Components of bottom-up gaze allocation in natural images

Robert J. Peters\textsuperscript{a,b,1}, Asha Iyer\textsuperscript{b}, Laurent Itti\textsuperscript{a} and Christof Koch\textsuperscript{b}

\textsuperscript{a}Computer Science and Neuroscience, University of Southern California, Los Angeles CA 90089
\textsuperscript{b}Computation and Neural Systems, Caltech, Pasadena, CA 91125

Abstract

Recent research (Parkhurst et al., 2002) showed that a model of bottom-up visual attention can account in part for the spatial locations fixated by humans while free-viewing complex natural and artificial scenes. That study used a definition of salience based on local detectors with coarse global surround inhibition. Here, we use a similar framework to investigate the roles of several types of non-linear interactions known to exist in visual cortex, and of eccentricity-dependent processing. For each of these, we added a component to the salience model, including richer interactions among orientation-tuned units, both at spatial short range (for clutter reduction) and long range (for contour facilitation), and a detailed model of eccentricity-dependent changes in visual processing. Subjects free-viewed naturalistic and artificial images while their eye movements were recorded, and the resulting fixation locations were compared with the models’ predicted salience maps. We found that the proposed interactions indeed play a significant role in the spatiotemporal deployment of attention in natural scenes; about half of the observed inter-subject variance can be explained by these different models. This suggests that attentional guidance does not depend solely on local visual features, but must also include the effects of interactions among features. As models of these interactions become more accurate in predicting behaviorally-relevant salient locations, they become useful to a range of applications in computer vision and human-machine interface design.

Key words: Salience; Attention; Eye movements; Contours.

To appear in Vision Research.

1 Introduction

Selective attention is the ubiquitous mechanism that regulates the bottleneck between the massively parallel world of sensation and the serial world of cognition (James, 1890). This is particularly true in the visual system of primates, where 50\% of the primary visual cortex is devoted to processing input from the central 2\% (10°) of the visual field (Wandell, 1995). In order to benefit from this non-uniform allocation of processing resources, the visual system relies on a combination of covert and overt attentional shifting mechanisms to efficiently bring behaviorally-relevant stimuli under the processing capabilities of central vision (Treue, 2003).

We used eye movements as an overt measure of where observers were directing their covert attention. This method is based on the pre-motor theory of attention (Rizzolatti et al., 1987), which suggests eye movements and attention shifts are driven by the same internal mechanisms. Links between eye movements and attention have been demonstrated by behavioral (Sheliga et al., 1994, 1995; Hoffman and Subramaniam, 1995; Kowler et al., 1995; Hafed and Clark, 2002) as well as physiological (Kustov and Robinson, 1996; Moore and Fallah, 2001; Moore et al., 2003; Moore and Fallah, 2004) and brain imaging (Nobre et al., 2000; Beauchamp et al., 2001) studies. A number of studies have taken information-theoretical or statistical approaches to eye movements, asking how fixated regions differ from non-fixated regions (Reinagel and Zador, 1999; Krieger et al., 2000; Privitera and Stark, 2000).
These studies have shown that fixated regions have high contrast and high variance (low correlation) (Reinagel and Zador, 1999), distinctive higher-order statistics (Krieger et al., 2000), and high local symmetry (Privitera and Stark, 2000). Yet while such results help characterize expected fixation locations, they do not explicitly include a mechanism by which biological vision might extract the relevant features from the input.

In the present study we use biologically-plausible quantitative models to test hypotheses regarding the links between brain and behavior. Each model variant operates in the basic framework of a model of bottom-up saliency-driven attention (Itti et al., 1998; Itti and Koch, 2001)—which we refer to here as the baseline salience model. This model comprises a number of parallel channels for processing different feature types, such as luminance, orientation, and color, and the outputs from each of the channels are combined to produce a single, feature-independent salience map. This salience map signals salient, i.e., conspicuous or interesting, locations in the visual scene. It has been shown that such salience maps can predict locations likely to be fixated by human observers with significantly better-than-chance accuracy (Parkhurst et al., 2002). Despite these results, covert and overt attentional fixation locations may sometimes be distinct (Posner and Cohen, 1984); nevertheless it is likely that overt and covert shifts of attention are closely related except in the presence of considerable effort to meet explicit instructions to the contrary (e.g., “don’t look at the stimulus, but keep on fixating”).

We asked whether, and to what extent, human fixation behavior is influenced by three putative physiological mechanisms. First, we considered short-range interactions among orientation-tuned units found in early visual cortical areas with retinotopically-overlapping receptive fields. This effect, known as cross-orientation suppression (Morrone et al., 1982; Deangelis et al., 1992), has traditionally been assumed to arise from local lateral connections within cortex (Worgotter and Koch, 1991; Somers et al., 1995; Crook et al., 1997; Deangelis et al., 1992), although it has also been proposed that thalamocortical inhibition could produce a functionally similar result (Carandini et al., 2002; Freeman et al., 2002). Regardless of the specific neuronal implementation, such divisive inhibition leads to contrast-enhancement and a sharpening of orientation tuning—similar to a center-surround operation, but operating in the orientation and frequency domains rather than in the spatial domain. Furthermore, divisive inhibition provides the gain control needed to work within the limited dynamic range of neurons (Heeger, 1992). A number of computational models for cross-orientation suppression have been proposed (Lauritzen et al., 2001; Kolesnik and Barlit, 2003); in particular, one of these (Lee et al., 1999) was shown to succinctly account for detection and discrimination thresholds in a range of psychophysical tasks involving isolated Gabor-like grating stimuli on a blank background. In the present study, we adapted this model to test the extent to which such local interactions may actually influence scanpaths over natural scenes.

Second, we considered long-range interactions among orientation-tuned units with non-overlapping receptive fields, which are thought to contribute to the visual system’s exquisite sensitivity to contours. The presence of such lateral interactions has been inferred from neuroanatomy and electrophysiology (Blasdel, 1992; Pettet and Gilbert, 1992; Das and Gilbert, 1999; Stettler et al., 2002) and from psychophysical studies demonstrating increased or decreased contrast detection thresholds at a central location depending on the presence and orientation of surround elements (Polat and Sagi, 1993, 1994a; Zenger and Sagi, 1996; Zenger et al., 2000). An appropriate arrangement of connection strengths (Polat and Sagi, 1994b; Li, 1998; Braun, 1999a; Li and Gilbert, 2002), involving facilitation between nearly collinear edge segments and inhibition between non-collinear parallel and orthogonal segments, has the effect of enhancing the activity of units that respond to the segments comprising an elongated contour such as the Gabor “snakes” described in Section 2.2. A number of such models exist (e.g., Li, 1998; Braun, 1999a; Tang et al., 2001); here, we adapted the model of Mundhenk and Itti (2002) to test whether contour-facilitation plays a significant role in determining fixation locations in complex images, and furthermore how that role depends on the relevance of contours to the behavioral task.

Third and last, we considered the cumulative effects of eccentricity-dependent processing through the early stages of vision. Anatomically, these effects begin in the retina with a strongly peaked distribution of cone photoreceptors and retinal ganglion cells near the fovea, along with correspondingly smaller receptive field sizes, and continue with a further expansion of foveal representation in primary visual cortex. One psychophysical manifestation is the influence of eccentricity on the relationship between contrast detection thresholds and spatial frequency; a similar relation exists for orientation discrimination thresholds (Virsu and Rovamo, 1979). There are essentially two effects when moving from the fovea to the periphery: discrimination thresholds become generally higher (reflecting overall poorer visual sensitivity), and the optimal spatial frequency becomes lower (reflecting larger receptive fields). Parkhurst et al. (2002) found that observers’ fixation locations followed a non-uniform spatial distribution favoring the center of stimulus images, while the baseline salience model predicted a uniform distribution of fixation locations. We designed a simple but efficient model of eccentricity-dependent effects in which the salience model’s intermediate feature maps are attenuated according to spatial frequency and eccentricity, in a manner quantitatively consistent with previously published contrast-detection and orientation-discrimination thresholds (Virsu and Rovamo, 1979). Using the distribution of fixation locations generated by observers, we compared the detailed model of eccentricity-dependent effects with simpler approximations (such as used by Parkhurst et al., 2002).

In each of these three cases, we find that the physiological mechanisms have a significant influence on the selection of fixation locations, at least to the extent that our coarse models capture the essence of these mechanisms. Importantly, we find this effect in tasks involving free viewing of crowded naturalistic scenes, including grayscale outdoor scenes (van Hateren and van der Schaaf, 1998), grayscale satellite imagery, and full-color fractals. Our aim was to cover a range of natural image types with a small number of categories. Although top-down factors based on emotional reaction or explicit memory can certainly play a significant
role in determining fixation locations (Yarbus, 1967), we deliber-
ately avoided images that would strongly trigger such factors
(such as close-up images of faces, familiar people, or well-known
locations) since we assume them to be partly outside the scope
of the bottom-up physiological mechanisms under consideration.
With that constraint, we selected outdoor scenes with elements
whose general types would be familiar (trees, grassy fields, streets
and sidewalks, campus buildings) but whose particular identi-
cies would be unknown to most observers. We also selected over-
head satellite imagery, involving scenes that are still interpretable
(roads, mountains, fields are easily identifiable), but which, in
contrast to outdoor scenes, are visually unfamiliar to most ob-
servers, due to the unusual overhead and wide-angle perspective.
Fractal images contain spatial frequency spectra similar to nat-
ural images, but contain no familiar elements. Finally we used a
set of images containing random arrangements of Gabor patches;
these served to specifically highlight the role of non-local inter-
actions.

The success of these models helps to support a quantitative link
between observers’ unconstrained overt behavior and the detailed
functional properties of individual neurons as inferred from single-
unit recordings and psychophysics experiments with constrained
stimuli and task conditions. This detailed computational model of
bottom-up, salience-based attention is useful for a range of appli-
cations from neuroscience to engineering. Machine vision systems
face the same difficulties as do biological vision systems, and so a
quantitative implementation of attentional selection can lead to
similar improvements for machine vision systems. Indeed, models
of bottom-up attention have been shown to improve the perfor-
ance of traditional computer vision object recognition systems,
both in the visual learning phase as well as in the subsequent
recognition phase (Miau and Itti, 2001; Walther et al., 2002;
Rutishauser et al., 2004). Accurate models of behavior also serve
a very practical goal in human-machine interface: particularly for
visual attention, there are many attention-demanding contexts
(e.g., driving, flying, surveillance, image analysis) in which even
a trained expert could occasionally benefit from an assistant sys-
image as before. However, when the image disappeared it was replaced by a full-screen uniform white mask. This was followed by a new response task that was trained to match the expert’s optimal behavior.

While on the other 50% of trials, a non-matching contour was selected trial to trial; since subjects could not predict the orientation, they were forced to consider the entire stimulus image, without being encouraged to focus on any particular aspect of the image. In the contour-detection task (b), each trial began with a fixation cross followed by a stimulus image as before. However, when the image disappeared it was replaced by a full-screen uniform white mask. This was followed by a new response screen containing a single schematic contour, and subjects made a 2-AFC as to whether there had been a matching contour at the same location in the just-seen image. On 50% of trials, there was in fact such a match, while on the other 50% of trials, a non-matching contour was selected from among the contours that matched other images in the same category.

tem that was trained to match the expert’s optimal behavior.

None of this denies the crucial roles of top-down, task-dependent attention in conscious vision (James, 1890; Koch, 2004), yet in the absence of detailed quantitative models, we have concentrated here on the contribution of bottom-up, salience-driven cues to fixation.

2 Methods

2.1 Subjects

Psychophysics subjects (ages 18-25) from the Caltech community participated as paid volunteers as follows. Four of these (“group A”) participated in the first set of free-viewing experiments involving outdoor photos, overhead satellite imagery, and fractals. Another four (“group B”) participated in a second free-viewing experiment involving Gabor snakes and Gabor arrays. Finally, seven subjects (“group C”) participated in a third experiment involving a comparison between the free-viewing and contour-detection tasks. Three of the subjects participated in more than one experiment, so the total number of individuals involved was twelve. Tables 2 and 4 give results from groups A (columns 1–3) and B (columns 4–5), while Table 3 gives results from group C. Informed consent was obtained from all subjects, and experimental procedures were approved by the California Institute of Technology’s Committee for the Protection of Human Subjects.

2.2 Stimuli

We used four classes of images (Figure 1), ranging in size from 1000 × 1000 to 1536 × 1024 pixels, subtending a visual angle of roughly 15.8° × 15.8° to 16.2° × 25°. The experiments reported here typically included about 100 images from each image class: grayscale 10-meter resolution “digital orthorectified” (DO110m) overhead satellite imagery; grayscale outdoor photographs; grayscale Gabor “snakes” and Gabor arrays containing arrays of randomly spaced and oriented Gabor elements generated with a previously-described algorithm (Braun, 1999b,a). The Gabor “snake” images included chains of Gabor elements that were properly aligned so as to induce a strong percept of a contour, even though element spacing and Gabor phase were otherwise random.

2.3 Free-viewing task

Images were presented to subjects in a free-viewing task (Figure 2a). Each trial began with a 1000ms fixation cross at the center of a blank screen, which subjects were instructed to fixate. This imposed some consistency on the initial conditions of the subsequent scanpaths, across different images and observers. Following the fixation cross, a target image was shown for 3000ms. Subjects were instructed to “look around the image” with no restrictions except the knowledge that they would have to provide a response, as follows. Immediately after the target image disappeared, a single line was presented at an arbitrary orientation bisecting the screen into two regions of equal size. The two regions were labeled as “1” and “2,” and subjects were required to make a button press indicating which region contained the location that they had found “most interesting” in the previous image. Our motivations for requiring this response were two-fold: (1) to encourage subjects to be vigilant in their task and engage in active eye movements (without a minimal task to motivate them, subjects might “efficiently” choose to make no eye movements at all); and (2) to avoid imposing any particular top-down bias on the task (such as would occur if subjects were asked to search for horizontal lines, or to judge the brightness of the image, or to name objects in the image), allowing direct comparisons with

---

2 From the National Geospatial-Intelligence Agency (NGA) (http://geengine.nga.mil/)
3 http://hlab.phys.rug.nl/imlib/
4 http://gnofract4d.sourceforge.net/
5 http://spanky.triumf.ca/
a model of bottom-up attention. Although no time limit was imposed on the responses, subjects were encouraged prior to the experiment not to dwell on the choice for too long, but rather to make their best guess if they felt unsure.

2.4 Contour-detection task

We used a second task to investigate the influence of contours on fixation locations (Figure 2b). The overall format of the task was similar to the free-viewing task, except (1) the image presentation time was shortened from 3000ms to 2500ms, (2) a full-screen uniform white mask was presented for 500ms immediately after each image to prevent subjects from relying on retinal afterimages to perform the task, and (3) a different response was required, as explained next. After the image and the mask, subjects were presented with a single schematic line-drawn contour, and responded with a key press to indicate whether that contour matched a contour that was present at the same location in the image they had just seen. On half of the trials, the contour was in fact a match to the preceding image, and on the other half, the contour was a non-match (selected from a pool of contours that matched other images in the experiment). The schematic contours were Bezier curves that closely approximated the shapes of hand-picked salient contours in the target images.

2.5 Eye tracking

Subjects were seated 75 cm from a CRT used for stimulus display, which subtended 26° x 19° of visual angle, and used a chinrest to minimize eye-tracking errors due to head movements. We used an infrared (IR) eye tracking system (ISCAN, Inc.) to sample and record subjects’ eye position at 120Hz. An illuminator and camera were placed ~65 cm from the subject, and his or her right eye was illuminated with a beam of low-intensity (~1 mW/cm²) invisible IR light (~850 nm). The camera recorded a close-up image of the eye, which was processed in real-time to extract the positions of two features: (1) p, the IR-dark spot at the center of the pupil, and (2) c, the IR-bright spot where the IR beam produces a peculiarity on the cornea. The vector difference \( \mathbf{v'} = \mathbf{p} - \mathbf{c} \) of these two positions gives a measure of eye position that is independent of head position. An empirical correspondence between \( \mathbf{v'} \) (in camera coordinates) and the subject’s real-world point-of-regard \( \mathbf{v} \) (in stimulus display coordinates) was established by a set of calibration trials in which the subject fixated a series of crosses shown at 25 different locations on an invisible 5 x 5 grid in the stimulus display. These \( \mathbf{v}-\mathbf{v'} \) pairings could then be used to interpolate the subject’s point-of-regard throughout the remainder of the session. Following each session, each lasting about 12 minutes, we re-recorded subjects’ eye positions at the 25 calibration locations in order to assess how much drift had occurred during the recording session. Across all eye-tracking sessions, the overall error was 0.54 ± 0.44° (mean ± s.d.) degrees of visual angle per calibration point.

2.6 Salience model

All of the models described here are based on the computational architecture of a salience model of bottom-up visual attention first proposed by Koch and Ullman (1985) and developed in detail by Itti et al. (1998) (see Figure 3). Each input image is processed in parallel through a number of feature channels (e.g., one each for color, luminance, orientation), and the outputs of these channels are ultimately combined to form a single salience map. This map ascribes a scalar value to each point in the input image, indicating how salient or “interesting” that location is, regardless of which features contributed to the salience.

The individual channels share a common architecture. In general, the input image is first passed through a series of linear filters at nine spatial scales to form a dyadic pyramid. These filter outputs are then subject to spatial competition via a center-surround operation, implemented as a difference between fine and coarse scales in the pyramid. Typically there are six feature maps generated by this center-surround operation, using center scales \( c \in \{2, 3, 4\} \) and surround scales at \( s = c + \delta \), with \( \delta \in \{3, 4\} \). The feature maps are summed across scales and passed through a nonlinear normalization operation designed to reduce or eliminate numerous weak local maxima in favor of a small number of stronger near-global maxima. This produces a single conspicuity map representing the output of the channel; these conspicuity maps are eventually summed across channels and renormalized to produce the final salience map.

The standard channels for static images include a luminance channel that responds to luminance contrast, an orientation channel (including filter outputs from multiple scales and orientations) that responds to orientation contrast, and a color channel that responds to opponent-color contrast. These reflect many of the fundamental computational operations thought to be performed in the early stages of the visual system (Marr, 1982; Wandell, 1995). Nevertheless, the modular architecture of the salience model allows other new channels to be included in parallel to the standard channels, or even to replace one or more of them. This is the approach we used in testing more detailed models of interactions among orientation-tuned units, as described next.

---

6 Although observers’ wearing of contact lenses or eyeglasses has been reported to lead to lower eye-tracking accuracy, we found no difference in the drift error between observers with corrected (n = 6) and uncorrected (n = 9) vision within our subject pool.

7 Source code for the iLab Neuromorphic Vision Toolkit (iNVT), including the salience model and each of the extensions described below, is freely available under the GNU General Public License (GPL) at http://ilab.usc.edu/toolkit/.
Fig. 3. Schematic diagram of the salience model (top) and salience maps (bottom row) corresponding to the four exemplar images in the left column of Figure 1. In the salience model, an input image is processed in parallel through multiple channels. In each channel (here for luminance, orientation, and color), the image is filtered at nine spatial scales, and the resulting feature maps pass through a center-surround operation to accentuate contrast (e.g., the map at scale 7 is subtracted from the map at scale 4). The center-surround maps are combined across spatial scales leading to one conspicuity map per channel; finally these conspicuity maps are combined across features to produce a single feature-independent salience map. Additional channels may be included in parallel to the three channels shown here; in our experiments, we tested a modified orientation channel that included short-range orientation interactions (Figure 4), a contour-facilitation channel based on long-range orientation interactions (Figure 5), and a model of eccentricity-dependent effects in which the luminance and orientation feature maps were attenuated as a function of eccentricity and spatial scale (Table 1).

2.7 Short-range orientation interactions

We adapted a model of interactions among overlapping orientation-tuned units (Lee et al., 1999; Itti et al., 2000) (see Figure 4) that could be substituted for the standard orientation channel in the salience model. In this enhanced orientation channel, orientation-sensitive units tuned to overlapping spatial locations, but to different orientations $\theta$ and spatial frequencies $\omega$, form an inhibitory pool. In the two-stage model, the feed-forward first-stage response $E_{\theta,\omega}$ is subject to self-excitation and suppression from the inhibitory pool. The result of these interactions is the non-linear second-stage response $R_{\theta,\omega}$, given by
Fig. 4. At top is a schematic diagram of the short-range orientation interactions model (figure adapted from Lee et al., 1999). In this model, an input image is passed through a set of linear filters tuned to different orientations and spatial frequencies. The linear outputs feed forward into a second stage, in which the set of filter outputs corresponding to a given spatial location form a pool that divisively inhibits each unit’s response at that location. As a result of this recurrent processing, the second stage output exhibits gain control and contrast enhancement relative to the first stage. We tested a modified version of the salience model from Figure 3 in which the standard orientation channel is replaced by one including short-range orientation interactions; the bottom row of images here shows the salience maps produced by such a modified model for the four exemplar images from the left column of Figure 1.

\[ R_{\theta,\omega} = \frac{(E_{\theta,\omega})^\gamma}{S^\delta + \sum_{\theta',\omega'} W_{\theta',\omega'} (E_{\theta',\omega'})^\gamma} \]

\[ \delta, \gamma : \text{power-law exponents} \]
\[ S : \text{semi-saturation constant} \]
\[ W_{\theta',\omega'} = e^{\frac{(\theta - \theta')^2}{2\sigma_{\theta}^2}} e^{\frac{(\omega - \omega')^2}{2\sigma_{\omega}^2}} \]
\[ \Sigma_{\theta}, \Sigma_{\omega} : \text{widths of inhibitory pool} \]

\[ E_{\theta,\omega} = Acse^{-\frac{(\theta - \theta')^2}{2\sigma_{\theta}^2}} e^{-\frac{(\omega - \omega')^2}{2\sigma_{\omega}^2}} + B \]

\[ c_s : \text{stimulus contrast} \]
\[ \theta_s : \text{stimulus orientation} \]
\[ \omega_s : \text{stimulus spatial frequency} \]
\[ \sigma_{\theta} : \text{sharpness of orientation tuning} \]
\[ \sigma_{\omega} : \text{sharpness of spatial frequency tuning} \]
\[ A : \text{contrast gain} \]
\[ B : \text{background activity level} \]

It should be noted that in the original model, the feed-forward responses \( E_{\theta,\omega} \) were calculated using ideal filters tuned for a given \( \theta \) and \( \omega \):
Fig. 5. At top is a schematic diagram of the long-range orientation interactions (contour-facilitation) model. A formal description is given in Appendix A.1.

In this model, an input image is passed through a series of filters tuned to 12 orientations (only 4 are depicted in this figure), all tuned to the same spatial scale. The output from these first-stage filters feeds into second-stage activation maps via a set of kernels that specify connection strengths as a function of relative spatial position and relative orientation tuning (see Figure 6). These connections are arranged so as to selectively enhance locations that form part of an elongated contour. The activation maps are summed across orientations and passed through a sigmoid non-linearity to yield the final output map. The model output evolves iteratively (three iterations were used in the present study); the second-stage maps recurrently excite their first-stage counterparts, and the output map recurrently modulates the strength of inhibition within the connection kernels to limit the dynamic range of the output. In practice, the model was instantiated at three spatial scales, but there were no interactions between scales at the intermediate stages; the outputs from each of the spatial scales were summed at the final stage to produce an overall output. The bottom row of images shows the salience maps produced by a modified salience model including a separate contour-facilitation channel in addition to the standard orientation channel (Figure 3), for the four exemplar images from the left column of Figure 1.

2.8 Long-range orientation interactions

We adapted a model of long-range orientation interactions from Mundhenk and Itti (2002) and Braun (1999a) (Figure 5) that was included as a new channel in the salience model. Briefly, this model relies on a set of weight matrices that determine how one orientation-tuned unit is influenced by other such units at different distances and orientations (Figure 6), in a manner reflecting the long-range axonal connections thought to be present in primary visual cortex (Blasdel, 1992). These matrices are sometimes described by their shape which resembles a “butterfly” or “bow-tie,” with wedges of excitatory connections leading from the central unit to other units that are similarly tuned and nearly collinear. Outside these wedge-shaped regions, there are inhibitory connections from the central unit leading to other similarly tuned units that are nearly parallel but not collinear. Our model did not include interactions among orthogonal or nearly orthogonal units. A formal mathematical description is given in Appendix A.1.

2.9 Eccentricity-dependent filtering

It has been reported that saccade targets tend to cluster around the current fixation location, rather than being uniformly distributed throughout the visual scene. That is, nearby targets are preferred over faraway ones. Although this effect has been fitted empirically with a Gaussian-decaying mask applied to the final salience map (Parkhurst et al., 2002), we asked whether a detailed model of eccentricity-dependent orientation-discrimination and contrast-detection thresholds would explain the behavior as well or better. We developed a model based on previously pub-
lution, and eccentricity do not map well onto the rough model of color processing contained in the baseline salience model's color channel, and partly because we found that a simple approximation to eccentricity effects, applied to the salience map as a whole, could already account well for observers' fixation locations.

Although interactions between color vision and eccentricity are widely reported (Hibino, 1992; Mullen and Losada, 1999; Lu et al., 1999; Anstis, 2002; Imhoff et al., 2004), we did not include a model of such interactions in our study. This was partly because the complex interactions between hue perception, isoluminance, ambient illumination, and eccentricity do not map well unto the rough model of color processing contained in the baseline salience model's color channel, and partly because we found that a simple approximation to eccentricity effects, applied to the salience map as a whole, could already account well for observers' fixation locations.

maps, the value of each unit was attenuated according to two factors: (1) \( x \), its eccentricity relative to the current fixation location, and (2) \( \omega \), the spatial frequency to which it responds. The attenuation coefficient \( m \) is given by \( m = c_\omega e^{-k_\omega x} \), where \( c_\omega \) and \( k_\omega \) are empirically-determined parameters depending on the spatial frequency \( \omega \); values for \( c_\omega \) and \( k_\omega \) are given in Table 1. This attenuation process is computationally efficient, since the attenuation values can be precomputed and stored as one “mask” for each spatial frequency; applying the masks when an input image is received requires just one array shift and multiplication per spatial frequency, rather than a convolution or Fourier transform.

Once the attenuation masks were applied to the internal feature maps in the luminance and orientation channels, the remainder of the salience algorithm proceeded as usual.

We also tested several approximations to this full model of

Fig. 6. Illustration of the weight matrices that connect neighboring units at different orientations in the contour model of Figure 5. See Appendix A.1 for details. (a) Each grid entry is a spatial array depicting the connection strengths between a central unit (unit 1) tuned to the orientation given by the column label, and a neighboring unit (unit 2) tuned to the orientation given by the row label. Within each grid entry, the spatial separation between unit 1 and unit 2 is represented by the \( x \)- and \( y \)-axes, and connection strength is represented by gray level: lighter pixels reflect regions of excitation, darker pixels reflect regions of inhibition, and gray pixels reflect the absence of any connection. The “butterfly” shape of the kernels reflects the symmetric cones of excitation connecting a central unit with neighbors whose position and orientation is such that the two units are “nearly collinear,” as well as the symmetric flanks of inhibition between units representing contour elements that are nearly parallel but non-collinear. (b) An enlargement of the \( 90^\circ /90^\circ \) kernel. Here, connection strength is represented by \( z \)-axis height as well as gray level, with values above and below 0.0 representing excitation and inhibition, respectively. (c) The orientation-independent component \( \rho \) is a function of the distance between the receptive field centers of units 1 and 2. (d) The orientation-dependent component \( \psi \) is a function of the angular difference between the mean orientation \( (\alpha) \) of units 1 and 2 and the orientation \( (\phi) \) of the line segment connecting the two units. The connection kernels are formed by the sum of an inhibitory component that depends only on distance (via \( \rho \)), and an excitatory component that depends on both distance (\( \rho \)) and orientation (\( \psi \)).

To apply the psychophysical thresholds to these internal feature
Values greater than zero suggest a greater correspondence than deviation. Next, the normalized salience values were extracted straightforward approach was as follows (Figure 7). Each salience of the constants fitted to empirical data from Virsu and Rovamo (1979). In this way, at each spatial location, the maximum possible salience was decreased by an factor that grew larger with increasing distance from the current center of fixation.

<table>
<thead>
<tr>
<th>spatial frequency ((\omega), cycles per degree)</th>
<th>luminance channel (c_\omega)</th>
<th>orientation channel (k_\omega)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.0</td>
<td>60.01</td>
<td>0.40</td>
</tr>
<tr>
<td>9.0</td>
<td>180.00</td>
<td>0.35</td>
</tr>
<tr>
<td>4.5</td>
<td>210.61</td>
<td>0.17</td>
</tr>
<tr>
<td>2.3</td>
<td>236.45</td>
<td>0.13</td>
</tr>
<tr>
<td>1.0</td>
<td>190.71</td>
<td>0.10</td>
</tr>
<tr>
<td>0.7</td>
<td>166.29</td>
<td>0.09</td>
</tr>
<tr>
<td>0.4</td>
<td>130.40</td>
<td>0.13</td>
</tr>
</tbody>
</table>

eccentricity-dependent filtering, in which only the final salience map was multiplied by a spatial mask. These masks decayed with eccentricity \(x\) either as \(ce^{-kx}\) or \(ce^{-kx^2}\), with varying values of the constants \(k\) and \(c\).

Finally, in all of the eccentricity-dependent filtering computations, the model’s “fixation location” (used as the center of the various spatial masks) was always yoked to the observer’s actual fixation location. Thus, the eccentricity-dependent filtering models were run once for each observer against which they were to be compared. This reflects that our models do not specify a mechanism for generating sequences of eye movements, but instead merely identify likely locations for an upcoming fixation, given the current fixation location.

2.10 Normalized scanpath salience (NSS)

Our analyses rest on the degree of correspondence between human fixation locations and model salience maps, taking into account the high inter-subject variability of eye movements. The most straightforward approach was as follows (Figure 7). Each salience map was linearly normalized to have zero mean and unit standard deviation. Next, the normalized salience values were extracted from each point corresponding to the fixation locations along a subject’s scanpath, and the mean of these values, or normalized scanpath salience (NSS), was taken as a measure of the correspondence between the salience map and scanpath. Due to the pre-normalization of the salience map, normalized scanpath salience values greater than zero suggest a greater correspondence than would be expected by chance between fixation locations and the salient points predicted by the model; a value of zero indicates no such correspondence, while values less than zero indicate an anti-correspondence between fixation locations and model-predicted salient points. Another benefit of the pre-normalization is that these measures could be compared across different subjects, image classes, and model variants; with such a data pool, statistical tests indicated whether the distribution of NSS values was different from the zero-mean distribution expected by chance.

This approach is similar to the one taken by Parkhurst et al. (2002) in that both rely on a linear transformation of salience values; however, our approach uses a variable dynamic range based on the variance of the salience values, while the alternate approach uses a fixed dynamic range based on the difference between the minimum and maximum values (which were rescaled to 0 and 100, respectively, in Parkhurst et al., 2002). In addition, our approach compares salience values at fixated locations to chance distributions unique to each image; the alternate approach compares salience values to a single chance distribution based on all images in a given image category. Our method was intended to accommodate the wide variety of salience distributions observed for different input images (for example, consider how a salience map with 100 points, 90 with value 1.0 and 10 with value 0.0, would be handled relative to a second salience map with 100 values spaced evenly between 0.0 and 1.0).

3 Results

A summary of all of the comparisons between models and human behavior in the free-viewing task is given in Tables 2 and 4. Each number gives the average NSS across all observers and images in that image class. In general, our data agree with previous results (Parkhurst et al., 2002) showing that the baseline salience model was significantly above chance \((p < 10^{-23})\) at predicting locations likely to be fixated by observers in a free-viewing task. As expected, this result was largely independent of image category for naturalistic images such as overhead imagery, outdoor photos, and fractals, but did not hold for more artificial images such as the Gabor arrays, for which the baseline salience model was virtually at chance in predicting fixation locations. Indeed, we chose to use the Gabor arrays for exactly this reason: nothing in the baseline model can “see” the contours, yet they are perceptually salient to human observers.

We used two pseudo-models as controls to estimate the theoretical minimum and maximum NSS values that could be expected of the salience models. First, the theoretical range of NSS values is bounded from below by the behavior of a random model, in which the “salience maps” simply contain noise drawn from a normal distribution\(^9\). The very nature of our analysis method requires that this random model should produce NSS values of 0, and indeed we found values that were nearly 0 (BSM in Table 2; slight differences from 0 are due to the finite size of our data set).

Second, the theoretical range of NSS values is bounded from above by the behavior of an inter-observer model in which the “salience

\(^9\) We obtained nearly identical results (NSS values close to 0) with another pseudo-model whose salience maps were obtained by a random spatial scrambling of the values in the actual salience map.
maps” are generated by the pooled fixation locations from all observers. For this, we constructed a spatial array containing a delta function peak at each fixation location from all observers, and blurred this array by convolving with a two-dimensional Gaussian, with half-width at half-height of \( \approx 1^\circ \) (see Figure 8). The blurring was intended to allow for variability in different observers’ fixation locations for the same target, and for spatial uncertainty from the eye tracking method. In practice, this process was slightly modified so that when predicting the fixation locations of observer A, the inter-observer model was based on data from all observers except A (i.e., a “leave-one-out” analysis). We found that this inter-observer model gave NSS values between 0.9 and 1.3, depending on the image type (Table 2).

One way to describe the performance of the salience model is to consider its performance as a percentage of the difference between the NSS scores of the random and inter-observer models. These values are shown in the bottom half of Table 2, and range from 39% to 57% for the natural image classes and from 9% to 15% for the Gabor arrays. Figure 9 gives a graphical depiction of these results.

Although the primary goal of the forced-choice task (“which of two regions was more interesting”) in the free-viewing experiment was to encourage observers to actively inspect the image without placing any particular top-down bias on their eye movements, observers’ responses to this task also offer an opportunity to compare an implicit measure of salience (i.e., observers’ fixation locations) with an explicit measure (i.e., their responses to the forced-choice task). Note that although during the free-viewing experiment (Figure 2a) observers did not know the orientation with which the response screen would be bisected until after they had viewed the image, we can retrospectively divide observers’ eye movements, as well as the models’ salience maps, according to this bisecting line for the purpose of subsequent analysis. We found no significant difference between the amount of time observers spent viewing the subsequently selected (mean ± s.e.m.: 1.42 ± 0.02s) and unselected (1.40 ± 0.02s) regions. There was also no significant difference between the average normalized salience in the subsequently selected (0 ± 0.006) and unselected (0 ± 0.006) regions. However, there was a significant tendency (\( p < 0.05 \), paired t-test) for the portion of the observers’ scanpath inside the selected region to have a higher NSS (0.52 ± 0.02) than the portion that fell inside the unselected region (0.47 ± 0.02). That is, observers tended to view more salient locations within the subsequently-selected region than in the unselected region, even though they spent equal amounts of time viewing both regions, and both regions had the same average salience.

3.1 Short-range orientation interactions

When the model of short-range orientation interactions was substituted for the standard orientation channel in the salience model, we observed a statistically significant 10–20% improvement (\( p < 0.05 \), paired t-test) in the NSS scores across all of the image classes, except for the Gabor snake and Gabor array

---

**Fig. 7.** Illustration of the method used to compare fixation locations obtained from eye tracking with salience maps obtained from various computational models. (a) A sample image is shown to both the human observer and the model. (b) The model generates a salience map (grayscale image), which is normalized to have zero-mean and unit standard deviation (see scalebar). A series of fixation locations is generated by the observer (connected dots), and the normalized salience value is extracted for each location (values are shown here next to the corresponding fixation locations). (c) The average normalized salience value across all fixation locations is taken as the normalized scanpath salience (NSS), and compared against the distribution of salience values across the entire salience map (gray histogram). For the scanpath shown here, the normalized scanpath salience indicates that, on average, the model-predicted salience at fixated locations was 1.304 standard deviations above chance level. Since the NSS is scale-free, it can be used to compare the degree of correspondence between observed and predicted behavior for different observers and images.

**Fig. 9.** Graphical presentation of the data from the bottom half of Table 2, illustrating the degree of correspondence (normalized scanpath salience, NSS) between observers’ fixation locations and the various models, as a percentage of the theoretical maximum NSS given by the inter-observer model.
3.2 Long-range orientation interactions

We added to the salience model a new channel for contour facilitation via long-range orientation interactions (BSM+CF in Table 2). This led to improved NSS scores over the baseline salience model by 19-36% for the three image classes, and by 30% for the Gabor arrays. Notably, only with long-range orientation interactions did the model's performance rise above chance levels for the Gabor arrays. In addition, for all image classes except the outdoor photos, the baseline model with contour facilitation had significantly higher NSS scores than did the baseline model with short-range orientation interactions. We also tested a model that included both short-range and long-range orientation interactions (BSM+SRI+CF in Table 2). This combined model bettered the individual models in just those cases where the individual models each led to a statistically significant improvement over the baseline model. Finally, turning again to the theoretical upper limit on model performance attained by the NSS attained by the inter-observer model, we found that the modified salience model including a contour-facilitation channel reached 36% to 74% of this maximum across the different image classes (Table 2).

We used a second experiment to specifically address the role of elongated contours in selecting fixation locations, by asking subjects to view the same images under two different task conditions: first, the standard free-viewing task, and subsequently, a contour-detection task. Table 3 shows the results of comparing models with behavior in these two tasks. The baseline model performance was worse in predicting fixation locations in the contour-detection task than in the free-viewing task; this is likely because performing the contour-detection task involves a greater top-down component, whereas the model mimics only bottom-up components. Nevertheless, there was an interaction between task and model (Table 3, bottom half): the relative improvement due to the contour-facilitation model over the baseline model was greater for the contour-detection task than for the free-viewing task, significantly so for the Gabor snake and Gabor array images. That is, the contour-facilitation model was better suited to the contour-detection task.

| Tree 2 | Results of comparing each model with scanpaths recorded during the free-viewing task (Figure 2a). Each number represents the average normalized scanpath salience (NSS) value, for a given model, across all of the fixation locations recorded while observers freely viewed images for 3000ms each. The normalized scanpath salience values were obtained by the method illustrated in Figure 7, in which salience maps were first normalized to have zero mean and unit standard deviation, and then for each scanpath the average normalized salience was computed for the fixation locations along the scanpath. Thus for the data shown here, a value of zero would indicate the absence of a correspondence between model predictions and observed fixation locations; a value of one would indicate that, on average, the model-predicted salience was one standard deviation above chance at each fixation location for all observers and all images in the given image category. The upper rows show these correspondences for salience maps predicted by (1) a random “model”; (2) the baseline salience model (BSM; see Figure 3), (3) a modified model including short-range orientation interactions (BSM+SRI; see Figure 4), (4) a second modified model including contour-facilitation (BSM+CF; see Figure 5), (5) a combined model including both short-range interactions and contour facilitation (BSM+SRI+CF), and (6) the control condition in which the “salience map” is derived from all observers’ scanpaths. This last condition quantifies how well the pooled fixation locations from all observers predict the specific fixation locations of individual observers; as such, it provides a theoretical upper limit for the performance of the models, since the models are not designed to account for inter-observer variability. Thus, the lower rows express the performance of each model as a percentage of the corresponding upper limit. Numbers with a * indicate models whose fit was significantly better than the corresponding baseline salience model (p < 0.05, paired t-test).

<table>
<thead>
<tr>
<th>Model Comparison</th>
<th>Outdoor</th>
<th>Fractal</th>
<th>Satellite</th>
<th>Gabor snake</th>
<th>Gabor array</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSM + Short-range interactions (SRI)</td>
<td>0.75 ± 0.03</td>
<td>0.56 ± 0.03</td>
<td>0.71 ± 0.03</td>
<td>0.11 ± 0.02</td>
<td>0.14 ± 0.02</td>
</tr>
<tr>
<td>BSM + Contour-facilitation (CF)</td>
<td>0.72 ± 0.03</td>
<td>0.60 ± 0.03</td>
<td>0.81 ± 0.03</td>
<td>0.41 ± 0.03</td>
<td>0.52 ± 0.02</td>
</tr>
<tr>
<td>BSM + SRI + CF</td>
<td>0.74 ± 0.03</td>
<td>0.66 ± 0.03</td>
<td>0.85 ± 0.03</td>
<td>0.40 ± 0.03</td>
<td>0.50 ± 0.02</td>
</tr>
<tr>
<td>Inter-observer</td>
<td>1.30 ± 0.04</td>
<td>1.13 ± 0.04</td>
<td>1.10 ± 0.04</td>
<td>1.15 ± 0.06</td>
<td>0.91 ± 0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model Comparison</th>
<th>NSS, mean ± s.e.m.</th>
<th>NSS, % of Inter-observer NSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random model</td>
<td>-0.01 ± 0.01</td>
<td>-2%</td>
</tr>
<tr>
<td>BSM salience model (BSM)</td>
<td>0.69 ± 0.03</td>
<td>2%</td>
</tr>
<tr>
<td>BSM + SRI</td>
<td>0.62 ± 0.03</td>
<td>1%</td>
</tr>
<tr>
<td>BSM + CF</td>
<td>0.10 ± 0.03</td>
<td>2%</td>
</tr>
<tr>
<td>BSM + SRI + CF</td>
<td>0.14 ± 0.02</td>
<td>15%</td>
</tr>
<tr>
<td>Inter-observer</td>
<td>0.72 ± 0.03</td>
<td>65%</td>
</tr>
</tbody>
</table>

We found that the modified salience model including a contour-facilitation channel reached 36% to 74% of this maximum across the different image classes (Table 2).
Fig. 8. Illustration of the inter-observer model, a control used to establish an upper bound on how well a model of bottom-up attention could be expected to predict observers’ eye movements. For each input image (left), observers’ scanpaths were recorded (center left column); each point in the scanpath represents a single sample from the 120Hz eye-tracking trace. From each scanpath, a heatmap was constructed (center right) by placing a Gaussian “blob” (half-width at half-height $\approx 1^\circ$) at the location of each sample from the eye movement trace. These blobs were summed across observers to produce a map (right) whose values represent how often observers were fixating in the vicinity of each location. As before, the bottom row of images corresponds to the four exemplar images from the left column of Figure 1.

### 3.3 Eccentricity-dependent filtering

Including eccentricity-dependent filtering in the salience model produced a large improvement in the ability to predict fixation locations. With the full implementation (in which internal feature maps were modulated according to eccentricity and spatial frequency) the average normalized scanpath salience values were: for outdoor images, 1.02 (versus 0.69 for the baseline model); for fractal images, 1.07 (versus 0.44); for satellite photos, 1.10 (versus 0.62). These represent ratios of 1.48, 2.43, and 1.77 for the three image classes, respectively, relative to the baseline model performance. In addition, the exponential approximation (in which only the final salience map was modulated by an exponential $e^{-x^2}$ decay with eccentricity $x$) produced results very similar to the full implementation (Table 4). Indeed, the normalized scanpath salience scores were 6% higher for the exponential approximation than for the full implementation. For comparison, we also implemented a Gaussian approximation in which the final salience map was modulated by an $e^{-x^2}$ decay with eccentricity, using the same specifications as in the modified model of Parkhurst et al. (2002). We found that although this Gaussian approxima-
Table 3
Results of comparing each model with eye-tracking data from the two different tasks (Figure 2). In the free-viewing task \((fv)\), subjects passively observed images, while in the contour-detection task \((cd)\), subjects were presented with a schematic contour following each image and were required to indicate whether that contour matched one that was present in the just-seen image. The top half shows the normalized scanpath salience \((NSS)\) metric described in Figure 7 and Table 2. (As in Table 2, all \(s.e.m.\) values were between 0.01 and 0.03, so for brevity these values have been omitted here.) Numbers marked with \(\ast\) indicate models whose \(NSS\) was significantly greater than the \(NSS\) of the baseline salience model. The bottom half shows the increment of each model’s \(NSS\) above the \(NSS\) of the baseline salience model. Numbers marked with \(\dagger\) indicate models for which this increment was significantly greater (paired \(t\)-test, \(p < 0.05\)) in the contour-detection task than in the the free-viewing task. The only model to significantly improve upon the baseline model \((\ast)\) and to perform significantly better in the contour-detection task \((\dagger)\) was the contour-facilitation model during the Gabor array and Gabor snake images.

<table>
<thead>
<tr>
<th>Model</th>
<th>Outdoor (fv)</th>
<th>Satellite (fv)</th>
<th>Gabor snake (fv)</th>
<th>Gabor array (fv)</th>
<th>Outdoor (cd)</th>
<th>Satellite (cd)</th>
<th>Gabor snake (cd)</th>
<th>Gabor array (cd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline salience model (BSM)</td>
<td>0.45</td>
<td>0.51</td>
<td>0.43</td>
<td>0.27</td>
<td>0.13</td>
<td>0.12</td>
<td>0.13</td>
<td>0.12</td>
</tr>
<tr>
<td>BSM + Short-range interactions</td>
<td>0.54*</td>
<td>0.60*</td>
<td>0.59*</td>
<td>0.44*</td>
<td>0.14</td>
<td>0.13</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>BSM + Contour-facilitation</td>
<td>0.52</td>
<td>0.59</td>
<td>0.66*</td>
<td>0.49*</td>
<td>0.51*</td>
<td>0.54*</td>
<td>0.49*</td>
<td>0.55*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Difference, (NSS) – Baseline salience (NSS)</th>
<th>Outdoor (fv)</th>
<th>Satellite (fv)</th>
<th>Gabor snake (fv)</th>
<th>Gabor array (fv)</th>
<th>Outdoor (cd)</th>
<th>Satellite (cd)</th>
<th>Gabor snake (cd)</th>
<th>Gabor array (cd)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline salience model (BSM)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>BSM + Short-range interactions</td>
<td>0.09</td>
<td>0.09</td>
<td>0.16</td>
<td>0.17</td>
<td>0.01</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.01†</td>
</tr>
<tr>
<td>BSM + Contour-facilitation</td>
<td>0.07</td>
<td>0.08</td>
<td>0.23</td>
<td>0.23</td>
<td>0.38</td>
<td>0.42†</td>
<td>0.36</td>
<td>0.44‡</td>
</tr>
</tbody>
</table>

4 Discussion

Our experiments were designed to explore the bottom-up physiological mechanisms that influence human behavior in an image-viewing task; we have disregarded important top-down contributions from attentional state, past experience, and other observer-specific factors, in order to assess how much can be attributed to bottom-up, stimulus-driven influences alone. In this respect, our method follows that of Parkhurst et al. (2002), and our results with the baseline salience model are in agreement with theirs: we found highly significant correspondences between model predictions and human fixation locations. However, the main focus of the present study was to extend this method to test, via more specific models, whether certain early vision mechanisms play a significant role in determining subjects’ fixation locations.

We rely on an assumption of a substantial overlap between the biological mechanisms responsible for covert attention shifts and overt eye movements; on this “pre-motor theory of attention” (Rizzolatti et al., 1987), pure attention shifts during fixation are essentially planned saccades whose motor execution is inhibited. This is supported by behavioral evidence showing that, despite motor inhibition, the spatial locus of attention exerts a small but detectable influence on the trajectories of subsequent saccades (Sheliga et al., 1994, 1995; Kustov and Robinson, 1996) or on the distribution of microsaccades during fixation (Hafed and Clark, 2002). These results suggest that computational models of attention and saccadic eye movements should be similar until the execution stage, where the dynamics would be expected to change due to the motor inertia of eye movements or the differing strengths of inhibition-of-return. Indeed, it is plausible that other modes of behavioral output, such as verbal report or finger-pointing, could be driven by the same core mechanisms. In ongoing work, we are using such approaches to further explore which computational elements are intrinsic to spatial attention, and which are specific to particular output modalities (Briand et al., 2000; Astafiev et al., 2003).

We found that a large overall fraction of the observed eye movement behavior could be attributed to the basic elements of early vision (luminance, orientation, color), according to the strong correspondence between the observed fixation locations and the predictions of the baseline salience model. Allowing that this general model of vision is not intended to account for inter-observer...
differences, an absolute upper limit on the performance of such models is given by the ability to predict one subjects’ behavior from the average behavior of the remaining subjects. The models we tested reached roughly 50% of this theoretical limit (Table 2, lower half); performance increased to 80–100% when eccentricity-dependent effects were accounted for (Table 4, lower half); as a crude measure, this suggests that the models could account for at least half of the variance in spatial positions of fixated locations, outside of inter-observer differences.

We tested three specific putative physiological mechanisms for their role in determining fixation locations. The first such mechanism that we tested was short-range inhibitory orientation interactions, also known as cross-orientation suppression. These interactions were modeled on the lateral inhibition that takes place within a hypercolumn in early visual cortical areas, which in turn is an abstraction of the concept that for a given receptive field in visual space, there is a confined population of cells in visual cortex that are tuned to all possible spatial scales and orientations. Lateral inhibition is a ubiquitous feature of sensory processing along spatial, temporal, and higher-order feature dimensions, as it decorrelates the input and maximizes information density (Simoncelli and Olshausen, 2001). Ultimately, this allows behaviorally relevant input to be represented in a more explicit and compact manner. Lee et al. (1999) used psychophysical experiments to validate a hypercolumn model, showing that changes in attentional state could be explained by changing the relative contributions of feed-forward excitatory and feed-back inhibitory connections. These connections determine, among other things, how easily an observer is able to identify a low-contrast grating in the presence of an overlapping grating of a different orientation. When we included these connections in our salience model, we found that the model’s salience maps predicted observers’ fixation locations significantly better. Thus, these connections, previously modeled with well-controlled minimalistic laboratory stimuli, also appear behaviorally relevant under less restrictive task conditions involving free-viewing natural scenes.

The second mechanism that we tested was long-range connections between different hypercolumns. Computationally, such connections or their equivalent have been introduced to explain the subjective salience of implicit contours like Gabor “snakes” that would otherwise be invisible to purely local processing (Li, 1998; Braun, 1999a). Indeed, without long-range connections, the salience model performed very poorly in predicting observers’ fixation locations in the Gabor arrays, since each individual Gabor element appears equally salient to a purely local mechanism. As we expected, the model performance increased dramatically (more than three-fold) when the long-range connections were included. However, somewhat unexpected was the fact that these connections lead to more modest improvements in predicting fixation locations in the natural image categories. This could be explained in one of two ways: either the model was not accurately identifying what observers’ considered to be “contours,” or the observers were giving relatively little weight to the contours that were present. To distinguish between these possibilities, we conducted a second psychophysics experiment in which observers viewed images under two different task conditions, one requiring them to specifically attend to contours, and one requiring only free viewing. If our model of contour facilitation based on long-range connections was simply inaccurate, then it should not have shown any additional benefit in predicting observers’ contour-detection behavior over their free-viewing behavior. Instead, we found that the improvement in model fit due to contour facilitation was greater when subjects performed the contour-detection task than when they performed the free-viewing task. Thus one possible conclusion is that, although our contour-facilitation model was accurately highlighting what would qualitatively be

Table 4
Results of comparing each model with the fixation locations recorded during the free-viewing task, identical to Table 2, except that each model now includes eccentricity-dependent filtering (EDF) in which salience values are increasingly attenuated at larger eccentricities (for eccentricity \( x \), the attenuation is proportional to \( e^{-x^2} \)). Numbers with a * indicate models whose fit was significantly better than the corresponding baseline salience model (\( p < 0.05 \), paired \( t \)-test).

<table>
<thead>
<tr>
<th>Model Configuration</th>
<th>Outdoor NSS, mean ± s.e.m.</th>
<th>Fractal NSS, mean ± s.e.m.</th>
<th>Satellite NSS, mean ± s.e.m.</th>
<th>Gabor snake NSS, mean ± s.e.m.</th>
<th>Gabor array NSS, mean ± s.e.m.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random model + EDF</td>
<td>0.01 ± 0.01</td>
<td>0.01 ± 0.01</td>
<td>0.01 ± 0.01</td>
<td>0.01 ± 0.01</td>
<td>0.01 ± 0.01</td>
</tr>
<tr>
<td>Baseline salience model (BSM) + EDF</td>
<td>1.00 ± 0.03</td>
<td>1.00 ± 0.03</td>
<td>1.00 ± 0.03</td>
<td>1.00 ± 0.03</td>
<td>1.00 ± 0.03</td>
</tr>
<tr>
<td>BSM + Short-range interactions (SRI) + EDF</td>
<td>0.25 ± 0.03</td>
<td>0.25 ± 0.03</td>
<td>0.25 ± 0.03</td>
<td>0.25 ± 0.03</td>
<td>0.25 ± 0.03</td>
</tr>
<tr>
<td>BSM + Contour-facilitation (CF) + EDF</td>
<td>1.20 ± 0.03</td>
<td>1.20 ± 0.03</td>
<td>1.20 ± 0.03</td>
<td>1.20 ± 0.03</td>
<td>1.20 ± 0.03</td>
</tr>
<tr>
<td>BSM + SRI + CF + EDF</td>
<td>1.24 ± 0.03</td>
<td>1.24 ± 0.03</td>
<td>1.24 ± 0.03</td>
<td>1.24 ± 0.03</td>
<td>1.24 ± 0.03</td>
</tr>
<tr>
<td>Inter-observer + EDF</td>
<td>1.44 ± 0.04</td>
<td>1.44 ± 0.04</td>
<td>1.44 ± 0.04</td>
<td>1.44 ± 0.04</td>
<td>1.44 ± 0.04</td>
</tr>
</tbody>
</table>

NSS, % of Inter-observer NSS

<table>
<thead>
<tr>
<th>Model Configuration</th>
<th>Outdoor</th>
<th>Fractal</th>
<th>Satellite</th>
<th>Gabor snake</th>
<th>Gabor array</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random model + EDF</td>
<td>-1%</td>
<td>-1%</td>
<td>1%</td>
<td>0%</td>
<td>-1%</td>
</tr>
<tr>
<td>Baseline salience model (BSM) + EDF</td>
<td>83%</td>
<td>79%</td>
<td>90%</td>
<td>81%</td>
<td>92%</td>
</tr>
<tr>
<td>BSM + Short-range interactions (SRI) + EDF</td>
<td>87%</td>
<td>86%</td>
<td>96%</td>
<td>77%</td>
<td>88%</td>
</tr>
<tr>
<td>BSM + Contour-facilitation (CF) + EDF</td>
<td>84%</td>
<td>86%</td>
<td>99%</td>
<td>91%</td>
<td>104%</td>
</tr>
<tr>
<td>BSM + SRI + CF + EDF</td>
<td>86%</td>
<td>89%</td>
<td>101%</td>
<td>89%</td>
<td>103%</td>
</tr>
<tr>
<td>Inter-observer + EDF</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
identified as "contours," observers' fixation locations were only weakly influenced by the presence of elongated contours in natural images where other salient image features were also present.

The third biological vision mechanism that we tested was the decay of sensitivity in peripheral relative to foveal vision. Anatomically, this decay is found throughout the visual system, including the decreasing density of photoreceptors and retinal ganglion cells away from the fovea, and "cortical magnification"—the over-representation in cortical surface area of central vision throughout visual cortex. This anatomical organization manifests itself behaviorally in increased contrast detection and orientation discrimination thresholds in the periphery, and in a non-uniform distribution of saccade targets with a disproportionate tendency toward the image center (Parkhurst et al., 2002). We used published reports of contrast-detection and orientation-discrimination thresholds (Virsu and Rovamo, 1979) to construct a detailed functional model of eccentricity-dependent effects, and asked whether this model could explain observers' non-uniform distribution of saccade targets within the context of the salience model. Indeed, we observed a strong increase in the model's predictive ability when it included eccentricity-dependent filtering, in agreement with Parkhurst et al. (2002). Furthermore, we found that the behavior of the full eccentricity-dependent model was matched by an approximation in which a single exponentially-decaying mask, centered at the current fixation location, is applied to the salience map. Such an exponentially-decaying mask gave a better fit to behavior than did a Gaussian-decaying mask as used in Parkhurst et al. (2002). It should be noted that since our experiments did not separately control covert attention shifts and overt eye movements, we cannot distinguish between mechanisms that might separately favor eye movements near the center of attention and the center of fixation. Along this line, future studies should explore how the functionally-defined shape of "central vision" might change with behavioral modalities such as covert attention, eye movements, finger-pointing, or verbal report.

In addition to building our understanding of biological vision, we have aimed to develop computational algorithms that are efficient enough to be useful in real-world machine vision applications. The models of short-range orientation interactions and eccentricity-dependent filtering described here have efficient implementations that do not significantly impact the execution time of the salience model, yet have significant effects on the model's ability to match human behavior. In contrast, the model of contour facilitation requires roughly an order of magnitude more processing time and is weakly relevant to behavior in some task conditions, but is also critically important in predicting behavior under other conditions such as the Gabor snakes that we tested, and also potentially in real-world tasks like road-finding in overhead imagery. Taken together, this suggests that a machine vision implementation might best compute an initial salience map based on local features alone, and secondarily perform more computationally intensive tasks like contour facilitation or object recognition within a restricted window selected by the first stage. Such systems will ultimately be useful both as stand-alone applications and as semi-automated assistants in tasks that rely on a human executor. The interface between biology and engineering is rich in research directions that will lead us closer not only toward understanding the inner workings of vision, but also toward building machines that assist, interact, collaborate, and synergize with real human visual systems.

A Appendix

A.1 Contour-facilitation model

The weights in the connection matrix (shown in Figure 6), following the model of Braun (1999a), are composed of two factors: (1) an orientation-independent factor, $\rho$, depending on the spatial positions $(x_a, y_a)$ and $(x_b, y_b)$ of the receptive field centers of two units, and (2) an orientation-dependent factor, $\psi$, depending on the units' preferred orientations, $\theta_a$ and $\theta_b$, and on the orientation of the line segment connecting the receptive field centers, $\phi_{ab}$. Four external parameters control this model: $d_{max}, k_{exc}, k_{inh}$, and $\beta$. The orientation-independent factor is a function of the Euclidean distance between the two positions:

$$d_{ab} = \left( (x_a - x_b)^2 + (y_a - y_b)^2 \right)^{1/2}$$

$$\rho_{ab} = \frac{d_{ab} \cdot \exp(1 - d_{ab})}{d_{max}}.$$  

The orientation-dependent factor relies on the following definitions. For a given angle $\theta$, we define $\tilde{\theta}$ such that $\theta = \tilde{\theta} + n \pi$ with $n$ being the integer for which $0 \leq \tilde{\theta} < \pi$. Then the canonical difference $\delta$ between two angles is defined as:

$$\delta(\theta_a, \theta_b) = \frac{\pi}{2} - |\tilde{\theta}_a - \tilde{\theta}_b| - \frac{\pi}{2}$$

or, equivalently:

$$\delta(\theta_a, \theta_b) = \begin{cases} |\tilde{\theta}_a - \tilde{\theta}_b| & \text{if } |\tilde{\theta}_a - \tilde{\theta}_b| < \pi/2, \\ \pi - |\tilde{\theta}_a - \tilde{\theta}_b| & \text{otherwise.} \end{cases}$$

Three angles are significant; these are the two units' preferred orientations $\theta_a$ and $\theta_b$ (as well as the average $\alpha_{ab} = (\theta_a + \theta_b)/2$), and the orientation of the line segment connecting the units' receptive field centers:

$$\phi_{ab} = \tan^{-1}\left(\frac{y_b - y_a}{x_b - x_a}\right).$$

From these angles, the orientation-dependent factor $\psi_{ab}$ is given as:

$$\psi_{ab} = \begin{cases} \exp(-\beta \cdot \delta(\alpha_{ab}, \phi_{ab})) & \text{if units } a \text{ and } b \text{ are "nearly collinear,"} \\ 0 & \text{otherwise.} \end{cases}$$

Two units are considered to be "nearly collinear" if the following conditions all hold:

$$\delta(\theta_a, \phi_{ab}) < \pi/4$$

$$\delta(\theta_b, \phi_{ab}) < \pi/4$$

Thus, $\psi$ will be large if the average of the two units' preferred orientations is similar to the orientation of the line segment connecting the two units, which is precisely the condition satisfied by a smooth contour.

Finally, the connection strength $w_{ab}$ between units $a$ and $b$ is given by a weighted sum of an excitatory part and an inhibitory part:

$$w_{ab} = w_{ab}^{exc} + w_{ab}^{inh}.$$
with
\[ w_{ab}^{\text{exc}} = k_{\text{exc}} \cdot \rho_{ab} \cdot \Psi_{ab} \]
\[ w_{ab}^{\text{inh}} = k_{\text{inh}} \cdot \rho_{ab} \cdot \Psi_{ab} \]

Note that the inhibitory nature of \( w_{ab}^{\text{inh}} \) is conferred by choosing a negative value for \( k_{\text{inh}} \), so overall excitatory or inhibitory connections are denoted by positive or negative values of \( w_{ab} \), respectively. The contour-facilitation algorithm can be pruned for CPU efficiency by setting \( k_{\text{inh}} = 0 \) when \( \delta(\theta_x, \theta_y) > \pi/4 \), so that \( w_{ab} = w_{ab}^{\text{exc}} = w_{ab}^{\text{inh}} = 0 \) under those conditions; this significantly reduces the number of computations that must be performed by the algorithm, at the price of a reduced ability to reject orthogonal line segments as unlikely contours. In order to allow algorithm parameters to be unit-less, the connection strength matrix is normalized by the maximum connection strength, so that after normalization the new maximum connection strength is 1. Sample \( w_{ab} \) matrices are illustrated in Figures 5 (“connection kernel”) and 6.

The iterative algorithm for contour-facilitation proceeds independently in three scale bands whose outputs are summed at the end of the process. Each scale band involves a network of several layers of units; some of these layers are triply indexed by \( x \) and \( y \) spatial positions as well as orientation \( \theta \), while others are doubly indexed by the spatial positions only, and finally the dynamic activity in several of the layers is tracked by a time-step counter \( t \): \( I(x, y, \theta) \): input given by oriented filtering of the original input image \( N^i(x, y, \theta) \): activation levels from interactions among units in \( I \)

\[ G^i(x, y) : \text{group-inhibition weights depending on time derivative of } \mathcal{E} \]

\[ S^i(x, y) : \text{leaky orientation-independent units driven by } N \]

\[ \mathcal{E}^i(x, y) : \text{output energy given by sigmoidal transformation of } S \]

Note that in the following description, symbols of the form \( k_{\ast} \) are external free parameters of the model. The immediate input to the contour-facilitation algorithm, \( I(x, y, \theta) \) (labeled as “filter output” in Figure 5), is given by applying the baseline salience model’s dyadic orientation-tuned pyramids (Itti et al., 1998) to the input image. The \( i \)-th entry in the activation matrix, \( N^i(x_i, y_i, \theta_i) \) (labeled as “activation maps” in Figure 5), is given by the transformation of the input \( I \) via the connection weights \( w_{ab} \):

\[ N^i(x_i, y_i, \theta_i) = \sum_j w_{ij} \cdot f(N^{i-1}(x_j, y_j, \theta_j)) \]

\[ \cdot g_{ij}^{i-1} \cdot I(x_i, y_i, \theta_i) \cdot I(x_j, y_j, \theta_j) \]

where \( [\cdot] \) represents rectification, \( f(N^{i-1}(x_j, y_j, \theta_j)) \) is a fast plasticity term that amplifies outgoing connections from units whose activity in the previous time step was high (labeled as “recurrent excitation” in Figure 5):

\[ f(N^{i-1}) = \begin{cases} 1 & \text{if } k_{\text{fast}} \cdot N^{i-1} < 1, \\ 5 & \text{if } k_{\text{fast}} \cdot N^{i-1} > 5, \\ k_{\text{fast}} \cdot N^{i-1} & \text{otherwise}, \end{cases} \]

and \( g_{ij} \) is a group-inhibition term that selectively modulates inhibitory connections (indicated by “modulation of inhibition” in Figure 5):

\[ g_{ij}^{i-1} = \begin{cases} 1 & \text{if } w_{ij} \geq 0, \\ G^i(x_j, y_j) & \text{otherwise}. \end{cases} \]

Then \( S^i(x, y) \) (indicated by the box containing a “Σ” in Figure 5) is given by

\[ S^i(x_i, y_i) = S^{i-1}(x_i, y_i) - k_{\text{leak}} + \sum_{\theta} N^i(x_i, y_i, \theta_i) \]

and \( \mathcal{E}^i(x, y) \) (labeled as “output” in Figure 5) is given by a sigmoidal transformation of \( S^i(x, y) \):

\[ \mathcal{E}^i(x, y) = \left( 1 + \exp \left( \frac{2 - 4 \cdot S^i(x, y)}{k_{\text{sigthresh}}} \right) \right)^{-1}. \]

Finally, the group-inhibition weights are updated based on a lowpass-filtered version of the change in output energy between time steps \( t-1 \) and \( t \):

\[ \Delta^i(x, y) = \text{lowpass} \left( \mathcal{E}^i(x, y) - \mathcal{E}^{i-1}(x, y) \right) \]

\[ G^i(x, y) = \mathcal{G}^{i-1}(x, y) + k_{\text{add}} \cdot \Delta^i(x, y) - k_{\text{sub}} \cdot [k_{\text{bottom}} - \Delta^i(x, y)] \]

with initial group-inhibition values at time \( t = 0 \) of \( \mathcal{G}^0(x, y) = 1 \). So, local inhibitory strength increases if the output energy is increasing at a rate faster than \( k_{\text{gtop}} \), and decreases if the output energy is increasing at a rate slower than \( k_{\text{bottom}} \).

Acknowledgements

This work was supported by the National Geospatial-Intelligence Agency (NGA, formerly known as the National Imagery and Mapping Agency, NIMA), the Sandia National Laboratories, the Engineering Research Centers (ERC) Program of the National Science Foundation under Award Number EEC-9402726, the Keck Foundation, and by a Predoctoral Fellowship from the Howard Hughes Medical Institute to R.J. Peters. The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof. The authors would like to thank Nathan Mundhenk for a reference implementation of the contour-facilitation model.
References


