Decision-Theoretic Saliency: Computational Principles, Biological Plausibility, and Implications for Neurophysiology and Psychophysics

Dashan Gao
dgao@ucsd.edu
Nuno Vasconcelos
nuno@ece.ucsd.edu
Statistical Visual Computing Laboratory, University of California San Diego, La Jolla, CA 92093, U.S.A.

A decision-theoretic formulation of visual saliency, first proposed for top-down processing (object recognition) (Gao & Vasconcelos, 2005a), is extended to the problem of bottom-up saliency. Under this formulation, optimality is defined in the minimum probability of error sense, under a constraint of computational parsimony. The saliency of the visual features at a given location of the visual field is defined as the power of those features to discriminate between the stimulus at the location and a null hypothesis. For bottom-up saliency, this is the set of visual features that surround the location under consideration. Discrimination is defined in an information-theoretic sense and the optimal saliency detector derived for a class of stimuli that complies with known statistical properties of natural images. It is shown that under the assumption that saliency is driven by linear filtering, the optimal detector consists of what is usually referred to as the standard architecture of V1: a cascade of linear filtering, divisive normalization, rectification, and spatial pooling. The optimal detector is also shown to replicate the fundamental properties of the psychophysics of saliency: stimulus pop-out, saliency asymmetries for stimulus presence versus absence, disregard of feature conjunctions, and Weber's law. Finally, it is shown that the optimal saliency architecture can be applied to the solution of generic inference problems. In particular, for the class of stimuli studied, it performs the three fundamental operations of statistical inference: assessment of probabilities, implementation of Bayes decision rule, and feature selection.

1 Introduction

The deployment of visual attention has long been believed to be driven by the interaction of two complementary components (James, 1981): a bottom-up, fast, stimulus-driven mechanism, and a top-down, slower, goal-driven mechanism. While many bottom-up (Koch & Ullman, 1985; Itti, Koch, &

Niebur, 1998; Li, 2002; Privitera & Stark, 2000; Malik & Perona, 1990; Kadir & Brady, 2001; Bruce & Tsotsos, 2006; Lowe, 1999; Sha'ashua & Ullman, 1988; Harris & Stephens, 1988; Heidemann, 2004; Förstner, 1994) and top-down (Gao & Vasconcelos, 2005a; Schiele & Crowley, 1996; Walker, Cootes, & Taylor, 1998; Wolfe, 1994; Fergus, Perona, & Zisserman, 2003; Borenstein & Ullman, 2004; Agarwal & Roth, 2002; Navalpakkam & Itti, 2007) saliency algorithms have been proposed in the computer and biological vision literature, there has been much less progress toward the development of a unified computational theory for the two saliency modes. In fact, little is known in terms of generic principles that could drive the design of both bottom-up and top-down saliency.

One hypothesis that we have been pursuing is that saliency is a discriminant process: salient visual attributes are those that best allow visual systems to decide between different hypotheses regarding the nature of the visual stimuli. We refer to this principle as that of discriminant saliency and proposed a top-down discriminant saliency algorithm for visual recognition problems (Gao & Vasconcelos, 2005a). In the recognition context, salient visual attributes were defined as the features whose response best distinguishes the visual concept (e.g., object) to recognize from all others that may be of possible interest (e.g., the set of all other object classes that compose the recognition problem). The importance of saliency for recognition stems from the fact that it enables learning from highly cluttered imagery, a task easily accomplished by biological vision, but quite difficult for classical computer vision algorithms. Discriminant saliency has so far been successfully applied to the design of object recognition systems, and the resulting saliency detector performs well in the presence of clutter (Gao & Vasconcelos, 2005a, 2005b; Hillel, Hertz, & Weinshall, 2005).

The application domain of the discriminant saliency principle is, however, not restricted to visual recognition or even top-down mechanisms. While top-down problems have the greatest practical interest for computer vision and enable a very objective comparison of different saliency detectors (through the measurement of recognition rates), they are less useful in what concerns determining the biological plausibility of a given saliency principle. This follows from the fact that much more is known about the bottom-up component of biological saliency in terms of both the neural circuits involved and the resulting subject behavior than its top-down counterpart. Saliency has, for example, been thoroughly studied in the visual search literature, where, while there appears to be little disagreement with respect to the bottom-up (or preattentive) component of subject behavior, most theories differ in the explanation of the top-down, or attentive, component (Treisman & Gelade, 1980; Treisman & Sato, 1990; Julesz, 1981; Wolfe, 1994; Verghese, 2001; Navalpakkam & Itti, 2007).

Bottom-up saliency is tightly connected to the the ubiquity of center-surround mechanisms in the early stages of biological vision (Kuffler, 1953; Enroth-Cugell & Robson, 1966; Hubel & Wiesel, 1965; Allman, Miezin, & McGuinness, 1985; Julesz, 1986; Cavanaugh, Bair, & Movshon, 2002;

Knierim & Van Essen, 1992; Nothdurft, 2000). A significant body of psychophysical evidence suggests that an important role of these mechanisms is to detect stimuli that are distinct from the surrounding background. In fact, it has long been established that the simplest visual concepts (e.g., bars) can be highly salient when viewed against a background of similar visual concepts (e.g., other bars) that differ from them only in terms of low-level properties such as color or orientation (Treisman & Gelade, 1980; Wolfe, 1998; Wolfe & Horowitz, 2004; Bravo & Nakayama, 1992; Found & Muller, 1995; Muller, Heller, & Ziegler, 1995; Nothdurft, 1993) (see the displays in Figure 5 for an example). This observation has been widely exploited for the design of saliency models (e.g., Itti et al., 1998) and is plausible under a decision-theoretic formulation of bottom-up saliency, where the background stimulus defines a null hypothesis, and salient visual features are those that best discriminate a foreground stimulus from that null hypothesis. This has motivated us to study the effectiveness of discriminant saliency as a driving principle for bottom-up saliency.

In this work, we report four contributions that resulted from this study. First, we show that the discriminant principle proposed by Gao and Vasconcelos (2005a) can be equally applied to bottom-up saliency, based on center-surround mechanisms. In particular, we propose a bottom-up saliency detector that, like its top-down counterpart (Gao & Vasconcelos, 2005a), is optimal in a decision-theoretic sense. This establishes a unified computational framework for bottom-up and top-down saliency. The second contribution, in the spirit of Attneave, Barlow, and others (Attneave, 1954; Barlow, 1961, 2001), is to show that by exploiting the regularities of the visual world, it is possible to implement the optimal solution in a computationally efficient manner. In particular, we show that under a widely used model of the statistics of natural image features, the generalized gaussian distribution, the optimal detector can be implemented with extreme computational simplicity. The third contribution is to show that discriminant saliency, under the constraint of computational parsimony, is biologically plausible. This is done in two ways. First, with respect to neurophysiology, it is shown that under the assumption that saliency is driven by linear filtering, there exists a one-to-one mapping between the optimal detector, in the decision-theoretic sense, and what is usually referred to as the standard architecture of V1: a cascade of linear filtering, divisive normalization, a quadratic nonlinearity, and spatial pooling (Carandini et al., 2005). Second, with respect to psychophysics, it is shown that the optimal solution, in the decision-theoretic sense, replicates the fundamental properties of preattentive saliency in visual search experiments: stimulus pop-out, saliency asymmetries, disregard of feature conjunctions, and Weber's law.

The combination of these three contributions provides a holistic functional justification for the standard architecture of V1: that it has the capability to optimally detect salient locations in the visual field, when optimality is defined in a decision-theoretic sense and sensible simplifications are allowed for the sake of computational parsimony. It is obviously not proposed that the whole of V1 is devoted to bottom-up saliency. The fourth contribution is to show that a minor extension of this architecture is capable of solving generic inference problems. More precisely, it is shown that under a minor extension of the currently prevalent simple cell model, V1 cells compute the fundamental operations of statistical inference (assessment of probabilities, implementation of decision rules, and feature selection) for visual processes that comply with the statistics of natural images. The specific computations are determined by the topology of the lateral connections of divisive normalization, and the architecture could, in principle, implement optimal decisions for many perceptual tasks other than saliency.

The letter is organized as follows. Section 2 introduces the decision-theoretic formulation of center-surround saliency and derives the optimal solution under the constraints of computational parsimony and tuning to natural image feature statistics. The biological plausibility of the optimal saliency detector is then discussed in section 3 in terms of both neurophysiology and psychophysics. Section 4 generalizes the decision-theoretic interpretation of V1 to the solution of arbitrary inference problems involving observations that comply with the statistics of natural images. Finally, section 5 provides some historical context for the work and discusses several possibilities for future research. The details of the implementation of the saliency detector used to produce all saliency results presented are provided in the appendix.

2 Discriminant Saliency

Discriminant saliency is rooted on a decision-theoretic interpretation of perception. Under this interpretation, perceptual systems evolve with the goal of producing 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: that the perceptual mechanisms should be as efficient as possible.

2.1 The Discriminant Hypothesis. Compatibility with decision-theoretic perception is possible if saliency is defined with respect to a null hypothesis, composed of stimuli that are not salient. Once this null hypothesis is available, the locations of the visual field that can be classified, with lowest expected probability of error, as not belonging to it are denoted salient. Mathematically, this is accomplished by (1) defining a binary classification problem that opposes the stimulus at the location to the null hypothesis, (2) finding the visual features that are most discriminating for this problem, and (3) equating the saliency of the location to the discriminant power of these features.

This definition has at least two interesting properties. First, it is applicable to a broad range of saliency problems. For example, different specifications of the null hypothesis enable its specialization to both top-down and bottom-up saliency. Second, the search for discriminant features is a well-defined and computationally tractable, problem that has been widely studied in the decision-theory literature. In previous work (Gao & Vasconcelos, 2005a), we have exploited these properties to derive an optimal top-down discriminant saliency detector for object recognition. In that case, the null hypothesis was the set of all objects other than that to recognize, and the salient features are those that best discriminate between the object of interest and this null set. In this work, we consider the problem of bottom-up, or preattentive, saliency.

2.2 Discriminant Bottom-Up Saliency. As was the case for the top-down pathway, we assume that bottom-up saliency is driven by linear filtering. The visual stimulus is first linearly decomposed into a set of feature responses and the saliency of each location inferred from a sample of these responses. We hypothesize that the goal of the preattentive visual system is to optimally drive the deployment of attention and that, in the absence of task-specific objectives, this reduces the saliency of each location to how distinct it is from the surrounding background. In decision-theoretic terms, it corresponds to (1) identifying the null hypothesis for the saliency of a location with the set of feature responses that surround it and (2) defining bottom-up saliency as optimal discrimination between the responses at location and surround. Optimality is defined in a minimum probability of error sense: the most salient locations are those that, based on the observed feature responses, can be discriminated from their surround with minimum expected probability of error. These are the locations that can be most confidently declared as different from surround by an ideal observer.[1]

The discriminant hypothesis can be seen as a rigorous mathematical formulation for more informal hypotheses that are frequently used in the literature. These include equating saliency to locations that are most different from surround or for which it is easiest to decide that stimuli at location and surround are different. The latter is, for example, subjacent to the visual search paradigm widely used in the study of human saliency. Decision theory offers a precise quantification of "easiest," equating it to "with smallest expected probability of error."

Mathematically, discriminant power is measured by introducing two windows, $\mathcal{W}_l^0$ and $\mathcal{W}_l^1$, at each location l of the visual field. $\mathcal{W}_l^1$ is an inner

[1] It is worth noting that this "ideal observer" is similar to the "ideal searcher" of Najemnik and Geisler (2005), in the sense that both are optimal under Bayes decision theory and use precise knowledge about the statistics of natural scenes. However, their tasks are fundamentally different: while the ideal searcher seeks a specific target in the visual field, which requires top-down guidance and an active search strategy to direct eye movements, the ideal observer for saliency seeks to identify the locations of the visual field that can be most confidently declared as different from surround. This is a strictly bottom-up, stimulus-driven, process that does not require a search strategy.

window that accounts for a center neighborhood and $\mathcal{W}_l^0$ an outer annulus that defines its surround. The responses of a predefined set of d features, henceforth referred to as *feature vectors*,[2] are measured at all image locations within the two windows and interpreted as observations drawn from a random process $\mathbf{X}(l) = (X_1(l), \ldots, X_d(l))$, of dimension d, conditioned on the state of a binary class label $Y(l) \in \{0, 1\}$. The feature vector observed at location j is denoted by $\mathbf{x}(j) = (x_1(j), \ldots, x_d(j))$, and feature vectors are independently drawn from the class-conditional probability densities $P_{\mathbf{X}(l)|Y(l)}(\mathbf{x} \mid i)$. Learning is supervised, in the sense that the assignment of feature vectors to classes is known: $\mathbf{x}(j)$ is drawn from class $Y(l) = 1$ when $j \in \mathcal{W}_l^1$ and from class $Y(l) = 0$ when $j \in \mathcal{W}_l^0$. For this reason, class $Y(l) = 1$ is denoted as the center class and class $Y(l) = 0$ as the surround class. The saliency at location l is quantified by the discriminant power of the features 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. Discriminant power is measured by the mutual information between features and class label:

$$I_l(\mathbf{X}; Y) = \sum_c \int p_{\mathbf{X}(l),Y(l)}(\mathbf{x}, c) \log \frac{p_{\mathbf{X}(l),Y(l)}(\mathbf{x}, c)}{p_{\mathbf{X}(l)}(\mathbf{x}) p_{Y(l)}(c)} \, d\mathbf{x}. \tag{2.1}$$

The overall computation is summarized by Figure 1. The l subscript emphasizes the fact that the mutual information is defined locally, within $\mathcal{W}_l$, and saliency detection consists of identifying the locations where equation 2.1 is maximal. These are the most informative locations with respect to the discrimination between center and surround. The overall process can be seen as a discriminant extension of the infomax principle of perceptual organization (Linsker, 1988) and draws on a long tradition of information-theoretic formulations for various levels of perception (Attneave, 1954; Watanabe, 1960; Barlow, 1961). From a purely information-theoretic perspective, it can also be interpreted as modeling brains as noisy communication channels, with features $\mathbf{X}$ as inputs and decision Y as output and the optimal neural representation as that which achieves channel capacity.

2.3 Computational Parsimony. The exact maximization of equation 2.1 is usually impractical, since it requires density estimates on a potentially high-dimensional feature space. The discriminant hypothesis is complemented by one of computational parsimony, which advises the search for

[2] We adopt the standard notation in machine learning, where a *feature* is one dimension of the space where a classification problem is defined, and *feature responses* or *feature vectors* are the observed sample points in that space. For example, the basis function associated with a Gabor filter is a feature, the convolution of the image with that filter produces a feature response at each image location, and the vector of responses of a set of features at a given location is a *feature vector*. Note that this is different from notation frequently used in psychophysics, where what we refer to a *feature* is denoted by *dimension* and what we refer to as *feature vector* is denoted by *feature*.

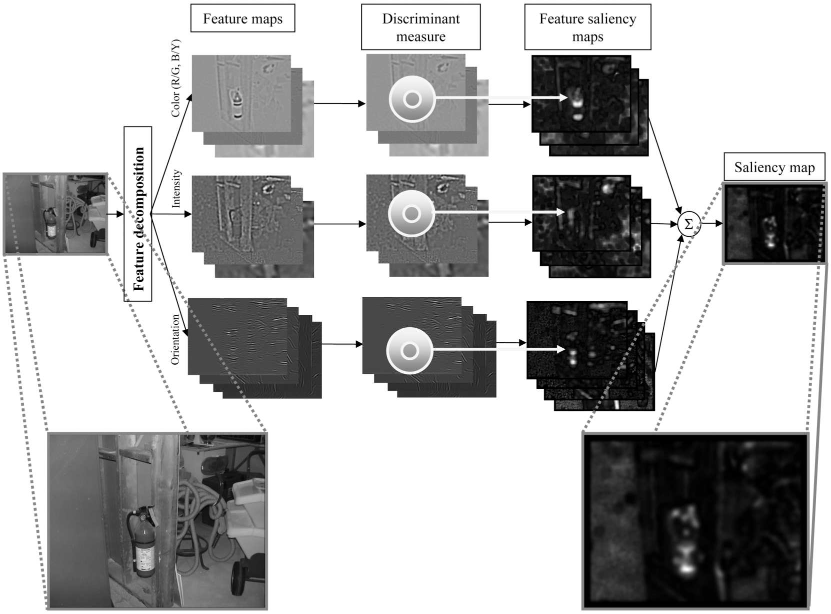

Figure 1: Decision-theoretic saliency. The visual field is projected into feature maps that account for color, intensity, orientation, scale, and so forth. Center and surround windows are then analyzed at each location to infer the discriminant power of each feature at that location. Feature saliency is defined as the power to discriminate between center and surround. Overall saliency is defined as the discriminant power of the entire feature set and (for natural scenes) can be approximated by the sum of all feature saliencies. (A color version of this figure is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/neco.2008.11-06-391.)

approximations that enable efficient computation. This can be achieved by exploiting a known property of the statistics of bandpass natural image features (e.g., Gabor or wavelet coefficients): that features in this class exhibit strongly consistent patterns of dependence across a wide range of imagery (Buccigrossi & Simoncelli, 1999; Huang & Mumford, 1999; Srivastava, Lee, Simoncelli, & Zhu, 2003; M. Vasconcelos & Vasconcelos, in press). These regularities are illustrated by Figure 2, which presents three images, the histograms of one coefficient of their wavelet decomposition, and the histograms of that coefficient conditioned on its parent. Although the drastically different visual appearance of the images affects the scale (variance) of the marginal distributions, their shape, or that of the conditional distributions between coefficients, is quite stable. The observation that these distributions follow a canonical (bow-tie) pattern, which is simply rescaled to match the marginal statistics of each image,

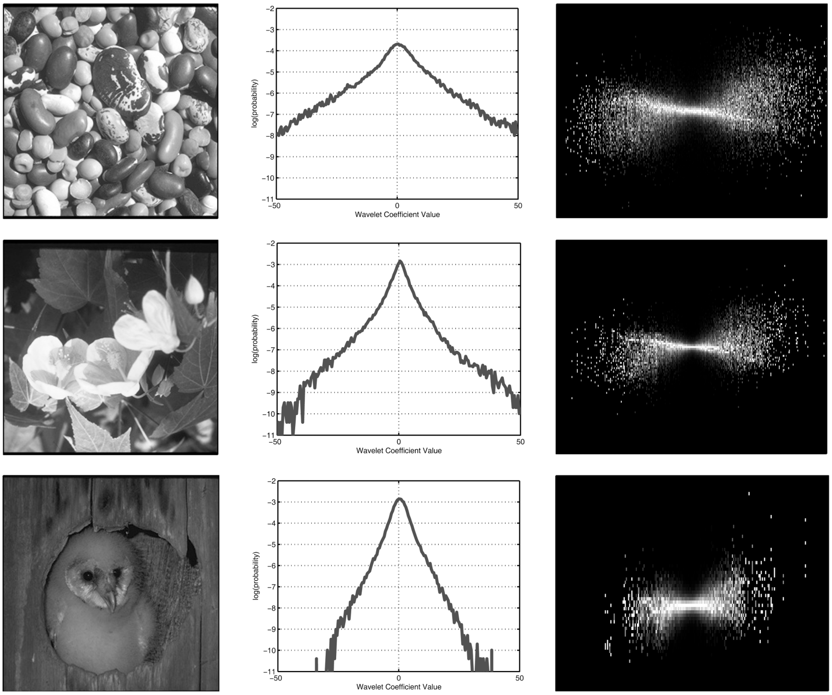

Figure 2: Constancy of natural image statistics. (Left) Three images. (Center) Each plot presents the histogram of the same coefficient from a wavelet decomposition of the image on the left. (Right) Conditional histogram of the same coefficient, conditioned on the value of its parent. Note the constancy of the shape of both the marginal and conditional distributions across image classes.

is remarkably consistent over the set of natural images. This consistency indicates that even though the fine details of feature dependence may vary from scene to scene, the coarse structure of such dependencies follows a universal statistical law that appears to hold for all natural scenes. This in turn suggests that feature dependencies are not greatly informative about the image class or, in the particular case of saliency, about whether observations originate in the center or surround. The following theorem shows that, when this is the case, equation 2.1 can be drastically simplified.

Theorem 1. *Let $\mathbf{X} = \{X_1, \ldots, X_d\}$ be a collection of features and Y the class label. If*

$$\frac{\sum_{i=1}^{d} [I(X_i; \mathbf{X}_{1,i-1}) - I(X_i; \mathbf{X}_{1,i-1} \mid Y)]}{\sum_{i=1}^{d} I(X_i; Y)} = 0, \tag{2.2}$$

where $\mathbf{X}_{1,i} = \{X_1, \ldots, X_i\}$*, then*

$$I(\mathbf{X}; Y) = \sum_{i=1}^{d} I(X_i; Y). \tag{2.3}$$

Proof. From the chain rule of mutual information (Cover & Thomas, 1991),

$$I(\mathbf{X}, Y) = \sum_{i=1}^{d} I(X_i; Y \mid \mathbf{X}_{1,i-1}).$$

Using the equality

$$\begin{aligned}
I(X; Y \mid \mathbf{Z}) &= E_{X,Y,\mathbf{Z}}\left[\log \frac{P_{X,Y|\mathbf{Z}}(x, y \mid \mathbf{z})}{P_{X|\mathbf{Z}}(x \mid \mathbf{z}) P_{Y|\mathbf{Z}}(y \mid \mathbf{z})}\right] \\
&= E_{X,Y,\mathbf{Z}}\left[\log \frac{P_{X,Y}(x, y)}{P_X(x) P_Y(y)} + \log \frac{P_{X,Y|\mathbf{Z}}(x, y \mid \mathbf{z}) P_Y(y)}{P_{X,Y}(x, y) P_{Y|\mathbf{Z}}(y \mid \mathbf{z})}\right. \\
&\quad \left. + \log \frac{P_X(x)}{P_{X|\mathbf{Z}}(x \mid \mathbf{z})}\right] \\
&= I(X; Y) + E_{X,Y,\mathbf{Z}}\left[\log \frac{P_{X|Y,\mathbf{Z}}(x \mid y, \mathbf{z})}{P_{X|Y}(x \mid y)}\right] - I(X; \mathbf{Z}) \\
&= I(X; Y) + E_{X,Y,\mathbf{Z}}\left[\log \frac{P_{X,\mathbf{Z}|Y}(x, \mathbf{z} \mid y)}{P_{X|Y}(x \mid y) P_{\mathbf{Z}|Y}(\mathbf{z} \mid y)}\right] - I(X; \mathbf{Z}) \\
&= I(X; Y) + I(X; \mathbf{Z} \mid Y) - I(X; \mathbf{Z})
\end{aligned}$$

with $X = X_i$ and $\mathbf{Z} = \mathbf{X}_{1,i-1}$, leads to

$$\begin{aligned}
I(\mathbf{X}, Y) &= \sum_{i=1}^{d} I(X_i; Y) + \sum_{i=1}^{d} [I(X_i; \mathbf{X}_{1,i-1}|Y) - I(X_i; \mathbf{X}_{1,i-1})] \\
&= \left(\sum_{i=1}^{d} I(X_i; Y)\right)\left(1 - \frac{\sum_{i=1}^{d} [I(X_i; \mathbf{X}_{1,i-1}) - I(X_i; \mathbf{X}_{1,i-1} \mid Y)]}{\sum_{i=1}^{d} I(X_i; Y)}\right).
\end{aligned}$$

The theorem follows.

The left-hand side of equation 2.2 measures the ratio between the information for discrimination contained in feature dependencies and that contained in the features themselves. While this ratio is usually nonzero, it is generally small for bandpass natural image features and smallest in the locations where the features are most discriminant. Hence, the approximation of equation 2.3 is best at the most salient locations, and the approximate definition of saliency,

$$S(l) = \sum_{i=1}^{d} I_l(X_i; Y), \tag{2.4}$$

is a sensible compromise between decision-theoretic optimality and computational parsimony. Note that this approximation does not assume that the features are independently distributed, but simply that their dependencies are not informative about the class. The function $S(l)$ is referred to as the *saliency map*, and salient locations are identified by searching for its local maxima.

2.4 Exploiting the Marginal Statistics of Natural Images. The computation of equation 2.4 requires empirical estimates of the mutual information $I_l(X_i; Y)$. These in turn require estimates of both the marginal probability densities of features X_i and their probability densities conditioned on the class Y. Extensive research on the statistics of natural images has shown that for bandpass features, all these densities are well approximated by generalized gaussian distributions (GGD) (Modestino, 1977; Farvardin & Modestino, 1984; Mallat, 1989; Clarke, 1985; Birney & Fisher, 1995; Do & Vetterli, 2002), of the form

$$P_X(x; \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left\{ -\left(\frac{|x|}{\alpha}\right)^{\beta} \right\}, \tag{2.5}$$

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 rate of decrease from the peak value and defines a subfamily of the GGD (e.g., the Laplacian family when $\beta = 1$ or the gaussian family when $\beta = 2$). As illustrated by the center column of Figure 2, the hallmark of the GGD when $\beta \approx 1$, an approximately linearly decreasing tail of the log-probability density function, is consistently observed for bandpass filter responses to natural images, independent of the image class itself. In this work, we assume that β is known, and the density estimation problem consists of estimating α. When the windows $\mathcal{W}_l^0$ and $\mathcal{W}_l^1$ are small, accurate estimates frequently require some form of regularization, which can be implemented with recourse to Bayesian procedures. The parameter α is considered a random variable and a distribution $P_\alpha(\alpha)$ introduced to account for prior beliefs in its configurations. Conjugate priors are a convenient choice that produces simple estimators, which enforce intuitive regularization. It turns out that for the GGD, it is easier to work with the inverse scale than the scale itself.

Lemma 1. *Let $\theta = \frac{1}{\alpha^\beta}$ be the inverse scale parameter of the GGD. The conjugate prior for θ is a gamma distribution,*

$$P_\theta(\theta) = Gamma\left(\theta, 1 + \frac{\eta}{\beta}, \nu\right) = \frac{\nu^{1+\eta/\beta}}{\Gamma(1+\eta/\beta)}\theta^{\eta/\beta}e^{-\nu\theta}, \tag{2.6}$$

whose shape and scale are controlled by hyperparameters η and ν, respectively. Under this prior, the maximum a posteriori (MAP) probability estimate of α, with respect to a sample $\mathcal{D} = \{x(1), \ldots, x(n)\}$ of independent observations drawn from equation 2.5, is

$$\hat{\alpha}_{MAP} = \left[\frac{1}{\kappa}\left(\sum_{j=1}^{n}|x(j)|^\beta + \nu\right)\right]^{1/\beta}, \tag{2.7}$$

with $\kappa = \frac{n+\eta}{\beta}$.

Proof. The likelihood of the sample $\mathcal{D} = \{x(1), \ldots, x(n)\}$ given θ is

$$P_{X|\theta}(\mathcal{D} \mid \theta) = \Pi_{j=1}^{n} P_{X|\theta}(x(j) \mid \theta) = \left(\frac{\beta\theta^{1/\beta}}{2\Gamma(1/\beta)}\right)^n \exp\left(-\theta\sum_{j=1}^{n} \mid x(j) \mid^\beta\right).$$

For the gamma prior, application of Bayes rule leads to the posterior

$$\begin{aligned} P_{\theta|X}(\theta \mid \mathcal{D}) &= \frac{P_{X|\theta}(\mathcal{D} \mid \theta)P_\theta(\theta)}{\int_\theta P_{X|\theta}(\mathcal{D} \mid \theta)P_\theta(\theta)\mathrm{d}\theta} \\ &= \frac{1}{Z}\theta^{(n+\eta)/\beta}\exp\left(-\left(\sum_{j=1}^{n} \mid x(j) \mid^\beta + \nu\right)\theta\right), \end{aligned}$$

where Z is a normalization constant that does not depend on θ. Since this is a gamma distribution, equation 2.6 is a conjugate prior for θ. Setting the derivative of $\log P_{\theta|X}(\theta \mid \mathcal{D})$ with respect to θ to zero,[3] it follows that the MAP estimate is

$$\hat{\theta}_{MAP} = \frac{n+\eta}{\beta}\left(\sum_{j=1}^{n}|x(j)|^\beta + \nu\right)^{-1}.$$

[3] It can also be shown that the second-order derivative is nonnegative and strictly positive for $\theta > 0$.

Applying the change of variable from θ to α leads to the MAP estimate of α,

$$\hat{\alpha}_{MAP} = \left[\frac{1}{\kappa} \left(\sum_{j=1}^{n} |x(j)|^{\beta} + \nu \right) \right]^{1/\beta}.$$

Note that the MAP estimate $\hat{\alpha}_{MAP}$ is equivalent to the maximum likelihood estimate of α, which would be obtained from an augmented sample with η additional observations of value $(\nu/\eta)^{1/\beta}$. Given this estimate, for each of the classes, estimates of the posterior class probabilities $P_{Y|X}(i \mid x), i \in \{0, 1\}$ can be computed as follows:

Lemma 2. *For a binary classification problem, with generalized gaussian class-conditional distributions $P_{X|Y}(x|c)$ of parameters (α_c, β_c), $c \in \{0, 1\}$, the posterior distribution for class $c = 0$ is*

$$P_{Y|X}(0|x) = s\left[\left(\frac{|x|}{\alpha_1} \right)^{\beta_1} - \left(\frac{|x|}{\alpha_0} \right)^{\beta_0} - K \right], \tag{2.8}$$

where

$$K = \log a + T, \tag{2.9}$$

$$a = \alpha_0/\alpha_1, \tag{2.10}$$

$T = \log(\frac{\beta_1 \pi_1 \Gamma(\frac{1}{\beta_0})}{\beta_0 \pi_0 \Gamma(\frac{1}{\beta_1})})$, $\pi_c = P_Y(c), c \in \{0, 1\}$, are the prior probabilities for the two classes, and $s(x) = (1 + e^{-x})^{-1}$ is a sigmoid.

Proof. Using Bayes rule and equation 2.5,

$$\begin{aligned} P_{Y|X}(0 \mid x) &= \frac{P_{X|Y}(x \mid 0) P_Y(0)}{P_{X|Y}(x \mid 0) P_Y(0) + P_{X|Y}(x \mid 1) P_Y(1)} \\ &= \frac{1}{1 + \frac{P_{X|Y}(x|1) P_Y(1)}{P_{X|Y}(x|0) P_Y(0)}} \\ &= \frac{1}{1 + \frac{\beta_1 \pi_1 \alpha_0 \Gamma\left(\frac{1}{\beta_0}\right)}{\beta_0 \pi_0 \alpha_1 \Gamma\left(\frac{1}{\beta_1}\right)} \frac{\exp\left\{ -\left(\frac{|x|}{\alpha_1}\right)^{\beta_1} \right\}}{\exp\left\{ -\left(\frac{|x|}{\alpha_0}\right)^{\beta_0} \right\}}} \\ &= \frac{1}{1 + \exp\left(\left(\frac{|x|}{\alpha_0}\right)^{\beta_0} - \left(\frac{|x|}{\alpha_1}\right)^{\beta_1} + K \right)}, \end{aligned} \tag{2.11}$$

where $K = \log a + T$, $a = \alpha_0/\alpha_1$, $T = \log(\frac{\beta_1\pi_1\Gamma(\frac{1}{\beta_0})}{\beta_0\pi_0\Gamma(\frac{1}{\beta_1})})$. The lemma follows from the definition of the sigmoid, $s(x) = (1+e^{-x})^{-1}$.

The combination of these two lemmas and some information-theoretic manipulation leads to the desired empirical estimate of the mutual information of equation 2.4:

Theorem 2. *For the binary classification problem with generalized gaussian class-conditional distributions $P_{X|Y}(x|c)$ of parameters (α_c, β_c), $c \in \{0, 1\}$, where β_c is known and α_c estimated, according to equation 2.7, from the center ($c = 1$) and surround ($c = 0$) windows $\mathcal{W}_l^c$ centered at l,*

$$I_l(X;Y) = H(Y) + \frac{1}{|\mathcal{W}_l|}\sum_{j\in\mathcal{W}_l}\phi\{g[x(j)]\}, \tag{2.12}$$

with $H(Y) = -\sum_{c=0}^{1} P_Y(c) log P_Y(c)$ the entropy of the class label, and $\mathcal{W}_l = \mathcal{W}_l^0 \cup \mathcal{W}_l^1$,

$$\phi(x) = s(x)\ log\ s(x) + s(-x)\ log\ s(-x), \tag{2.13}$$

$s(x) = (1+e^{-x})^{-1}$ a sigmoid, and

$$g[x(j)] = \psi[x(j);\Phi_0] - \psi[x(j);\Phi_1] + K_l, \tag{2.14}$$

where

$$\psi[x(j);\Phi_c] = \frac{|x(j)|^{\beta_c}}{\xi_c}, \tag{2.15}$$

$$\xi_c = \frac{1}{\kappa_c}\left(\nu_c + \sum_{k\in\mathcal{W}_l^c}|x(k)|^{\beta_c}\right), \tag{2.16}$$

$\Phi_c = (\kappa_c, \nu_c)^T$ is the vector of prior hyperparameters of class c, as defined in lemma 1, and K_l is given by equation 2.9.

Proof. We start by using some well-known results from information theory (Cover & Thomas, 1991) to rewrite

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y \mid X) \\ &= E_X\left[H(Y) + E_{Y|X}\left[\log P_{Y|X}(c \mid x)\right]\right] \\ &= E_X\left[H(Y) + \sum_{c=0}^{1} P_{Y|X}(c \mid x)\log P_{Y|X}(c \mid x)\right], \end{aligned} \tag{2.17}$$

where $H(Y) = -\sum_{c=0}^{1} P_Y(c)\log P_Y(c)$ is the entropy of Y and $H(Y|X) = -E_{X,Y}\left[\log P_{Y|X}(c|x)\right]$ the conditional entropy of Y given X. Given a location l, the corresponding center $\mathcal{W}_l^1$ and surround $\mathcal{W}_l^0$ windows, and the set of associated feature responses $x(j)$, $j \in \mathcal{W}_l = \mathcal{W}_l^0 \cup \mathcal{W}_l^1$, equation 2.17 can be estimated by replacing expectations with sample means,

$$I_l(X;Y) = \frac{1}{|\mathcal{W}_l|}\sum_{j\in\mathcal{W}_l}\left[H(Y) + \sum_{c=0}^{1} P_{Y|X}(c\,|\,x(j))\log P_{Y|X}(c\,|\,x(j))\right], \quad (2.18)$$

where the outer summation pools all locations in the center and surround windows. Combining with equation 2.8 and using MAP estimates for α_0 and α_1 in equation 2.7 leads, after some algebra, to equations 2.12 to 2.16.

3 Plausibility of Discriminant Saliency

For the values of a_l typical of natural image patches ($a_l \approx 1$), the computations of theorem 2 can be implemented with the network of Figure 3. In this section, we show that this is consistent with a number of well-known properties of the neurophysiology and psychophysics of preattentive vision.

3.1 Neurophysiological Plausibility. Early vision occurs mostly in the primary visual cortex (V1), where cells are usually classified as simple and complex (Hubel & Wiesel, 1962; Skottun et al., 1991; Carandini et al., 2005). Classical studies focused on stimuli incident on the cell's receptive field, and simple cells were modeled as cascades of a linear filter and a rectifying nonlinearity (Movshon, Thompson, & Tolhurst, 1978; Jones & Palmer, 1987). More recently, it has been noted that important properties of simple cell behavior, such as gain control, require an additional stage of divisive normalization of the cell response by that of others (Heeger, 1992; Carandini, Heeger, & Movshon, 1997). Complex cells are frequently modeled as units that pool squared and half-rectified outputs of linear units with similar orientation, the energy model proposed by Adelson & Bergen (1985). We refer to the combination of complex and, divisively normalized, simple cells as the *standard V1 architecture* (Carandini et al., 2005).

It follows from theorem 2 that the optimal solution of discriminant saliency is fully compatible with this architecture. The theorem decomposes the computation of saliency, into three basic operations: equation 2.15 divisively normalizes each feature response by the responses of the feature in the neighborhood $\mathcal{W}_l^c$, equation 2.14 computes the differential between the responses divisively normalized by the center and surround neighborhoods, and equation 2.12 pools this differential response across the window $\mathcal{W}_l$, after application of the nonlinearity of equation 2.13. As shown in Figure 4, this nonlinearity is very close to a hard-limited version

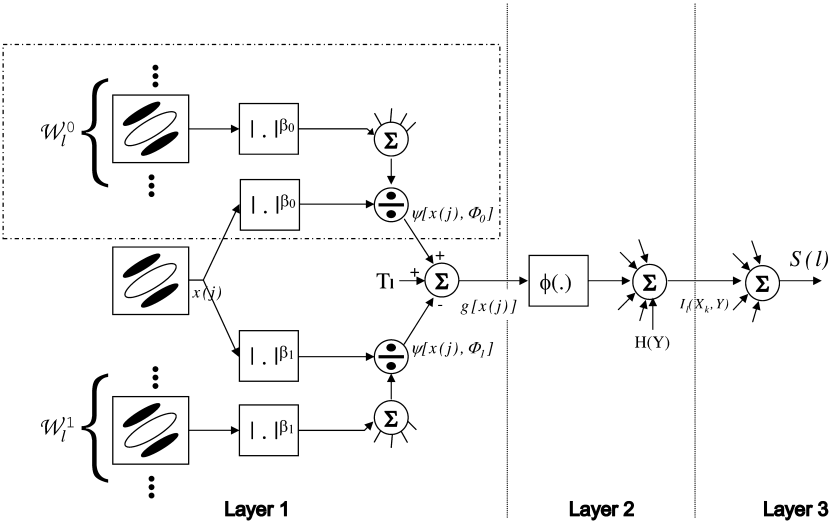

Figure 3: Three-layer saliency network. Layer 1: Simple (Gabor) cells. Cell at location l is rectified and laterally connected to cells in center $\mathcal{W}_l^1$ and surround $\mathcal{W}_l^0$ windows. Lateral connections implement differential divisive normalization. Layer 2: Complex cells. Cell at l pools simple cells in $\mathcal{W}_l$ after rectification by $\phi(x)$. Complex cells maintain the scale and orientation tuning of the pooled simple cells but are location invariant. Layer 3: Pools complex cells associated with each l so as to compute its saliency (see equation 2.4). Suggests organization of complex cells by feature type in cortical columns. Note that the j subscript refers to any location within $\mathcal{W}_l$, and the pooling of layer 2 is across all such locations.

of the quadratic function,

$$\tilde{\phi}(x) = 0.07x^2 - \log(2). \tag{3.1}$$

This quadratic form conforms to the quadratic nonlinearity advocated by the energy model of complex cells[4] (Adelson & Bergen, 1985).

If the step of equation 2.14 is omitted, these are really just the computations of the standard V1 architecture. This is probably best understood by considering the first two layers of the network of Figure 3 and momentarily

[4] Note that $\phi(x)$ is negative simply because we have elected to leave $H(Y)$ as a free parameter, which is added to the pooling stage of layer 2. The sum $H(Y) + \phi\{g[x(j)]\}$ is always in the range $[0, \log(2)]$, because $\phi\{g[x(j)]\} = -H(Y|X(j))$. It follows that $H(Y) + \phi\{g[x(j)]\}$ is well approximated by the hard-limited version of $0.07g^2[x(j)]$; that is, it is always nonnegative and compliant with the energy model of complex cells. In Figure 4, we have assumed that $P_Y(0) = P_Y(1)$, which justifies the $-\log(2)$ factor in equation 3.1.

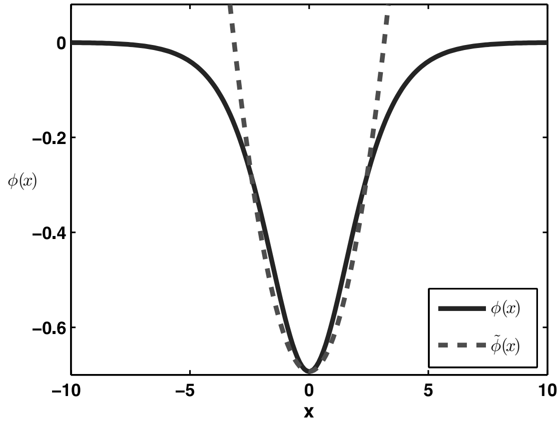

Figure 4: Complex cell nonlinearity. $\phi(x)$ and its approximation by a quadratic function $\tilde{\phi}(x)$.

disregarding the top branch (dashed box), which accounts for the contribution of the surround window $\mathcal{W}_l^0$. The first two layers of the remaining network are exactly the standard V1 architecture: a stage of Gabor simple cells, divisively normalized by the outputs of their peers, subject to rectification by $\phi(\cdot)$ and pooled, in a manner akin to the classical energy model of complex cells. The implementation of the complete network simply requires the replacement of the divisively normalized simple cell by a cell that is differentially divisively normalized by the outputs of the cells in the center and surround. The addition of a third layer, which pools across features, produces the saliency measure of equation 2.4.

3.2 Consistency with Psychophysics. While biological plausibility is interesting, the ultimate test for a saliency model is whether it can explain the psychophysics data available in the saliency literature. To address this question, we start by recalling that there is extensive psychophysical evidence in support of a functional decomposition of vision in two stages: a preattentive (or bottom-up) stage, totally stimulus driven and very efficient, and an attentive stage that requires (top-down) feedback from higher cortical areas. While various, sometimes conflicting, theories have been proposed to explain both stages (Treisman & Gelade, 1980; Julesz, 1981; Wolfe, 1994), there is wide agreement on the fundamental properties of bottom-up processing.

Most can be traced to Treisman's influential feature integration theory (FIT) (Treisman & Gelade, 1980) and the study of visual search (Wolfe, 1998), where subjects are asked to detect target objects embedded in distractor fields. By measuring response time versus display size, it is possible to infer which tasks are solved preattentively. Most theories predict that the visual stimulus is projected into feature maps that encode properties like color, orientation, or motion (Wolfe, 1998; Wolfe & Horowitz, 2004). Feature maps are then combined into a master, or saliency (Koch & Ullman, 1985), map that drives attention, allowing top-down processing to concentrate on a small region of the visual field. The saliency map is scalar and registers only the degree of relevance of each location to the search, not which features are responsible for it. Hence, only target features that are absent in the distractors can be found preattentively, in which case the target pops out (see Figure 5). When target-distractor differences are due to a conjunction of features, the search cannot be resolved, at least without top-down processing (Treisman & Sato, 1990; Wolfe, 1994, 1998). An extensive history of search experiments has produced a thorough quantitative characterization of preattentive vision. In particular, Treisman and colleagues documented the importance of search asymmetries (Treisman & Souther, 1985; Treisman & Gormican, 1988): while the presence in the target of a feature absent from the distractors produces pop-out, the reverse (pop-out due to absence, in the target, of a distractor feature) does not hold. In fact, Treisman and Gormican (1988) showed that in addition to presence or absence, there are asymmetries between weaker and stronger responses, and they presented evidence for the hypothesis that these asymmetries satisfy Weber's law.

All these properties are consistent with discriminant saliency. This is illustrated by Figure 5, which presents the output of an implementation of the network of Figure 3 (whose details are described in the appendix) to classical visual search displays. In particular, the figure shows that the network replicates the human percept of pop-out when target-distractors differ by a single feature, inability to detect target-distractor differences that involve feature conjunctions, and asymmetry of response to permutations of feature presence or absence. In addition to this qualitative evidence, it is also possible to replicate quantitative predictions from psychophysics, such as the compliance of asymmetries with Weber's law. This compliance follows from the fact that up to pooling and rectification, saliency is determined by $g[x(j)]$, in equation 2.14. For sensible sizes of the center and surround windows, this function is dominated by $\psi[x(j); \Phi_0]$, which, from equation 2.15, is (up to the regularization by ν_0, which simply prevents unbounded responses) exactly Treisman's proposal: the normalization of the feature response at j by the distractor response from the surround window. Figure 6 demonstrates Weber's law for the saliency network of Figure 3 by replicating the classic experiment conceived by Treisman to demonstrate the law in the context of visual search (Treisman & Gormican, 1988).

3.3 Implications. We have seen so far that discriminant saliency is biologically plausible and consistent with the psychophysics of saliency. However, the discussion is also interesting in the sense of providing a unified justification for a number of disjoint observations from neurophysiology

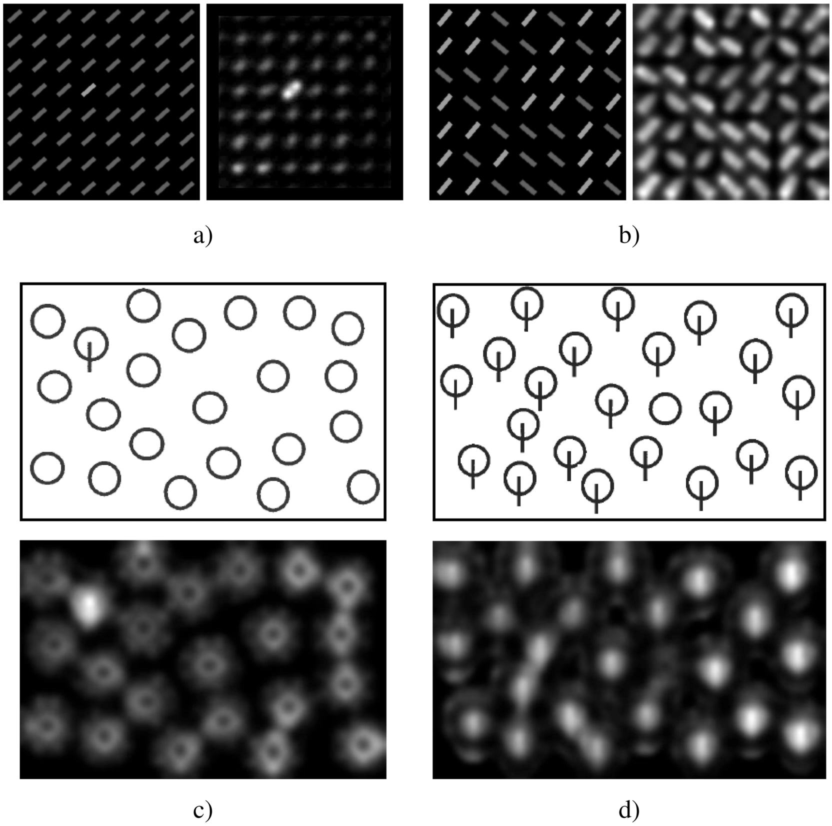

Figure 5: Classical visual search displays and associated output of the network of Figure 3. (a) Left: Example where target differs from distractors by a single feature (intensity). Right: Saliency map has strong peak at target location, justifying the percept of pop-out. (b) Left: Example where target (bar in third row and third column) differs from distractors by a conjunction of two features (intensity and orientation). Right: Target saliency is small, justifying the absence of pop-out. (c, d) Examples of pop-out asymmetry. (c) A target that differs from distractors by the presence of a feature (vertical bar) is very salient. (d) A target that differs from distractors by absence of the same feature has much smaller saliency. Note that the contrast of each saliency map has been adjusted to facilitate its visualization. This implies that absolute saliency values are not comparable across displays, but only within each saliency map. (A color version of this figure is available online at http://www.mitpressjournals.org/doi/suppl/10.1162/neco.2008.11-06-391.)

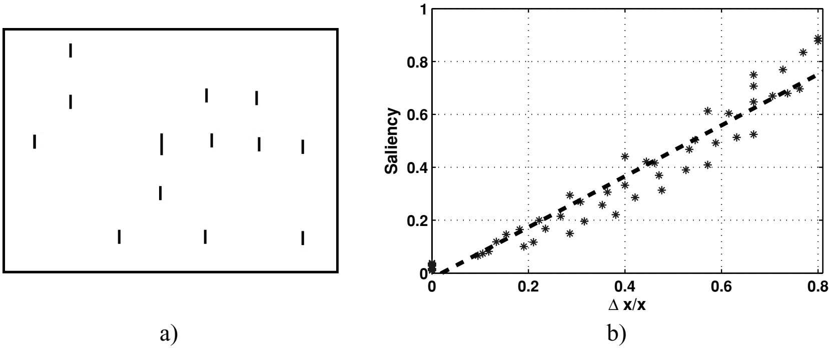

Figure 6: Performance of the saliency network of Figure 3 on the experiment devised by Treisman to show that visual search follows Weber's law (experiment 1a in Treisman & Gormican, 1988). In this experiment subjects were presented displays, such as the one shown in *a*, where target and distractors differed in terms of a single continuous variable: the target length (distractor length constant within each display). The output of the saliency network, at the target location, to a collection of displays of this type is shown in *b*, as a function of the ratio between the difference in target or distractor lengths (Δx) and the distractor length (x). The dashed line is the least-squares fit of Weber's law (a line) to the saliency data.

and psychophysics. In this context, we start by noting that the parallel between the network of Figure 3 and the standard architecture of V1 provides an interesting interpretation of V1 itself. It shows that V1 has the capability to optimally detect salient locations in the visual field, when optimality is defined in a decision-theoretic sense and certain (sensible) approximations are allowed for the sake of computational parsimony. The most significant among these approximations, that of the mutual information by a sum of marginal mutual information in equation 2.4, results from the assumption that feature dependencies are not informative for discrimination of natural image classes. It is interesting that, at least in the context of visual search (see Figure 5), preattentive vision is equally aggressive at disregarding feature dependencies (conjunctions).

While this disregard is widely acknowledged in the literature, we are aware of no previous computational explanation of why the preattentive visual system would choose to do so. The combined goals of decision-theoretic optimality and computational parsimony provide such an explanation: to the degree that equation 2.2 holds under natural scene statistics, restricting parallel search to the analysis of individual features has no loss of optimality. We have tested the importance of feature dependencies to image classification and found that accounting for dependencies between feature pairs can be beneficial, but there appears to be little gain in considering larger conjunctions (N. Vasconcelos & Vasconcelos, 2004; M. Vasconcelos & Vasconcelos, in press). While noticeable, the gains of pair-wise conjunctions over single features are not overwhelming, even for full-blown image classification. In the case of preattentive vision, by definition subject to tighter timing constraints, evolution could have simply deemed the gains of processing conjunctions unworthy of the inherent complexity.

Another interesting connection is between divisive normalization and saliency asymmetries. These are, in some sense, the central components of the neurophysiology of V1 and the psychophysics of visual search. Divisive normalization explains a rich set of neural behaviors that cannot be accommodated by the classic model of linear filtering plus nonlinearity, search asymmetries are one of the most heavily studied properties of visual search. Discriminant saliency provides a unified functional justification to these observations: optimal decision making, which exploits the statistical structure of natural images to achieve computational efficiency and is possible with biological hardware.

4 Statistical Inference in V1

So far we have shown that the decision-theoretic formulation of saliency, when combined with the constraint of computational parsimony, offers a holistic functional justification for V1. Obviously it is not likely that the whole of V1 would be uniquely devoted to saliency, let alone bottom-up saliency. This raises the question of whether the computational architecture discussed so far could be applied to the solution of generic inference problems. Answering this question, in the most general form, requires the derivation of a functional justification for the building blocks (cells) that compose V1. In what follows, we show that such a justification is indeed possible but requires a minor extension of the current simple cell model. We show, however, that under this extension, the cells of the standard V1 architecture perform the fundamental operations of statistical inference for processes that conform to the statistics of natural images. We then discuss some interesting consequences of this finding and relate it to previous proposals for the organization of perceptual systems.

4.1 Extended Simple Cell Model. In the discussion above, the optimality of the standard V1 architecture for the maximization of equation 2.4 requires $a_l \approx 1$ in equation 2.10. While this approximation is acceptable for the saliency problem, it is possible to make the statistical interpretation of the saliency network of Figure 3 exact. In fact, this requires only absorbing the two components of $\log a_l$ into $\psi[x(j);\Phi_0]$ and $\psi[x(j);\Phi_1]$, that is, redefining these quantities as

$$\tilde{\psi}[x(j);\Phi_c] = \frac{|x(j)|^{\beta_c}}{\xi_c} + \log \alpha_c. \tag{4.1}$$

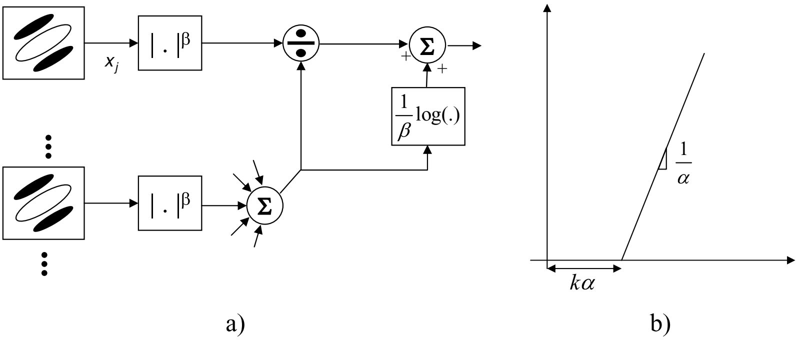

Figure 7: Extension of the standard simple cell model that makes the probabilistic interpretation of the standard V1 architecture, summarized by Table 1, exact. (a) The log of the contrast α that (divisively) normalizes the cell response is added to it. (b) The cell's curve of response has slope proportional to $1/\alpha$ and a shift to the right that is approximately linear in α.

Combining equations 2.5, 2.7, and 2.15, it is straightforward to show that for generalized gaussian stimuli, equation 4.1 is, up to a normalization constant, the estimate of

$$-\log P_{X|Y}[x(j) \mid c] \tag{4.2}$$

resulting from the MAP estimation of the scale parameter α_c. Physiologically, the implementation of equation 4.1 requires a slight extension of the current standard simple cell model, which is depicted in Figure 7. This extension consists of adding the log of the normalizing contrast α_c to the output of the cell, complementing the gain modulation of divisive normalization with a rightward shift of the response curve by $\alpha_c(\log 1/\alpha_c)^{1/\beta_c}$. For the (small) values of α_c typically found in natural scenes, this shift is approximately linear in α_c. This extension is compatible with existing cell recording data (Holt & Koch, 1997; Chance, Abbott, & Reyes, 2002; Doiron, Longtin, Berman, & Maler, 2000), and there is even evidence that when adaptation is considered, a shift occurs and is indeed proportional to the normalizing contrast (constant shifts of log contrast for multiplicative contrast increases) (Ohzawa, Sclar, & Freeman, 1985).

4.2 Fundamental Operations of Statistical Inference. The existence of a one-to-one mapping between equations 4.1 and 4.2 is significant in the sense of showing that simple cells can be interpreted as probabilistic inference units, tailored to the statistics of natural stimuli. In fact, revisiting equation

Table 1: V1 Cells Implement the Atomic Computations of Statistical Inference Under the Assumption of GGD Statistics.

Cell type	Computation	Function	Description
Simple	$\tilde{\psi}[x(j);\,\Phi_c]$	$-\log P_{X\|Y}[x(j)\|c]$	Negative log likelihood
Simple differential	$\tilde{\psi}[x(j);\,\Phi_0] - \tilde{\psi}[x(j);\,\Phi_1]$	$\log \frac{P_{X\|Y}[x(j)\|1]}{P_{X\|Y}[x(j)\|0]}$	Log likelihood ratio
Complex	$H(Y) - \langle \phi\{g[x(j)]\}\rangle_{\mathcal{W}_l}$	$I(X;\,Y)$	Mutual information

Notes: All operations are based on empirical probability estimates derived from the regions used for divisive normalization. The computations are exact for the extended simple cell model of Figure 7.

2.12 after this modification reveals that all components of the standard V1 architecture have a statistical interpretation, and this interpretation covers the three fundamental operations of statistical inference: probability inference, decision rules, and feature selection. The fundamental operation of statistical learning, parameter estimation, is also performed within the architecture, through the divisive normalization subjacent to all computations.

The statistical role of the different cell types is summarized in Table 1, which suggests a clear functional distinction between simple and complex cells. While simple cells assess probabilities, differential simple cells implement decision rules, and complex cells are feature detectors. Physiologically, this is consistent with most aspects of the existing simple or complex cell dichotomy (e.g., the lack of location and polarity sensitivity of complex cells) but suggests a novel refinement of simple cells into two subclasses: simple cells and differential simple cells. Simple cells conform to the currently accepted model, which is well known to explain most aspects of cell response within the classical receptive field (CRF) (Schwartz & Simoncelli, 2001; Cavanaugh et al., 2002). Differential simple cells include additional divisive normalization from a region external to the CRF. They could explain the well-documented observation that many cells are modulated by stimuli that fall outside this region (Sillito, Grieve, Jones, Cudeiro, & Davis, 1995; Sengpiel, Sen, & Blakemore, 1997; Levitt & Lund, 1997; Cavanaugh et al., 2002). Note, in particular, that the subtraction of $\tilde{\psi}[x(j);\,\Phi_1]$ from $\tilde{\psi}[x(j);\,\Phi_0]$ can be either excitatory or inhibitory, depending on the stimulus contrasts inside and outside the CRF. The availability of two independent mechanisms to control the responses from the two regions appears necessary to explain the recordings from cells that exhibit this behavior. We intend to investigate this issue in detail in future research.

Overall, the taxonomy of Table 1 assigns much more credit to simple cells than simply performing signal processing operations, such as filtering and gain control. In fact, it suggests that the central operation for learning within V1 is the divisive normalization that takes place in these cells, either in the log-likelihood form of equation 4.1 or the log-likelihood ratio form of equation 2.14. The coincidence that divisive normalization also solves the signal processing challenge of gain control is an extremely fortunate one, arguably too fortunate for evolution to pass on by. At a more generic level, the taxonomy of Table 1 also makes a compelling argument for the interpretation of brains as Bayesian inference engines, tuned to the statistics of the natural world. Note, in particular, that the exact shapes of the probability distributions of Table 1 are determined by the MAP estimates of their parameters. These estimates are, in turn, defined by the center and surround regions $\mathcal{W}_l^0$ and $\mathcal{W}_l^1$, specified by the lateral connections of divisive normalization. It follows that all probabilities could be computed with respect to distributions defined by arbitrary regions of the visual field by simply relying on alternative topologies for these connections. Furthermore, since all computations are in the log domain, operations such as Bayes rule or the chain rule of probability can be implemented through simple pooling. Hence, in principle, the architecture could implement optimal decisions for many other perceptual tasks.

5 Discussion

We close by relating our work to previous proposals for the organization of perceptual systems, comparing with other bottom-up saliency models, and identifying some open questions for future research.

5.1 Connections to Previous Work. In terms of prior literature, our work is obviously inspired by Barlow's hypothesis that neurons are optimally tuned to the statistics of natural stimuli. However, we depart from Barlow's initial suggestion that this optimization aims for coding efficiency (in the sense of redundancy reduction) (Barlow, 1961), or even more recent suggestions for sparseness (Olshausen & Field, 1996), or independence (Bell & Sejnowski, 1997; Schwartz & Simoncelli, 2001). While these hypotheses explain the Gabor-like decompositions found in V1 (Olshausen & Field, 1996; Bell & Sejnowski, 1997), and can even be used to derive probabilistic models for cell processing (Wainwright, Schwartz, & Simoncelli, 2002; Wainwright & Simoncelli, 2000), they have limited reach as overall principles for neural organization. For example, although a constraint like sparseness improves generalization and can be applied to top-down decision rules, it does not, per se, offer a solution to top-down problems like object recognition. This is opposite to the goal of decision-theoretic optimality, which, as shown here and in our previous work (Gao & Vasconcelos, 2005a), can be equally applied to bottom-up or top-down processing.

Overall, although redundancy reduction, sparseness, or independence may be desirable constraints to guarantee generalization, efficient representation, or tractable energy consumption, they are unlikely to be the end goal of a brain, which is (first and foremost) a decision-making device. Barlow himself has acknowledged this point, recasting his ideas in terms of redundancy exploitation rather than reduction. Recently Barlow (2001) essentially proposed what we are also suggesting: "To determine the best . . . way for an animal to behave . . . , its brain should decide what hypotheses about the world around the animal hold true at that time. The brain must therefore derive the probabilities of hypotheses being true from the evidence provided by its senses, and this is what Bayes' expression tells one how to do." He fell short, however, of showing how this could be implemented with biologically plausible computations or combined with his previous proposal of neural tuning to stimulus statistics. The discussion above shows that statistically tuned decision-making optimality could indeed appear as early as in V1.

5.2 Comparison to Previous Bottom-Up Saliency Models. Many bottom-up saliency models have been previously proposed in both the computer and biological vision literatures (Itti et al., 1998; Koch & Ullman, 1985; Li, 2002; Wolfe, 1994; Privitera & Stark, 2000; Malik & Perona, 1990; Kadir & Brady, 2001; Bruce & Tsotsos, 2006; Sha'ashua & Ullman, 1988; Harris & Stephens, 1988; Heidemann, 2004). Unlike this work, none of these models is derived from a generic principle for neural organization. With a few exceptions (e.g., Harris & Stephens, 1988), it is not clear if they are optimal in a well-defined sense, whether that optimality is subject to any type of constraints (e.g., sparseness, computational parsimony), or whether they have any connection to the statistics of perceptual stimuli. Typically these models simply propose a collection of image processing operations that mimic what is known about preattentive vision. This could be based on either psychophysics or neurophysiology but rarely accounts for both. As far as we know, only two previous proposals have tried to combine biological and psychophysical plausibility (Itti et al., 1998; Li, 2002).

The first, and probably most commonly used, computational saliency model, was proposed by Itti et al. (1998). It is similar to the one derived in this work in a number of aspects, including the front end (biologically inspired features, such as "color-double-opponent" channels and Gabor filtering), which we have mostly replicated (see the appendix), and a "center-surround" definition of saliency. On the other hand, it defines saliency as simple feature subtraction, as opposed to the explicit optimization for image locations of maximum discriminant power, which we now propose. Although simple feature differences can replicate some basic aspects of perceptual saliency, such as pop-out and conjunctive search, they do not explain some less trivial properties, such as the asymmetries of visual search, or provide any links to the underlying neural mechanisms of visual perception (such as divisive normalization). In this sense, the saliency mechanisms now proposed provide more insights and richer connections to both the psychophysics and neurophysiology literature. On the other hand, Itti's model also implements a biologically plausible mechanism (neural network with a layer of integrate-and-fire neurons and a winner-take-all "inhibition of return" stage), for the prediction of eye movements and search time in visual search experiments. We have so far not considered this problem.

The second was proposed by Li (2002), who argued that the preattentive computational mechanisms of primary visual cortex (V1) produce a saliency map and implemented a biologically motivated saliency model, featuring layer 2–3 pyramidal cells, interneurons, and horizontal intracortical connections. This model has been shown to reproduce not only the known excitatory and inhibitory contextual effects observed in physiological studies of V1 cells, but also various human psychophysical behaviors in visual search, such as basic features versus conjunction search, search asymmetries, and the effect of background homogeneity on search difficulty. While the model successfully links the psychophysical results to V1 physiology and anatomy, it does not explain the latter. It also does not account for some widely accepted V1 properties, such as the classical division into simple and complex cells, or more recent aspects such as divisive normalization. Finally, it does not provide a holistic functional justification for V1, as is the case for the work presented in this letter.

It is important to emphasize that the discriminant saliency architecture now derived is not a model but the optimal solution to the saliency problem, under the hypothesis that bottom-up saliency is the result of optimal decision making, for center-surround discrimination, exploiting the regularities of natural image statistics to achieve parsimony. Philosophically, this offers an explanation as to why V1 is organized as it is rather than just mimicking it. In practice, it avoids the main difficulty of model building, which is to determine exact values for all the parameters involved. Typically this difficulty leads to simplified architectures, which can replicate some, but not all, of the psychophysics. For this reason, a number of important aspects of saliency, such as asymmetries, Weber's law, and other nonlinearities, are frequently unaccounted for by prior saliency models.

It is also interesting to note that unlike most previous models, the optimal architecture relies heavily on divisive normalization. The latter is shown to be intrinsically connected to MAP parameter estimation and the computation of the probabilities required by optimal decisions. Without it, only suboptimal saliency judgements would be possible. Finally, the optimal solution equates divisive normalization to the adoption of a nonlinear and asymmetric measure of stimulus similarity, the Kullback-Leibler (KL) divergence, which underlies the mutual information used to quantify discrimination. This suggests that the KL divergence is the "right" measure of stimulus similarity, at least for discriminant tasks. Its asymmetry is responsible for the asymmetry of the saliency judgements exhibited by the architecture now derived. The latter is a hallmark of the psychophysics of early vision, which linear approaches (e.g., those based on amplitude differences) cannot replicate.

5.3 Future Work. The discussion above suggests a number of possibilities for future work. We have already mentioned that one question that merits further investigation is the relationship between the differentially normalized simple cell model of Table 1 and the response of simple cells to stimulation outside the CRF (Allman et al., 1985; Sillito et al., 1995; Sengpiel et al., 1997; Levitt & Lund, 1997; Cavanaugh et al., 2002). Another interesting possibility is to investigate in greater detail the role of the scale prior of equation 2.6. In the absence of any goals, it makes sense to simply adopt the interpretation of the prior as a regularizer, in which case the conjugate form of equation 2.6 is an acceptable choice. There is, however, no reason that the prior could not be controlled by feedback from higher-level cortical areas. From this point of view, the connection between adaptation and the shift of the simple cell response by the MAP estimate of the scale, discussed in section 4.1, could be a consequence of a process of hierarchical Bayesian inference that integrates top-down priors with bottom-up observations, as proposed in Lee and Mumford (2003), Lee (2003), and Lee, Yang, Romero, and Mumford (2002). This could, in turn, provide a coherent justification to the diversity of responses to stimulation of the CRF that have been reported in the literature (Sillito et al., 1995; Sengpiel et al., 1997; Levitt & Lund, 1997; Cavanaugh et al., 2002).

One property of the GGD to which we have so far not devoted much attention is the shape parameter β. We acknowledge some ambivalence with respect to the importance of this parameter. On one hand, it defines the heavy-tailed nature of the statistics of natural image features that justify the adoption of GGD-type of models. In this sense, it is important that β be constrained to a range of values that guarantees leptokurtic distributions. On the other hand, the precise value of β does not appear to be critical as long as it falls within this range. This is visible in Figures 3 and 7, where β simple changes the exponent of $|x|$. In the range of values consistent with leptokurtic behavior (β in the vicinity of 1), this does not seem to make a tremendous difference. Our experimentation with different values of this parameter (see the appendix) has also not produced qualitatively significant changes in the resulting saliency maps. It would nevertheless be interesting to investigate how all the equations would change if this parameter were estimated in a Bayesian manner and whether the result would be of any neurophysiological or psychophysical significance.

Finally, it would be interesting to extend some of the ideas presented above to other areas of cortical processing. For example, it is known that in addition to divisive normalization in V1, gain modulation appears in a number of more general contexts, including gaze direction, attention, coordinate transformations, and object recognition (Salinas & Thier, 2000; Chance et al., 2002). A natural question is whether these contexts are amenable to decision-theoretic formulation, similar to that performed above. In the particular context of attention, the ability of this formulation to account for both bottom-up and top-down saliency, appears to us as a definite plus. We intend to investigate this issue in the future.

Appendix: Bottom-Up Discriminant Saliency Detector

All saliency results presented in this work were obtained with a discriminant saliency detector implemented as follows. The image was first subject to a feature decomposition that mimics the three major neural pathways of primary visual cortex. Its implementation follows the work of Itti et al. (1998) and is based on a five-channel decomposition of the visual stimulus into an intensity map and four broadly tuned color channels,

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

where $\tilde{r} = r/I$, $\tilde{g} = g/I$, $\tilde{b} = b/I$, and $\lfloor x \rfloor_+ = \max(x, 0)$. The four color channels were in turn combined into two color opponent channels, $R - G$ for red/green and $B - Y$ for blue/yellow opponency. These and the intensity map were then convolved with three Mexican hat wavelet filters, centered at spatial frequencies 0.04, 0.08, and 0.16 cycle per pixel, to generate nine feature channels. All of these channels, plus a Gabor decomposition of the intensity map, constitute the feature space $\mathcal{X}$. The Gabor decomposition was implemented with a dictionary of zero-mean Gabor filters at three spatial scales (centered at frequencies of 0.08, 0.16, and 0.32 cycle per pixel) and four directions (evenly spread from 0 to π). Its algorithmic implementation followed the work of Manjunath and Ma (1996).

The saliency map was computed with the three-layer saliency network of Figure 3, with the shape parameter, β, of the GGD estimated through the method of moments (Huang & Mumford, 1999). The saliency detection performance does not depend critically on this parameter; for example, arbitrarily setting $\beta = 1$ produced qualitatively similar results. The choice of sizes for the center and surround windows was guided by available evidence from psychophysics and neurophysiology, where it is known that human percepts of saliency depend on the density and size of the items in the display (Nothdurft, 2000; Knierim & Van Essen, 1992), and the strength of neural response is a function of the stimulus that falls in the center and surround areas of the receptive field of a neuron (Knierim & Van Essen, 1992; Allman et al., 1985; Cavanaugh et al., 2002; Li & Li, 1994).

In particular, we mimicked the common practice of making the size of the display items comparable to that of the CRF (see, e.g., Treisman & Gelade, 1980; Hubel & Wiesel, 1965), by setting the size of the center window to a value comparable to the size of the display items (e.g., 1/10 of the image width for all displays shown in Figure 5).

With respect to the surround, it is known that pop-out occurs only when this area covers enough display items (Nothdurft, 2000), and there is a limit on the spatial extent of the underlying neural connections (Knierim & Van Essen, 1992; Allman et al., 1985; Cavanaugh et al., 2002; Li & Li, 1994). Considering this biological evidence, the surround window was made six times larger than that of the center at all image locations. Informal experimentation with these parameters has shown that the saliency results are not significantly affected by them. For example, setting the surround to the complement of the center window (i.e., the remaining display area) (Bruce & Tsotsos, 2006) did not produce any qualitatively noticeable changes. A precise characterization of the impact of center and surround window size on the saliency output, as well as connections to physiological data, are subjects that we intend to address in future research.

Finally, to improve their intelligibility, the saliency maps of Figure 5 were subject to smoothing, contrast enhancement (by squaring), and a normalization that maps the saliency value to the interval [0, 1].

Acknowledgments

We thank the anonymous reviewers for insightful comments that considerably strengthened the presentation of this work. The research was supported by NSF Career award IIS-0448609, NSF grant IIS-0534985, and a gift from Google Inc.

References

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

Agarwal, S., & Roth, D. (2002). Learning a sparse representation for objection detection. In *Proc. European Conference on Computer Vision* (Vol. 4, pp. 113–130) Berlin: Springer.

Allman, J., Miezin, F., & McGuinness, E. (1985). Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual Review Neuroscience, 8*, 407–430.

Attneave, F. (1954). Informational aspects of visual perception. *Psychological Review, 61*, 183–193.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Barlow, H. B. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems, 12*, 241–253.

Bell, A. J., & Sejnowski, T. J. (1997). The "Independent components" of natural scenes are edge filters. *Vision Research, 37*(23), 3327–3338.

Birney, K. A., & Fisher, T. R. (1995). On the modeling of DCT and subband image data for compression. *IEEE Transactions on Image Processing, 4*, 186–193.

Borenstein, E., & Ullman, S. (2004). Learn to segment. In *Proc. European Conference on Computer Vision* (pp. 315–328). Berlin: Springer.

Bravo, M., & Nakayama, K. (1992). The role of attention in different visual search tasks. *Perception and Psychophysics, 51*, 465–472.

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.* 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.

Carandini, M., Demb, J., Mante, V., Tolhurst, D., Dan, Y., Olshausen, B., et al. (2005). Do we know what the early visual system does? *Journal of Neuroscience, 25*, 10577–10597.

Carandini, M., Heeger, D., & Movshon, A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience, 17*, 8621–8644.

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

Chance, F., Abbott, L., & Reyes, A. (2002). Gain modulation from background synaptic input. *Neuron, 35*, 773–782.

Clarke, R. (1985). *Transform coding of images*. San Diego, CA: Academic Press.

Cover, T., & Thomas, J. (1991). *Elements of information theory*. New York: Wiley.

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

Doiron, B., Longtin, A., Berman, N., & Maler, L. (2000). Subtractive and divisive inhibition: Effect of voltage-dependent inhibitory conductances and noise. *Neural Computation, 13*, 227–248.

Enroth-Cugell, C., & Robson, J. G. (1966). The contrast sensitivity of retinal ganglion cells of the cat. *Journal of Physiology, 187*, 517–522.

Farvardin, N., & Modestino, J. W. (1984). Optimum quantizer performance for a class of non-gaussian memoryless sources. *IEEE Transactions on Information Theory, 30*(3), 485–497.

Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 264–271). Washington, DC: IEEE Computer Society.

Förstner, W. (1994). A framework for low level feature extraction. In *Proc. European Conference on Computer Vision* (pp. 383–394). Berlin: Springer.

Found, A., & Muller, H. J. (1995). Searching for unknown feature targets on more than on dimension: Further evidence for a "dimension weighting" account. *Perception and Psychophysics, 58*(1), 88–101.

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

Gao, D., & Vasconcelos, N. (2005b). Integrated learning of saliency, complex features, and object detectors from cluttered scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (pp. 282–287). Washington, DC: IEEE Computer Society.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Alvey Vision Conference* (pp. 147–151). Manchester, UK: University of Manchester.

Heeger, D. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience, 9*, 181–197.

Heidemann, G. (2004). Focus-of-attention from local color symmetries. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(7), 817–830.

Hillel, A. B., Hertz, T., & Weinshall, D. (2005). Efficient learning of relational object class models. In *Proc. IEEE International Conference on Computer Vision* (pp. 1762–1769). Washington, DC: IEEE Computer Society.

Holt, G., & Koch, C. (1997). Shunting inhibition does not have a divisive effect on firing rates. *Neural Computation, 9*, 1001–1013.

Huang, J., & Mumford, D. (1999). Statistics of natural images and models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (pp. 541–547). Washington, DC: IEEE Computer Society.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive field, binocular interaction, and functional architecture of in the cat's visual cortex. *Journal of Physiology, 160*, 106–154.

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

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.

James, W. (1981). *The principles of psychology*. Cambridge, MA: Harvard University Press (Original work published 1890)

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

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

Julesz, B. (1986). Texton gradients: The texton theory revisited. *Biological Cybernetics, 54*, 245–251.

Kadir, T., & Brady, M. (2001). Scale, saliency and image description. *International Journal of Computer Vision, 45*, 83–105.

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. *Human Neurobiology, 4*, 219–227.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology, 16*, 37–68.

Lee, T. S. (2003). Computations in the early visual cortex. *Journal of Physiology, 97*, 121–139.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A, 20*(7), 1434–1448.

Lee, T. S., Yang, C., Romero, R., & Mumford, D. (2002). Neural activity in early visual cortex reflects behavioral experience and higher order perceptual saliency. *Nature Neuroscience, 5*(6), 589–597.

Levitt, J., & Lund, J. (1997). Contrast dependence of contextual effects in primate visual cortex. *Nature, 387*, 73–76.

Li, C., & Li, W. (1994). Extensive integration field beyond the classical receptive field of cat's striate cortical neurons—classification and tuning properties. *Vision Research, 34*(18), 2337–2355.

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

Linsker, R. (1988). Self-organization in a perceptual network. *IEEE Computer, 21*(3), 105–117.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proc. IEEE International Conference on Computer Vision* (pp. 1150–1157). Washington, DC: IEEE Computer Society.

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A, 7*(5), 923–932.

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.

Manjunath, B. S., & Ma, W. Y. (1996). Texture feature for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(8), 837–842.

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

Movshon, J. A., Thompson, I. D., & Tolhurst, D. J. (1978). Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *Journal of Physiology, 283*, 53–77.

Muller, H. J., Heller, D., & Ziegler, J. (1995). Visual search for singleton feature targets within and across feature dimensions. *Perception and Psychophysics, 57*(1), 1–17.

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*, 387–391.

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron, 53*(4), 605–617.

Nothdurft, H. C. (1993). The role of features in preattentive vision: Comparison of orientation, motion and color cues. *Vision Research, 33*(14), 1937–1958.

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

Ohzawa, I., Sclar, G., & Freeman, R. (1985). Contrast gain control in the cat's visual system. *Journal of Neurophysiology, 54*, 651–667.

Olshausen, B., & Field, D. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*, 607–609.

Privitera, C., & Stark, L. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*, 970–982.

Salinas, E., & Thier, P. (2000). Gain modulation: A major computational principle of the central nervous system. *Neuron, 27*, 15–21.

Schiele, B., & Crowley, J. (1996). Where to look next and what to look for. In *Intelligent Robots and Systems (IROS)* (pp. 1249–1255). Singapore: World Scientific.

Schwartz, O., & Simoncelli, E. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience, 4*, 819–825.

Sengpiel, F., Sen, A., & Blakemore, C. (1997). Characteristics of surround inhibition in cat area 17. *Experimental Brain Research, 116*, 216–228.

Sha'ashua, A., & Ullman, S. (1988). Structural saliency: The detection of globally salient structures using a locally connected network. In *Proc. IEEE International Conference on Computer Vision* (pp. 321–327). Washington, DC: IEEE Computer Society.

Sillito, A., Grieve, K., Jones, H., Cudeiro, J., & Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature, 378*, 492–496.

Skottun, B., Valois, R. D., Grosof, D., Movshon, J., Albrecht, D., & Bonds, A. (1991). Classifying simple and complex cells on the basis of response modulation. *Vision Research, 31*, 1079–1086.

Srivastava, A., Lee, A., Simoncelli, E., & Zhu, S. (2003). On advances in statistical modeling of natural images. *Journal of Mathematical Imaging and Vision, 18*, 17–33.

Treisman, A., & Gelade, G. (1980). A feature-integration 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.

Treisman, A., & Sato, S. (1990). Conjunction search revisited. *Journal of Experimental Perception and Performance, 16*, 459–478.

Treisman, A., & Souther, J. (1985). Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General, 114*, 285–310.

Vasconcelos, M., & Vasconcelos, N. (in press). Natural image statistics and low complexity feature selection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.

Vasconcelos, N., & Vasconcelos, M. (2004). Scalable discriminant feature selection for image retrieval and recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 770–775). Washington, DC: IEEE Computer Society.

Verghese, P. (2001). Visual search and attention: A signal detection theory approach. *Neuron, 31*, 523–535.

Wainwright, M. J., Schwartz, O., & Simoncelli, E. P. (2002). Natural image statistics and divisive normalization: Modeling nonlinearities and adaptation in cortical neurons. In R. Rao, B. Olshausen, & M. Lewicki (Eds.), *Probabilistic models of the brain: Perception and neural function* (pp. 203–222). Cambridge, MA: MIT Press.

Wainwright, M. J., & Simoncelli, E. P. (2000). Scale mixtures of gaussians and the statistics of natural images. In S. A. Solla, T. K. Leen, K.-R. Müller (Eds.), *Advances* *in neural information processing systems, 12* (pp. 855–861). Cambridge, MA: MIT Press.

Walker, K., Cootes, T., & Taylor, C. (1998). Locating salient object features. In *Proc. British Machine Vision Conference* (pp. 557–566). Malvern, UK: British Machine Vision Association.

Watanabe, S. (1960). Information-theoretical aspects of Inductive and Deductive Inference. *IBM Journal of Research and Development*4, 208–231.

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

Wolfe, J. M. (1998). Visual Search. In H. Pashler (Ed.), *Attention* (pp. 13–74). Hove, U.K. Psychology Press.

Wolfe, J. M., & Horowitz, T. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience, 5*, 495–501.

Received November 6, 2006; accepted April 8, 2008.