Center-surround patterns emerge as optimal predictors for human saccade targets

Wolf Kienzle
Max Planck Institute for Biological Cybernetics, Empirical Inference Department, Tübingen, Germany

Matthias O. Franz
Max Planck Institute for Biological Cybernetics, Empirical Inference Department, Tübingen, Germany, &
University of Applied Sciences, Cognitive Systems Group, Konstanz, Germany

Bernhard Schölkopf
Max Planck Institute for Biological Cybernetics, Empirical Inference Department, Tübingen, Germany

Felix A. Wichmann
Max Planck Institute for Biological Cybernetics, Empirical Inference Department, Tübingen, Germany,
Technical University of Berlin, Modelling of Cognitive Processes Group, Berlin, Germany, &
Bernstein Center for Computational Neuroscience, Berlin, Germany

The human visual system is foveated, that is, outside the central visual field resolution and acuity drop rapidly. Nonetheless much of a visual scene is perceived after only a few saccadic eye movements, suggesting an effective strategy for selecting saccade targets. It has been known for some time that local image structure at saccade targets influences the selection process. However, the question of what the most relevant visual features are is still under debate. Here we show that center-surround patterns emerge as the optimal solution for predicting saccade targets from their local image structure. The resulting model, a one-layer feed-forward network, is surprisingly simple compared to previously suggested models which assume much more complex computations such as multi-scale processing and multiple feature channels. Nevertheless, our model is equally predictive. Furthermore, our findings are consistent with neurophysiological hardware in the superior colliculus. Bottom-up visual saliency may thus not be computed cortically as has been thought previously.

Keywords: visual saliency, eye movements, receptive field analysis, classification images, kernel methods, support vector machines, natural scenes

Citation: Kienzle, W., Franz, M. O., Schölkopf, B., & Wichmann, F. A. (2009). Center-surround patterns emerge as optimal predictors for human saccade targets. *Journal of Vision, 9*(5):7, 1–15, http://journalofvision.org/9/5/7/, doi:10.1167/9.5.7.

Introduction

The human visual system scans the world by directing the center of gaze from one location to another via rapid eye movements, called saccades. In the period between saccades the direction of gaze is held fixed for a few hundred milliseconds (fixations). It is principally during fixations that information enters the visual system (Matin, 1974). Remarkably, however, we constantly perceive a coherent, high-resolution scene despite the visual acuity of the eye quickly decreasing away from the center of gaze. This suggests that saccade targets are not chosen at random, but according to an effective strategy.

It has been known for a long time that cognitive, or top-down effects, such as the observer's task, thoughts, or intentions have an effect on saccadic selection (Hopfinger, Buonocore, & Mangun, 2000; Oliva, Torralba, Castelhano, & Henderson, 2003; Yarbus, 1967). Another well-known fact is that the incoming image itself can have properties that attract the saccadic system. As an example, a bright spot in a dark scene is likely to attract our attention, regardless of top-down effects. Today there is a considerable amount of evidence that such bottom-up cues

doi: 10.1167/9.5.7 Received January 21, 2008; published May 13, 2009 ISSN 1534-7362 © ARVO

influence saccadic targeting (Baddeley & Tatler, 2006; Bruce & Tsotsos, 2006; Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; Li, 2002; Mannan, Ruddock, & Wooding, 1997; Parkhurst, Law, & Niebur, 2002; Parkhurst & Niebur, 2003; Privitera & Stark, 2000; Raj, Geisler, Frazor, & Bovik, 2005; Rajashekar, Cormack, Bovik, & Geisler, 2002; Reinagel & Zador, 1999; Renninger, Coughlan, Verghese, & Malik, 2005; Tatler, Baddeley, & Gilchrist, 2005), for reviews see Henderson (2003), Itti and Koch (2001), and Krauzlis, Liston, and Carello (2004). A prominent study is that of Reinagel and Zador (1999), who showed that the local contrast (i.e., the local standard deviation of intensities) tends to be larger at the center of gaze. Krieger et al. (2000) found regularities also in higher order statistics. Many studies established connections to the underlying physiology by assuming biologically plausible image filters (Baddeley & Tatler, 2006; Bruce & Tsotsos, 2006; Itti, Koch, & Niebur, 1998; Tatler et al., 2005) and using statistical tests to prove their relevance. Perhaps the most popular biologically inspired model is due to Itti et al. (1998), which combines contrast, orientation, and color features, as suggested in Koch and Ullman (1985). Parkhurst et al. (2002) tested this model against real eye movement data and found that it is capable of explaining a significant amount of the variance in fixation locations.

Much of the ongoing work in the field is devoted to improving the predictivity of existing models by extending them in various directions, e.g., by modeling the influence of global scene statistics (Harel, Koch, & Perona, 2007; Peters, Iyer, Itti, & Koch, 2005; Torralba, Oliva, Castelhano, & Henderson, 2006). Unfortunately, despite the variety of existing models—or perhaps because of it—a precise description of the typical spatial structure of saccade targets remains elusive. This is partly due to the fact that most plausible image features are correlated with each other (in particular, with contrast), and so, with enough test examples, many different models can be shown to have a significant effect on saccadic targeting.

In this work we derive a description of typical saccade targets directly from eye movement data via system identification. Unlike previous studies where the relevant structure is determined manually—e.g. selecting Gabors as visual filters—we do not make any assumptions in this regard, but numerically infer them from data. This approach is more common in neurophysiology when modeling response properties of neurons (Victor, 2005). There, a generic model (e.g., linear or quadratic) is fitted to experimental spike data such that it describes the stimulus-response relationship of the neuron. Insight about the neuron's functionality is gained by analyzing the fitted model in terms of relevant input patterns. Here, we apply the same idea but to saccades instead of spikes: we model the relationship between spatial intensity patterns in natural images and the response of the saccadic system. This allows us to identify the most relevant image patterns that guide the bottom-up component of the saccadic selection system, which we refer to here as *perceptive fields*. Perceptive fields are analogous to receptive fields but at the psychophysical level (Jung & Spillmann, 1970; Neri & Levi, 2006; Wichmann, Graf, Simoncelli, Bülthoff, & Schölkopf, 2005).

The main result of this work is to show that the perceptive fields of the saccadic targeting system are simple center-surround patterns of a single spatial scale, and that a very simple model with these receptive fields has the same predictive power as much more complicated models. Besides the simplicity of our model, a substantial difference from previous results lies in the fact that our model emerges from data through numerical optimization, instead of being designed by hand. Thus, it provides for the first time a view of the saccadic selection system which is unbiased with respect to design choices that are otherwise unavoidable.

Methods

Analysis overview

The aim of this work is to find the functional relationship between the appearance of a local image region and its visual saliency, i.e., how likely it is to become the target of a saccadic eye movement during free-viewing. The basic idea of our approach to this problem is that of fitting a linear model $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ to eye movement data, i.e., such that $f(\mathbf{x})$ describes the visual saliency of a local image region whose visual appearance is represented by a vector $\mathbf{x} \in \mathbb{R}^n$ (here, $\mathbf{x}$ holds the pixel luminances from an image region). The fitted weights $\mathbf{w}$ represent exactly the optimal stimulus of the model (here, a luminance pattern), and can therefore be interpreted as the characteristic visual pattern that drives visual saliency.

This analysis is more generally known as system identification and is commonly used for identifying biological systems (Victor, 2005), e.g., neuronal receptive fields (Jones & Palmer, 1987) ("reverse correlation"). It is also used in psychophysics, under the name of "classification images." For example, Rajashekar, Bovik, and Cormack (2006) used this approach to identify structural cues that human observers use to perform visual search, e.g., they had a subject search for a horizontal edge in noise and found that the classification image (the weights w of a fitted linear model) is also a horizontal edge. The authors then concluded that during the search, the saccadic targeting system was driven by the identified edge pattern. In this work we want to arrive at a similar, but more general result, namely we want to identify the characteristic luminance patterns for a *free-viewing* task on *natural images*. In other words, we want to find characteristic patterns that drive bottom-up visual saliency.

Unfortunately, unlike for specific search tasks as in Rajashekar et al. (2006), a linear model is not appropriate for describing visual saliency. One reason for this is that linear models cannot describe systems that yield a positive response to both a pattern $\mathbf{x}_0$ and to its inverse $-\mathbf{x}_0$, since $f(\mathbf{x}_0) = -f(-\mathbf{x}_0)$. This property is extremely restrictive. As an example, if we merely extend the target in the visual search task from Rajashekar et al. (2006) from a horizontal edge to that same edge with both polarities allowed (i.e., also upside down), a linear model would not be valid anymore, since its output on a horizontal edge is exactly the negative of the output on the same edge upside down. In practice, if a linear model is fitted to such data, the complementary data samples will essentially cancel each other, resulting in an unstructured, or at least very noisy classification image.

Here, we therefore use a nonlinear generalization of the linear approach: we fit a nonparametric model $f(\mathbf{x}) = \sum \alpha_i \varphi_i(\mathbf{x})$, where $\mathbf{x}$ is an image patch and φ_i are nonlinear basis functions. In this model, the fitted parameters are the weights α_i. An advantage of this approach to nonlinearity is that the model is still linear in the fitted parameters (α_i), and yet implements a nonlinear relationship through its nonlinear basis functions $\varphi_i(\mathbf{x})$.

In this work we use Gaussian radial basis functions $\varphi_i(\mathbf{x}) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}||^2)$, centered at examples of recorded salient and non-salient image regions $\mathbf{x}_i$. In this case, our nonparametric model takes the form of a *support vector machine* (Schölkopf & Smola, 2002). This particular choice of nonlinearity brings two advantages. First, the resulting model is very general in the sense that it is able to capture stimulus-response relationships of any order. Second, Gaussian radial basis functions satisfy a positive definiteness property (see Extracting perceptive fields section), which means that the resulting nonlinear analysis is indeed a straightforward generalization of the traditional linear approach. More precisely, we show below that due to this property the concept of the optimal stimulus being just the weight vector $\mathbf{w}$ in a dot-product with the input ($\mathbf{w}^\top \mathbf{x}$) directly translates to the nonlinear case.

The proposed approach is summarized by the cartoon in Figure 1: to see this, consider the space of image patches, i.e. the vector space in which each dimension corresponds to a pixel value (luminance) within an image patch. For illustration purposes, we assume that there are only two pixels in a patch, and that the image plane is the space of image patches. Now, the dots in panel (a) denote the recorded examples x_i of salient (white) and non-salient

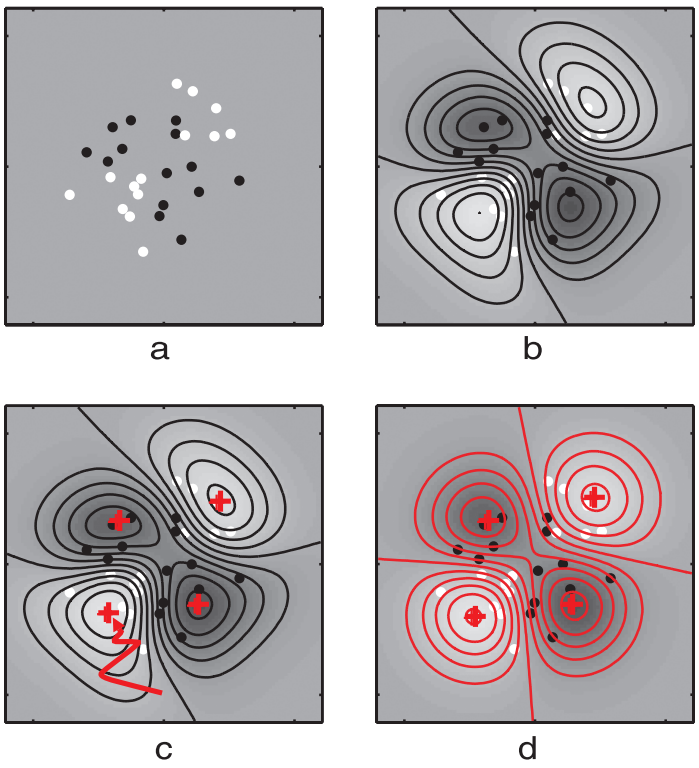

Figure 1. A cartoon illustrating the computation of nonlinear perceptive fields. (a) denotes saccade targets (white dots) and non-targets (black dots) in the space of image patches (here, the image plane). (b) shows the fitted kernel model (black lines are level curves, brighter areas denote higher target probability), a weighted sum of Gaussian radial basis functions, centered at the data points. (c) the perceptive fields are the maximally excitatory and inhibitory stimuli (red plus signs). They are found by gradient search (red zigzag line). (d) the saliency model: a Gaussian radial basis function is placed at each of the four perceptive fields, resulting in the function represented by the red level curves.

(black) image regions, i.e. two-dimensional vectors that describe the local luminance values at locations in natural scenes where people did (white) or did not (black) look in our experiment. The initial step of our analysis consists of fitting a nonlinear real-valued function $f(\mathbf{x})$ to the data, such that $f(\mathbf{x})$ takes on larger values if $\mathbf{x}$ is salient, and smaller (negative) values otherwise. In this paper, we use a nonparametric model for f taking the form of a sum of weighted Gaussian bumps $\varphi_i(\mathbf{x}) = \exp(-\gamma||\mathbf{x}_i - \mathbf{x}||^2)$. The fitted model is illustrated in panel (b) by the black level curves and shaded background. Note that there are four extremal points in this example (denoted by the red plus signs in panel (c)), namely two maxima and two minima. The image patches at these locations correspond to the optimal stimuli of the fitted model f (and hopefully of the underlying system as well), since it is at these locations where the value of f, the saliency, is either extremely high or low. The key step in our analysis is that from the fitted nonlinear model f, the optimal stimuli can be determined via gradient search (red zigzag line). In this paper we refer to the optimal stimuli as the *nonlinear perceptive fields,* stressing the fact that these luminance patterns are similar to receptive fields, but stem from a psychophysical experiment, not from neurophysiological recordings. The cartoon here is realistic in the sense that f has two maxima and two minima. This is also true for our actual data set, the four perceptive fields are shown in Figure 3. Our analysis concludes with the proposition of a simple bottom-up saliency model based on this result (Figure 4), i.e., a radial basis function network with only four basis functions, centered on the perceptive fields (red level curves in Figure 1, panel d). We show that this simple model, being purely data-driven, is as predictive as the more complex models based on "biologically plausible" intuition.

Eye movement recording

A database of 200 natural images was recorded using a 12 bit Canon EOS 1Ds Digital SLR and a number of professional L-grade Canon lenses (24–70/f2.8, 70–200/f2.8, 135/f2) in several zoos in Southern Germany (Supplementary Figure 1). To minimize the photographer's bias for scene selection, we took 1626 images and then randomly selected a subset of 200 of them. To remove the centering bias, scenes were photographed at very high resolution (4064 × 2704) and then cropped to 1024 × 768, centered at a random position. The images were then converted to 8 bit grayscale. Each of our 14 subjects viewed all 200 scenes (4 sessions with 50 trials) on a linearized 19″ Iiyama CRT at 60 cm distance (1024 × 768 full screen resolution, 100 Hz refresh rate). All subjects were paid for the participation and were naive with respect to the purpose of our study. The order of presentation was different and random for each subject. Each trial started with a fixation cross at a random location on the screen on a gray background at the mean intensity of all images. The cross was shown for a random duration, drawn from a Gaussian distribution with a mean of 3 s and a standard deviation of 1 s, however with the restriction that no duration was less than 1 s. Then the scene was displayed for a random duration (mean 2 s, standard deviation 0.5 s, minimum 1 s). Subjects were instructed to merely "look around in the scene" while keeping the head still; subjects' heads were stabilized using a chin rest. Eye movements were recorded with an Eyelink II video eye-tracker using pupil tracking at 250 Hz, calibrated with the standard software (9 point grid). All subjects had calibration errors below 0.5 degrees with an estimated average measurement error of 0.40 (±0.14 *SD*) degrees. We classified all eye movements with a speed above 26.8 degrees per second (>3 pixels per sample) as saccades. Saccade targets were extracted from the images at the median position of consecutive scene samples between two saccades.

We took great care to monitor and compensate for drift in the recording equipment during the experiment by displaying an 4 × 3 calibration grid at the start of each session as well as after every 10 trials. Subjects were instructed to fixate the calibration grid. In a postprocessing step, a separate affine transformation compensating for measurement errors was fit to the grid data using least squares, and then linearly interpolated over the entire session. The calibration data were also used to compute leave-one-out estimates of the measurement error which was linearly interpolated between two calibration steps. All trials with an estimated leave-one-out error above 1.0 deg were discarded, along with trials where the subject had missed the initial fixation target or the calibration grid. Saccade targets closer than 0.5 degrees to the image boundary were discarded, too, to avoid problems in the subsequent patch extraction (Figure 2a). This yielded 18,065 fixations with an estimated leave-one-out error of 0.54 (±0.19 *SD*) degrees. The mean saccade length was 7.0 (±5.3 *SD*) degrees. Fixations lasted for 250 (±121 *SD*) milliseconds.

Non-target locations

An equal number (18,065) of non-target locations was generated using the actual target locations, but from different images. This is a common procedure in the eye movement literature (Reinagel & Zador, 1999) and ensures that the locations of targets and non-targets are identically distributed. In this way learning artifacts due to possible variations in the local statistics of different image regions are prevented. This is illustrated in Figure 2b.

Patch extraction

At each target and non-target location, we extracted a 13 × 13 image patch and stored it in a 169-dimensional

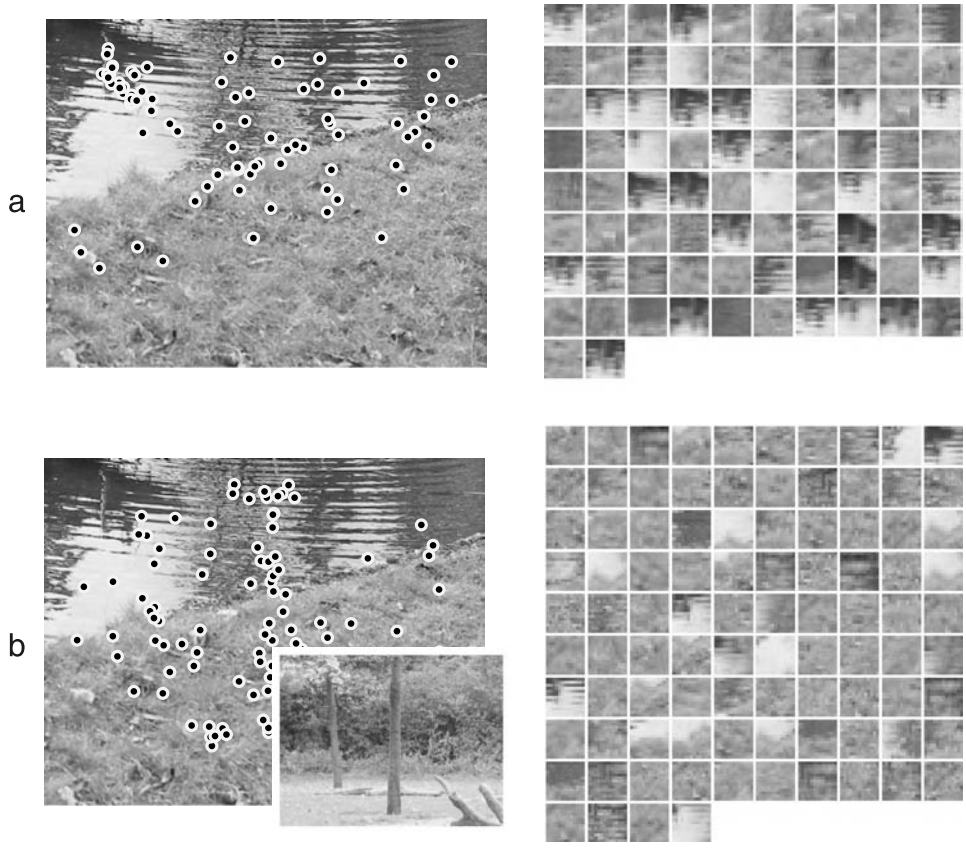

Figure 2. Extraction of local image patches from recorded eye movements. (a) The dots denote fixation locations from all 14 subjects on a natural image. Each location yielded an entry in our database of target patches, shown to the right. (b) Non-target patches were extracted from the same image, but using fixation locations that were recorded on a different, unrelated scene (small image). Note how contrast is visibly increased in the target patches. Differences in the spatial structure, however, are hardly noticeable.

vector $\mathbf{x}_i$. A label variable $y_i \in \{1, -1\}$ was associated with every patch, denoting target or non-target, respectively.

The size of the local neighborhood is a crucial parameter to our analysis. If it is chosen smaller than the relevant feature size the resulting model will fail. If it is too large, the number of dimensions which carry relevant structure will be small compared to the noninformative, noisy features. While some studies use patch sizes related to the 2–3 degree size of the fovea, e.g., Reinagel and Zador (1999), we instead optimized the patch size directly within the fitting framework. The rationale behind this is that saliency may be optimally predicted by a pattern that extends over much larger (or smaller) areas than the fovea. For example, it is not clear how much spatial context (e.g. dark background) a fovea-sized bright spot may need to be regarded as salient by the visual system. To this end, we built 11 complete data sets with different patch sizes, covering the full range of possibly reasonable patch sizes, i.e. 11 different sizes, equally spaced on a logarithmic scale ranging between 0.47 degrees (corresponding to 13×13 pixels at full resolution, and roughly equal to the standard deviation of the measurement noise) and 27 degrees (the height of the screen).

Another important parameter in our method is the spatial resolution of the image patches, as it places a lower bound on the granularity of the identified structure. If spatial resolution is chosen too low, the analysis might miss small relevant features. If chosen too high, the amount of data will not be sufficient to estimate the perceptive fields robustly. Here, we chose a spatial resolution of 13×13 pixels, regardless of the patch size. 13×13 patches were extracted by subsampling the image, using a Gaussian low pass filter to reduce aliasing effects. The 13×13 choice is reasonable as we confirmed in a number of control experiments: after the optimal patch size of 5.4 degrees was found (as described in the next section), we generated additional control data sets, all with a patch size of 5.4 degrees, but at higher resolutions up to 41×41. We found that the perceptive fields and the predictivity of the model did not change at resolutions above 13×13, suggesting that high frequency information above the Nyquist limit of the 13×13 patch (1.2 cpd) are not necessary for predicting visual saliency.

We subtracted the mean intensity from each patch to reduce irrelevant variance in the data. This amounts to the assumption that the DC component of an image region

does not contribute to its visual saliency. We verified in a control experiment that the models' performance indeed drops slightly (but not significantly) when the mean is not subtracted. In that case, the obtained perceptive fields are identical. As expected, however, their DC component is not zero, but very close to the average DC component over the entire data set.

Data splits

Dependencies between the local image structure at two different locations in the data set (regardless whether target or non-target) can lead to artifacts or overestimation of our model's predictivity. There are two main sources of such dependencies. First, two locations in the same image may be closer than half the patch size, i.e., the extracted patches overlap. Second, two target locations in the same image may be generated by the same subject within one trial.

The following two measures were taken to minimize such dependencies. First, the data (targets and non-targets) were divided into a training (two thirds) and a test set (one third). This was done such that both sets contained data from all 200 images, but never from the same subject on the same image. Second, whenever the model's predictivity was estimated, we used 8-fold cross-validation estimates with the folds split image-wise, i.e., such that no validation or test fold contained any data from images in the corresponding training fold.

Fitting the nonparametric model

A model

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \exp(-\gamma ||\mathbf{x}_i - \mathbf{x}||^2), \quad (1)$$

was fitted to the training split of the data (m = 24,370 patches) using the support vector algorithm (Schölkopf & Smola, 2002), which minimizes the regularized risk

$$R(f) = \sum_{i=1}^{m} \max(0, 1 - y_i f(\mathbf{x}_i)) + \frac{\lambda}{2} ||f||^2, \quad (2)$$

with respect to the weights α_i. The first term in Equation 2 denotes the data fit. It is zero, indicating a perfect fit, whenever $y_i f(\mathbf{x}_i) \geq 1$. It attempts to push $f(\mathbf{x}_i)$ to values ≥ 1 if $y_i = 1$, and to values ≤ -1 if $y_i = -1$. If successful, this will result in a *margin* of separation between the two classes, with f taking values in $[-1, 1]$. The number of points falling inside this margin will depend on the strength of the regularization, measured by $||f||^2$. The smaller $||f||^2$, the smoother the solution f. The tradeoff between data fit and smoothness is controlled by the parameter λ. The model is nonparametric, and its descriptive power grows with m, the number of data points it is fitted to. In fact, with the choice of Gaussian radial basis functions (Equation 1), it is sufficiently flexible to fit any smooth stimulus-response relationship in the data (Steinwart, 2001). Figures 1a and 1b illustrate a fitted model with Gaussian radial basis functions.

While it is in principle possible to use a range of different regularizers $||f||^2$, a convenient choice is the one employed by support vector machines. By means of a nonlinear mapping induced by the *kernel* $\exp(-\gamma ||\mathbf{x}_i - \mathbf{x}||^2)$, it represents the function f as a vector in a high-dimensional space, and then uses the squared length of that vector as a regularizer. Moreover, the decision function in that space is linear, and the problem of finding it can be reduced to a so-called *quadratic program.* A support vector machines is but one example of a *kernel method,* a class of methods which have recently also gained popularity as models in psychology (Jäkel, Schölkopf, & Wichmann, 2007). They all deal with nonlinearity by employing kernels that correspond to dot products in high-dimensional spaces, allowing for the construction of geometric algorithms in such spaces that correspond to nonlinear methods in the input domain.

In addition to the weights α_i, there are three design parameters that have to be set: γ, λ, and the patch size d. These were determined by maximizing cross-validation estimates of the model's accuracy, using an eight fold, images-wise split of the training set. We conducted an exhaustive search on an $11 \times 9 \times 13$ grid with the grid points equally spaced on a log scale such that d = 0.47,…, 27 degrees, $\gamma = 5 \cdot 10^{-5}$…, $5 \cdot 10^{3}$, and $\lambda = 10^{-3}$,…, 10^{4}, resulting in the optimal values $\lambda = 1$, $\sigma = 1$, d = 5.4 degrees. Performance was relatively stable with respect to changes of d in the range from 2.5 to 8.1 degrees, and changes of λ and γ up to a factor of 3 and 10, respectively. Note that finding the optimal weights α_i for a given set of the three design parameters is a convex problem and has therefore no local minima. As a result, an exhaustive grid search optimizes all parameters in our model globally and jointly.

Extracting perceptive fields

The fitting method described above belongs to the family of *kernel methods*. In that framework the type of basis functions $\varphi_i(\mathbf{x}) = \exp(-\gamma ||\mathbf{x}_i - \mathbf{x}||^2)$ is referred to as the *kernel* $k(\mathbf{x}_i, \mathbf{x})$. An essential feature of kernel methods is that suitable kernels—such as the Gaussian radial basis function employed in our model—must satisfy a positive definiteness property (Schölkopf & Smola, 2002), in which case it can be shown that

$$k(\mathbf{x}_i, \mathbf{x}) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \rangle, \quad (3)$$

i.e., the kernel corresponds to a dot product in some implicit feature space $\mathcal{F}$, induced by a feature mapping Φ. Φ is implicitly defined by the kernel and is usually nonlinear. By virtue of Equation 3, however, a standard linear fitting algorithm can be used, such as ridge regression, logistic regression, Fisher discriminant, or support vector machines. In effect, while the resulting kernel model $f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i k(\mathbf{x}_i, \mathbf{x})$ is nonlinear in its input $\mathbf{x}$, the theoretical and practical benefits of linear methods are retained.

Here, we show that the same property also provides a straightforward nonlinear generalization of linear perceptive (or receptive) field analysis. The proposed approach is based on the *preimage problem* in kernel methods (Scholkopf et al., 1999). Due to Equation 3, the fitted kernel model $f(\mathbf{x})$ is linear in the implied feature space $\mathcal{F}$,

$$f(\mathbf{x}) = \langle \Psi, \Phi(\mathbf{x}) \rangle, \tag{4}$$

where $\Psi = \sum_{i=1}^{m} \alpha_i \Phi(\mathbf{x}_i)$. Thus, in $\mathcal{F}$, Ψ is the linear perceptive field of f. In order to visualize Ψ, we exploit the fact that the feature mapping Φ maps image patches to vectors in $\mathcal{F}$. The goal is to invert the feature mapping at Ψ, which yields an image patch $\mathbf{z} = \Phi^{-1}(\Psi)$ corresponding to the receptive field. Since not every vector in $\mathcal{F}$ has such a preimage in the space of image patches, $\mathbf{z}$ is defined as the patch whose image in $\mathcal{F}$ is closest to Ψ, i.e., $\mathbf{z} = \arg\min_{\mathbf{x}} ||\Psi - \Phi(\mathbf{x})||^2$. In case of a Gaussian radial basis kernel this amounts to solving

$$\mathbf{z} = \arg\max_{\mathbf{x}} \langle \Psi, \Phi(\mathbf{x}) \rangle, \tag{5}$$

(Scholkopf et al., 1999). Interestingly, this definition of a nonlinear perceptive field coincides with that of the maximally excitatory stimulus, since the argmax argument is actually just $f(x)$ (see Equation 4). This not only provides an alternative interpretation of $\mathbf{z}$, but shows that we can solve the optimization problem (Equation 5), without having to compute the dot product in the (potentially high dimensional) feature space $\mathcal{F}$. By minimizing instead of maximizing Equation 5, we find the maximally inhibitory stimulus. In the feature space $\mathcal{F}$ this corresponds to the vector which is closest to Ψ, but points in the opposite direction. Note that due to the nonlinear nature of Φ, $\mathbf{z}$ is in general not unique, i.e., there can be multiple perceptive fields. For illustration, Figure 1c shows the optimal stimuli as red pluses.

To compute the perceptive fields, we solved Equation 5 using the method of steepest descent/ascent (Figure 1c, red zigzag line). Note that the f in Equation 1 defines a smooth function, and, since the Gaussian radial basis function is bounded, so is f, and hence all minima and maxima exist. Initial values for the gradient search were random patches with pixels drawn from a normal distribution with zero mean and standard deviation 0.11, the mean value in the training data. As mentioned above, the result of the gradient search is not unique. Thus, in order to find all perceptive fields, we solved the optimization problem many times with different initial values. This could be intractable, since f could have a large number of extremal points. In our case, however, we found that this was not a problem. After running the search 1,000 times, we found only 4 distinct solutions. This was verified by clustering the 1,000 optima using k-means. The number of clusters k was found by increasing k until the clusters were stable. Interestingly, the clusters for both minima and maxima were already highly concentrated for $k = 2$, i.e., within each cluster, the average variance of a pixel was less than 0.03% of the pixel variance of its center patch. This result did not change if initial values were random natural patches (standard deviation 0.11) or the training examples $\mathbf{x}_i$.

As an aside, note that our method can be interpreted as "inverting a neural network." This technique was also used by Lehky, Sejnowski, and Desimone (1992) to characterize neurons in the monkey striate cortex inverting a multi-layer perceptron. Further system identification methods based on neural networks can be found in Lau, Stanley, and Dang (2002) and Prenger, Wu, David, and Gallant (2004). The use of kernel methods for neuronal modeling was proposed by Wu, David, and Gallant (2006), and first steps toward perceptive field analysis in psychophysics using kernel methods were made by Wichmann et al. (2005).

The saliency model

The perceptive fields represent the most excitatory or inhibitory regions in stimulus space (Figure 1c, red pluses), the number of which is determined by the complexity of the underlying system. Interestingly, as described in the Results section, we found that the saccadic system can be modeled with four perceptive fields only. Motivated by this observation, we constructed a simple saliency model by placing radial basis functions centered at the perceptive fields, i.e., a simple feed-forward network

$$s(\mathbf{x}) = \sum_{i=1}^{4} \beta_i \varphi_i(\mathbf{x}), \tag{6}$$

with four radial basis units $\varphi_i = \exp(-\gamma ||\mathbf{z}_i - \mathbf{x}||^2)$, centered at the patterns $\mathbf{z}_1 \ldots \mathbf{z}_4$ (Figures 3a–3d). The network weights β_i were fitted by optimizing the same objective as before (Equation 2), using the optimal values for γ, λ, and d reported above. This yielded $\beta_1 = 0.94$ and $\beta_2 = 1.70$ for the excitatory units and $\beta_3 = -1.93$, $\beta_4 = -1.82$ for the inhibitory units. Figure 1d illustrates this procedure.

To compare our model to the model by Itti et al. (1998), we used the publicly available Matlab code (http://www.saliencytoolbox.net). All parameters were set to their

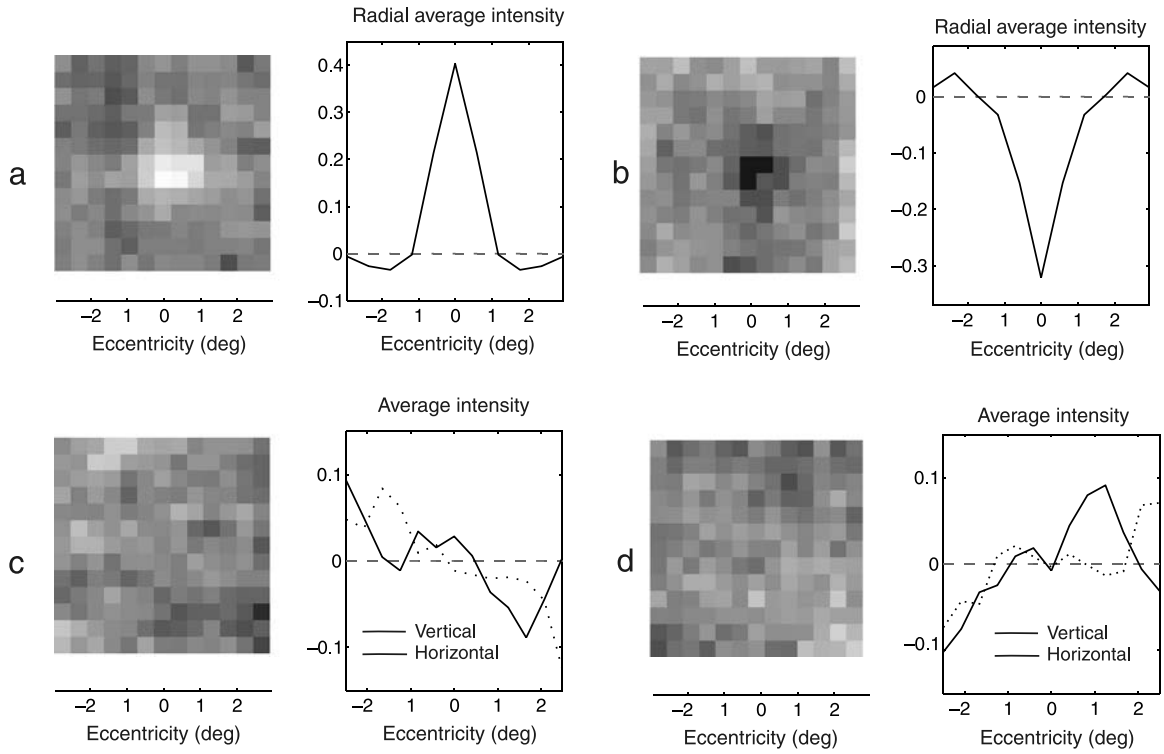

Figure 3. The four nonlinear perceptive fields revealed by our analysis. They represent the image structure which is most (a, b) or least (c, d) likely to become a saccade target. To the right of the two excitatory perceptive fields (a, b), their radial profiles (averaged over all directions) are shown. They both have center-surround structure. The patterns that do not draw saccades ("inhibitory") (c, d) are plotted together with their average horizontal and vertical profiles. The signal-to-noise ratio in the inhibitory patches is not as high as in the excitatory patches (note the different scales on the vertical axes), making the latter more difficult to interpret. The profiles suggest low-frequency, ramp-like structures and may allow shadows to be ignored.

default values with the following two exceptions. First, we disabled color features, since our stimuli were grayscale only. Second, we chose the normalization scheme "standard," which lead to the highest predictivity among all choices (curiously, the more recent normalization scheme, "iterative", performed worse, regardless of the number of iterations). The reported ROC scores in the Results section are averages over eight folds of the test data.

Results

Eye movement data

Eye movements of 14 human subjects were recorded during free viewing of static grayscale natural scenes, which were displayed on a 19-inch Iiyama CRT at 60 cm viewing distance. Every subject viewed 200 images, each one for on average three seconds. After each saccade, an image patch around the targeted location in the scene was extracted (see Methods section for details). We also generated an equally sized control set of non-target patches selected by using the locations of fixations in a unrelated images. In total, 36,130 target and non-target patches were extracted. Figure 2 shows examples of our data set (see also Supplementary Figure 1).

As can be seen from Figure 2, target patches typically have higher contrast than non-target patches. Averaged over the entire data set of 200 natural scenes, the root mean-squared (RMS) contrast (the standard deviation of pixels in a patch) in the target patches was 1.26-fold higher than in the non-target patches, 0.120 ($\pm 5.7 \cdot 10^{-4}$ *SEM*) versus 0.095 ($\pm 5.4 \cdot 10^{-4}$ *SEM*). The relevance of RMS contrast has been a well-known result at least since Reinagel and Zador's work (Reinagel & Zador, 1999). In contrast, finding characteristic differences in the spatial structure of the patches is a much harder problem, as Figure 2 suggests. This difficulty does not change if the two sets are compared in terms of their principal components (Rajashekar et al., 2002) or their independent components (Bell & Sejnowski, 1997; Olshausen & Field, 1996) (see Supplementary Figure 1).

Nonlinear system identification

As described above, the kernel model can have multiple perceptive fields, both "excitatory" in the sense that

eye-movements are more likely to be directed toward image patches resembling the excitatory perceptive field and "inhibitory" in the sense that eye-movements are less likely to be directed toward image patches resembling the inhibitory perceptive field. The number of which is determined by the data, i.e., the saccadic selection process. Here, the analysis yielded two excitatory and two inhibitory perceptive fields, which are depicted in Figure 3.

The excitatory perceptive fields (Figures 3a and 3b) have a clear center-surround structure, the inhibitory perceptive fields (Figures 3c and 3d) show a very flat, ramp-like structure. This indicates that the saccadic selection system is maximally attracted by center-surround patterns, while regions with very slowly varying intensity have few saccades to them. The spatial extent of center regions in the excitatory patterns was estimated by fitting Difference-of-Gaussians using the parameterization from Croner and Kaplan (1995). The on-center pattern led to a center diameter of 1.40 degrees, for the off-center pattern we found 1.42 degrees—a near perfect on-off-center pair.

Existing hypotheses on saccadic selection and bottom-up saliency implicate orientated excitatory features such as Gabor filters (Baddeley & Tatler, 2006; Bruce & Tsotsos, 2006; Itti et al., 1998; Mannan, Ruddock, & Wooding, 1996). Our perceptive fields are not edge-like, however, but center-surround: in terms of information processing in the mammalian visual systems center-surround receptive fields are typically found earlier, i.e. in the retina, the LGN and mid-brain structures such as the superior colliculus (SC), whereas orientated receptive fields are predominantly found cortically.

Control experiments

To ensure that the center-surround patterns and no artifacts emerge from the data we designed two control experiments.

First, we checked whether the center-surround patterns can emerge trivially, only due to the uncertainty in the gaze position measurements, e.g. by a subtle blurring or averaging effect. In particular, we wanted to know if our center-surround result can be generated from qualitatively different perceptive fields, such as orientated edges. In other words, we wanted to know if our result can be generated blurred version of a very different true perceptive field. Thus, we designed a control experiment which is exactly the same analysis method, only that the data are not recorded from human subjects, but simulated using a known perceptive field (Supplementary Figure 3): we generated two synthetic eye movement data sets with known underlying features, namely a contrast (squared DoG) filter and an edge (squared Gaussian derivative) filter, respectively. The simulated gaze positions were corrupted with random "measurement" noise, and the resulting perceptive fields were computed. This experiment was repeated for different detector scales and noise levels ranging from zero to about five times the standard deviation of the estimated true measurement noise, which was 0.54 (±0.19 *SD*) degrees. In addition, different spatial scales for the true features were tested. We found that the perceptive fields either showed the true underlying structure, or no structure at all if the "measurement" noise was too high (above roughly twice the estimated true measurement noise level). This indicates that our method does not generate spurious structure, regardless of the level of measurement uncertainty or the scale of the true feature. In particular, the perceptive fields computed from edge data were never center-surround or vice versa (see Supplementary Figure 3). In addition, this experiment shows that the frequency components of the center-surround patterns in Figure 3 are not significantly affected by the measurement noise: while the uncertainty in the position measurements (standard deviation 0.4 deg) suggests that no frequencies above about 1 cpd can be resolved, the passband of our center-surround patches is one order of magnitude below this limit (around 0.15 cpd), and hence unlikely to be influenced by this effect. Furthermore, the center surround perceptive fields Supplementary Figure 3 (8% feature size, 100% noise) have a passband at roughly the double frequency (0.3 cpd), and are still correctly identified by our method.

In a second control experiment, we checked for a possible bias due to the choice of our image database (cf. Figure 2). While natural scenes are an arguably reasonable choice, it is unclear whether the center-surround structure of our perceptive fields changes for a different stimulus set. To test this hypothesis, we recorded a second set of (real) eye movements using stimuli which contained mostly man-made objects, such as office scenes. This yielded 19,036 saccade targets. Here, the mean saccade length was 6.9 (±5.6 *SD*) degrees, fixations lasted for 243 (±118 *SD*) milliseconds on average. Reassuringly, this also yielded only two excitatory perceptive fields with center-surround structure (on- and off-center), despite the fact that the local structure in that data set is governed by somewhat different features, e.g. long and sharp edges (see Supplementary Figure 4).

Thus we conclude that the center-surround structure of the perceptive fields in Figures 3a and 3b indeed reflects the behavior of the visual system, and is not an artifact due to measurement error, bias in our method, or the particular choice of stimuli presented to the subjects.

A simple saliency model

Motivated by the above results, we constructed a simple computational model for visual saliency. Although the term *saliency* was originally introduced for allocation of both covert and overt visual attention (Koch & Ullman, 1985), it has become common practice to use it as a quantity monotonically related to the "probability of

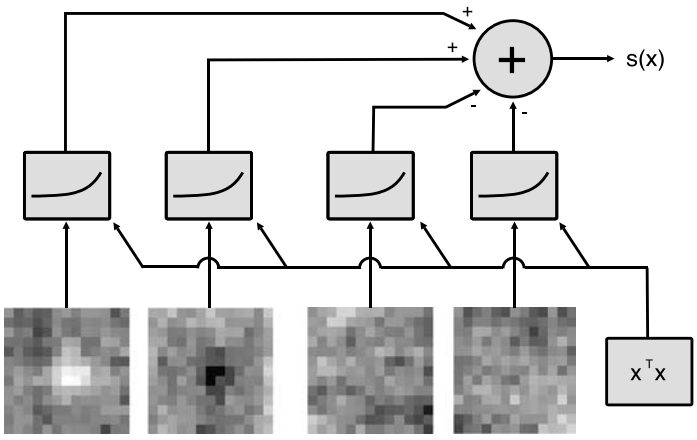

Figure 4. The proposed saliency model. It computes the visual saliency of an image location based on local image structure. The input is linearly filtered by the four kernels shown on the bottom (these are the perceptive fields from Figure 3), and the outputs are fed into an exponential point nonlinearity. The local signal energy, $\mathbf{x}^\top\mathbf{x}$, inhibits the inputs of the nonlinearities, which results in a tradeoff between pure contrast and structural cues (see also Figure 5). The nonlinearly transformed signals are weighted according to their excitatory or inhibitory nature and summed into a saliency value $s(\mathbf{x})$.

looking somewhere" (Henderson, 2003; Itti & Koch, 2001). It is in that sense that we use the term here.

The proposed saliency model is shown in Figure 4. It consists of a simple feed-forward network with four radial basis units $\varphi_i = \exp(-\gamma||\mathbf{z}_i - \mathbf{x}||^2)$, centered at the perceptive field patterns $\mathbf{z}_1 \ldots \mathbf{z}_4$ (Figures 3a–3d). The weights β_i were fit to the data to maximize predictivity (see Methods section).

Note that this is not a pure linear-nonlinear-linear model. The perceptive fields inside the radial basis functions $\varphi_i(\mathbf{x})$ are not only linear filters corresponding to relevant subspaces. Rather, they define excitatory ($\beta_i > 0$) or inhibitory ($\beta_i < 0$) *regions* in the space of image patches. A connection to linear-nonlinear-linear models can be made, however, by expanding the square in the radial basis function $||\mathbf{z}_i - \mathbf{x}||^2 = \mathbf{z}_i^\top\mathbf{z}_i + \mathbf{x}^\top\mathbf{x} - 2\mathbf{z}_i^\top\mathbf{x}$. Here, $\mathbf{z}_i^\top\mathbf{z}_i$ is a constant, $\mathbf{x}^\top\mathbf{x}$ is the signal energy of the input patch $\mathbf{x}$, and $-\mathbf{z}_i^\top\mathbf{x}$ is a linear filter. Thus, we can write the radial basis units as $\exp(a_i\mathbf{z}_i^\top\mathbf{x} + b)$, with a positive constant a_i and an offset b which depends only on the signal energy (b acts akin to a contrast gain-control mechanism trading-off pure contrast and local image structure). In particular, for any fixed energy, the perceptive fields in our model indeed behave like linear filters, followed by an exponential nonlinearity.

To assess the predictivity of this model, we used a set of 200 independent test images, divided into 8 sets of 25 to compute error bars. A standard measure for the predictivity of saliency models is the Wilcoxon-Mann-Whitney statistic (the probability that a randomly chosen target patch receives higher saliency than a randomly chosen negative one), which for our model was 0.64 (±0.011 *SEM*). We also tested the saliency model by Itti et al. (1998) on our data set and found its performance to be 0.62 (±0.022 *SEM*). Furthermore, tested on the office stimuli of our second control experiment, our model—while trained on natural images—still led to 0.62 (±0.010 *SEM*), whereas the model by Itti et al. yielded 0.57 (±0.024 *SEM*). Two important conclusions can be drawn from these results: first, our model is at least as predictive on natural scenes as the best existing models. Second, even if we disregard the admittedly not dramatic differences in predictivity, the models differ substantially in terms of complexity. The model by Itti et al. implements contrast and orientation features at multiple scales with lateral inhibition, while our model uses merely four features at a single scale within a simple feed-forward network. The good performance of our model on office scenes, on which the model was not trained, indicates that our model does not overfit. Rather, due to its simplicity, it seems to be more robust than Itti et al.'s model, yielding stable results if the stimulus type varies.

The above results are perhaps not surprising, since also in the model by Itti et al., the most stable feature seems to be luminance contrast, computed with Difference-of-Gaussian operators: Parkhurst et al. (2002) found that for natural scenes, luminance contrast it is more predictive than color or edge contrast. Also, Itti (2006) reported that under increased simulation realism (e.g., foveation, wide screen, etc.), the predictivity of the color and orientation channels degraded, whereas contrast remained the most stable feature in his model. This is consistent with our proposition that the most basic bottom-up feature for saccadic selection is one-scale center-surround.

Model behavior

In order to characterize the behavior of our model in a realistic environment, we illustrate its response to actual natural image patches. To this end, we randomly collected 50,000 patches from randomly chosen natural images and sorted them according to the saliency output predicted by our model $s(\mathbf{x})$. The most and least salient 100 patches are shown in Figure 5. Patches in the left panel (a) are the most salient, those in shown in the center panel (b) are the least salient ones.

We should like to stress two observations: First, there is a visual increase in contrast toward the salient patches, which is in agreement with the well-known result that local contrast is an essential relevant feature (Reinagel & Zador, 1999). Indeed, the average RMS contrast was 0.136 (±0.002 *SEM*) in the 100 most salient patches and 0.044 (±0.001 *SEM*) in the 100 least salient patches. The second observation is that RMS contrast alone should not be equated with visual saliency. To illustrate this, the 100 least salient patches from panel (b) are plotted again in the right panel (c) of Figure 5, this time with their RMS

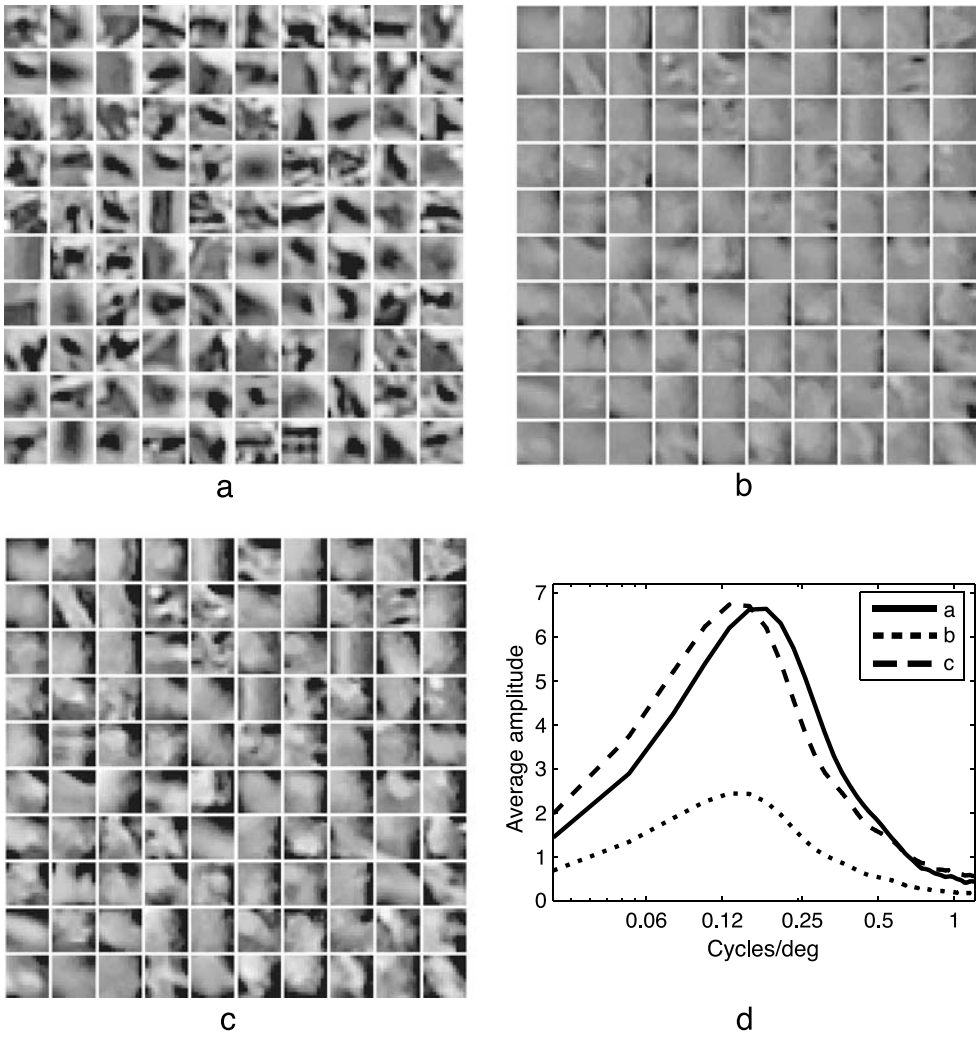

Figure 5. Image patches sorted by visual saliency. This figure was generated by feeding 50,000 randomly selected natural image patches through our model (Figure 4) and sorting them according to their predicted saliency. Panel (a) shows the 100 most salient, (b) the least salient patches. Apparently, contrast is much higher in salient patches, which is in agreement with previous work. But there are also structural differences. Panel (c) shows the 100 least salient patches again, but with their r.m.s. contrast scaled to that of the 100 most salient patches in panel (a). Panel (d) shows the frequency components of the patches in panels a, b, and c (averaged over all spatial directions and over all 100 patches, respectively). We observe that the structure of salient patches is typically corner-like, and localized around the patch center, while the structure of the non-salient patches is more ramp-like and has stronger low-frequency components.

contrast adjusted to that of the most salient patches in panel (a). This shows that the structure of the most salient patches tends to be of medium spatial frequency and localized at the patch centers. The structure of the least salient stimuli, on the other hand, is more ramp-like, i.e., not localized at the centers, but at the edges or corners of the patches, and has stronger low-frequency components, as shown in panel (d). Note that this behavior is not surprising, but reflects the structural differences between the excitatory and inhibitory perceptive fields in the model (Figure 4). In summary, we arrive at a similar conclusion as Krieger et al. (2000), who analyzed higher-order structure in saccade targets. They concluded that "the saccadic selection system avoids image regions which are dominated by a single orientated structure. Instead it selects regions containing different orientations, like occlusions, corners, etc." (p. 208, first paragraph).

Discussion

A fair comparison between saliency models is difficult to achieve for several reasons. First, there are strong correlations between plausible image features. For example, most conceivable image features are correlated with RMS contrast: since this quantity is increased at saccade

targets, most saliency measures are correlated with contrast, and so seemingly different models are all roughly equally predictive (Baddeley & Tatler, 2006; Bruce & Tsotsos, 2006; Itti et al., 1998; Reinagel & Zador, 1999; Tatler et al., 2005). Another dilemma is that basically all studies have used different stimulus sets. Even if the stimulus class is identical in two studies, a particular set of images can still be "easy" or "difficult" (in the sense of how obvious salient regions are). As a result, it has hitherto not been possible to reliably quantify differences between models, in particular, the effect that different types of image features have on predictivity (and hence, plausibility).

What distinguishes our approach and results from those obtained previously is not a significantly higher predictivity on natural scenes, but a different critical aspect: our nonlinear system identification technique does not make any assumptions about the shape, scale, or number of the perceptive fields. Rather, they *result from the data* as those patterns which maximize predictivity: our perceptive fields are the optimal predictors for saccade targets. This is in contrast to previous studies which either assumed relevant structure by the choice of image features (Gabor or others), or used linear identification methods (Rajashekar et al., 2006; Tavassoli, van der Linde, Bovik, & Cormack, 2007).

An interesting question is how fine the spatial structure of perceptive fields can in principle be resolved, given the spatial uncertainty introduced by the measurement error of the eye tracker. Namely, in a linear setting, the observed measurement standard deviation of 0.4 deg would (under the assumption of the measurements being Gaussian distributed) effectively filter out all frequency components in the perceptive field above around 1 cpd. This value is suspiciously close to the Nyquist frequency of our optimal perceptive fields (1.2 cpd), which might lead to the conclusion that our approach ignores potentially relevant high frequency components of visual saliency. However, this is not the case, since the significant frequency components of the perceptive fields are one order of magnitude lower, around 0.15 cpd (Figure 3). Moreover, note that due to the nonlinearity of our model, the above mentioned equivalence between position uncertainty and blur does not hold anymore, i.e., the spatial detail of the perceptive fields is not limited by the measurement noise, but only by the patch resolution. That the optimal patch resolution (1.2 cpd) is only incidentally related to the cutoff frequency corresponding to the measurement error (1 cpd) is further supported by the fact that the predictivity of the model does not increase with patch resolution: as described in Patch extraction section, increasing the resolution up to 41 × 41 (allowing frequencies of up to 3.8 cpd to enter the model) does not change the model's predictivity or its perceptive fields. In summary, our nonlinear model is indeed able to pick up high frequency components well beyond the limit suggested by the measurement error, however, these components are not part of the optimal solution to predicting eye movements.

While our results are consistent with the majority of previous studies (Itti, 2006; Krieger et al., 2000; Parkhurst et al., 2002; Reinagel & Zador, 1999), it is interesting to note that they seem to contradict the results from the recent studies by Tatler et al. (2005) and Baddeley and Tatler (2006) which are very similar to ours: there, the authors found that high frequency edges (2.7 cpd) are the most predictive feature for visual saliency. As we argued above, this contradictory finding cannot be attributed to an inherent insensitivity of our method to high-frequency details. Indeed, Baddeley and Tatler (2006) use a similar architecture to ours, with a different type of nonlinearity that could be, in principle, closely approximated by our network. In contrast to our approach, the authors tested a set of six fixed filter types chosen a priori among which a high-frequency oriented bandpass performed best. All of the six filter types investigated operate in the frequency range between 0 and 3.8 cpd tested in our study and thus constitute possible outcomes of our method. However, whereas our recovered perceptive fields are safely within the range of stable reconstruction, we cannot exclude the possibility that the high-frequency structure found by Baddeley et al. might be too sensitive to noise to be reconstructed by our method. Within the stable frequency range, we did not observe any tendency in performance toward the higher frequencies. We currently do not have a plausible explanation for this obvious discrepancy which remains a subject for further study.

The question where in the brain saliency computations take place has recently become a focus of research interest (Henderson, 2003; Itti & Koch, 2001; Treue, 2003). Our center-surround perceptive fields result from a psychophysical experiment but are strikingly similar to physiological receptive fields in the primate visual system. This is not an uncommon phenomenon: Neri and Levi (2006) review a number of examples where related experiments in psychophysics and physiology show similarities between measurements in human observers in single neurons.

The size of the on-/off-centers—1.40 and 1.42 degrees—is very similar to that reported for receptive fields in superior colliculus (SC) of monkey by Cynader and Berman (1972): at 7.0 degrees eccentricity—the average saccade length in our data—they found center sizes of 1 to 2 degrees in the superficial layer and around 3 degrees in the intermediate and deep layers. The role of SC in allocating (as opposed to generating) saccadic eye movements has been known for a long time (Goldberg & Wurtz, 1972), and has received increased empirical support recently (Basso & Wurtz, 1997; Krauzlis & Dill, 2002; Kustov & Robinson, 1996; McPeek & Keller, 2002, 2004). A short review by Krauzlis et al. (2004) concludes *"Although the SC is best known for its role in the motor control of saccades, it appears to serve a more general function related to evaluating possible targets, defining the goal for orienting movements, and in updating the representation of the goal as the movement is executed. (p. 1450)"*. Thus the collicular pathway for eye-movement generation is actively involved in fixation

target selection and, verily likely, saccade initiation regulation. Thus the fact that our psychophysical perceptive fields not only resemble physiological receptive fields but match important size and tuning properties of SC cells may not only be a coincidence but be taken as evidence for a role of SC in bottom-up visual saliency computations. We speculate that a substantial part of bottom-up saliency computations might be carried out sub-cortically, perhaps directly in the superior colliculus. Many previous models explicitly or tacitly—by the choice of oriented filters—assumed that visual saliency is computed in visual cortex. Our results suggest that bottom-up saliency driven eye-movements may be controlled and executed via a fast pathway involving the SC and that cognitively controlled top-down eye-movements may be computed cortically.

In summary, we have presented a novel nonlinear system identification technique with which we can derive perceptive fields from human saccade targets. Our technique should work with most—if not all—stimulus sets. In our case, we used natural images as input. Based on this technique we derived a nonlinear model that:

1. is extremely simple compared to previously suggested models;
2. predicts human saccade targets in natural scenes at least as well as previously suggested models;
3. generalizes to a novel image set better than previously suggested models;
4. is free of prior assumptions regarding the shape, scale, or number of filters;
5. can be implemented with optimal filters resembling those in the SC in shape, size and spatial frequency tuning, suggesting that bottom-up visual saliency may be computed sub-cortically in SC.

Acknowledgments

Parts of this work have been previously presented at the Neural Information Processing Systems (NIPS) Meeting 2006 (Kienzle, Wichmann, Schölkopf, & Franz, 2007c) and the Computational and Systems Neuroscience (COSYNE) Meeting 2007 (Kienzle, Franz, Macke, Wichmann, & Schölkopf, 2007a; Kienzle, Wichmann, Schölkopf, & Franz, 2007b). We would like to thank Matthias Bethge, Bruce Henning, Frank Jäkel and Jakob Macke for many helpful discussions. This work was funded, in part, by the Bernstein Computational Neuroscience Program of the German Federal Ministry of Education and Research.

Commercial relationships: none.
Corresponding author: Felix A. Wichmann.
Email: felix.wichmann@tu-berlin.de.
Address: Technical University of Berlin, Modelling of Cognitive Processes Group, Sekr. FR 6-4, Franklinstr. 28-29, 10587 Berlin, Germany.

References

Baddeley, R. J., & Tatler, B. W. (2006). High frequency edges (but not contrast) predict where we fixate: A bayesian system identification analysis. *Vision Research, 46,* 2824–2833. [PubMed]

Basso, M. A., & Wurtz, R. H. (1997). Modulation of neuronal activity by target uncertainty. *Nature, 389,* 66–69. [PubMed]

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

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

Croner, L. J., & Kaplan, E. (1995). Receptive fields of P and M ganglion cells across the primate retina. *Vision Research, 35,* 7–24. [PubMed]

Cynader, M., & Berman, N. (1972). Receptive-field organization of monkey superior colliculus. *Journal of Neurophysiology, 35,* 187–201. [PubMed]

Goldberg, M. E., & Wurtz, R. H. (1972). Activity of superior colliculus in behaving monkey. II. Effect of attention and neural responses. *Journal of Neurophysiology, 35,* 560–574. [PubMed]

Harel, J., Koch, C., & Perona, P. (2007). Graph-based visual saliency. In *Advances in neural information processing systems, 19*. [Article]

Henderson, J. M. (2003). Human gaze control during real-world scene perception. *Trends in Cognitive Science, 7,* 498–504. [PubMed]

Hopfinger, J. B., Buonocore, M. H., & Mangun, G. R. (2000). The neural mechanisms of top-down attentional control. *Nature Neuroscience, 3,* 284–291. [PubMed]

Itti, L. (2006). Quantitative modeling of perceptual salience at human eye position. *Visual Cognition, 14,* 959–984. [Article]

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews, Neuroscience, 2,* 194–203. [PubMed]

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,* 1254–1259. [Article]

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2007). A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology, 51,* 343–358. [Article]

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional Gabor filter model of simple

receptive fields in cat striate cortex. *Journal of Neurophysiology, 58,* 1233–1258. [PubMed]

Jung, R., & Spillmann, L. (1970). Receptive-field estimation and perceptual integration in human vision. In F. A. Young & D. B. Lindsley (Eds.), *Early experience and visual information processing in perceptual and reading disorders* (pp. 181–197).

Kienzle, W., Franz, M. O., Macke, J. H., Wichmann, F. A., & Schölkopf, B. (2007a). Nonlinear receptive field analysis: Making kernel methods interpretable. In *Computational and Systems Neuroscience Meeting (COSYNE).*

Kienzle, W., Wichmann, F. A., Schölkopf, B., & Franz, M. O. (2007b). A nonparametric approach to bottom-up visual saliency. In *Advances in Neural Information Processing Systems (NIPS) 19.* [Article]

Kienzle, W., Wichmann, F. A., Schölkopf, B., & Franz, M. O. (2007c). Center-surround filters emerge from optimizing predictivity in a freeviewing task. In *Computational and Systems Neuroscience Meeting (COSYNE).*

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4,* 219–227. [PubMed]

Krauzlis, R. J., Liston, D., & Carello, C. D. (2004). Target selection and the superior colliculus: Goals, choices and hypotheses. *Vision Research, 44,* 1445–1451. [PubMed]

Krauzlis, R. J., & Dill, N. (2002). Neural correlates of target choice for pursuit and saccades in the primate superior colliculus. *Neuron, 35,* 355–363. [PubMed] [Article]

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision, 3,* 201–214. [PubMed]

Kustov, A. A., & Robinson, D. L. (1996). Shared neural control of attentional shifts and eye movements. *Nature, 384,* 74–77. [PubMed]

Lau, B., Stanley, G. B., & Dang, Y. (2002). Computational subunits of visual cortical neurons revealed by artificial neural networks. *Proceedings of the National Academy of Sciences of the United States of America, 99,* 8974–8979. [PubMed] [Article]

Lehky, S. R., Sejnowski, T. J., & Desimone, R. (1992). Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. *Journal of Neuroscience, 12,* 3568–3581. [PubMed] [Article]

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

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision, 10,* 165–188. [PubMed]

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation patterns made during brief examination of two-dimensional images. *Perception, 26,* 1059–1072. [PubMed]

Matin, E. (1974). Saccadic suppression: A review and an analysis. *Psychological Bulletin, 81,* 899–917. [PubMed]

McPeek, R. M., & Keller, E. L. (2002). Saccade target selection in the superior colliculus during a visual search task. *Journal of Neurophysiology, 88,* 2019–2034. [PubMed] [Article]

McPeek, R. M., & Keller, E. L. (2004). Deficits in saccade target selection after inactivation of superior colliculus. *Nature Neuroscience, 7,* 757–763. [PubMed]

Neri, P., & Levi, D. M. (2006). Receptive versus perceptive fields from the reverse-correlation viewpoint. *Vision Research, 46,* 2465–2474. [PubMed]

Oliva, A., Torralba, A. B., Castelhano, M. S., & Henderson, J. M. (2003). Top-down control of visual attention in object detection. In *ICIP* (vol. 1, pp. 253–256).

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

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

Parkhurst, D. J., & Niebur, E. (2003). Scene content selected by active vision. *Spatial Vision, 16,* 125–154. [PubMed]

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

Prenger, R., Wu, M. C., David, S. V., & Gallant, J. L. (2004). Nonlinear V1 responses to natural scenes revealed by neural network analysis. *Neural Networks, 17,* 663–679. [PubMed]

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

Raj, R., Geisler, W. S., Frazor, R. A., & Bovik, A. C. (2005). Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 22,* 2039–2049. [PubMed]

Rajashekar, U., Bovik, A. C., & Cormack, L. K. (2006). Visual search in noise: Revealing the influence of structural cues by gaze-contingent classification image analysis. *Journal of Vision, 6*(4):7, 379–386,

http://journalofvision.org/6/4/7/, doi:10.1167/6.4.7. [PubMed] [Article]

Rajashekar, U., Cormack, L. K., Bovik, A. C., & Geisler, W. S. (2002). Image properties that draw fixations. In *Vision Sciences Society, Second Annual Meeting*.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the center of gaze. *Network, 10,* 341–350. [PubMed]

Renninger, L. W., Coughlan, J., Verghese, P., & Malik, J. (2005). An information maximization model of eye movements. *Advances in Neural Information Processing Systems, 17,* 1121–1128. [PubMed]

Scholkopf, B., Mika, S., Burges, C. J. C., Knirsch, P., Muller, K.-R., Rätsch, G., et al. (1999). Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks, 10,* 1000–1017. [PubMed]

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels*. Cambridge, MA: MIT Press.

Steinwart, I. (2001). On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research, 2,* 67–93.

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

Tavassoli, A., van der Linde, I., Bovik, A. C., & Cormack, L. K. (2007). An efficient technique for revealing visual search strategies with classification images. *Perception & Psychophysics, 69,* 103–112. [PubMed] [Article]

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,* 766–786. [PubMed]

Treue, S. (2003). Visual attention: The where, what, how and why of saliency. *Current Opinion in Neurobiology, 13,* 428–432. [PubMed]

Victor, J. D. (2005). Analyzing receptive fields, classification images and functional images: Challenges with opportunities for synergy. *Nature Neuroscience, 8,* 1651–1656. [PubMed] [Article]

Wichmann, F. A., Graf, A. B., Simoncelli, E. P., Bülthoff, H. H., & Schölkopf, B. (2005). Machine learning applied to perception: Decision-images for gender classification. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems* (vol. 17, pp. 1489–1496). Cambridge, MA, USA: MIT Press. [Article]

Wu, M. C. K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience, 29,* 477–505. [PubMed]

Yarbus, A. (1967). *Eye movements and vision*. New York: Plenum Press.