On the plausibility of the discriminant center-surround hypothesis for visual saliency

Dashan Gao Vijay Mahadevan Nuno Vasconcelos
Statistical Visual Computing Laboratory,
Department of Electrical and Computer Engineering,
University of California San Diego,
9500 Gilman Drive, La Jolla, California, 92093
{dgao, vmahadev, nuno}@ucsd.edu

Abstract

It has been suggested that saliency mechanisms play a role in perceptual organization. This work evaluates the plausibility of a recently proposed generic principle for visual saliency: that all saliency decisions are optimal in a decision-theoretic sense. The discriminant saliency hypothesis is combined with the classical assumption that bottom-up saliency is a center-surround process, to derive a (decision-theoretic) optimal saliency architecture. Under this architecture, the saliency of each image location is equated to the discriminant power of a set of features with respect to the classification problem that opposes stimuli at center and surround. The optimal saliency detector is derived for various stimulus modalities, including color, orientation, and motion, and shown to make accurate quantitative predictions of various psychophysics of human saliency, for both static and motion stimuli. These include some classical non-linearities of orientation and motion saliency, and a Weber law that governs various types of saliency asymmetries. The discriminant saliency detectors are also applied to various saliency problems of interest in computer vision, including the prediction of human eye fixations on natural scenes, motion-based saliency in the presence of ego-motion, and background subtraction in highly dynamic scenes. In all cases, the discriminant saliency detectors outperform previously proposed methods, from both the saliency and the general computer vision literatures.

Introduction

An important goal of any perceptual system is to organize the various pieces of visual information that land on the retina. This organization requires both the grouping of distinct pieces into coherent units, to be perceived as objects, and the segregation of objects from their surroundings ("figure/ground" segregation). Both problems are simplified by a preliminary step of localized processing, known as bottom-up saliency, that highlights the regions of the visual field which most differ from their surround. These saliency mechanisms appear to rely on measures of local contrast (dissimilarity) of elementary features, like intensity, color, or orientation, into which the visual

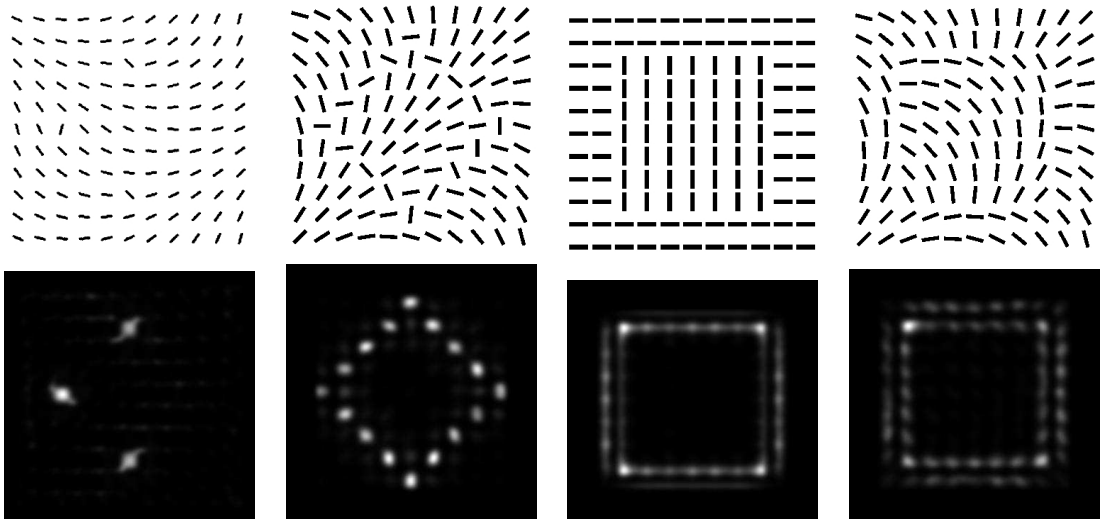


Figure 1. Four displays (top row) and saliency maps produced by the algorithm proposed in this article (bottom row). These examples show that saliency analysis facilitates aspects of perceptual organization, such as grouping (left two displays), and texture segregation (right two displays).

stimulus is first decomposed. It is well known that such contrast measures can reproduce perceptual phenomena such as texture segmentation (Beck, 1966b, 1972; Julesz, 1975, 1984; Olson & Attneave, 1970), target pop-out (Treisman, 1985; Treisman & Gormican, 1988; Nothdurft, 1991a), or even grouping (Beck, 1966a; Sagi & Julesz, 1985). For example, (Nothdurft, 1992) has shown that upon the brief inspection of a pattern such as that depicted in the leftmost display of Figure 1, subjects report the global percept of a "triangle pointing to the left". This percept is quite robust to the amount of (random) variability of the distractor bars, and to the orientation of the bars that make up the vertices of the triangle. In fact, these bars do not even have to be oriented in the same direction: the triangle percept only requires that they have sufficient orientation contrast with their neighbors. Another example of this type of perceptual grouping, as well as some examples of texture segregation, are shown in Figure 1 . Below each display we present the saliency maps produced by the saliency detector proposed in this work. Clearly, the saliency maps are informative of either the boundary regions or the elements to be grouped.

Computational modeling of saliency

The mechanisms of visual saliency, their neurophysiological basis, and psychophysics, have been extensively studied during the last decades. In result of these studies, it is now well known that saliency mechanisms exist for a number of elementary dimensions of visual stimuli (henceforth denoted as features), including color, orientation, depth, and motion, among others. More recently, there has been an increasing interest in computational models for saliency, in both biological and computer vision. The overwhelming majority of these models is inspired by, or aims to replicate, known properties of either the psychophysics or physiology of pre-attentive vision (Wolfe, 1994; Itti, Koch, & Niebur, 1998; Rosenholtz, 1999; Li, 2002; Bruce & Tsotsos, 2006; Harel, Koch, & Perona, 2007; Kienzle, Wichmann, Schölkopf, & Franz, 2007). These models all compute

a *saliency map* (Koch & Ullman, 1985), through either the combination of intermediate feature-specific saliency maps (e.g., Itti et al., 1998; Wolfe, 1994; Itti & Koch, 2001), or the direct analysis of feature interactions (e.g., Li, 2002).

What distinguishes these models is mostly the computational measure of saliency. In what is perhaps the most popular model for bottom-up saliency, Itti et al. (1998) measures contrast as the difference between the stimulus at a location and the stimulus in its neighborhood, in a center-surround fashion. This model has been shown to successfully replicate many observations from psychophysics (Itti & Koch, 2000; Parkhurst, Law, & Niebur, 2002; Peters, Iyer, Itti, & Koch, 2005), for both static and motion stimuli, and applied to the design of computer vision algorithms for robotics, and video compression (Itti, 2004; Walther & Koch, 2006; Shic & Scassellati, 2007). In the Guided Search model, Wolfe (1994) has, on the other hand, emphasized the modulation of the bottom-up activation maps by top-down, goal-dependent, knowledge. Li (2002) has argued that saliency maps are a direct product of the pre-attentive computations of primary visual cortex (V1), and implemented a saliency model inspired by the basic properties of the neural structures found in V1. This has also been shown to reproduce many psychophysical traits of human saliency, establishing a direct link between psychophysics and the physiology of V1. While many of these early saliency models aimed to reproduce various known properties of biological vision, they lacked a formal justification for their image processing steps in terms of a unifying computational principle for saliency. Some more recent models have tried to address this problem, by deriving saliency mechanisms as optimal implementations of generic computational principles, such as the maximization of self-information (Bruce & Tsotsos, 2006), or "surprise" (Itti & Baldi, 2005). It is not yet clear how closely these models comply with the classical psychophysics, since existing evaluations have been limited to the prediction of human eye fixation data.

In this work, we study the effectiveness of an alternative, and currently less popular, hypothesis that all saliency decisions are *optimal in a decision-theoretic sense*. This hypothesis is denoted as *discriminant saliency*, and was first proposed by Gao and Vasconcelos (2005), in a computer vision context. While initially posed as an explanation for top-down saliency, of interest mostly for object recognition, the hypothesis of decision theoretic optimality is much more general, and indeed applicable to any form of center-surround saliency. This has motivated us to test its ability to explain the psychophysics of human saliency. Since these are better documented for the bottom-up neural pathway than for its top-down counterpart, we derive a bottom-up saliency detector which is optimal in a decision-theoretic sense. In particular, we hypothesize that, in the absence of high-level goals, the most salient locations of the visual field are those that enable the discrimination between center and surround with smallest expected probability of error. This is referred to as the *discriminant center-surround hypothesis* and, by definition, produces saliency measures that are optimal in a classification sense. We derive optimal mechanisms for a number of saliency problems, ranging from static spatial saliency, to motion-based saliency in the presence of ego-motion or even complex dynamic backgrounds. The ability of these mechanisms to both reproduce the classical psychophysics of human saliency, and solve saliency problems of interest for computer vision, is then evaluated. From the psychophysics point of view, it is shown that, for both static and moving stimuli, discriminant saliency not only explains all observations previously replicated by existing models, but also makes quantitative predictions (for non-linear aspects of human saliency) which are beyond their reach. From the computer vision standpoint, it is shown that the saliency algorithms now proposed can predict human eye fixations with greater accuracy than previous approaches, and outperform state-of-the-art algorithms for background subtraction. In particular, it is shown that, by

simply modifying the probabilistic models employed in the discriminant saliency measure - from well known models of natural image statistics, to the statistics of simple motion features, to more sophisticated dynamic texture models - it is possible to produce saliency detectors for either static or dynamic stimuli, which are insensitive to background image variability due to texture, ego-motion, or scene dynamics.

Discriminant center-surround saliency

Discriminant saliency

Discriminant saliency is rooted in a decision-theoretic interpretation of perception. Under this interpretation, perceptual systems evolve to produce decisions about the state of the surrounding environment that are *optimal in a decision-theoretic sense*, e.g. that have minimum probability of error. This goal is complemented by one of *computational parsimony*, i.e. that the perceptual mechanisms should be as efficient as possible. Discriminant saliency is defined with respect to two classes of stimuli: a class of *stimuli of interest*, and a *null* hypothesis, composed of all the stimuli that are not salient. Given these two classes, the locations of the visual field that can be classified, with *lowest expected probability of error*, as containing stimuli of interest are denoted as salient. Mathematically, this is accomplished by 1) defining a binary classification problem that opposes stimuli of interest to the null hypothesis, and 2) equating the saliency of each location in the visual field to the discriminant power (with respect to this problem) of the visual features extracted from that location. This definition of saliency is applicable to a broad set of problems. For example, different specifications of stimuli of interest and null hypothesis enable its specialization to both top-down and bottom-up saliency. From a computational standpoint, the search for discriminant features is a well-defined, and tractable, problem that has been widely studied in the literature. These properties have been exploited, by Gao and Vasconcelos (2005), to derive an optimal top-down saliency detector, which equates stimuli of interest to an object class and null hypothesis to all other object classes. In this work, we consider the problem of bottom-up saliency.

Discriminant center-surround saliency

Due to the ubiquity of "center-surround" processing in the early stages of biological vision (Hubel & Wiesel, 1965; Knierim & Van Essen, 1992; Cavanaugh, Bair, & Movshon, 2002), it is commonly assumed that bottom-up saliency is determined by how distinct the stimulus at each location of the visual field is from the stimuli in its surround. This "center-surround" hypothesis can be naturally formulated as a classification problem, as required by discriminant saliency, and illustrated in Figure 2. This consists of defining, at each image location l,

- *stimuli of interest:* observations within a neighborhood $\mathcal{W}_l^1$ of l (henceforth referred to as the ***center***), and
- *null hypothesis:* observations within a surrounding window $\mathcal{W}_l^0$ (henceforth referred to as the ***surround***).

All observations are responses, to the visual stimulus, of a pre-defined set of features $\mathbf{X}$. The saliency of location l is equated to the power of $\mathbf{X}$ to discriminate between the *center* and *surround* of l, based on the distributions of the feature responses estimated from the two regions.

Mathematically, the feature responses within the two windows, $\mathcal{W}_l^0$ and $\mathcal{W}_l^1$ are observations from a random process $\mathbf{X}(l) = (X_1(l), \ldots, X_d(l))$, of dimension d, drawn conditionally on the state of a hidden variable $Y(l)$. The feature vector observed at location j is denoted as $\mathbf{x}(j) =$

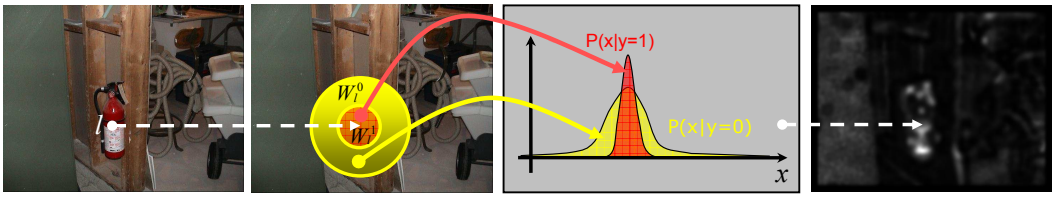


Figure 2. Illustration of discriminant center-surround saliency.

$(x_1(j), \ldots, x_d(j))$. Feature vectors $\mathbf{x}(j)$ such that $j \in \mathcal{W}_l^c, c \in \{0, 1\}$ are drawn from class c according to conditional densities $P_{\mathbf{X}(l)|Y(l)}(\mathbf{x}|c)$. Vectors drawn with $Y(l) = c$ are referred to as belonging to the *center class* if $c = 1$ and the *surround class* if $c = 0$. The saliency of location l, $S(l)$, is equal to the discriminant power of $\mathbf{X}$ for the classification of the observed feature vectors $\mathbf{x}(j), \forall j \in \mathcal{W}_l = \mathcal{W}_l^0 \cup \mathcal{W}_l^1$, into *center* and *surround*. This is quantified by the mutual information between features, $\mathbf{X}$, and class label, Y,

$$\begin{aligned} S(l) &= I_l(\mathbf{X}; Y) \\ &= \sum_c \int p_{\mathbf{X}(l),Y(l)}(\mathbf{x}, c) \log \frac{p_{\mathbf{X}(l),Y(l)}(\mathbf{x}, c)}{p_{\mathbf{X}(l)}(\mathbf{x}) p_{Y(l)}(c)} d\mathbf{x}. \end{aligned} \quad (1)$$

The l subscript emphasizes the fact that both the classification problem and the mutual information are defined locally, within $\mathcal{W}_l$. The function $S(l)$ is referred to as the ***saliency map***. Note that (1) defines the discriminant saliency measure in a very generic sense, independently of the stimulus dimension under consideration, or any specific feature sets. In fact, (1) can be applied to any type of stimuli, and any type of local features, as long as the probability densities $P_{\mathbf{X}(l)|Y}(\mathbf{x}|c)$ can be estimated from the center and surround neighborhoods. In what follows, we derive the discriminant center-surround saliency for a variety of features, including intensity, color, orientation, motion, and even more complicated dynamic texture models.

Discriminant saliency detection in static imagery

We start by deriving the optimal saliency detector for static stimuli, whose building blocks are illustrated in Figure 3.

Extraction of visual features

The first stage, feature decomposition, follows the proposal of Itti and Koch (2000), which closely mimics the earliest stages of biological visual processing. The image to process is first subject to a feature decomposition into an intensity map (I), and four broadly-tuned color channels (R, G, B, and Y),

$$\begin{aligned} I &= (r + g + b)/3, \\ R &= \lfloor \tilde{r} - (\tilde{g} + \tilde{b})/2 \rfloor_+, \\ G &= \lfloor \tilde{g} - (\tilde{r} + \tilde{b})/2 \rfloor_+, \\ B &= \lfloor \tilde{b} - (\tilde{r} + \tilde{g})/2 \rfloor_+, \\ Y &= \lfloor (\tilde{r} + \tilde{g})/2 - |\tilde{r} - \tilde{g}|/2 \rfloor_+, \end{aligned}$$

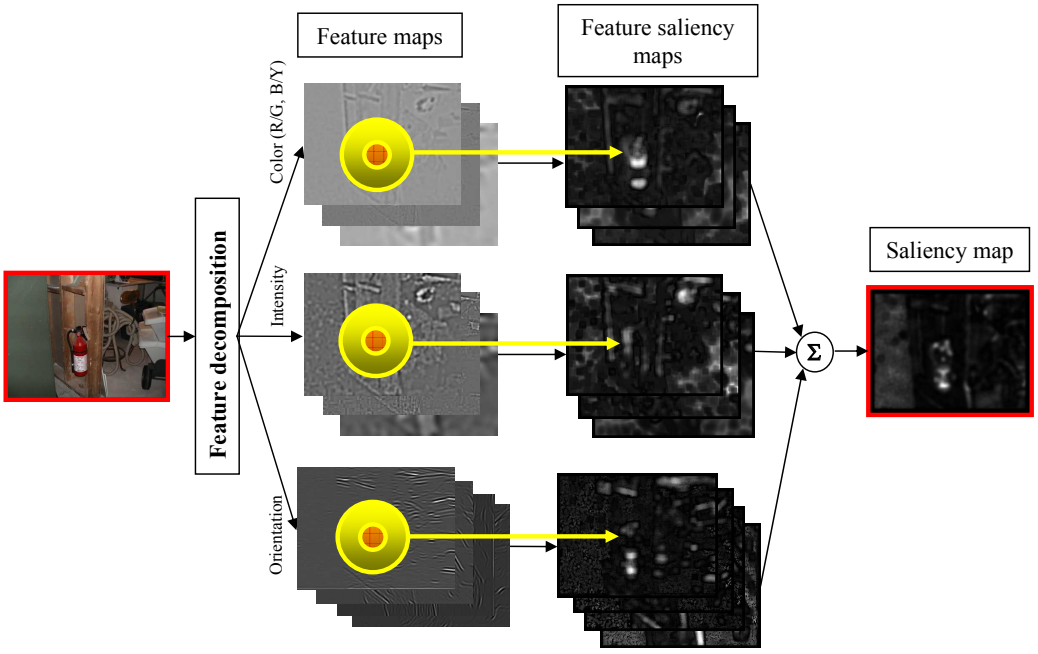


Figure 3. Bottom-up discriminant saliency detector.

where $\tilde{r} = r/I, \tilde{g} = g/I, \tilde{b} = b/I$, and $\lfloor x \rfloor_+ = \max(x, 0)$. The four color channels are, in turn, combined into two color opponent channels, $R - G$ for red/green and $B - Y$ for blue/yellow opponency. The two opponency channels, together with the intensity map, are convolved with three Mexican hat wavelet filters, centered at spatial frequencies 0.02, 0.04 and 0.08 cycle/pixel, to generate nine feature channels. The feature space $\mathcal{X}$ consists of these nine channels, plus a Gabor decomposition of the intensity map, implemented with a dictionary of zero-mean Gabor filters at 3 spatial scales (centered at frequencies of 0.08, 0.16, and 0.32 cycle/pixel) and 4 directions (evenly spread from 0 to π).

Leveraging natural image statistics

The second stage of the detection involves estimating the mutual information of (1), at each image location, for the center-surround classification problem. This is, in general, impractical since it requires density estimates on a potentially high-dimensional feature space. A known statistical property of band-pass natural image features, such as Gabor or wavelet coefficients, can nevertheless be exploited to drastically reduce complexity. This property is that band-pass features exhibit strongly *consistent* patterns of dependence across a very wide range of natural image classes (Buccigrossi & Simoncelli, 1999; Huang & Mumford, 1999). For example, Buccigrossi and Simoncelli (1999) have shown that, when a natural image is subject to a wavelet decomposition, the conditional distribution of any wavelet coefficient, given the state of the co-located coefficient of immediately coarser scale (known as its "parent"), invariably has a bow-tie shape. This implies that, while the coefficients are statistically dependent, their dependencies carry little information about the image class (Vasconcelos, 2004; Buccigrossi & Simoncelli, 1999). In the particular case of saliency, feature dependencies are not greatly informative about whether the observed feature vectors originate

in the center or the surround. Experimental validation of this hypothesis (Vasconcelos, 2003, 2004) has shown that, for natural images, (1) is well approximated by the sum of marginal mutual informations between individual features and class label[1]

$$S(l) \;=\; \sum_{i=1}^{d} I_l(X_i; Y). \tag{2}$$

This is a sensible compromise between decision theoretic optimality and computational parsimony.

Since (2) only requires estimates of marginal densities, it has significantly less complexity than (1). This complexity can be further reduced by exploiting the well known fact that marginal densities of band-pass features are accurately modeled by a generalized Gaussian distribution (GGD) (Modestino, 1977; Clarke, 1985; Mallat, 1989),

$$P_X(x; \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{ -\left(\frac{|x|}{\alpha} \right)^{\beta} \right\}, \tag{3}$$

where $\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} \mathrm{d}t$, $t > 0$, is the Gamma function, α a *scale* parameter, and β a *shape* parameter. The parameter β controls the decay rate from the peak value, and defines a sub-family of the GGD (e.g., the Laplacian family when $\beta = 1$ or the Gaussian family when $\beta = 2$). When the class conditional densities, $P_{X|Y}(x|c)$, and the marginal density, $P_X(x)$, follow a GGD, the mutual information of (2) has a closed form (Do & Vetterli, 2002)

$$I(X; Y) = \sum_c P_Y(c) KL\left[P_{X|Y}(x|c) || P_X(x) \right], \tag{4}$$

with

$$KL[P_X(x; \alpha_1, \beta_1) || P_X(x; \alpha_2, \beta_2)] = \log\left(\frac{\beta_1 \alpha_2 \Gamma(1/\beta_2)}{\beta_2 \alpha_1 \Gamma(1/\beta_1)} \right) + \left(\frac{\alpha_1}{\alpha_2} \right)^{\beta_2} \frac{\Gamma((\beta_2 + 1)/\beta_1)}{\Gamma(1/\beta_1)} - \frac{1}{\beta_1}, \tag{5}$$

where $KL[p||q] = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$ is the Kullback-Leibler (K-L) divergence between $p(x)$ and $q(x)$. Hence, the discriminant saliency measure only requires the estimation of the α and β parameters, for the center and surround widows, and the computation of (4), (5), and (2).

Gao and Vasconcelos (2007) have shown that, for maximum a posteriori estimation of the parameters (α_c and β_c, $c \in \{0, 1\}$) with conjugate (Gamma) priors, there is a one-to-one mapping between the discriminant saliency detector and a neural network that replicates the standard architecture of V1: a cascade of linear filtering, divisive normalization, quadratic non-linearity and spatial pooling. In the implementation presented in this article, we have instead adopted the method of moments for all parameter estimation, because it is computationally more efficient on a non-parallel computer. Under the method of moments, α and β are estimated through the relationships

$$\sigma^2 = \frac{\alpha^2 \Gamma(\frac{3}{\beta})}{\Gamma(\frac{1}{\beta})} \quad \text{and} \quad \kappa = \frac{\Gamma(\frac{1}{\beta})\Gamma(\frac{5}{\beta})}{\Gamma^2(\frac{3}{\beta})}, \tag{6}$$

[1] Note that this approximation *does not* assume that the features are independently distributed, but simply that their dependencies are not informative about the class.

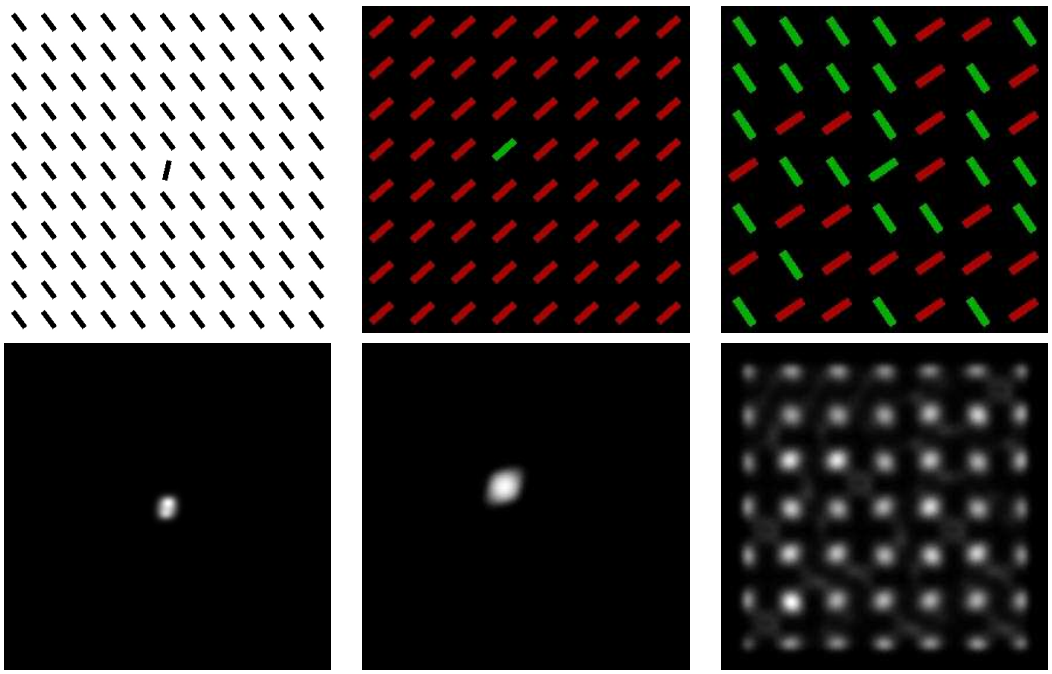


Figure 4. Discriminant saliency output (bottom row) for displays (top row) where target and distractors differ in terms of single features ((a) orientation and (b) color), or (c) feature conjunctions (color and orientation). Brightest regions are most salient. The strong saliency peaks at the targets of (a) and (b) indicate a strong pop-out effect. The lack of distinguishable saliency variations between the target (fourth line and fourth column) and distractors of (c) indicates that the target does not pop-out.

where σ^2 and κ are, respectively, the variance and kurtosis of X

$$\sigma^2 = E_X[(X - E_X[X])^2], \text{ and } \kappa = \frac{E_X[(X - E_X[X])^4]}{\sigma^4}.$$

In summary, parameter estimation only requires sample moments of the feature responses within the center and surround windows, and is very efficient. The method of moments has also been shown to produce good fits to natural images (Huang & Mumford, 1999).

For all experiments reported in this work, the choice of center and surround window sizes was guided by studies from psychophysics and neurophysiology (e.g., Nothdurft, 2000; Cavanaugh et al., 2002). For the psychophysics experiments of the following section, we followed the common practice (e.g., Treisman & Gelade, 1980; Hubel & Wiesel, 1965) of setting the size of the center window to a value *comparable* to that of the display items, and the size of the surround window to 6 times that value. Informal experimentation with these parameters has shown that the saliency results are not significantly affected by variations around the parameter values adopted. To improve intelligibility, the saliency maps shown in this article were subject to smoothing, contrast enhancement (by squaring), and a normalization of the saliency value to the interval $[0, 1]$. This implies that absolute saliency values are not comparable across displays, but only within each saliency map.

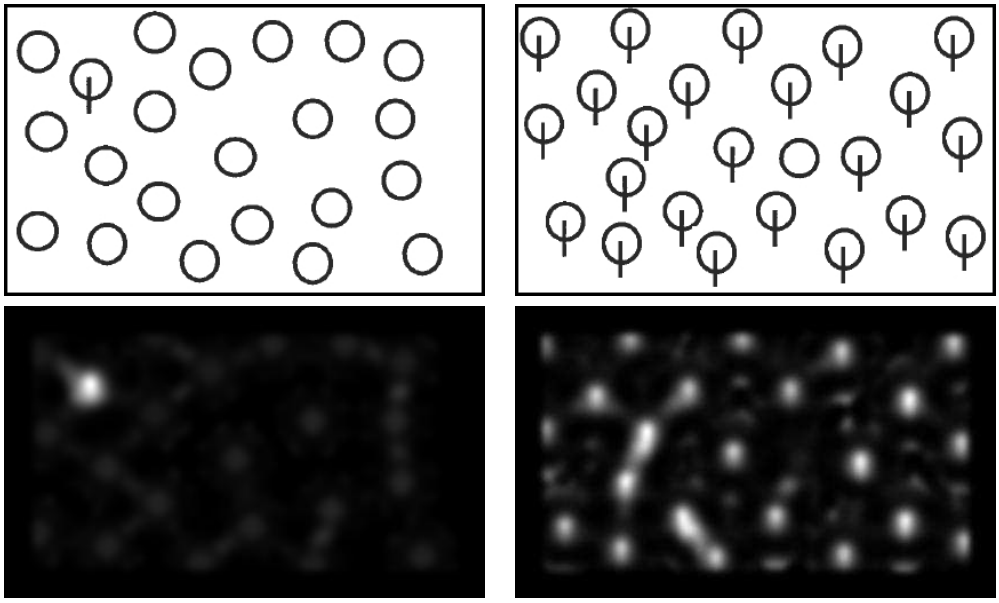


Figure 5. Example of pop-out asymmetry (discriminant saliency maps shown below each display). (left) A target ("Q") defined by the presence of a feature that the distractors ("O") lack produces a strong pop-out effect. (right) The reverse does not lead to noticeable pop-out.

Consistency with psychophysics

To evaluate the compliance of discriminant saliency with psychophysics, we start with a series of displays used in classical studies of visual attention (Treisman & Gelade, 1980; Treisman & Gormican, 1988; Nothdurft, 1993). These displays are commonly used in the literature to investigate whether saliency detectors reproduce the fundamental properties of human saliency (Itti et al., 1998; Rosenholtz, 1999). This is the case of discriminant saliency, which replicates the percept of pop-out for single feature search (e.g., Figure 4 (a), (b)), disregard of feature conjunctions (e.g., Figure 4 (c)), and saliency asymmetries for feature presence vs. absence (e.g., Figure 5), in addition to various grouping and segmentation percepts (e.g., Figure 1). Although interesting, this type of evaluation is purely *qualitative*, and therefore anecdotal. Given the simplicity of the displays, it is not hard to conceive of other center-surround operations that could produce similar results. To address this problem, we introduce an alternative evaluation strategy, based on the comparison of *quantitative predictions*, made by the saliency detectors, and available human data. It is our belief that quantitative predictions are essential for an *objective* comparison of different saliency principles. We show that this process can be useful, by performing various objective comparisons between discriminant saliency and the popular saliency model of Itti and Koch (2000)[2].

In the first experiment, we examine the ability of the saliency detectors to predict a well known nonlinearity of human saliency. While it has long been known that local feature contrast affects percepts such as target pop-out and texture segregation, most early studies in the psychophysics of saliency pursued the threshold at which these events occur. Examples includes the threshold at

[2]Results obtained with the MATLAB implementation by Walther and Koch (2006).

which a (previously non-salient) target pops-out (Foster & Ward, 1991; Nothdurft, 1991b), two formerly indistinguishable textures segregate (Landy & Bergen, 1991; Julesz, 1981), a "serial" visual search becomes "parallel", or vice versa (Treisman & Gelade, 1980; Wolfe, Friedman-Hill, Stewart, & O'Connell, 1992; Moraglia, 1989). In the context of objective evaluation, these studies are less interesting than a posterior set, which also measured the saliency of pop-out targets above the detection threshold (Nothdurft, 1993; Motoyoshi & nishida, 2001; Regan, 1995). In particular, Nothdurft (1993) characterized the saliency of pop-out targets due to orientation contrast, by comparing the conspicuousness of orientation defined targets and luminance defined ones, and using luminance as a reference for relative target salience. He showed that the saliency of a target increases with orientation contrast, but in a non-linear manner, exhibiting both threshold and saturation effects: 1) there exists a threshold below which the effect of pop-out vanishes, and 2) above this threshold saliency increases with contrast, saturating after some point. The overall relationship has a *sigmoidal* shape, with lower (upper) threshold t_l (t_u).

The results of this experiment are illustrated in Figure 6. The figure presents plots of saliency strength as a function of the target orientation contrast. The human data collected by Nothdurft (1993) is presented in (a), while the predictions of the discriminant saliency detector are shown in (b). Note that the latter closely predicts the strong threshold and saturation effects of the former, suggesting that $t_l \approx 10^\circ$ and $t_u \approx 40^\circ$. These predictions are consistent with the human data. The same experiment was repeated for the model of Itti and Koch (2000) which, as illustrated by Figure 6 (c), exhibited no quantitative compliance with human performance.

A second experiment addressed the ability of the saliency detectors to make quantitative predictions regarding classical saliency asymmetries: while the presence in the target of some feature absent from the distractors produces pop-out, the reverse (pop-out due to the absence, in the target, of a distractor feature) does not hold (Treisman & Gormican, 1988). The qualitative results of Figure 5 show that discriminant saliency has the ability to reproduce these asymmetries. We investigated if it could also make objective predictions for the strength of this asymmetry. For this, we relied on data collected in visual search experiments (Treisman & Gormican, 1988), which showed that asymmetries occur not only for the existence and absence of a feature, but also for quantitatively weaker and stronger responses along one feature dimension. In fact, through a series of experiments involving displays in which the target differs from distractors only in terms of length, Treisman showed that the asymmetries follow Weber's law. Figure 7 (a) presents one example of the displays used in this experiment, where the target (a vertical bar at the center of the display) has a different length from the distractors (a set of vertical bars). The discriminant saliency detector was applied to these displays, and the results are presented, in Figure 7 (b). The figure shows the saliency predictions obtained at the target location, across the set of displays, as a scatter plot. The dashed line shows the best fit to Weber's law: target saliency is approximately linear in the ratio between the difference of target/distractor length (Δx) and distractor length (x). For comparison, Figure 7 (c) presents the corresponding scatter plot for the model of Itti and Koch (2000), which does not replicate human performance.

Motion saliency

An important property of human saliency is its ubiquity: saliency mechanisms have been observed for various cues, including orientation, color, texture and motion (Treisman & Gelade, 1980; Nothdurft, 1991a). It has also been suggested that orientation and motion saliency could be encoded by similar mechanisms (Nothdurft, 1993; Ivry & Cohen, 1992). Since (1) can be applied to

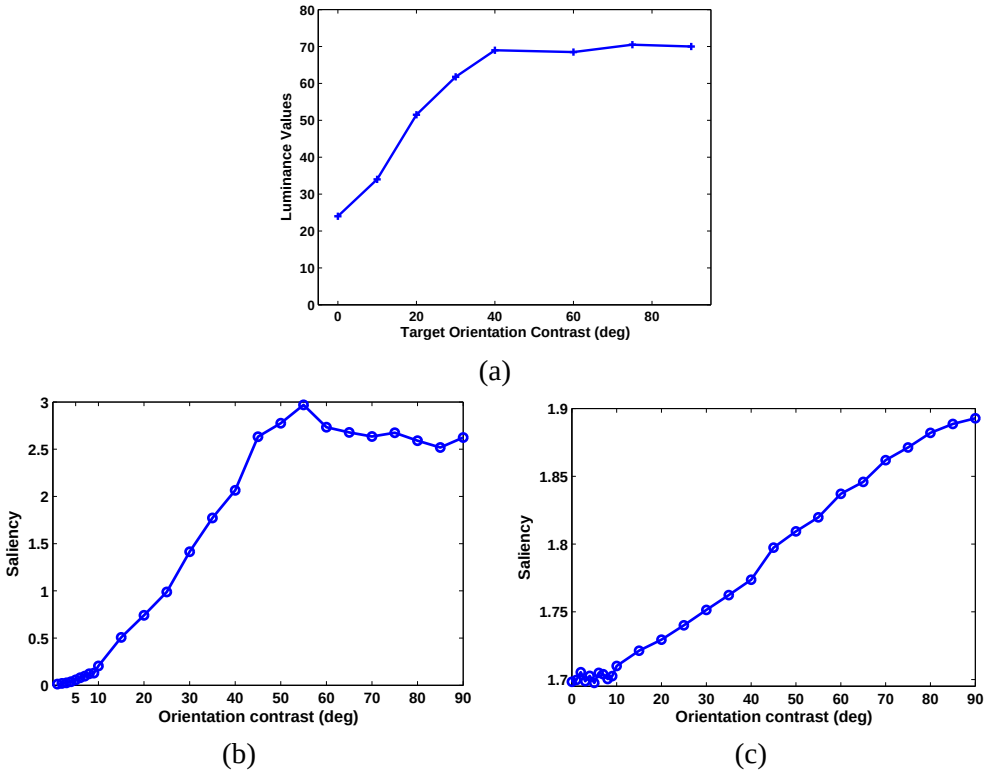


Figure 6. The nonlinearity of human saliency responses to orientation contrast (reproduced from Figure 9 of Nothdurft (1993)) (a) is replicated by discriminant saliency (b), but not by the model of Itti & Koch (2000) (c).

any type of stimuli and features this is, in principle, possible to replicate with discriminant saliency. In this section we verify this hypothesis, by deriving the discriminant saliency detector for motion stimuli, and provide evidence of its ability to predict human psychophysics.

Motion-based discriminant saliency detector

To compute motion information from video sequences, we adopt the spatiotemporal filtering approach of Adelson and Bergen (1985), and Heeger (1988). Spatiotemporal filtering is a biologically plausible mechanism for motion estimation, and has been shown to comply with the physiology and psychophysics of the early stages of the visual cortex (Adelson & Bergen, 1985). Since spatiotemporal orientation is equivalent to velocity, a set of 3-D Gabor (spatiotemporal) filters, tuned to a specific orientation in space and time, is used to extract the motion energy associated with different velocities. The algorithmic implementation of the spatiotemporal filters used in this work was based on the separable spatiotemporal filters of Heeger (1988). We considered only one spatial scale, and the spatial frequency of each Gabor filter was fixed to $.25$ cycles per pixel. Three temporal scales (temporal frequencies of 0,$\pm.25$ cycles per frame) and 4 spatial orientations ($0, \pi/4, \pi/2$ and $3\pi/4$) were used, in a total of 12 filters. The standard deviation of the spatial Gaussian was set to 1, and that of the temporal Gaussian to 2. This set of filter parameters were chosen for simplicity, we have not experimented thoroughly with them. We have also only considered the intensity of the input

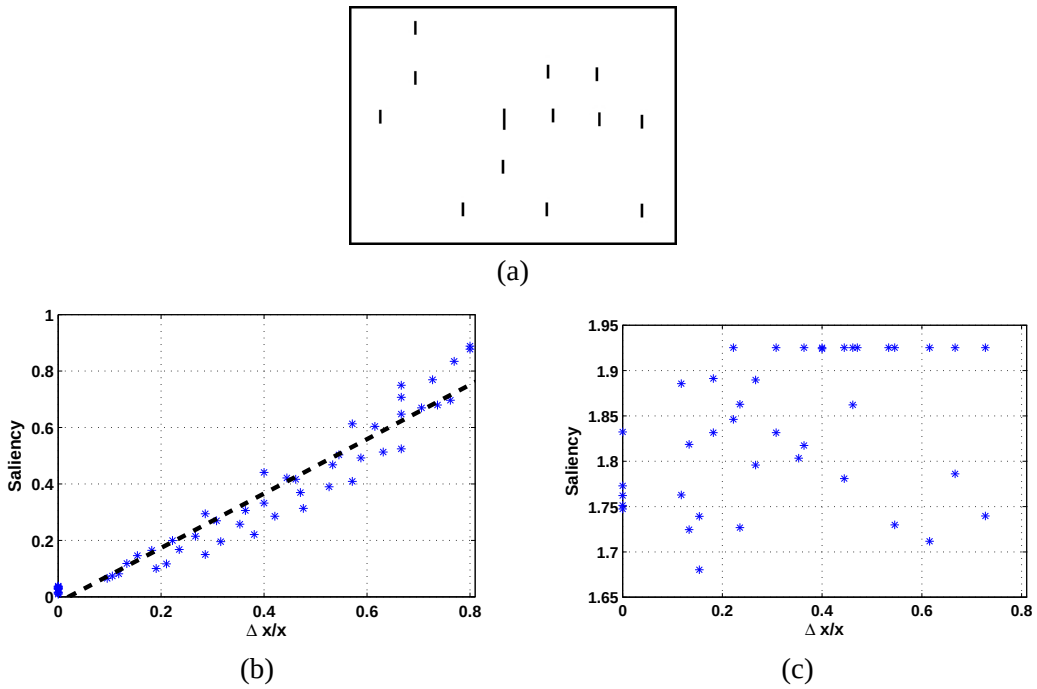


Figure 7. An example display (a) and performance of saliency detectors (discriminant saliency (b) and the model of Itti & Koch (2000) (c)) on Treisman's Weber's law experiment.

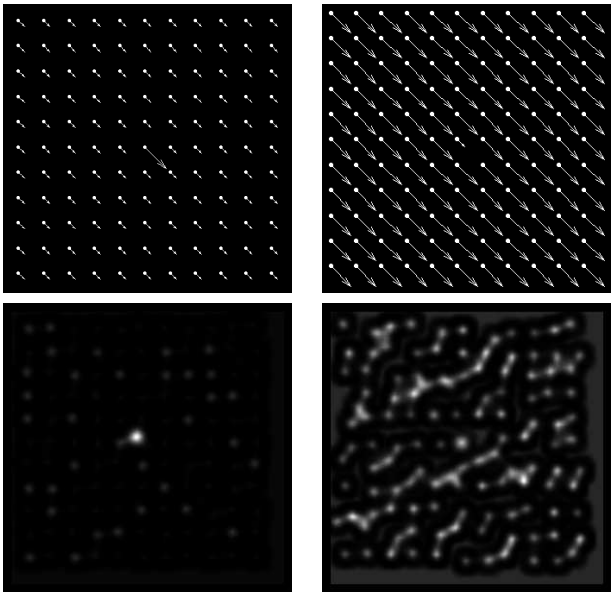
video frames, and all color information was discarded. These intensity maps were convolved with the 12 spatiotemporal filters, to produce the feature maps used by the saliency algorithm. Saliency was then computed as in the static case, using (2)-(5).

Consistency with psychophysics of motion perception

To evaluate the compliance of the discriminant saliency detector with the psychophysics of human motion saliency (Nothdurft, 1993; Ivry & Cohen, 1992), we start with some qualitative observations[3]. Ivry and Cohen (1992) showed that search asymmetries also hold for moving stimuli. For example, searching for a fast-moving target among slowly-moving distractors is easier than the reverse. We applied the motion-based discriminant saliency detector to a set of sequences used to demonstrate the asymmetries of motion pop-out (Ivry & Cohen, 1992), with the results illustrated in Figure 8. The figure presents quiver plots of the motion stimuli, under the two conditions, and one frame of the resulting discriminant saliency map. The conspicuous saliency peak at the target in Figure 8 (a) shows a strong pop-out effect when the target speed is greater than that of the distractors. No noticeable pop-out effect is observed in Figure 8 (b), where the distractor speed is greater than that of the target. This shows that the discriminant saliency detector can replicate the asymmetries of motion saliency.

As was the case for static stimuli, we complemented this qualitative observation with a quantitative analysis of the saliency predictions made by the discriminant detector. Nothdurft (1993) found that human saliency responses to motion are very similar to those observed for orientation:

[3] All motion stimuli sequences in the experiments were generated using the Psychtoolbox (Brainard, 1997).



(a) (b)

Figure 8. Discriminant saliency detector output for (a) a fast-moving target among slowly-moving distractors, and (b) a slowly-moving target among fast-moving distractors. Top row shows quiver plots of the stimuli (the direction of motion is specified by the arrow whose length indicates the speed), and bottom row the corresponding saliency maps.

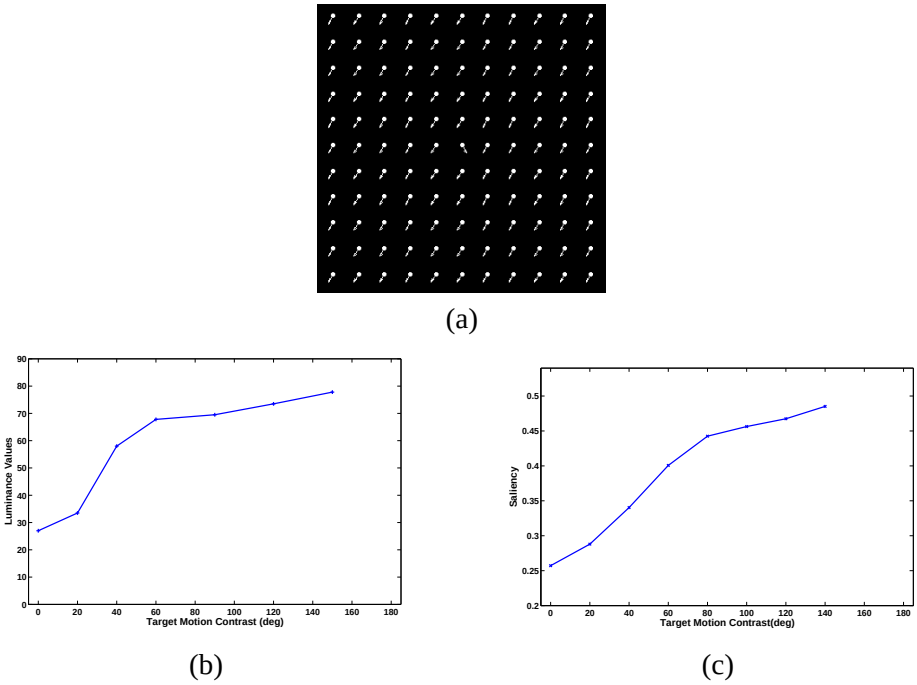
the perception of saliency of moving targets increases nonlinearly with motion contrast, and shows significant saturation and threshold effects. To test the compliance of discriminant saliency with this nonlinearity, we applied it to the motion displays of Nothdurft (1993). An example is shown in plot (a) of Figure 9, where (b) shows a plot of the human saliency data, reproduced from the original figure of Nothdurft (1993), and (c) presents the predictions made by discriminant saliency. The two plots are very similar, both exhibiting threshold and saturation effects.

Applications in computer vision

The ability of discriminant saliency to make accurate predictions of the psychophysics of human saliency, for both static and motion stimuli, encouraged us to examine its performance as a solution for computer vision problems. We considered the problems of predicting human eye fixations, detecting salient moving objects in the presence of ego-motion, and background subtraction from highly dynamic scenes. In all cases, the output of the discriminant saliency detector was compared to either human performance, or state-of-the-art solutions from the computer vision literature.

Prediction of eye fixations on natural scenes

We started by testing the ability of the static discriminant saliency detector to predict the location of human eye fixations. For this, we compared the discriminant saliency maps obtained from a collection of natural images to the eye fixation locations recorded from human subjects, in a free-viewing task. The experimental set-up and protocol closely followed that of Bruce and Tsotsos (2006), using eye fixations from 20 subjects and 120 different natural color images, depicting urban



(a)

(b)

(c)

Figure 9. The nonlinearity of human saliency responses to motion contrast (reproduced from Figure 9 of Nothdurft, 1993) (b) is replicated by discriminant saliency (c). A quiver plot of one instance of the motion display used in the experiment (with background contrast (bg)=0, target contrast (tg)=60) is illustrated in (a). The direction of motion is specified by the arrow, whose length indicates the speed.

scenes (both indoor and outdoor). All images were presented, in random order, to each subject for 4 seconds, with a mask inserted between consecutive presentations. Subjects were given no instructions, and there were no pre-defined initial fixations. The distance between subjects and a $21''$ CRT monitor was set to 75cm, and a standard non-head-mounted gaze tracking device was applied to record the eye movements.

Saliency maps were first quantized into a binary mask that classified each image location as either a fixation or non-fixation (Tatler, Baddeley, & Gilchrist, 2005). Using the measured human fixations as ground truth, a receiver operator characteristic (ROC) curve was generated by varying the quantization threshold. Perfect prediction corresponds to an ROC area (area under the ROC curve) of 1, while chance performance occurs at an area of 0.5. Since the metric makes use of all saliency information in both the human fixations and the saliency detector output, it has been adopted in various recent studies (Bruce & Tsotsos, 2006; Harel et al., 2007; Kienzle et al., 2007). The predictions of discriminant saliency were compared to those of the methods of Itti and Koch (2000) and Bruce and Tsotsos (2006). Table 1 presents average ROC areas for all detectors, across the entire image set[4], as well as the "inter-subject" ROC area. This measures the consistency of the fixations between human subjects, as proposed in (Harel et al., 2007): for each subject, a "human saliency map" is derived from the fixations of all other subjects, by convolving these fixations with

[4]It should be noted that the results of Itti and Koch (2000) and Bruce and Tsotsos (2006), for both this table and all subsequent figures, were optimized for this particular image set, by tuning of model parameters. This was not done for discriminant saliency, whose results were produced with the parameter settings of the previous section.

Saliency model	Discriminant	Itti & Koch (2000)	Bruce & Tsotsos (2006)	Inter-subject
ROC area	0.7694	0.7287	0.7547	0.8766

Table 1: ROC areas for different saliency models with respect to all human fixations.

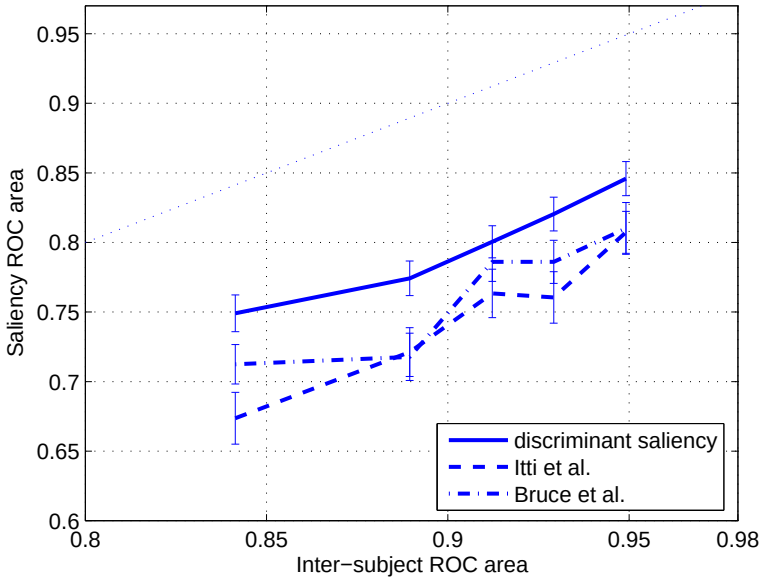


Figure 10. Average ROC area, as a function of inter-subject ROC area, for the saliency algorithms discussed in the text.

a circular, decaying, kernel (2-D Gaussian) of properly selected decay constant (which matches the decreasing density of photoreceptors in the retina). The "inter-subject" ROC area is measured by comparing the fixations of the subject to this saliency map, and averaging across subjects and images. It is clear that discriminant saliency has the best performance among the three saliency detectors.

Nevertheless, there is still a non-negligible gap to human performance. In an attempt to characterize this gap, we studied in greater detail the relationship between saliency maps and subjects' fixations. We focused on the first two fixations which, as suggested by Tatler et al. (2005), are more likely to be driven by bottom-up mechanisms than later fixations. Figure 10 presents the ROC areas of the three detectors as a function of the "inter-subject" ROC area, for these fixations. Again, discriminant saliency exhibited the strongest correlation with human performance, at all levels of inter-subject consistency. Furthermore, the gains of discriminant saliency were largest when inter-subject consistency was strongest. In this region, the performance of discriminant saliency (0.85) was close to 90% of that of humans (0.95), while the other two detectors only achieved close to 85% (0.81).

Discriminant saliency on motion fields

Motion saliency is of importance for various computer vision applications. For example, a robot could benefit from a motion saliency module to identify objects approaching it. However,

Figure 11. Saliency in the presence of ego-motion. (a) representative frames from a video sequence shot with a moving camera, (b) the saliency map produced by the motion-based discriminant saliency detector, and (c) the "surprise" maps by the model of Itti & Baldi (2005).

motion saliency is not trivial to implement when there is ego-motion. If the robot is moving itself, the optical flow due to the moving objects is easily confounded with that originated by background variation due to the robot's motion. This is illustrated by Figure 11, which shows several frames (top row) from a video sequence shot with a moving camera. The sequence depicts a leopard running in a grassland. The camera motion introduces significant variability in the background, making the detection of foreground motion (the leopard) a difficult task. This can be confirmed by analyzing the saliency predictions of algorithms previously proposed in the literature. One example is the "surprise" model of Itti and Baldi (2005) [5]. Although it is one of the best saliency detectors for these types of sequences, the "surprise" maps generated by this algorithm (bottom row of the figure) frequently assign more saliency to the motion of the background than to that of the leopard.

The saliency maps produced by motion-based discriminant saliency are shown in the middle row of the figure. They are clearly superior to those produced by the surprise model, disregarding the background and concentrating all saliency on the animal's body. This example shows that motion-based discriminant saliency is very robust to the presence of ego-motion. This is due to the fact that discriminant saliency is based on a measure of motion contrast. While there is variability in the background optical flow (due to a combination of camera motion and a mostly static scene) this is usually much smaller than the variability of the object's optical flow (especially for non-rigid objects). Hence, the object region has larger motion contrast and is deemed more salient. This is similar to the grouping examples of Figure 1, where feature contrast plays an important role in grouping and segmentation percepts.

Discriminant Saliency for dynamic scenes

One further source of complexity is the possibility that the scene is itself dynamic, e.g. a background consisting of water waves, or tree leaves moving with the wind. In this case, the variability of background optical flow can be larger than that of the object optical flow, for any object. This problem is so complex that, even though background subtraction is a classic problem in computer

[5] Results were generated using the iLab Neuromorphic Vision Toolkit available from http://ilab.usc.edu/toolkit

vision, there has been relatively little progress for these types of scenes (e.g., see Sheikh and Shah (2005) for a review). In order to capture the motion patterns characteristic of these backgrounds, it is necessary to rely on reasonably sophisticated probabilistic models, such as the dynamic texture model (Doretto, Chiuso, Wu, & Soatto, 2003). A dynamic texture (DT) is an autoregressive, generative model for video. It models the spatial component of the video and the underlying temporal dynamics as two stochastic processes. The video is represented as a time-evolving state process $x_t \in \mathbb{R}^n$, and the appearance of a frame $y_t \in \mathbb{R}^m$ is a linear function of the current state vector and observation noise. The system equations are

$$\begin{aligned} x_t &= Ax_{t-1} + v_t \\ y_t &= Cx_t + w_t \end{aligned} \tag{7}$$

where $A \in \mathbb{R}^{n \times n}$ is the state transition matrix, $C \in \mathbb{R}^{m \times n}$ is the observation matrix. The state and observation noise are given by $v_t \sim_{iid} \mathcal{N}(0, Q,)$ and $w_t \sim_{iid} \mathcal{N}(0, R)$, respectively. Finally, the initial condition is distributed as $x_1 \sim \mathcal{N}(\mu, S)$.

Due to the probabilistic nature of the dynamic texture model, it can be easily incorporated on a center-surround discriminant saliency detector. Given a sequence of images, the parameters of the dynamic texture are learned for the center and surround regions at each image location, using algorithms discussed by Doretto et al. (2003), and Chan and Vasconcelos (In Press). Saliency is then computed with the mutual information of (2), using as $P_{X(l),Y(l)}(x, c)$ the probabilistic representation of the center and surround linear dynamic systems. In this case, the discriminant saliency measure becomes a measure of contrast between the compliance of the center and surround regions with the dynamic texture assumption. Since this assumption tends to be accurate for dynamic natural scenes, but not necessarily for objects, the result is a background subtraction algorithm applicable to complex dynamic scenes.

This can be seen in Figures 13-15, which depict the saliency maps produced by the dynamic texture-based discriminant saliency (DTDS) detector for three video sequences. The first (Water-Bottle from Zhong and Sclaroff (2003)) depicts a bottle floating in water in rain, and is shown in Figure 13 (a). The second sequence, Surfer, containing a surfer moving in water, is shown in 14 (a). This sequence is more challenging, as the water surface displays a lower frequency sweeping wave interspersed with high frequency components due to turbulent wakes (created by the surfer, and crest of the sweeping wave). The third, Cyclists (Figure 15 (a)), shows a pair of cyclists moving across a field. The resolution of the clip is poor, and there is considerable background movement, making it difficult to extract the foreground reliably. We compared the output of the DTDS detector with a state-of-the-art background subtraction algorithm from computer vision, based on a Gaussian mixture model (GMM) (Stauffer & Grimson, 1999; Zivkovic, 2004), as well as the "surprise" model (Itti & Baldi, 2005).

Figures 13 (b), 14 (b), and 15 (b) show the saliency maps produced by discriminant saliency detector, DTDS, for the three sequences. The DTDS detector performs well in all cases, detecting the foreground objects while ignoring the movement in the background. As can be seen in Figures 13 (c) and (d), Figures 14 (c) and (d), and Figures 15 (c) and (d), the foreground detection of the other methods is very noisy, and cannot adapt to the highly dynamic nature of the background. The "surprise" maps of the early frames are especially noisy, since a training phase is required to learn the model parameters, a limitation that does not affect DTDS. Highly stochastic spatiotemporal stimuli, such as the sweeping wave crest or the very fast moving background field, create serious difficulties to both the GMM and the surprise detector. Unlike the saliency maps of

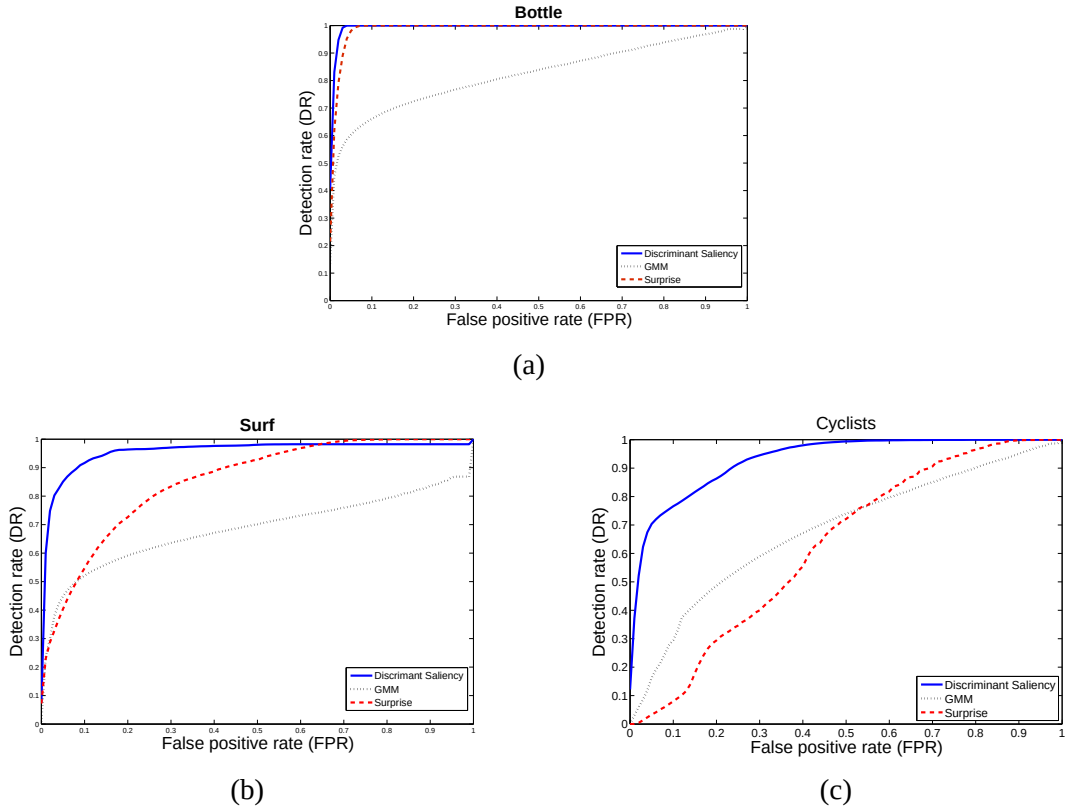


Figure 12. Performance of background subtraction algorithms on: (a) Water-Bottle, (b)Surfer, and (c) Cyclists.

DTDS, the resulting saliency maps contain substantial energy in regions of the background, sometimes completely missing the foreground objects. These saliency maps would be difficult to analyze by subsequent vision (e.g. object tracking) modules. To produce a quantitative comparison of the saliency maps, these were thresholded at a range of values. The results were compared with manually annotated ground-truth foreground masks, and an ROC curve produced for each algorithm. The results are shown in Figure 12. DTDS clearly outperforms both the GMM based background model and the "surprise" model.

Conclusion

In this work, we have evaluated the plausibility of a recently proposed hypothesis for bottom-up saliency: that it is the result of optimal decision making, under constraints of computational parsimony. It was shown that this hypothesis can be applied to various stimulus modalities, and optimal saliency detectors were derived for color, orientation, and motion. These detectors were shown to replicate quantitative psychophysics aspects of human saliency, for both static and moving stimuli. Application of the detectors to problems of interest in computer vision, including the prediction of human eye fixations on natural scenes, motion-based saliency in the presence of ego-motion, and background subtraction in highly dynamic scenes, also revealed better performance than existing solutions to these problems.

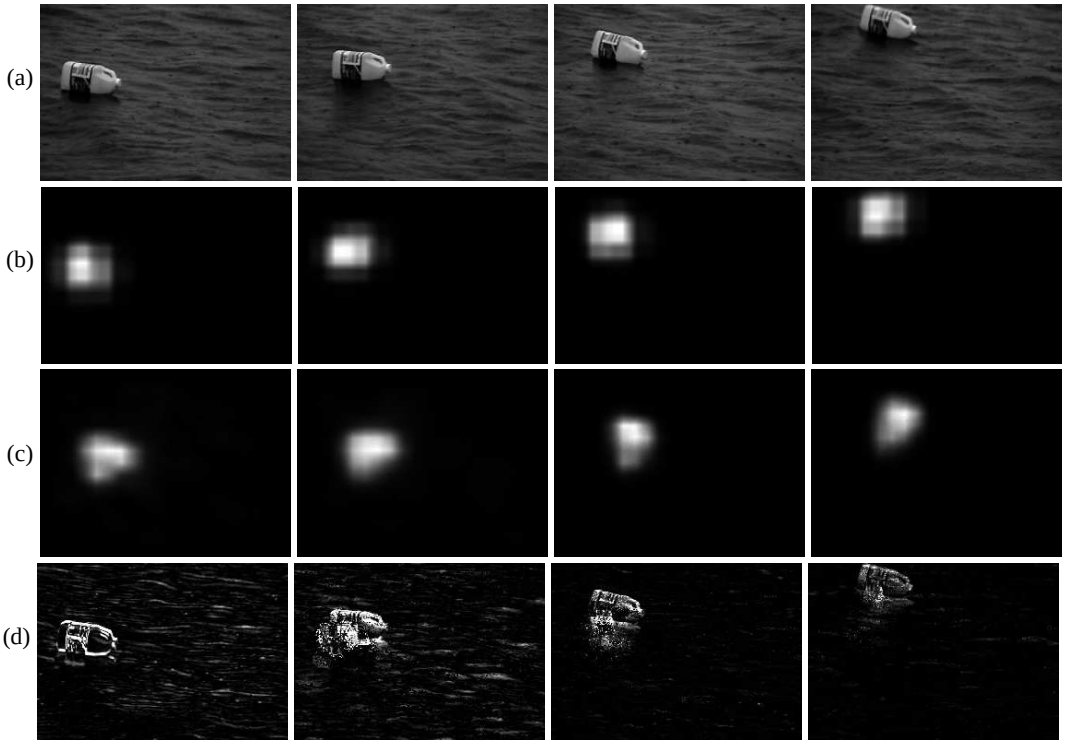


Figure 13. Results on Bottle: (a) original; (b) DTDS; (c) surprise; and (d) GMM model.

Acknowledgement

The authors thank Neil Bruce for kindly sharing the eye fixation data, and saliency predictions of (Bruce & Tsotsos, 2006). This research was supported by NSF awards IIS-0448609, and IIS-0534985.

References

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

Beck, J. (1966a). Effect of orientation and of shape similarity on perceptual grouping. *Perception & Psychophysics*, *1*, 300-302.

Beck, J. (1966b). Perceptual grouping produced by changes in orientation and shape. *Science*, *154*, 538-540.

Beck, J. (1972). Similarity grouping and peripheral discriminability under uncertainty. *American Journal of Psychology*, *85*, 1-19.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in neural information processing systems 18* (p. 155-162). Cambridge, MA: MIT Press.

Buccigrossi, R., & Simoncelli, E. (1999). Image compression via joint statistical characterization in the wavelet domain. *IEEE Transactions on Image Processing*, *8*, 1688-1701.

Cavanaugh, J., Bair, W., & Movshon, J. (2002). Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *J. Neurophysiol.*, *88*, 2530–2546.

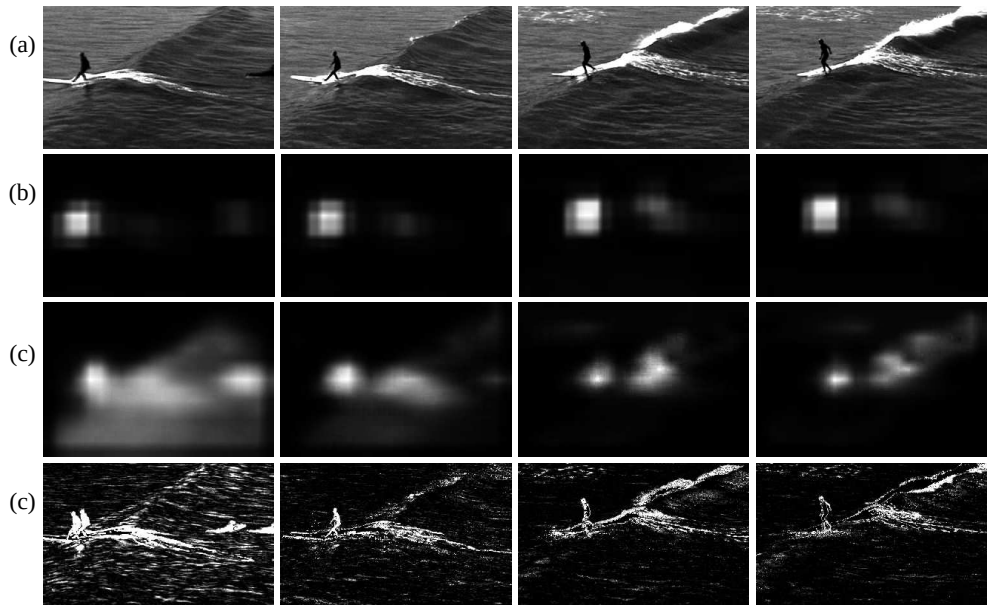


Figure 14. Results on Surfer: (a) original; (b) DTDS; (c) surprise; and (d) GMM model.

Chan, A. B., & Vasconcelos, N. (In Press). Modeling, clustering, and segmenting video with mixtures of dynamic textures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.

Clarke, R. (1985). *Transform coding of images*. Academic Press.

Do, M. N., & Vetterli, M. (2002). Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance. *IEEE Trans. on Image Processing*, *11*(2), 146-158.

Doretto, G., Chiuso, A., Wu, Y. N., & Soatto, S. (2003). Dynamic textures. *International Journal of Computer Vision*, *51*(2), 91-109.

Foster, D. H., & Ward, P. A. (1991). Asymmetries in oriented-line detection indicate two orthogonal filters in early vision. In *Proceedings: Biological sciences* (Vol. 243, p. 75-81).

Gao, D., & Vasconcelos, N. (2005). Discriminant saliency for visual recognition from cluttered scenes. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems 17* (p. 481-488). Cambridge, MA: MIT Press.

Gao, D., & Vasconcelos, N. (2007). Decision-theoretic saliency: computational principle, biological plausibility, and implications for neurophysiology and psychophysics. submitted to *Neural Computation*.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (pp. 545–552). Cambridge, MA: MIT Press.

Heeger, D. (1988). Optical flow from spatiotemporal filters. *International Journal of Computer Vision*, *1*(4), 279-302.

Huang, J., & Mumford, D. (1999). Statistics of Natural Images and Models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (p. 541-547).

Hubel, D. H., & Wiesel, T. N. (1965). Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophysiol.*, *28*, 229-289.

Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing*, *13*(10), 1304-1318.

Itti, L., & Baldi, P. (2005). A principled approach to detecting surprising events in video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (p. 631-637). San Diego, CA.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention.

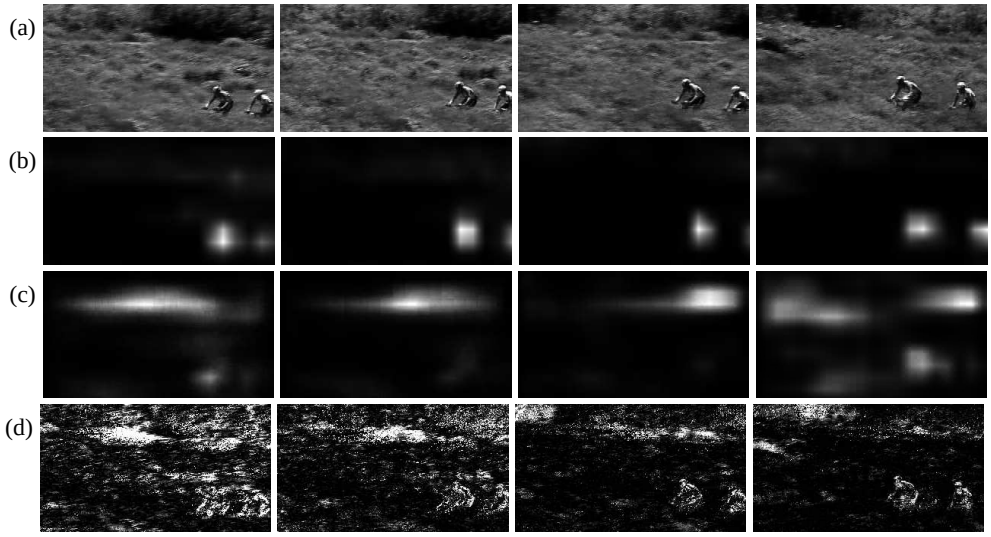


Figure 15. Results on Cyclists: (a) original; (b) DTDS; (c) surprise; and (d) GMM model.

Vision Research, *40*, 1489-1506.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention,. *Nature Rev. Neurosci.*, *2*(3), 194-203.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254-1259.

Ivry, R. B., & Cohen, A. (1992). Asymmetry in visual search for targets defined by differences in movement speed. *J Exp Psychol Hum Percept Perform*, *18*, 1045-1057.

Julesz, B. (1975). Experiments in the visual perception of texture. *Scientific American*, *232*(4), 34-43.

Julesz, B. (1981). A theory of preattentive texture discrimination based on first order statistics of textons. *Biology and Cybernetics*, *41*, 131-138.

Julesz, B. (1984). A brief outline of the texton theory of human vision. *Trends in Neuroscience*, *7*, 41-45.

Kienzle, W., Wichmann, F. A., Schölkopf, B., & Franz, M. O. (2007). A nonparametric approach to bottom-up visual saliency. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing systems 19* (pp. 689–696). Cambridge, MA: MIT Press.

Knierim, J. J., & Van Essen, D. C. (1992). Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *Journal of Neurophysiology*, *67*(4), 961-980.

Koch, C., & Ullman, S. (1985). Shift in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.*, *4*, 219-227.

Landy, M. S., & Bergen, J. R. (1991). Texture segregation and orientation gradient. *Vision Research*, *31*, 679-691.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, *6*(1), 9-16.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 674-693.

Modestino, J. W. (1977). Adaptive nonparametric detection techniques. In P. Papantoni-Kazakos & D. Kazakos (Eds.), *Nonparametric methods in communications* (p. 29-65). New York: Marcel Dekker.

Moraglia, G. (1989). Display organization and the detection of horizontal line segments. *Perception and psychophysics*, *45*, 265-272.

Motoyoshi, I., & nishida, S. (2001). Visual response saturation to orientation contrast in the perception of texture boundary. *Journal of the Optical Society of America A*, *18*(9), 2209-2219.

Nothdurft, H. C. (1991a). The role of local contrast in pop-out of orientation, motion and color. *Investigative Ophthalmology and Visual Science*, *32*(4), 714.

Nothdurft, H. C. (1991b). Texture segmentation and pop-out from orientation contrast. *Vision Research*, *31*(6), 1073-1078.

Nothdurft, H. C. (1992). Feature analysis and the role of similarity in preattentive vision. *Perception and Psychophysics*, *52*(4), 355-375.

Nothdurft, H. C. (1993). The conspicuousness of orientation and motion contrast. *Spatial Vision*, *7*, 341-363.

Nothdurft, H. C. (2000). Salience from feature contrast: variations with texture density. *Vision Research*, *40*, 3181-3200.

Olson, R. K., & Attneave, F. (1970). What variables produce similarity grouping? *American Journal of Psychology*, *83*, 1-21.

Parkhurst, D. J., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*(1), 107-123.

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

Regan, D. (1995). Orientation discriminant for bars defined by orientation texture. *Perception*, *24*, 1131-1138.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Research*, *39*, 3157-3163.

Sagi, D., & Julesz, B. (1985). "where" and "what" in vision. *Science*, *228*, 1217-1219.

Sheikh, Y., & Shah, M. (2005). Bayesian modeling of dynamic scenes for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(11), 1778-92.

Shic, F., & Scassellati, B. (2007). A behavioral analysis of computational models of visual attention. *Journal International Journal of Computer Vision*, *73*, 159-177.

Stauffer, C., & Grimson, W. (1999). Adaptive background mixture models for real-time tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (p. 246-252).

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, *45*, 643-659.

Treisman, A. (1985). Preattentive processing in vision. *Computer vision, Graphics, & Image Processing*, *31*, 156-177.

Treisman, A., & Gelade, G. (1980). A feature-integratrion theory of attention. *Cognitive Psychology*, *12*(1), 97-136.

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, *95*, 14-58.

Vasconcelos, N. (2003). Feature selection by maximum marginal diversity: optimality and implications for visual recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 1, p. 762-769).

Vasconcelos, N. (2004). Scalable discriminant feature selection for image retrieval and recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 2, p. 770-775).

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

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202-238.

Wolfe, J. M., Friedman-Hill, S. R., Stewart, M. I., & O'Connell, K. M. (1992). The role of categorization in visual search for orientation. *Journal of Experimental Psychology: Human Perception & Performance*, *18*, 34-49.

Zhong, J., & Sclaroff, S. (2003). Segmenting foreground objects from a dynamic textured background via a robust Kalman filter. In *Proc. of IEEE International Conference on Computer Vision* (Vol. 1, p. 44).

Zivkovic, Z. (2004). Improved adaptive Gaussian mixture model for background subtraction. In *Proc. of International Conference on Pattern Recognition* (Vol. 2, p. 28-31).