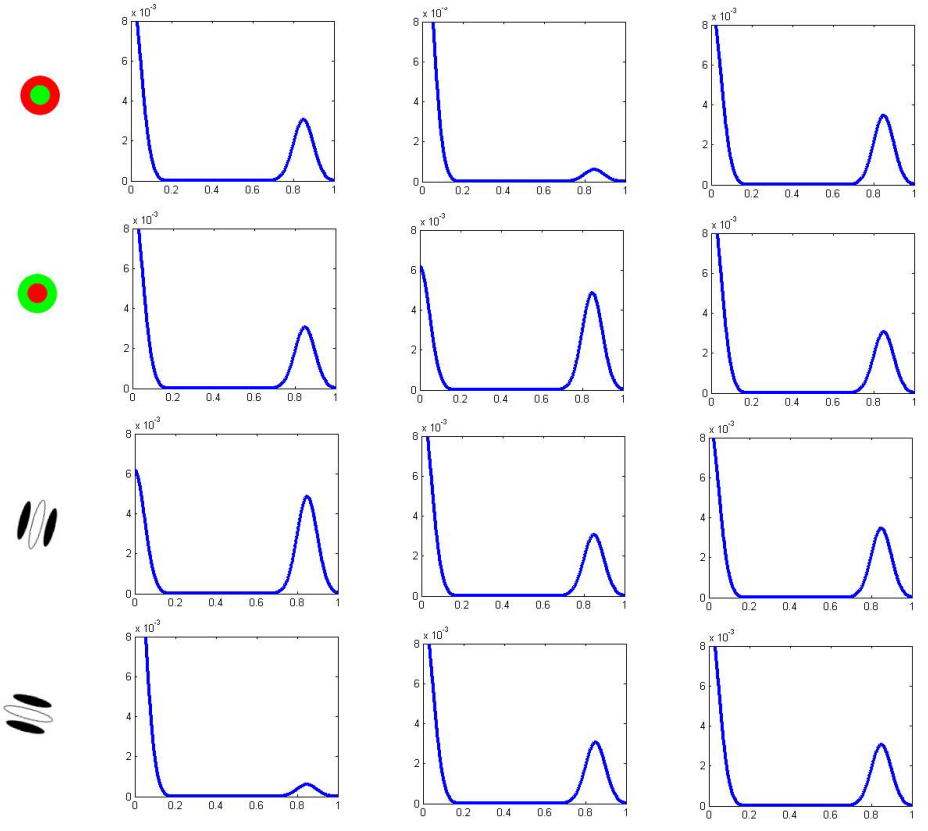

Fig. 4. Hypothetical probability densities associated with the response of four types of units. Shown is examples based on idealized units for the stimulus in question and crafted to exemplify how the responses of the units in question give rise to the observed effects.)

a priori determination of the difficulty of any given search task. That being said, it may be possible to determine a coarse ordering for different types of search on the basis of the coarse approximation of early visual coding we have learned. It is interesting to note that within an information theoretic interpretation, the observed behavior supports both the extreme view of FIT in the event that a single cell type exists that is perfectly tuned to each of the variations in target and distractor, and a continuum of difficulties between these extremes in more involved cases in which target and distractors are coded by a complex population of neurons.

3.2 Target-Distractor Similarity

Two factors that appear to be critical in determining the difficulty of search tasks are the similarity between target and distractors [9,10], and the heterogeneity of

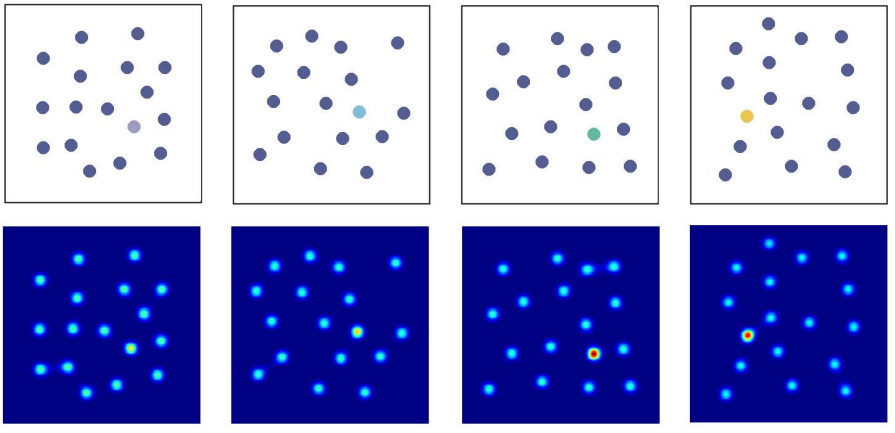

Fig. 5. Four stimulus patterns displaying increasing target-distractor distance in feature space (Top: left to right) and their associated saliency (Bottom: left to right)

distractors [9] (e.g. distractors at a single orientation versus several orientations). As the target becomes more similar to the distracting elements, the search task becomes more difficult as is the case for increased distractor heterogeneity. Existing explanations for this behavior appeal to signal detection theory, treating the difference between the target and distractors as the signal, and the distractor-distractor difference as noise. Generally these models are tailored specifically to addressing the issue of stimulus similarity in visual search. The appropriate behavior is inherent in AIM without the need to appeal to a more specialized model. Consider the stimulus shown in figure 5 (based on example shown in [11]). The basic result in this case is that the task of locating the target becomes progressively easier as the distance between target and distractor in feature space increases. So for example, the case shown top left in figure 5 is the most difficult, with cases becoming somewhat easier from left to right. A very important consideration in addressing whether the model yields appropriate behavior, is that beyond a certain distance in feature space, the effect of a further shift in feature space on search difficulty is negligable as observed in [9]. That is, the difficulty associated with finding the target in the top right stimulus example is equivalent to that of finding the target in the stimulus pane second from right. It is interesting to note that these results may be seen as consistent with the notion of an inhibitory surround in feature space as observed in [12] and as predicted in [13].

It is interesting to consider how each of these considerations correspond to the behaviour exhibited by AIM. The output of the model reveals that indeed a shift of target away from distractors in feature space renders an increase in search efficiency to a certain extent and at some point levels out as demonstrated in figure 5 (bottom row). The effect can be summarized as follows: The unit

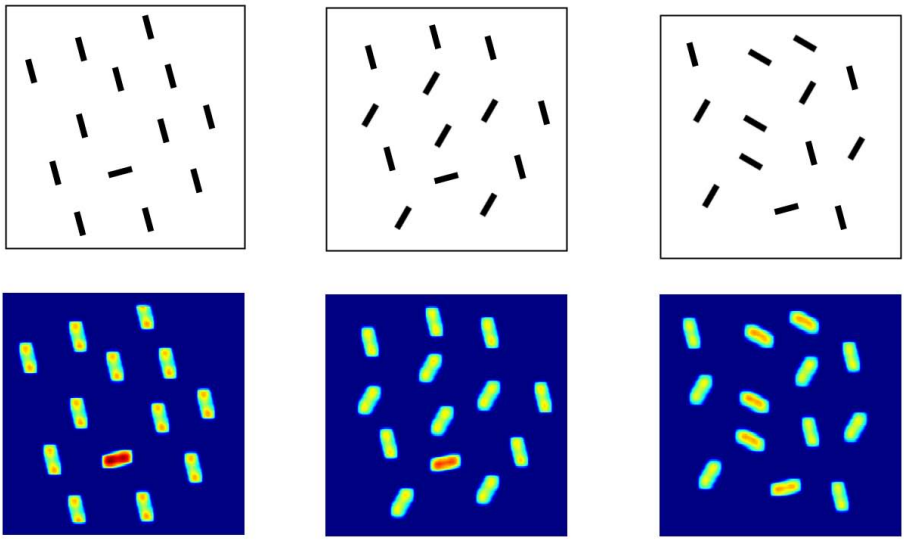

Fig. 6. An example of the effect of increasing distractor heterogeneity (Top: left to right) and saliency maps resulting from the output of AIM (Bottom: left to right)

whose optimal response corresponds most closely to the properties of the target item also elicits a response to the background elements. The strength of this response is inversely proportional to the distance in feature space between target and distractors. As such, distractor items similar to the target translate to an increased observation likelihood of features associated with the target leading to a decreased information value associated with the target. Outside of a certain distance in feature space, the distracting elements no longer elicit a response from the cell tuned to the target features.

3.3 Distractor Heterogeneity

Having addressed the effect of similarity between target and distractor, it is natural to also question the role of distractor-distractor similarity on visual search behaviour. The central result in this domain, is that an increase in distractor heterogeneity leads to an increase in search difficulty. This is exemplified by the stimulus patterns appearing in the top row of figure 7. In the top left case, the singleton item yields a pop-out effect which is diminished by increasing the spread of orientations present in distracting elements. The output of AIM demonstrating the predicted saliency of stimulus items appears in the bottom row, demonstrating the predicted output in agreement with the results presented in [9]. In this case there are two effects of increasing distractor heterogeneity, one of which is guaranteed for any ensemble of cells, and the other depending on the specific tuning properties of the cells in question. Splitting the distractor elements across two or more dimensions has the effect of lowering the observation likelihood of features associated with any given distractor thus rendering

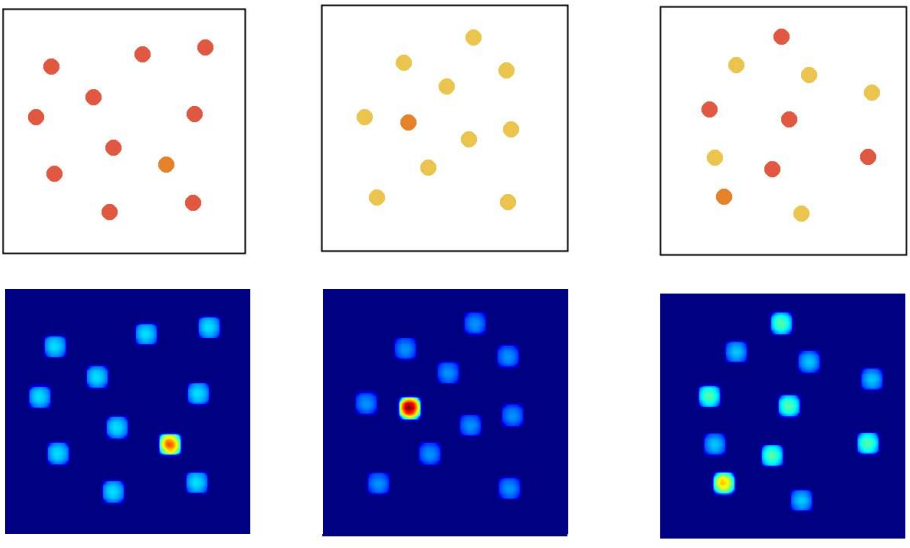

Fig. 7. Increased distractor heterogeneity in color space (top) and corresponding saliency maps (bottom)

them more salient. As a result the ratio of target relative to distractor saliency is diminished yielding a more difficult search. In the example shown, there is also a secondary effect of broad tuning curves on the orientation selective cells. As such, the distractors may increase the observation likelihood of the target item, and also there exists distractor-distractor interaction. This latter effect would presumably be eliminated given an encoding with more specific selectivity in the orientation domain.

3.4 Search "Asymmetries"

Apparent asymmetries in visual search paradigms have gained interest as an important consideration for models to address. Rosenholtz reveals that many of these asymmetries arise from asymmetric experiment design and thus are not truly search asymmetries [16]. For example, a pink circle among red circles may be easier to spot than a red circle among pink. However, changing the background saturation can reverse this effect as described in [14]. An example stimulus based on these experiments is shown in figure 8. Rosenholtz proposes a model of saliency based on the Mahalanobis distance between a target feature vector and the mean of the distractor distribution within some feature space. Rosenholtz' model is able to account for the behavior arising from asymmetric experiment design within a symmetric model. However, it is unclear how a model of this kind may generalize to account for some of the search behaviors described thus far such as the distinction between efficient and inefficient search tasks. The behavior observed in these experiments is intrinsic to the more general formulation of AIM as revealed by the output of the algorithm appearing in the

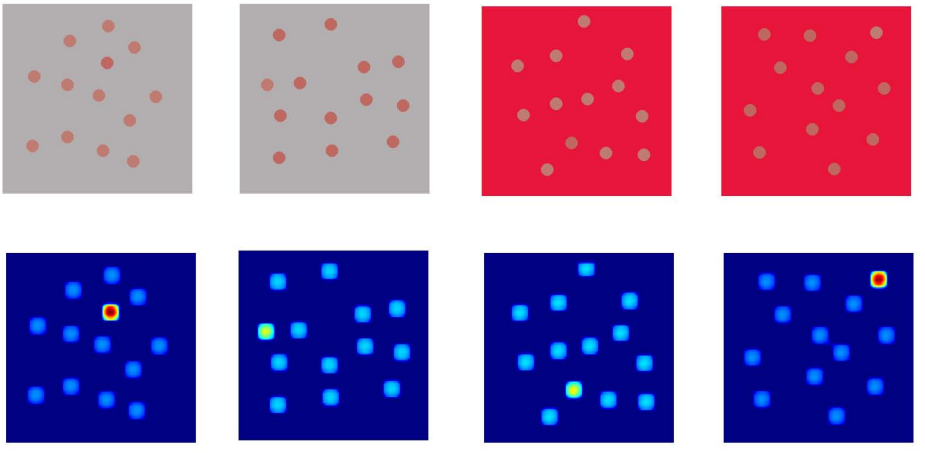

Fig. 8. An example of a visual search paradigm in which switching the background color reverses the difficulty ranking of the two conditions

bottom row of figure 8. The pink-target red-distractor search is more difficult than the converse, however a change in background color reverses the relative difficulty. This effect is due to the role that the background content plays in the likelihood estimate on target and distractor features. That is, increasing the likelihood associated with observations corresponding to the target or distractors respectively.

There are a few search asymmetries which have not yet been placed in the class of asymmetric experiment design, most notably the difference between the detectability of a moving target among stationary distractors versus a stationary target among coherently moving distractors. Consideration of this case in the context of AIM makes evident that this should also be classed as an asymmetric experimental design for the same reason as the color tasks. In the case of a moving target, motion selective neurons will respond strongly to the target, but not to the distractors and background. For the coherently moving distractors, motion selective units will respond to the distractors, and will produce no response for both the target and the background. As such, the target is easily discriminated in the moving target case, but not so in the moving distractor case. This of course relies on certain assumptions about the underlying spatiotemporal basis. This consideration generalizes to any apparent asymmetry where the presence of a feature results in pop-out while its absence results in an inefficient search. Additional examples include a Q among O's or a + among -'s. An example of this is depicted in figure 9 along with the output of AIM on these stimuli. It is interesting to note that the distinction typically made in the psychophysics literature between "true" asymmetries as in [15], and those resulting from poor experimental design [16] is moot when examined in the context of the behavior of AIM. In all cases, it is the role that activity in non-stimulus locations has on the

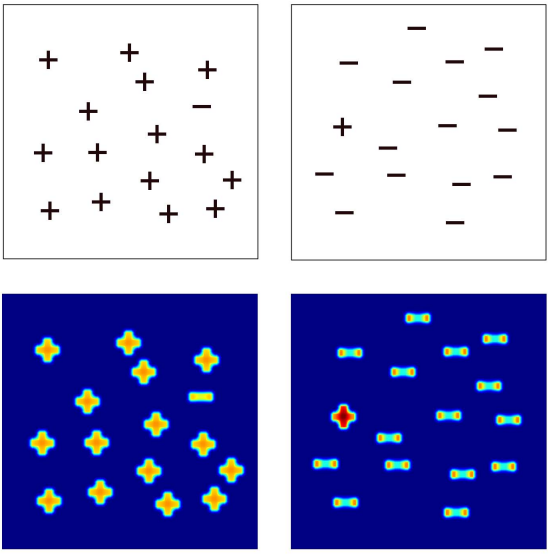

Fig. 9. An example of an asymmetry which results from the presence versus the absence of a feature

perceived saliency. This is an important point in particular for those models that posit properties derived from preattentive segmentation of target and distractor elements.

4 Discussion

In this paper, we considered the extent to which a definition of saliency motivated by information theory is in agreement with a large body of existing psychophysics results. Analysis reveals that the proposal is capable of addressing a wide range of behaviors including some which heretofor have only been observed in more specialized models. As a whole the results provide a compelling case for an information based definition in the determination of visual saliency and visual search behavior adding to the existing body of fixation based support for the proposal described in [1]. Future work will include a deeper analysis of some of the observed behaviors and drawing explicit connections to neural circuitry. Preliminary analysis reveals considerable similarity between the behavior of the model, and cortical gain control mechanisms (e.g. [17]) which we expect to reveal specific connections between primate neuroanatomy and the role of *information* in determining visual saliency.

Acknowledgments. The authors gratefully acknowledge the support of NSERC in supporting this work. John Tsotsos is the NSERC Canada Research Chair in Computational Vision.

References

1. Bruce, N., Tsotsos, J.K.: Saliency Based on Information Maximization. Advances in Neural Information Processing Systems 18, 155–162 (2006)
2. Olshausen, B.A., Field, D.J.: Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607–609 (1996)
3. Bell, A.J., Sejnowski, T.J.: The 'Independent Components' of Natural Scenes are Edge Filters. Vision Research 37(23), 3327–3338 (1997)
4. Cardoso, J.F.: High-order contrasts for independent component analysis. Neural Computation 11(1), 157–192 (1999)
5. Itti, L., Baldi, P.: Bayesian Surprise Attracts Human Attention. Advances in Neural Information Processing Systems 18, 547–554 (2006)
6. Li, Z.: A saliency map in primary visual cortex. Trends in Cognitive Sciences 6(1), 9–16 (2002)
7. Treisman, A., Gelade, G.: A feature integration theory of attention. Cognitive Psychology 12, 97–136 (1980)
8. Wolfe, J.M.: What Can 1,000,000 Trials Tell Us About Visual Search? Psychological Science 9(1) (1998)
9. Duncan, J., Humphreys, G.W.: Visual search and stimulus similarity. Psychol. Rev. 433, 433–458 (1989)
10. Pashler, H.: Target-distractor discriminability in visual search. Perception & Psychophysics 41, 285–292 (1987)
11. Wolfe, J.M., Horowitz, T.S.: What attributes guide the deployment of visual attention and how do they do it? Nature Reviews Neuroscience 5, 1–7 (2004)
12. Tombu, M.N., Tsotsos, J.K.: Attentional inhibitory surrounds in orentation space. Journal of Vision 5(8), 1013, 1013a (2005)
13. Tsotsos, J.K., Culhane, S., Yan Kei Wai, W., Lai, Y., Davis, N., Nuflo, F.: Modeling visual attention via selective tuning. Artificial intelligence 78, 507–545 (1995)
14. Rosenholtz, R., Nagy, A.L., Bell, A.R.: The effect of background color on asymmetries in color search. Journal of Vision 4(3), Article 9, 224–240 (2004)
15. Treisman, A., Gormican, S.: Feature analysis in early vision: evidence from search asymmetries. Psychol Rev. 95(1), 15–48 (1988)
16. Rosenholtz, R.: Search asymmetries? What search asymmetries? Perception & Psychophysics 63(3), 476–489 (2001)
17. Schwartz, O., Simoncelli, E.: Natural signal statistics and sensory gain control. Nature Neuroscience 4(8), 819–825 (2001)