SUN: A Bayesian Framework for Saliency Using Natural Statistics

Lingyun Zhang Tim K. Marks Matthew H. Tong
Honghao Shan Garrison W. Cottrell

Dept. of Computer Science and Engineering
University of California, San Diego
9500 Gilman Dr., La Jolla, CA92037

December 6, 2007

Abstract

We propose a definition of saliency by considering what the visual system is trying to optimize when directing attention. The resulting model is a Bayesian framework from which bottom-up saliency emerges naturally as the self-information of visual features, and overall saliency (incorporating top-down information with bottom-up saliency) emerges as the pointwise mutual information between the features and the target when searching for a target. An implementation of our framework demonstrates that our model's bottom-up saliency maps perform as well as or better than existing algorithms in predicting people's fixations in free viewing. Unlike existing saliency measures, which depend on the statistics of the particular image being viewed, our measure of saliency is derived from natural image statistics, obtained in advance from a collection of natural images. For this reason, we call our model SUN (Saliency Using Natural statistics). A measure of saliency based on natural image statistics, rather than based on a single test image, provides a straightforward explanation for many search asymmetries observed in humans; the statistics of a single test image lead to predictions that are not consistent with these asymmetries. In our model, saliency is computed locally, which is consistent with the neuroanatomy of the early visual system and results in an efficient algorithm with few free parameters.

Keywords: saliency, attention, eye movements.

1 Introduction

The surrounding world contains a tremendous amount of visual information, which the visual system cannot fully process. The visual system thus faces the problem of how to allocate its processing resources to focus on important aspects of a scene. Despite the limited amount of visual information the system can handle, sampled by discontinuous fixations, we experience a seamless, continuous world. Yet humans, as well as most other animals, thrive using this heavily downsampled visual information. Visual attention as reflected in eye movements discloses the sampling strategy of the visual system and is of great research interest as an essential component of visual cognition. Psychologists have investigated visual attention for many decades using psychophysical experiments, such as visual search tasks, with carefully controlled stimuli. Sophisticated mathematical models have been built to account for the wide variety of human performance data (e.g. Treisman & Gelade, 1980; Wolfe, Cave, & Franzel, 1989; Bundesen, 1990).

With the development of affordable and easy-to-use modern eye-tracking systems, the locations that people fixate when they perform certain tasks can be explicitly recorded and can provide insight into how people allocate their attention when viewing complex natural scenes. The proliferation of eye-tracking data over the last two decades has led to a number of computational models attempting to account for the data and addressing the question of what attracts attention. Most models have focused on bottom-up attention, where the subjects are free-viewing a scene and salient objects attract attention. Many of these saliency models are based on findings from psychology and neurobiology and explain the mechanisms guiding attention allocation (Koch & Ullman, 1985; Itti, Koch, & Niebur, 1998; Wolfe et al., 1989). More recently, a number of models attempt to explain attention based on more mathematically motivated principles, addressing the goal of the computation (Chauvin, Herault, Marendaz, & Peyrin, 2002; Kadir & Brady, 2001; Renninger, Coughlan, Verghese, & Malik, 2004; Harel, Koch, & Perona, 2007; Gao & Vasconcelos, 2004, 2007; Bruce & Tsotsos, 2006; Oliva, Torralba, Castelhano, & Henderson, 2003; Torralba, Oliva, Castelhano, & Henderson, 2006; Zhang, Tong, & Cottrell, 2007). Both types of models tend to rely solely on the statistics of the current test image when it comes to computing the saliency of a point in the image. We argue here that natural statistics (the statistics of visual features in natural scenes, which an organism would learn through experience) must also play an important role in this process.

In this paper, we make an effort to address the underlying question: what is the goal of the computation performed by the attentional system? Our model starts from the simple assumption that an important goal of the visual system is to find potential targets, and builds up a Bayesian probabilistic framework of what the visual system should calculate to optimally achieve this goal. In this framework, bottom-up saliency emerges naturally as *self-information*. When searching for a particular target, top-down effects from a known target emerge in our model as a log-likelihood term in the Bayesian formulation. The model also dictates how to combine bottom-up and top-down information, leading to *pointwise mutual information* as a measure of overall saliency. We develop a bottom-up saliency algorithm that performs as well as or better than state-of-the-art saliency algorithms at predicting human fixations when free viewing images. Whereas existing bottom-up saliency measures are defined solely in terms of the image currently being viewed, ours is instead defined based on natural statistics (collected from a set of natural images), to represent the visual experience an organism would acquire during development. This difference is most notable when comparing with models that also use a Bayesian formulation (e.g. Torralba et al. (2006)) or self-information (e.g. Bruce and Tsotsos (2006)). For this reason, we call our saliency model SUN (Saliency Using Natural statistics). As a result of using natural statistics, SUN provides a straightforward account of many human search asymmetries that cannot be explained based on the statistics of the test image alone. Unlike many models, our measure of saliency only involves local computation on images. This makes the SUN algorithm not only more efficient, but also more biologically plausible, as long-range connections are scarce in the lower levels of the visual system.

The rest of the paper is organized as follows. Section 2 reviews previous work. In Section 3, we assume an intuitive goal for the visual system, and infer what measure of saliency would have evolved to achieve this goal. In Section 4, we describe how we implement SUN for computing bottom-up saliency maps based on the proposed definition of saliency. Results are presented and evaluated using human eye-tracking data in Section 5 along with a discussion of evaluation methods. Section 6 concludes the paper with a discussion. This includes a comparison, in the context of visual search asymmetries, of the predictions of saliency based on natural statistics versus saliency based on the statistics of the current image. Saliency based on natural statistics explains many observations in the visual search paradigm, such as the search asymmetry between feature presence versus absence, between prototypes versus non-prototype exemplars, and between other-race versus same-race faces (Zhang et al., 2007).

2 Previous Work

In this section we discuss previous saliency models, which have achieved good performance in predicting human fixations in viewing images. The motivation for these models has come from psychophysics and biological systems (Itti et al., 1998; Itti & Koch, 2001), classification optimality (Gao & Vasconcelos, 2004, 2007), the task of looking for a target (Oliva et al., 2003; Torralba et al., 2006), or information maximization (Bruce & Tsotsos, 2006).

Itti and Koch's saliency model (Itti et al., 1998; Itti & Koch, 2000, 2001) is one the earliest and the most used for comparison in later work. The model is an implementation of and expansion on the basic ideas first proposed by Koch and Ullman (1985). The model is inspired by the visual attention literature, such as feature integration theory (Treisman & Gelade, 1980), and care is taken in the model's construction to ensure that the model is neurobiologically plausible. The model takes an image as input, which is then decomposed into three channels: intensity, color, and orientation. A center-surround operation, implemented by taking the difference of the filter responses from two scales, yields a set of feature maps. The feature maps for each channel are then normalized and combined across scales and orientations, creating conspicuity maps for each channel. The conspicuous regions of these maps are further enhanced by normalization, and the channels are linearly combined to form the overall saliency map. This process allows locations to vie for conspicuity within each feature dimension, but has separate feature channels contribute to saliency independently; this is consistent with the feature integration theory. This model has been shown to be successful in predicting human fixations and to be useful in object detection (Itti et al., 1998; Itti & Koch, 2001; Parkhurst, Law, & Niebur, 2002). However, it can be criticized as being ad hoc, partly because the overarching goal of the system (i.e., what it is designed to optimize) is not specified, and it has many parameters that need to be hand-selected.

Several saliency algorithms are based on measuring the complexity of a local region (Yamada

& Cottrell, 1995; Chauvin et al., 2002; Kadir & Brady, 2001; Renninger et al., 2004). Yamada and Cottrell (1995) measure the variance of 2D Gabor filter responses across different orientations. Kadir and Brady (2001) measure the entropy of the local distribution of image intensity. Renninger et al. (2004) measure the entropy of local line orientation histograms, and the most salient point at any given time is the one that provides the greatest information gain conditioned on the knowledge obtained during previous fixations. All of these saliency-as-variance/entropy models are based on the idea that the entropy of a feature distribution over a local region measures the richness and diversity of that region (Chauvin et al., 2002), and intuitively a region should be salient if it contains features with many different orientations and intensities. A common critique of these models is that highly textured regions are always salient regardless of their context. For example, human observers find an egg in a nest highly salient, but local-entropy-based algorithms find the nest to be much more salient than the egg (Gao & Vasconcelos, 2004; Bruce & Tsotsos, 2006).

Gao and Vasconcelos (2004, 2007) proposed a specific goal for saliency: classification. That is, a goal of the visual system is to classify each stimulus as belonging to a class of interest (or not) and saliency should be assigned to useful locations for that task. This was first used for object detection (Gao & Vasconcelos, 2004), where a set of features are selected to best discriminate the class of interest (e.g., faces or cars) from all other stimuli, and saliency is defined as the weighted sum of feature responses for the set of features that are salient for that class. This forms a definition that is inherently top-down and goal directed, as saliency is defined for a particular class. Gao and Vasconcelos (2007) define bottom-up saliency using the idea that locations are salient if they differ greatly from their surroundings. They use difference of Gaussians (DoG) filters and Gabor filters, measuring the saliency of a point as the Kullbach-Liebler (KL) divergence between the histogram of filter responses at the point and the histogram of filter responses in the surrounding region. This addresses a previously mentioned problem commonly faced by complexity-based models (as well as some other saliency models that use linear filter responses as features): these models always assign high saliency scores to highly textured areas. In Section 6.3, we will discuss a way that the SUN model could address this problem, by using non-linear features that model complex cells or neurons in higher levels of the visual system.

Oliva and colleagues proposed a probabilistic model for visual search tasks (Oliva et al., 2003; Torralba et al., 2006). When searching for a target in an image, the probability of interest is the joint probability that the target is present in the current image, together with the target's location (if the target is present), given the observed features. This can be calculated using Bayes rule:

$$p(O=1,L|F,G) = \underbrace{\frac{1}{p(F|G)}}_{\substack{\text{bottom-up saliency} \\ \text{(as defined by Oliva et al.)}}} p(F|O=1,L,G)\,p(L|O=1,G)\,p(O=1|G) \tag{1}$$

where $O = 1$ denotes the event that the target is present in the image, L denotes the location of the target when $O = 1$, F denotes the local features at location L, and G denotes the global features of the image. The global features of G represent the scene gist. Experiments show that the gist of a scene can be quickly determined, and the focus of their work largely concerns how this gist affects eye movements. The first term on the right side of (1) is independent of the target and is defined as bottom-up saliency; Oliva and colleagues approximate this conditional probability distribution using the current image's statistics. The remaining terms on the right side of (1) respectively address the distribution of features for the target, the likely locations for the target, and the probability of the target's presence, all conditioned on the scene gist. As we will see in Section 3, our use of Bayes' rule to derive saliency is reminiscent of this approach. However, the probability of interest in the work of Oliva and colleagues is whether or not a target is present anywhere in the test image, whereas the probability we are concerned with is the probability that a target is present *at each point* in the visual field. In addition, Oliva and colleagues condition all of their probabilities on the values of global features. Conditioning on global features/gist affects the meaning of all terms in equation (1), and justifies their use of current image statistics for bottom-up saliency. In contrast, SUN focuses on the effects of an organism's prior visual experience.

Bruce and Tsotsos (2006) define bottom-up saliency based on maximum information sampling. Information, in this model, is computed as Shannon's self-information, $-\log p(F)$, where F is a vector of the visual features observed at a point in the image. The distribution of the features is estimated from a neighborhood of the point, which can be as large as the entire image. When the neighborhood of each point is indeed defined as the entire image of interest, as implemented in (Bruce & Tsotsos, 2006), the definition of saliency becomes identical to the bottom-up saliency term in equation (1) from the work of Oliva and colleagues (Oliva et al., 2003; Torralba et al., 2006). It is worth noting, however, that the feature spaces used in the two models are different. Oliva and colleagues use biologically-inspired linear filters of different orientations and scales. These filter responses are known to correlate with each other; for example, a vertical bar in the image will activate a filter tuned to vertical bars but will also activate (to a lesser degree) a filter tuned to 45-degree-tilted bars. The joint probability of the entire feature vector is estimated using multivariate Gaussian distributions (Oliva et al., 2003) and later multivariate generalized Gaussian distributions (Torralba et al., 2006). Bruce and Tsotsos (2006), on the other hand, employ features that were learned from natural images using independent component analysis (ICA). These have been shown to resemble the receptive fields of neurons in primary visual cortex (V1), and their responses have the desired property of sparsity. Furthermore, the features learned are approximately independent, so the joint probability of the features is just the product of each feature's marginal probability, simplifying the probability estimation without making unreasonable independence assumptions.

3 A Bayesian framework for saliency

We propose that one goal of the visual system is to find potential targets that are important for survival, such as food and predators. To achieve this, the visual system must actively estimate the probability of a target at every location given the visual features observed. We propose that this probability is visual saliency.

To formalize this, let z denote a point in the visual field. A point here is loosely defined; in the implementation described in Section 4, a point corresponds to a single image pixel. (In other contexts, a point could refer other things, such as an object (Zhang et al., 2007).) We let the binary random variable C denote whether or not a point belongs to a target class, let the random variable L

denote the location (i.e., the pixel coordinates) of a point, and let the random variable F denote the visual features of a point. Saliency of a point z is then defined as $p(C = 1|F = f_z, L = l_z)$, where f_z represents the feature values observed at z and l_z represents the location (pixel coordinates) of z. This probability can be calculated using Bayes' rule:

$$\begin{aligned} s_z &= p(C = 1|F = f_z, L = l_z) \\ &= \frac{p(F = f_z, L = l_z|C = 1)p(C = 1)}{p(F = f_z, L = l_z)}. \end{aligned} \tag{2}$$

We assume for simplicity that features and location are independent and conditionally independent given $C = 1$:

$$p(F = f_z, L = l_z) = p(F = f_z)p(L = l_z), \tag{3}$$

$$p(F = f_z, L = l_z|C = 1) = p(F = f_z|C = 1)p(L = l_z|C = 1). \tag{4}$$

This entails the assumption that the distribution of a feature does not change with location. For example, (3) implies that a point in the left visual field is just as likely to be green as a point in the right visual field. Furthermore, (4) implies (for instance) that a point on a target in the left visual field is just as likely to be green as a point on a target in the right visual field[1]. With these independence assumptions, (2) can be rewritten as:

$$s_z = \frac{p(F = f_z|C = 1)p(L = l_z|C = 1)p(C = 1)}{p(F = f_z)p(L = l_z)} \tag{5}$$

$$= \frac{p(F = f_z|C = 1)}{p(F = f_z)} \cdot \frac{p(L = l_z|C = 1) \cdot p(C = 1)}{p(L = l_z)} \tag{6}$$

$$= \underbrace{\frac{1}{p(F = f_z)}}_{\substack{\text{Independent}\\ \text{of target}\\ \text{(bottom-up saliency)}}} \cdot \underbrace{\underbrace{p(F = f_z|C = 1)}_{\text{Likelihood}} \cdot \underbrace{p(C = 1|L = l_z)}_{\text{Location prior}}}_{\substack{\text{Dependent on target}\\ \text{(top-down knowledge)}}} \tag{7}$$

To compare this probability across locations in an image, it suffices to estimate the log probability (since logarithm is a monotonically increasing function). For this reason, we take the liberty of using the term saliency to refer both to s_z and to $\log s_z$, which is given by:

$$\log s_z = \underbrace{-\log p(F = f_z)}_{\text{Self-information}} + \underbrace{\log p(F = f_z|C = 1)}_{\text{Log likelihood}} + \underbrace{\log p(C = 1|L = l_z)}_{\text{Location prior}}. \tag{8}$$

The first term on the right side of this equation, $-\log p(F = f_z)$, depends only on the visual features observed at the point, and is independent of any knowledge we have about the target class.

[1] These independence assumptions do not generally hold (they could be relaxed in future work). For example, illumination is not invariant to location: as sunshine normally comes from above, the upper part of the visual field is likely to be brighter. But illumination contrast features, such as the responses to DoG (Difference of Gaussians) filters, will be more invariant to location changes.

In information theory, $-\log p(F = f_z)$ is known as the *self-information* of the random variable F when it takes the value f_z. Self-information increases when the probability of a feature decreases—in other words, rarer features are more informative. We have discussed self-information in Section 2 in the context of previous work, but as we will see later, SUN's use of self-information differs from that of previous approaches.

The second term on the right side of (8), $\log p(F = f_z|C = 1)$, is a log-likelihood term that favors feature values that are consistent with our knowledge of the target. For example, if we know that the target is green, then the log-likelihood term will be much larger for a green point than for a blue point. This corresponds to the top-down effect when searching for a known target, consistent with the finding that human eye movement patterns during iconic visual search can be accounted for by a maximum likelihood procedure for computing the most likely location of a target (Rao, Zelinsky, Hayhoe, & Ballard, 2002).

The third term in (8), $\log p(C = 1|L = l_z)$, is independent of visual features and reflects any prior knowledge of where the target is likely to appear. It has been shown that if the observer is given a cue of where the target is likely to appear, the observer attends to that location (Posner & Cohen, 1984). For simplicity and fairness of comparison with (Itti & Koch, 2001; Bruce & Tsotsos, 2006; Gao & Vasconcelos, 2007), we assume location invariance (no prior information about the locations of potential targets) and omit the location prior; in Section 5, we will further discuss the effects of the location prior.

After omitting the location prior from (8), the equation for saliency has just two terms, the self-information and the log-likelihood, which can be combined:

$$\begin{aligned}
\log s_z &= \underbrace{-\log p(F{=}f_z)}_{\substack{\text{Self-information} \\ \text{(bottom-up saliency)}}} + \underbrace{\log p(F{=}f_z|C{=}1)}_{\substack{\text{Log likelihood} \\ \text{(top-down knowledge)}}} && (9) \\
&= \log \frac{p(F{=}f_z|C{=}1)}{p(F{=}f_z)} && (10) \\
&= \underbrace{\log \frac{p(F{=}f_z, C{=}1)}{p(F{=}f_z)p(C{=}1)}}_{\substack{\text{Pointwise mutual information} \\ \text{(overall saliency)}}} \quad . && (11)
\end{aligned}$$

The resulting expression, which is called the *pointwise mutual information* between the visual feature and the presence of a target, is a single term that expresses overall saliency. Intuitively, it favors feature values that are more likely in the presence of a target than in a target's absence.

When the organism is not actively searching for a particular target (the *free viewing* condition), the organism's attention should be directed to any *potential* targets in the visual field, despite the fact that the features associated with the target class are unknown. In this case, the log-likelihood term in (8) is unknown, so we omit this term from the calculation of saliency (this can also be thought of as assuming that for an unspecified target, the likelihood distribution is uniform over feature values). In this case, the overall saliency reduces to just the self-information term: $\log s_z = -\log p(F{=}f_z)$. We take this to be our definition of bottom-up saliency. It implies that the rarer a feature is, the more it will attract our attention.

Note that all of the probability distributions described here should be learned by the visual system through experience. Because the goal of the SUN model is to find potential targets in the surrounding environment, the probabilities should reflect the natural statistics of the environment and the learning history of the organism, rather than just the statistics of the current image. (This is especially obvious for the top-down terms, which require learned knowledge of the targets.)

In summary, calculating the probability of a target at each point in the visual field leads naturally to the estimation of information content. In the free-viewing condition, when there is no specific target, saliency reduces to the self-information of a feature. This implies that when the attention is directed only by bottom-up saliency, moving one's eyes to the most salient points in an image can be regarded as maximizing information sampling, which is consistent with the basic assumption of Bruce and Tsotsos (2006). When a particular target is being searched for, on the other hand, our model implies that the best features to attend to are those that have the most mutual information with the target. This has been shown to be very useful in object detection with objects such as faces and cars (Ullman, Vidal-Naquet, & Sali, 2002).

In the rest of this paper, we will concentrate on bottom-up saliency for static images. This corresponds to the free viewing condition, when no particular target is of interest. In Section 4, we provide a simple and efficient algorithm for bottom-up saliency that (as we demonstrate in Section 5) produces state-of-the-art performance in predicting human fixations. In addition, we discuss a surprising result concerning the location prior in Section 5.1.

4 Implementation

In this section, we develop an algorithm based on our SUN model that takes color images as input and calculates their saliency maps (the saliency at every pixel in an image). Given a probabilistic formula for saliency, such as the one we derived in Section 3, there are two key factors that affect the final results of a saliency model when operating on an image. One is the feature space, and the other is the probability distribution over the features.

In most existing saliency algorithms, the features are calculated as responses of biologically plausible linear filters, such as DoG (difference of Gaussians) filters and Gabor filters (e.g., Itti et al., 1998; Itti & Koch, 2001; Oliva et al., 2003; Torralba et al., 2006; Gao & Vasconcelos, 2004, 2007). In (Bruce & Tsotsos, 2006), the features are calculated as the responses to filters learned from natural images using independent component analysis (ICA). In this paper, we conduct experiments with both kinds of features.

Below, we describe the SUN algorithm for estimating the bottom-up saliency that we derived in Section 3, $-\log p(F = f_z)$. Here, a point z corresponds to a pixel in the image. For the remainder of the paper, we will drop the subscript z for notational simplicity. In this algorithm, F is a random vector of filter responses, $F = [F_1, F_2, \ldots]$, where the random variable F_i represents the response of the ith filter at a pixel, and $f = [f_1, f_2, \ldots]$ are the values of these filter responses at this location.

4.1 Method 1: Difference of Gaussians filters

As noted above, many existing models use a collection of DoG (difference of Gaussians) and/or Gabor filters as the first step of processing the input images. These filters are popular due to their resemblance to the receptive fields of neurons in the early stages of the visual system, namely the lateral geniculate nucleus of the thalamus (LGN) and the primary visual cortex (V1). DoGs, for example, give the well-known "Mexican hat" center-surround filter. Here, we apply DoGs to the intensity and color channels of an image.

Let r, g and b denote the red, green, and blue components of an input image pixel. The intensity (I), red/green (RG), and blue/yellow (BY) channels are calculated as:

$$I = r + g + b, \qquad RG = r - g, \qquad BY = b - \frac{r+g}{2} - \frac{\min(r,g)}{2}. \tag{12}$$

The DoG filters are generated by[2]

$$DoG(x,y) = \frac{1}{\sigma^2}\exp\left(-\frac{x^2+y^2}{\sigma^2}\right) - \frac{1}{(1.6\sigma)^2}\exp\left(-\frac{x^2+y^2}{(1.6\sigma)^2}\right). \tag{13}$$

where (x,y) is the location in the filter. These filters are convolved with the intensity and color channels (I, RG, and BY) to produce the filter responses. We use four scales of DoG ($\sigma = 4, 8, 16$ or 32 pixels) on each of the three channels, leading to 12 feature response maps. The filters are shown in Figure 1, top.

By computing these feature response maps on a set of 138 images of natural scenes (photographed by the first author), we obtained an estimate of the probability distribution over the observed values of each of the 12 features. To parameterize this estimated distribution for each feature F_i, we used an algorithm proposed by Song (2006) to fit a zero-mean generalized Gaussian distribution, also known as an exponential power distribution, to the filter response data:

$$p(f;\sigma,\theta) = \frac{\theta}{2\sigma\Gamma(\frac{1}{\theta})}\exp\left(-\left|\frac{f}{\sigma}\right|^{\theta}\right). \tag{14}$$

In this equation, Γ is the gamma function, θ is the shape parameter, σ is the scale parameter, and f is the filter response. This resulted in one shape parameter, θ_i, and one scale parameter, σ_i, for each of the 12 filters: $i = 1, 2, ..., 12$. Figure 1 shows the distributions of the four DoG filter responses on the intensity (I) channel across the training set of natural images, and the fitted generalized Gaussian distributions. As the figure shows, the generalized Gaussians provide an excellent fit to the data.

Taking the logarithm of (14), we obtain the log probability over the possible values of each feature:

$$\log p(F_i = f_i) = \log\theta_i - \log 2 - \log\sigma_i - \log\Gamma\left(\frac{1}{\theta_i}\right) - \left|\frac{f_i}{\sigma_i}\right|^{\theta_i} \tag{15}$$

$$= -\left|\frac{f_i}{\sigma_i}\right|^{\theta_i} + const., \tag{16}$$

[2] Equation (13) is adopted from the function filter_DOG_2D, from *Image Video toolbox for Matlab* by Piotr Dollar. The toolbox can be found at http://vision.ucsd.edu/∼pdollar/toolbox/doc/.

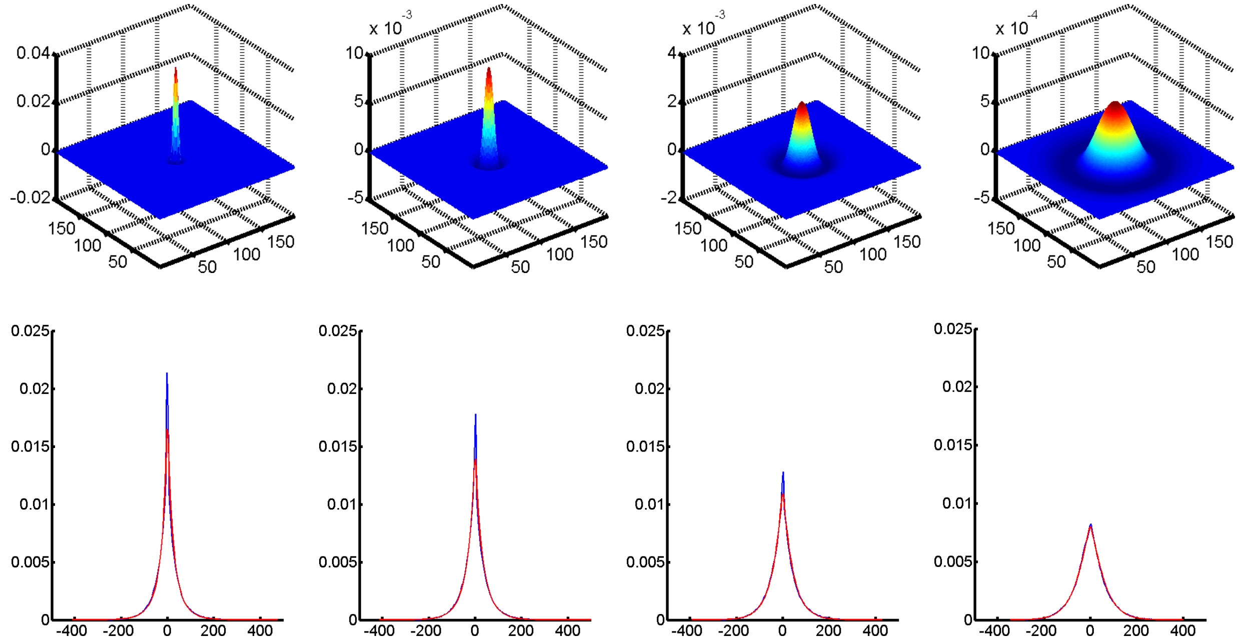

Figure 1: Four scales of difference of Gaussians (DoG) filters were applied to each channel of a set of 138 images of natural scenes. *Top:* The four scales of difference of Gaussians (DoG) filters that were applied to each channel. *Bottom:* The graphs show the probability distribution of filter responses for these four filters (with σ increasing from left to right) on the intensity (I) channel collected from the set of natural images (blue line), and the fitted generalized Gaussian distributions (red line). Aside from the natural statistics in this training set being slightly sparser, the generalized Gaussian distributions provide an excellent fit.

where the constant term does not depend on the feature value. To simplify the computations, we assume that the 12 filter responses are independent. Hence the total bottom-up saliency of a point takes the form:

$$\log s = -\log p(F = f) = \sum_{i=1}^{12} \left| \frac{f_i}{\sigma_i} \right|^{\theta_i} + const. \tag{17}$$

4.2 Method 2: Linear ICA Filters

In SUN's final formula for bottom-up saliency (17), we assumed independence between the filter responses. However, this assumption does not always hold. For example, a bright spot in an image will generate a positive filter response for multiple scales of DoG filters. In this case the filter responses, far from being independent, are highly correlated. It is not clear how this correlation affects the saliency results when a weighted sum of filter responses is used to compute saliency (as in Itti et al., 1998; Itti & Koch, 2001) or when independence is assumed in estimating probability (as in our case). Torralba et al. (2006) used a multivariate generalized Gaussian distribution to fit

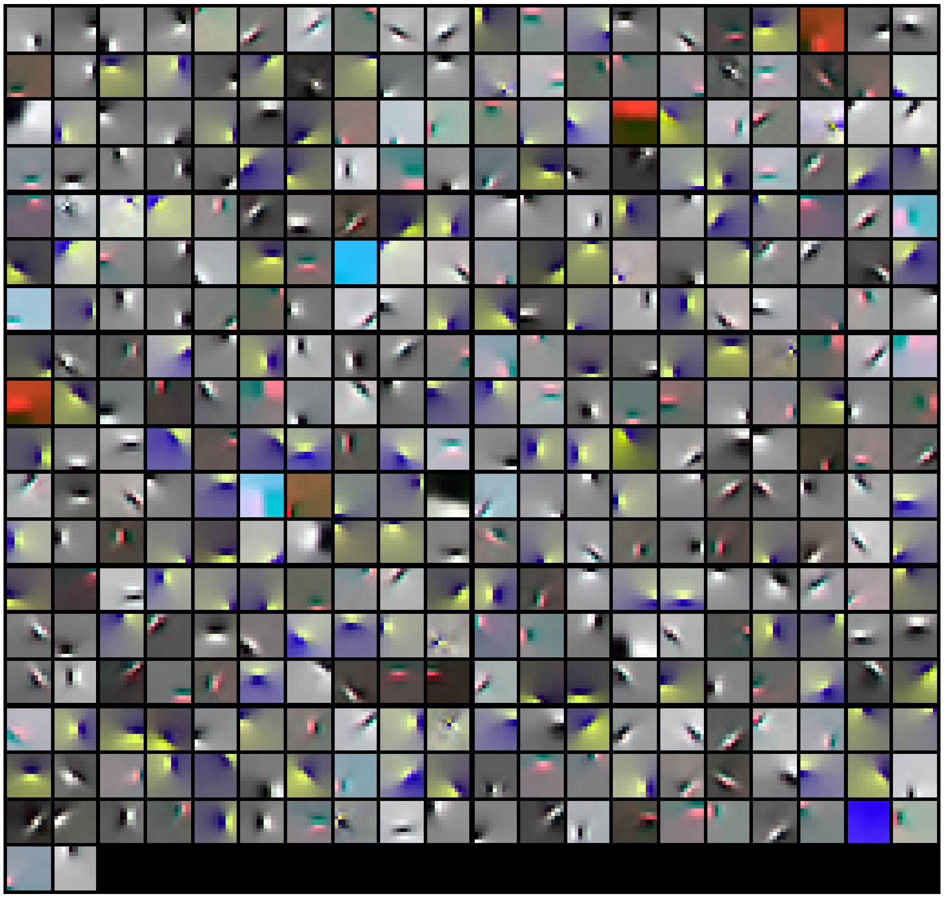

Figure 2: The 362 linear features learned by applying a complete independent component analysis (ICA) algorithm to 11×11 patches of color natural images from the Kyoto dataset.

the joint probability of the filter responses. However, although the response of a single filter has been shown to be well fitted by a univariate generalized Gaussian distribution, it is less clear that the joint probability follows a multivariate generalized Gaussian distribution. Also, much more data is necessary for a good fit of a high-dimensional probability distribution than for one-dimensional distributions. It has been shown that estimating the moments of a generalized Gaussian distribution has its limitations even for the one-dimensional case (Song, 2006), and it is much less likely to work well for the high-dimensional case.

To obtain the linear features used in their saliency algorithm, Bruce and Tsotsos (2006) applied independent component analysis (ICA) to a training set of natural images. This has been shown to yield features that qualitatively resemble those found in the visual cortex (Olshausen & Field, 1996; Bell & Sejnowski, 1997). Although the linear features learned in this way are not entirely independent, they have been shown to be independent up to third-order statistics (Wainwright, Schwartz, & Simoncelli, 2002). Such a feature space will provide a much better match for the independence assumptions we made in (17). Thus, in this method we follow (Bruce & Tsotsos,

2006) and derive complete ICA features to use in SUN. It is worth noting that although Bruce and Tsotsos (2006) use a set of natural images to train the feature set, they determine the distribution over these features solely from a single test image when calculating saliency.

We applied the FastICA algorithm (Hyvarinen & Oja, 1997) to 11-pixel × 11-pixel color natural image patches drawn from the Kyoto image dataset (Wachtler, Doi, Lee, & Sejnowski, 2007). This resulted in $11 \cdot 11 \cdot 3 - 1 = 362$ features[3]. Figure 2 shows the linear ICA features obtained from the training image patches.

Like the DoG features from Section 4.1, the ICA feature responses to natural images can be fitted very well using generalized Gaussian distributions, and we obtain the shape and scale parameters for each ICA filter by fitting its response to the ICA training images. The formula for saliency is the same as in Method 1 (equation 17), except that the sum is now over 362 ICA features (rather than 12 DoG features).

Some examples of bottom-up saliency maps computed using the algorithms from Methods 1 and 2 are shown in Figure 3. Each row displays an original test image (from Bruce & Tsotsos, 2006), the same image with human fixations overlaid as red crosses, and the saliency maps on the image computed using Method 1 and Method 2. For Method 1, we applied the DoG filters to 511 × 681 images; for computational efficiency of Method 2, we downsampled the images by a factor of 4 before applying the ICA-derived filters. Figure 3 is included for the purpose of qualitative comparison; the next section provides a detailed quantitative evaluation.

5 Results

5.1 Evaluation method and the center bias

5.1.1 ROC area

Several recent publications (Bruce & Tsotsos, 2006; Harel et al., 2007; Gao & Vasconcelos, 2007; Kienzle, Wichmann, Schlkopf, & Franz, 2007) use the ROC area metric proposed by Tatler, Baddeley, and Gilchrist (2005) to evaluate eye fixation prediction. Using this method, the saliency map is treated as a binary classifier on every pixel in the image; pixels with larger saliency values than threshold are classified as fixated while the rest are classified as non-fixated. Human fixations are used as ground truth. By varying the threshold, an ROC curve can be drawn and the area under the curve indicates how well the saliency map predicts actual human eye fixations. This measurement has the desired characteristic of transformation invariance, in that the area under the ROC curve does not change when applying any monotonically increasing function (such as logarithm) to the saliency measure.

Assessing performance in this manner runs into problems because most human fixation data sets collected with head-mounted eye tracking systems have a strong center bias. This bias is partly due to factors related to the setup of the experiment, such as subjects being centered with respect

[3] The training image patches are considered as $11 \cdot 11 \cdot 3 = 363$-dimensional vectors, z-scored to have zero mean and unit standard deviation, then processed by principal component analysis (where one dimension is lost due to mean subtraction).

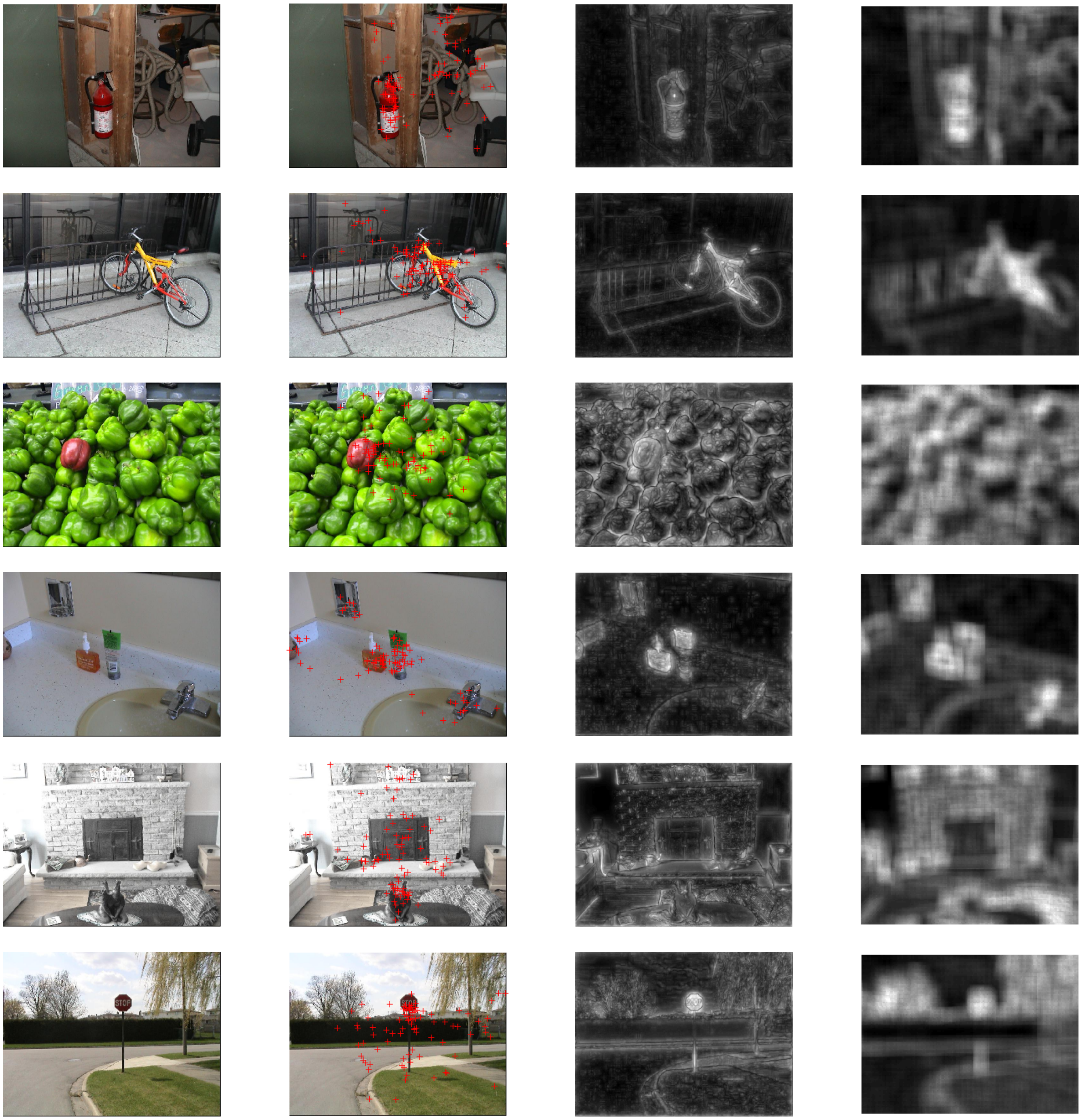

Figure 3: Examples of saliency maps for qualitative comparison. Each row contains, from left to right: An original test image; the same image with human fixations (from Bruce & Tsotsos, 2006) shown as red crosses; the saliency map produced by our SUN algorithm with DoG filters (Method 1); and the saliency map produced by SUN with ICA features (Method 2).

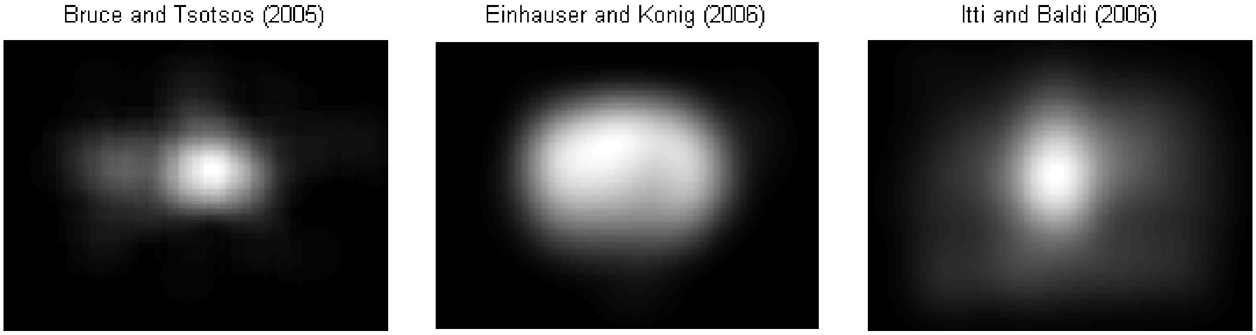

Figure 4: Plots of all human eye fixation locations in three data sets. *Left:* Subjects viewing color images (Bruce & Tsotsos, 2006); *Middle:* Subjects viewing grey images (W. Einhauser & Konig, 2006); *Right:* Subjects viewing color videos (Itti & Baldi, 2006).

to the center of the screen and framing effects caused by the monitor, but also reflects the fact that human photographers tend to center objects of interest (Parkhurst & Niebur, 2003; Tatler et al., 2005). Figure 4 shows the strong center bias of human eye fixations from free-viewing color static images (data from Bruce & Tsotsos, 2006), gray static images (data from W. Einhauser & Konig, 2006) and videos (data from Itti & Baldi, 2006). In fact, simply using a Gaussian blob centered in the middle of the image as the saliency map produces excellent results. For example, on the data set collected in (Bruce & Tsotsos, 2006), a Gaussian blob fitted to the human eye fixations for that set has an ROC area of 0.80, exceeding the reported results of 0.75 (in Bruce & Tsotsos, 2006) and 0.77 (in Gao & Vasconcelos, 2005) on this data set.

5.1.2 KL divergence

Itti and colleagues make use of the Kullback-Leibler (KL) divergence between the histogram of saliency sampled at eye fixations and that sampled at random locations as the evaluation metric for their dynamic saliency (Itti & Baldi, 2005, 2006). If a saliency algorithm performs significantly better than chance, the saliency computed at human-fixated locations should be higher than that computed at random locations, leading to a high KL divergence between the two histograms. This KL divergence, similar to the ROC measurement, has the desired property of transformation invariance—applying a continuous monotonic function (such as logarithm) to the saliency values would not affect scoring (Itti & Baldi, 2006). In (Itti & Baldi, 2005, 2006), the random locations are drawn from a uniform spatial distribution over each image frame. Like the ROC performance measurement, the KL divergence awards excellent performance to a Gaussian blob due to the center bias of the human fixations. The Gaussian blob discussed earlier, trained on the (Bruce & Tsotsos, 2006) data, yields a KL divergence of 0.44 on the data set of Itti and Baldi (2006), exceeding their reported result of 0.24. Thus, both the ROC and KL measurements are strongly sensitive to the effects of the center bias.

5.1.3 Edge effects

These findings imply that models which make use of a location prior (discussed in Section 3) would better model human behavior. Since all of these models (Itti et al., 1998; Itti & Koch, 2000; Bruce & Tsotsos, 2006; Gao & Vasconcelos, 2007) calculate saliency at each pixel without regard to the pixel's location, it would appear that both the ROC measurement and the KL divergence provide a fair comparison between models since no model takes advantage of this additional information.

However, both measures are corrupted by an edge effect due to variations in the handling of invalid filter responses at the borders of images. When an image filter lies partially off the edge of an image, the filter response is not well defined and various methods are used to deal with this problem. Figure 5 shows the average of all of the image saliency maps using each of the algorithms of (Itti & Koch, 2001; Bruce & Tsotsos, 2006; Gao & Vasconcelos, 2007) on the data set of Bruce and Tsotsos (2006). It is clear from Figure 5 that all three algorithms have borders with decreased saliency, but to varying degrees. These border effects introduce an implicit center bias on the saliency maps; "cool borders" result in the bulk of salience being located at the center of the image. Because different models are affected by these edge effects to varying degrees, it is difficult to determine using the previously described measures whether the difference in performance between models is due to the models themselves, or merely due to edge effects[4].

Figure 6 illustrates the impact that varying amounts of edge effects can have on the ROC area evaluation score by examining the performance of dummy saliency maps that are all 1's except for a border of 0's. The map with a four-pixel border yields an ROC area of 0.62, while the map with an eight-pixel border has an area of 0.73. All borders are small relative to the 120 × 160 pixel saliency map and for these measurements, we assume that the border points are never fixated by humans, which corresponds well with actual human fixation data. A dummy saliency map of all 1's with no border has a baseline ROC area of 0.5.

The KL measurement, too, is quite sensitive to how the filter responses are dealt with at the edges of images. Since the human eye fixations are rarely near the edges of the test images, the edge effects primarily change the distribution of saliency of the random samples. For the dummy saliency maps used in Figure 6, the baseline map (of all 1's) gives a KL divergence of 0, the four-pixel-border map gives a KL divergence of 0.12, and the eight-pixel-border map gives a KL divergence of 0.25.

While this dummy example presents a somewhat extreme case, we have found that in comparing algorithms on real data sets (using the ROC area, the KL divergence, and other measures), the differences between algorithms are dwarfed by differences due to how borders are handled.

5.1.4 Eliminating border effects

Parkhurst and Niebur (2003) and Tatler et al. (2005) have pointed out that random locations should be drawn from the distribution of actual human eye fixations. In this paper, we measure the KL divergence between two histograms: the histogram of saliency at the fixated pixels of a test image,

[4] When comparing different feature sets within the same model, edge effects can also make it difficult to assess which features are best to use; larger filters result in a smaller valid image after convolution, which can artificially boost performance.

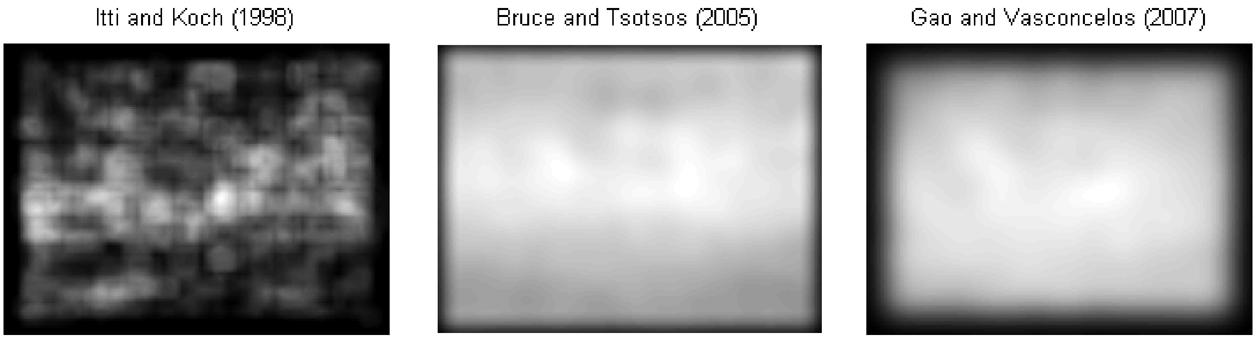

Figure 5: The average saliency maps of three recent algorithms on the stimuli used in collecting human fixation data by Bruce and Tsotsos (2006). Averages were taken across the saliency maps for the 120 color images. The algorithms used are, from left to right: Itti et al. (1998), Bruce and Tsotsos (2006), and Gao and Vasconcelos (2007). All three algorithms exhibit decreased saliency at the image borders, an artifact of the way they deal with filters that lie partially off the edge of the images.

and the histogram of saliency at the same pixel locations but of a randomly chosen image from the test set (effectively shuffling the saliency maps with respect to the images). This method of comparing models has several desired properties. First, it avoids the aforementioned problem that a static saliency map (such as a centered Gaussian blob) can receive a high score even though it is completely independent of the input image. By shuffling the saliency maps, any static saliency map will give a KL divergence of zero—for a static saliency map, shuffling has no effect, and the salience values at the human-fixated pixels are identical to those from the same pixel locations at a random image. Secondly, shuffling saliency maps also diminishes the effect of variations in how borders are handled since few eye fixations are located near the edges.

The potential problem with the shuffling method is that because photos taken by humans are often centered on interesting objects, the center is often genuinely more salient than the periphery. As a result, shuffling saliency maps can bias the random samples to be at more salient locations, which leads to an underestimate of a model's performance (Carmi & Itti, 2006). However, this does not affect the validity of this evaluation measurement for comparing the relative performance of different models, and its properties make for a fair comparison that is free from border effects.

5.2 Performance

We evaluate our bottom-up saliency algorithm on human fixation data from (Bruce & Tsotsos, 2006). Data were collected from 20 subjects free-viewing 120 color images for 4 seconds each. As described in Section 4, we calculated saliency maps for each image using DoG filters (Method 1) and linear ICA features (Method 2). We also obtained saliency maps for the same set of images using the algorithms of Itti et al. (1998, obtained from Bruce and Tsotsos)[5], Bruce and Tsotsos

[5] The saliency maps that produce the score for Itti et al. in Table 1 come from Bruce and Tsotsos (2006) and were calculated using the online Matlab saliency toolbox (http://www.saliencytoolbox.net/index.html) using the parameters

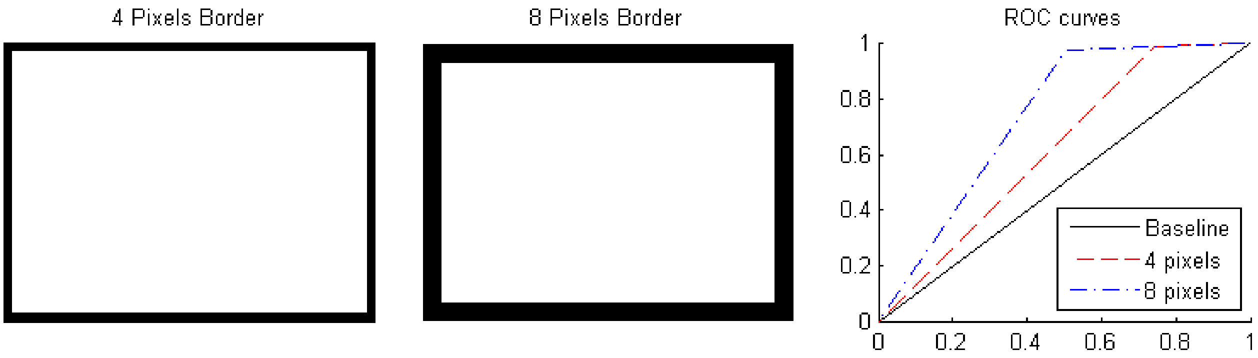

Figure 6: Illustration of edge effects on performance. *Left:* A saliency map of size 120 × 160 that consists of all 1's except for a four-pixel-wide border of 0's. *Center:* A saliency map of size 120 × 160 that consists of all 1's except for an eight-pixel-wide border of 0's. *Right:* The ROC curves of these two dummy saliency maps, as well as for a baseline saliency map (all 1's). The ROC areas for these two curves are 0.62 and 0.73, respectively. (The baseline ROC area is 0.5.)

(2006, implemented by the original authors)[6], and Gao and Vasconcelos (2007, implemented by the original authors). The performance of these algorithms evaluated using the KL measure described in Section 5.1.4 is summarized in Table 1. For each algorithm, the shuffling of the saliency maps is repeated 100 times. Each time, KL divergence is calculated between the histograms of unshuffled saliency and shuffled saliency on human fixations. The mean and the standard errors are reported in the table.

Table 1: Performance in predicting human eye fixations when viewing color images. Comparison of our SUN algorithm (Method 1 using DoG filters and Method 2 using linear ICA features) with previous algorithms. The KL divergence metric measures the divergence between the saliency distributions at human fixations and at randomly shuffled fixations (see text for details); higher values therefore denote better performance.

Model	**KL (std. error)**
Itti et al. (1998)	0.1130 (0.00115)
Bruce and Tsotsos (2006)	0.2029 (0.00173)
Gao and Vasconcelos (2007)	0.1535 (0.00155)
SUN: Method 1 (DoG filters)	0.1723 (0.0122)
SUN: Method 2 (linear ICA filters)	0.2097 (0.00157)

The results show that SUN with DoG filters (Method 1) significantly outperforms Itti and

that correspond to (Itti et al., 1998). Using the default parameters of this online toolbox generates inferior binary-like saliency maps that give a KL score of 0.1095 (0.00140).

[6] The results reported in (Bruce & Tsotsos, 2006) used ICA features of size 7 × 7. The results reported here, obtained from Bruce and Tsotsos, used features of size 11 × 11, which the authors say achieved better performance.

Koch's algorithm ($p < 10^{-57}$) and Gao and Vasconcelos' algorithm ($p < 10^{-14}$), where significance was measured with a two-tailed t-test over different random shuffles. Between Method 1 (DoG features) and Method 2 (ICA features), the ICA features work significantly better ($p < 10^{-32}$). There are further advantages to using ICA features: efficient coding has been proposed as one of the fundamental goals of the visual system (Barlow, 1994), and linear ICA has been shown to generate receptive fields akin to those found in primary visual cortex (V1) (Olshausen & Field, 1996; Bell & Sejnowski, 1997). In addition, generating the feature set using natural image statistics means that both the feature set and the distribution over features can be calculated simultaneously. However, it is worth noting that the online computations for Method 1 (using DoG features) take significantly less time since only 12 DoG features are used, compared to 362 ICA features in Method 2. There is thus a trade off between efficiency and performance in our two methods.

SUN with linear ICA features (Method 2) performs significantly better than Bruce and Tsotsos' algorithm ($p = 0.0035$) on this data set, though the KL divergence scores are numerically quite close. This similarity in performance is not surprising, for two reasons. First, since both algorithms construct their feature sets using ICA, the feature sets are qualitatively similar. Second, although SUN uses the statistics learned from a training set of natural images whereas Bruce and Tsotsos (2006) calculate these statistics using only the current test image, the response distribution for a low-level feature on a single image of a complex natural scene will generally be close to overall natural scene statistics. In any case, SUN is more efficient than the algorithm of Bruce and Tsotsos (2006). The reason is that in our SUN algorithm, the probability distributions of features are precomputed offline from the training set, while in their algorithm the probability distributions have to be estimated for every image.

6 Discussion

In this paper, we have derived a theory of saliency from the simple assumption that a goal of the visual system is to find potential targets such as prey and predators. Based on a probabilistic description of this goal, we proposed that bottom-up saliency is the self-information of visual features and that overall saliency is the pointwise mutual information between the visual features and the desired target. Here, we have focused on the bottom-up component. The use of self-information as a measure of bottom-up saliency provides a surface similarity between our SUN model and some existing models (Bruce & Tsotsos, 2006; Oliva et al., 2003; Torralba et al., 2006), but this belies fundamental differences between our approach and theirs. In Section 6.1, we explain that the core motivating intuitions behind SUN lead to a use of different statistics, which better accounts for a number of human search assymetries.

6.1 Test image statistics vs. natural scene statistics

6.1.1 Comparison with previous work

All of the existing bottom-up saliency models described in Section 2 compute saliency by comparing the feature statistics at a point in a test image with either the statistics of a neighborhood of the point or the statistics of the entire test image. When calculating the saliency map of an image, these models only consider the statistics of the current test image. In contrast, SUN's definition of saliency (derived, as explained in Section 3, from a simple intuitive assumption about a goal of the visual system) compares the features observed at each point in a test image to the statistics of natural scenes. An organism would learn these natural statistics through a lifetime of experience with the world; in the SUN algorithm, we obtained them from a collection of natural images (see Section 4).

SUN's formula for bottom-up saliency is similar to the one in the work of Oliva and colleagues (Oliva et al., 2003; Torralba et al., 2006) and the one in (Bruce & Tsotsos, 2006) in that they are all based on the notion of self-information. However, the differences between image statistics and natural statistics lead to radically different kinds of self-information. Briefly, the motivation for using self-information with current image statistics is that a foreground object is likely to have features that are distinct from the features of the background. The idea that the saliency of an item is dependent on its deviation from the average statistics of the image can find its roots in the visual search model proposed in (Rosenholtz, 1999), which accounted for a number of motion pop-out phenomena, and can be seen as a generalization of the center-surround-based saliency found in Koch and Ullman (1985). SUN's use of natural statistics for self-information, on the other hand, corresponds to the intuition that since targets are observed less frequently than background during an organism's lifetime, rare features are more likely to indicate targets. The idea that infrequent features attract attention has its origin in findings that novelty attracts the attention of infants (Fantz, 1964; Caron & Caron, 1968; Fagan, 1970; Friedman, 1972) and that novel objects are faster to find in visual search tasks (see (Wolfe, 2001) for a review). This fundamental difference in motivation between SUN and existing saliency models leads to very different predictions about what attracts attention.

In Section 6.1.2, below, we show that by using natural image statistics, SUN provides a simple explanation for a number of psychophysical phenomena that are difficult to account for using the statistics of either a local neighborhood in the test image or the entire test image. In addition, since natural image statistics are computed well in advance of the test image presentation, in the SUN algorithm the estimation of saliency is strictly local and efficient.

6.1.2 Visual search asymmetry

When the probability of a feature is based on the distribution of features in the current test image, as in previous saliency models, a straightforward consequence is that if all items in an image are identical except for one, this odd item will have the highest saliency and thus attract attention. For example, if an image consists of a number of vertical bars with one bar that is slightly tilted from the vertical, the tilted bar "pops out" and attracts attention almost instantly (Treisman & Gormican, 1988); see Figure 7, left, for an illustration. If, on the other hand, an image consists of a number of slightly-tilted-from-vertical bars with one vertical bar, saliency based on the statistics of the current image predicts the same pop-out effect for the vertical bar. However, this simply is not the case, as humans do not show the same pop-out effect: it requires more time and effort for humans to find a vertical bar within a sea of tilted bars (Treisman & Gormican, 1988); see Figure 7, right, for an illustration. This is known in the visual search literature as a *search asymmetry*, and this

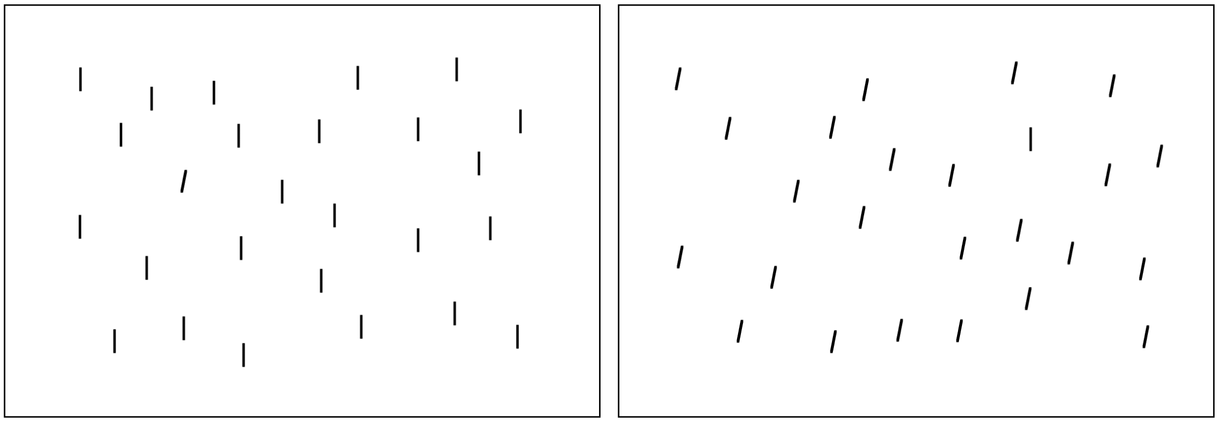

Figure 7: Illustration of the “prototypes do not pop out” visual search asymmetry (Treisman & Gormican, 1988). *Left:* A tilted bar in a sea of vertical bars pops out—the tilted bar can be found almost instantaneously. *Right:* A vertical bar in sea of tilted bars does not pop out. The bar with the odd orientation in this case requires more time and effort for subjects to find than in the case illustrated in the image on the left.

particular example corresponds to findings that “prototypes do not pop out” because the vertical is regarded as a prototypical orientation (Treisman & Souther, 1985; Treisman & Gormican, 1988; Wolfe, 2001).

Unlike saliency measures based on the statistics of the current image or a neighborhood in the current image, saliency based on natural statistics readily predicts this search asymmetry. The vertical orientation is prototypical because it occurs more frequently in natural images than the tilted orientation (Schaaf & Hateren, 1996). As a result, the vertical bar will have smaller salience than the surrounding tilted bars, so it will not attract attention as strongly.

Another visual search asymmetry exhibited by human subjects involves long and short line segments. Saliency measures based on test image statistics or local neighborhood statistics predict that a long bar in a group of short bars (illustrated on the left in Figure 8) should be as salient as a short bar in a group of long bars (illustrated on the right in Figure 8). However, it has been shown that humans find a long bar among short bar distractors much more quickly than they find a short bar among long bars (Treisman & Gormican, 1988). Saliency based on natural statistics readily predicts this search asymmetry, as well. Due to scale invariance, the probability distribution over the lengths of line segments in natural images follows the power law (Ruderman, 1994). That is, the probability of the occurrence of a line segment of length v is given by $p(V = v) \propto \frac{1}{v}$. Since longer line segments have lower probability in images of natural scenes, the SUN model implies that longer line segments will be more salient.

Visual search asymmetry is also observed for higher-level stimuli such as roman letters, Chinese characters, animal silhouettes, and faces. For example, people are faster to find a mirrored letter in normal letters than the reverse (Frith, 1974). People are also faster at searching for an inverted animal silhouette in a sea of upright silhouettes than the reverse (Wolfe, 2001), and faster

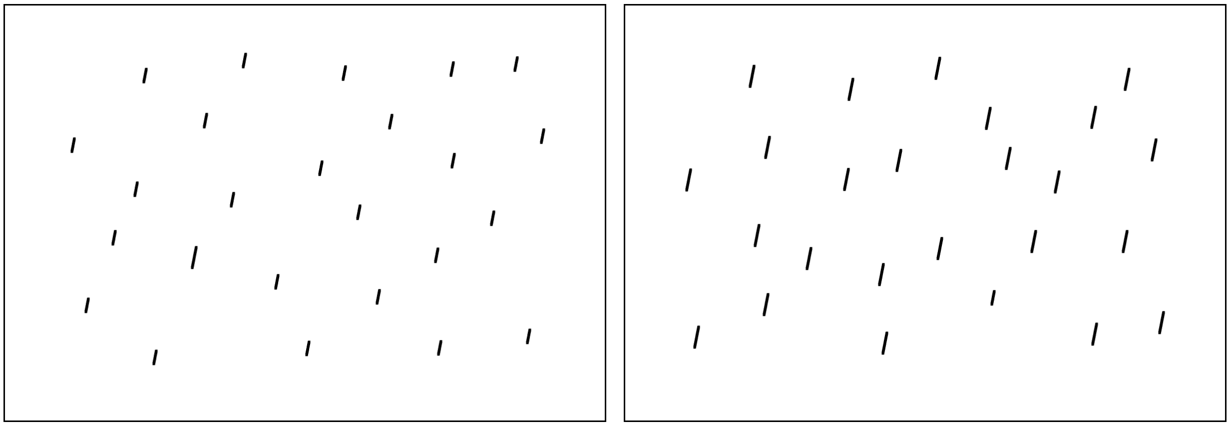

Figure 8: Illustration of a visual search asymmetry with line segments of two different lengths (Treisman & Gormican, 1988). *Left:* A long bar is easy to locate in a sea of short bars. *Right:* A short bar in a sea of long bars is harder to find.

at searching for an inverted face in a group of upright faces than the reverse (Nothdurft, 1993). These phenomena have been referred to as "the novel target is easier to find." Here, "novel" means that subjects have less experience with the stimulus, indicating a lower probability of encounter during development. This corresponds well with SUN's definition of bottom-up saliency, as items with novel features are more salient by definition.

If the saliency of an item depends upon how often it has been encountered by an organism, then search asymmetry should vary among people with different experience with the items involved. This seems to indeed be the case. Modified/inverted Chinese characters in a sea of real Chinese characters are faster to find than the reverse situation for Chinese readers, but not for non-Chinese readers (Wang, Cavanagh, & Green, 1994; Shen & Reingold, 2001). Levin found an "other-race advantage" as American Caucasians are faster to search for an African-American face among Caucasian faces than to search for a Caucasian face among African-American faces (Levin, 1996). This is consistent with what SUN would predict for American Caucasian subjects that have more experience with Caucasian faces than with African-American faces. In addition, Levin found that Caucasian basketball fans who are familiar with many African-American basketball players do not show this other-race search advantage (Levin, 2000). These seem to provide direct evidence that experience plays an important role in saliency (Zhang et al., 2007), and the statistics of the current image alone cannot account for these phenomena.

6.2 Efficiency comparison with existing saliency models

Table 2 summarizes some computational components of several algorithms discussed in Section 2 and compared in Section 5. Computing feature statistics in advance using a data set of natural images allows the SUN algorithm to compute saliency quickly compared with algorithms that require calculations of statistics on the current image. In addition, SUN requires strictly local

Table 2: Some computational components of the algorithms. Notably, our SUN algorithm requires only offline probability distribution estimation and no global computation over the test image in calculating saliency.

Model	**Statistics calculated using**	**Global operations**	**Statistics calculated on image**
Itti et al. (1998)	N/A	Sub-map normalization	None
Bruce and Tsotsos (2006)	Current image	Probability estimation	Once for each image
Gao and Vasconcelos (2007)	Local region of current image	None	Twice for each pixel
SUN	Training set of natural images (pre-computed offline)	None	None

operations, which is consistent with implementation in the low levels of the visual system.

6.3 Higher-order features

The range of visual search asymmetry phenomena described in Section 6.1.2 seem to suggest that the statistics of observed visual features are estimated by the visual system at many different levels, including basic features such as color and local orientation as well as higher-level features. The question of exactly what feature set is employed by the visual system is beyond the scope of this paper. In the current implementation of SUN, explained in Section 4, we only consider linear filter responses as features for computational efficiency. This use of linear features (DoG or linear ICA features) causes highly-textured areas to have high saliency, a characteristic shared with complexity-based algorithms (Yamada & Cottrell, 1995; Chauvin et al., 2002; Kadir & Brady, 2001; Renninger et al., 2004). In humans, however, it is often not the texture itself but a change in texture that attracts attention. Saliency algorithms that use local region statistics, such as (Gao & Vasconcelos, 2007), address this problem explicitly.

Our SUN model could resolve this problem implicitly by using a non-linear feature space. Whereas linear ICA features learned from natural images respond to discontinuities in illumination or color, higher-order non-linear ICA features are found to respond to discontinuity in textures (Shan, Zhang, & Cottrell, 2007; Karklin & Lewicki, 2003; Osindero, Welling, & Hinton, 2005). Figure 9 shows an image of birthday candles, the response of a linear DoG filter to that image, and the response of a non-linear feature inspired by the higher-order ICA features learned in (Shan et al., 2007). Perceptually, the white hole in the image attracts attention (Bruce & Tsotsos, 2006). Whereas the linear feature has zero response to this hole, the higher-order feature responds strongly

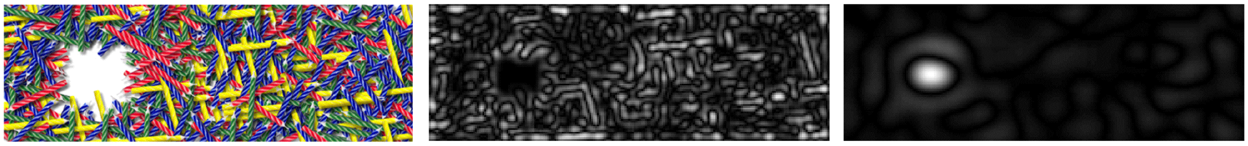

Figure 9: Demonstration that non-linear features could capture discontinuity of textures without using a statistical model that explicitly measures the local statistics. *Left:* The input image, adapted from (Bruce & Tsotsos, 2006). *Middle:* The response of a linear DoG filter. *Right:* The response of a non-linear feature. The non-linear feature is constructed by applying a DoG filter, then non-linearly transforming the output before another DoG is applied. (See (Shan et al., 2007) for details on the non-linear transformation.) Whereas the linear feature has zero response to the white hole in the image, the non-linear feature responds strongly in this region, consistent with the white region's perceptual salience.

in this region. We will explore the use of such features in future work.

6.4 Conclusion

Based on the intuitive assumption that a goal of the visual system is to find potential targets, we derived a definition of saliency in which overall visual saliency is the pointwise mutual information between the observed visual features and the presence of a target, and bottom-up saliency is the self-information of the visual features. Using this definition, we developed a simple algorithm for bottom-up saliency that can be expressed in a single equation (17). We applied this algorithm using two different set of features, difference of Gaussians (DoG) and ICA-derived features, and compared the performance to several existing bottom-up saliency algorithms. Not only does SUN perform as well as or better than the state-of-the-art algorithms, but it is also more computationally efficient. In its use of self-information to measure bottom-up saliency, SUN is similar to the algorithms in (Oliva et al., 2003; Torralba et al., 2006; Bruce & Tsotsos, 2006), but stems from a different set of intuitions and is calculated using different statistics. In SUN, the probability distribution over features is learned from natural statistics (which corresponds to an organism's visual experience over time), whereas the previous saliency models compute the distribution over features from each individual test image. We explained that several search asymmetries that may pose difficulties for models based on test image statistics can be accounted for when feature probabilities are obtained from natural statistics.

In future work, we intend to incorporate the higher-level non-linear features discussed in Section 6.3. In addition, our definition of overall saliency includes a top-down term that captures the features of a target. Although this is beyond the scope of the present paper, we plan to examine top-down influences on saliency in future research; preliminary work with images of faces shows promise. We also plan to extend the implementation of SUN from static images into the domain of video.

Acknowledgements

The authors would like to thank Neil D. Bruce and John K. Tsotsos for sharing their human fixation data and simulation results; Dashan Gao and Nuno Vasconcelos for sharing their simulation results; Itti and Koch for sharing their human fixation data; and Einhauser and colleagues for sharing their human fixation data. We would also like to thank Dashan Gao, Dan N. Hill, Piotr Dollar, and Michael C. Mozer for helpful discussions, and everyone in GURU (Gary's Unbelievable Research Unit) and PEN (Perceptual Expertise Network) for comments. This work is supported by NIMH grant MH57075 to GWC, and NSF Science of Learning Center grant to the Temporal Dynamics of Learning Center #SBE-0542013 (GWC, PI).

References

Barlow, H. (1994). What is the computational goal of the neocortex? In C. Koch (Ed.), *Large scale neuronal theories of the brain* (pp. 1–22). Cambridge, MA: MIT Press.

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

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.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*(4), 523–547.

Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *Journal of Vision*, *6*(9), 898–914.

Caron, R. F., & Caron, A. J. (1968). The effects of repeated exposure and stimulus complexity on visual fixation in infants. *Psychonomic Science*, *10*, 207–208.

Chauvin, A., Herault, J., Marendaz, C., & Peyrin, C. (2002). Natural scene perception: visual attractors and image processing. *Progress in Neural Processing*, 236–248.

Fagan, J. F. (1970). Memory in the infant. *Journal of Experimental Child Psychology*, *9*, 217–226.

Fantz, R. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, *146*(3644), 668.

Friedman, S. (1972). Habituation and recovery of visual response in the alert human infant. *Journal of Experimental Child Psychology*, *13*, 339–349.

Frith, U. (1974). A curious effect with reversed letters explained by a theory of schema. *Perception and Psychophysics*, *16*, 113–116.

Gao, D., & Vasconcelos, N. (2004). 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. (2005). Integrated learning of saliency, complex features, and object detectors from cluttered scenes. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - volume 2* (pp. 282–287). Washington, DC, USA: IEEE Computer Society.

Gao, D., & Vasconcelos, N. (2007). Bottom-up saliency is a discriminant process. In *Ieee international conference on computer vision.* Rio de Janeiro, Brazil.

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In *Advances in Neural Information Processing Systems 19.* Cambridge, MA: MIT Press.

Hyvarinen, A., & Oja, E. (1997). A fast fixed-point algorithm for independent component analysis. *Neural Computation*, *9*(7), 1483-1492.

Itti, L., & Baldi, P. (2005). A principled approach to detecting surprising events in video. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - volume 1* (pp. 631–637). Washington, DC, USA: IEEE Computer Society.

Itti, L., & Baldi, P. (2006). Bayesian surprise attracts human attention. In Y. Weiss, B. Schölkopf, & J. Platt (Eds.), *Advances in Neural Information Processing Systems 18* (p. 1-8). Cambridge, MA: MIT press.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10-12), 1489-1506.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience*, *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.

Kadir, T., & Brady, M. (2001). Saliency, scale and image description. *International Journal of Computer Vision*, *45*(2), 83-105.

Karklin, Y., & Lewicki, M. (2003). Learning higher-order structures in natural images. *Network: Computation in Neural Systems*, *14*(3), 483–499.

Kienzle, W., Wichmann, F. A., Schlkopf, 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.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, *4*(4), 219–27.

Levin, D. (1996). Classifying faces by race: The structure of face categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1364–1382.

Levin, D. (2000). Race as a visual feature: Using visual search and perceptual discrimination tasks to understand face categories and the cross-race recognition deficit. *Journal of Experimental Psychology: General*, *129*(4), 559–574.

Nothdurft, H. (1993). Faces and facial expressions do not pop out. *Perception*, *22*(11), 1287-1298.

Oliva, A., Torralba, A., Castelhano, M., & Henderson, J. (2003). Top-down control of visual attention in object detection. In *Proceedings of International Conference on Image Processing* (p. 253-256). Barcelona, Catalonia: IEEE press.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Osindero, S., Welling, M., & Hinton, G. (2005). Topographic product models applied to natural scene statistics. *Neural Computation*, *18*(2), 381–414.

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

Parkhurst, D., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision*, *16*(2), 125-154.

Posner, M. I., & Cohen, Y. (1984). Components of attention. In H. Bouma & D. G. Bouwhuis (Eds.), *Attention and Performance X* (p. 55-66). Erlbaum.

Rao, R., Zelinsky, G., Hayhoe, M., & Ballard, D. (2002). Eye movements in iconic visual search. *Vision Research*, *42*(11), 1447–1463.

Renninger, L. W., Coughlan, J. M., Verghese, P., & Malik, J. (2004). An information maximization model of eye movements. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17* (p. 1121-1128). Cambridge, MA: MIT Press.

Rosenholtz, R. (1999). A simple saliency model predicts a number of motion popout phenomena. *Vision Res*, *39*(19), 3157–63.

Ruderman, D. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, *5*(4), 517–548.

Schaaf, A. van der, & Hateren, J. H. van. (1996). Modelling the power spectra of natural images: statistics and information. *Vision Research*, *36*(17), 2759–70.

Shan, H., Zhang, L., & Cottrell, G. W. (2007). Recursive ICA. In B. Schölkopf, J. Platt, & T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19* (pp. 1273–1280). Cambridge, MA: MIT Press.

Shen, J., & Reingold, E. (2001). Visual search asymmetry: The influence of stimulus familiarity and low-level features. *Perception & Psychophysics*, *63*(3), 464–75.

Song, K. (2006). A globally convergent and consistent method for estimating the shape parameter of a generalized gaussian distribution. *IEEE Transactions on Information Theory*, *52*(2), 510-527.

Tatler, B., Baddeley, R., & Gilchrist, I. (2005). Visual correlates of fixation selection: effects of scale and time. *Vision Research*, *45*(5), 643–59.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological Review*, *113*(4), 766–786.

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

Treisman, A., & Gormican, S. (1988). Feature analysis in early vision: evidence from search asymmetries. *Psychological Review*, *95*(1), 15-48.

Treisman, A., & Souther, J. (1985). Search asymmetry: a diagnostic for preattentive processing of separable features. *Journal Experimental Psychology: General*, *114*(3), 285-310.

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. *Nature Neuroscience*, *5*(7), 682-687.

Wachtler, T., Doi, E., Lee, T., & Sejnowski, T. (2007). Cone selectivity derived from the responses of the retinal cone mosaic to natural scenes. *Journal of Vision*, *7*(8), 1-14.

Wainwright, M., Schwartz, O., & Simoncelli, E. (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. 203–22). MIT Press.

Wang, Q., Cavanagh, P., & Green, M. (1994). Familiarity and pop-out in visual search. *Perception & Psychophysics*, *56*(5), 495–500.

W. Einhauser, K.-P. H., W. Kruse, & Konig, P. (2006). Differences of monkey and human overt attention under natural conditions. *Vision Research*, *46*(8-9), 1194–1209.

Wolfe, J. (2001). Asymmetries in visual search: An introduction. *Perception & Psychophysics*, *63*(3), 381–389.

Wolfe, J., Cave, K., & Franzel, S. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 419-33.

Yamada, K., & Cottrell, G. W. (1995). A model of scan paths applied to face recognition. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (p. 55-60). Pittsburgh, PA: Mahwah: Lawrence Erlbaum.

Zhang, L., Tong, M. H., & Cottrell, G. W. (2007). Information attracts attention: a probabilistic account of the cross-race adavantage in visual search. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (p. 749-754). Nashville, Tennessee.