PERGAMON

Vision Research 42 (2002) 107–123

Vision Research

www.elsevier.com/locate/visres

# Modeling the role of salience in the allocation of overt visual attention

Derrick Parkhurst [a,d], Klinton Law [b,d], Ernst Niebur [c,d,*]

[a] The Department of Psychology, The Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA
[b] The Department of Biomedical Engineering, The Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA
[c] The Department of Neuroscience, The Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA
[d] The Zanvyl Krieger Mind/Brain Institute, The Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218, USA

Received 12 July 2000; received in revised form 5 March 2001

## Abstract

A biologically motivated computational model of bottom-up visual selective attention was used to examine the degree to which stimulus salience guides the allocation of attention. Human eye movements were recorded while participants viewed a series of digitized images of complex natural and artificial scenes. Stimulus dependence of attention, as measured by the correlation between computed stimulus salience and fixation locations, was found to be significantly greater than that expected by chance alone and furthermore was greatest for eye movements that immediately follow stimulus onset. The ability to guide attention of three modeled stimulus features (color, intensity and orientation) was examined and found to vary with image type. Additionally, the effect of the drop in visual sensitivity as a function of eccentricity on stimulus salience was examined, modeled, and shown to be an important determiner of attentional allocation. Overall, the results indicate that stimulus-driven, bottom-up mechanisms contribute significantly to attentional guidance under natural viewing conditions. © 2002 Elsevier Science Ltd. All rights reserved.

*Keywords:* Eye movements; Natural images; Visual attention; Computational model; Salience

## 1. Introduction

The amount of incoming information to the primate visual system is much greater than that which can be fully processed. It is well known that only part of this information is processed in full detail while the remainder is left relatively unprocessed (for reviews see Desimone & Duncan, 1995; Egeth & Yantis, 1997; Niebur & Koch, 1998, chap. 9). For example, from the high-resolution foveal representation in the retina where most processing resources are allocated to the central 5° of the visual field, to late stages of visual cortical processing where receptive fields invariably grow to encompass the fovea, the neural architecture disproportionately represents the central visual field. Additionally, dynamic mechanisms of selective attention focus the processing resources of the visual system by functioning as an information gating mechanism. Together, attentional mechanisms and neural architecture determine what visual information is or is not fully processed.

In order that behaviorally relevant visual information is appropriately selected, efficient mechanisms must be in place. Two major attentional mechanisms are known to control this selection process. First, bottom-up attentional selection is a fast, and often compulsory, stimulus-driven mechanism. There is now clear evidence indicating that attention can be *captured* under the right stimulus conditions. For example, highly salient feature singletons (Bacon & Egeth, 1994; Treisman & Gelade, 1980) or abrupt onsets of new perceptual objects (Yantis & Hillstrom, 1994; Yantis & Jonides, 1984) automatically attract attention. The other mechanism, top-down attentional selection, is a slower, goal-directed mechanism where the observer's expectations or intentions influence the allocation of attention. Observers can volitionally select regions of space (Posner, 1980) or individual objects (Rock & Gutman, 1981; Duncan, 1984; Tipper, Weaver, Jerreat, & Burak, 1994) to attend.

The degree to which these two mechanisms play a role in determining attentional selection under natural

[*] Corresponding author. Tel.: +1-410-516-8643; fax: +1-410-516-8648.
*E-mail address:* niebur@jhu.edu (E. Niebur).

PII: S0042-6989(01)00250-4

viewing conditions has been for a long time under debate. Much of the research relevant to this question has focused on the way in which people make eye movements while viewing complex natural scenes. The logic of this approach rests on the assumption that eye movements and attention are associated. This assumption is a reasonable one given that both eye movements and attention are related to the selection of the most important parts of the visual input. Although the locations of attention and fixation can be dissociated, psychophysical evidence indicates that focal attention at the location of a pending eye movement is a necessary precursor for that movement (Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995; Kowler, Anderson, Dosher, & Blaser, 1995; McPeek, Maljkovic, & Nakayama, 1999; Schneider & Deubel, 1995; Shepherd, Findlay, & Hockey, 1986).

Some of the earliest studies showed that observers preferentially look at people and faces, suggesting a significant role for top-down influences (Buswell, 1935; Yarbus, 1967). Later, more quantitative analyses indicated that observers look at regions which are deemed to be informative (Antes, 1976; Mackworth & Morandi, 1967). Unfortunately, from these studies it is unclear to what degree semantic features (top-down) as compared to visual features (bottom-up) influenced the informativeness ratings. Other evidence suggests that top-down semantic influences do affect attentional guidance, leading to longer and more frequent fixations on items that are inconsistent with scene context (DeGraef, Christiaens, & d'Ydewalle, 1990; Henderson, Weeks, & Hollingsworth, 1999; Henderson & Hollingsworth, 1998; Loftus & Mackworth, 1978). Furthermore, individual observers exhibit idiosyncratic scanpaths upon repeated viewings of the same stimulus (Noton & Stark, 1971), suggesting an that internal representation is created on initial viewing that guides later reviewings. While later studies (Brant & Stark, 1997; Ellis, 1986; Stark & Ellis, 1981) have supported this theory, other results (Mannan, Ruddock, & Wooding, 1997) using complex natural scenes rather than simple line drawings find little or no evidence for repetitive scanpaths. Finally, further evidence of top-down control comes from the effect of task instructions. Both the patterns of fixation locations (Yarbus, 1967) and the spatio-temporal dynamics of eye movements (Andrews & Coppola, 1999) can vary with task.

On the other side of the debate, the strong evidence for stimulus-driven attentional capture indicates that bottom-up selection can influence attentional allocation in simple experimental paradigms, but there is little research examining the extent of bottom-up attentional allocation under more natural viewing conditions. Only recently have studies quantitatively examined the similarity between extracted image features in natural scenes and the fixation locations made when participants free view these scenes. In general, only measures of edge density and local contrast tend to be greater at the points of fixation than at other locations (Krieger, Rentschler, Hauske, Schill, & Zetzsche, 2000; Mannan, Ruddock, & Wooding, 1996; Mannan et al., 1997; Reinagel & Zador, 1999).

We address the extent to which bottom-up, stimulus-driven factors influence the allocation of attention by examining the correlation between stimulus salience, as determined by a biologically plausible computational model of bottom-up selective attention, and human eye movements obtained while viewing complex natural and artificial scenes. The model, shown in Fig. 1, functionally implements many of the processes important in early vision (e.g. center-surround organization, lateral inhibition and multiscale interactions). In a purely bottom-up fashion, the model takes as input an image and processes it in three parallel feature channels using a range of spatial scales. The resulting topographic feature maps are then combined across scales and channels to form a "saliency map" (Koch & Ullman, 1985). The saliency map indicates the most salient, or visually important, regions in the image. If attention is stimulus-driven under normal viewing conditions, there should be a positive correlation between the locations of fixation and the salience of the stimulus at those locations. This logic rests on the assumption that eye movements and attention are associated. As mentioned earlier, this assumption is valid for a number of reasons. First, both eye movements and attention can serve the same goal, selecting the instantaneously most important parts of the visual input. Second, our examination of attentional allocation is limited to natural viewing conditions, when constraints that might dissociate eye movements and attention are limited. Finally, psychophysical evidence indicates that focal attention at the location of a pending eye movement is a necessary precursor for that movement (Deubel & Schneider, 1996; Hoffman & Subramaniam, 1995; Kowler et al., 1995; McPeek et al., 1999; Schneider & Deubel, 1995; Shepherd et al., 1986).

To test the correlation between stimulus salience and fixation locations, an experiment was conducted where eye movements were recorded while participants free viewed four different types of images (see Fig. 2 for examples). The images ranged in degree of realism from computer generated fractals, where we suspected eye movements to be the most stimulus-driven, to home interiors and building and city scenes, where we suspected top-down influences to be much more common. Participants were required to free view each image for five seconds. The free viewing task was chosen in order to avoid introducing task dependent top-down effects on eye movements. In addition, the free viewing task was chosen because it most closely approximates natural viewing conditions.

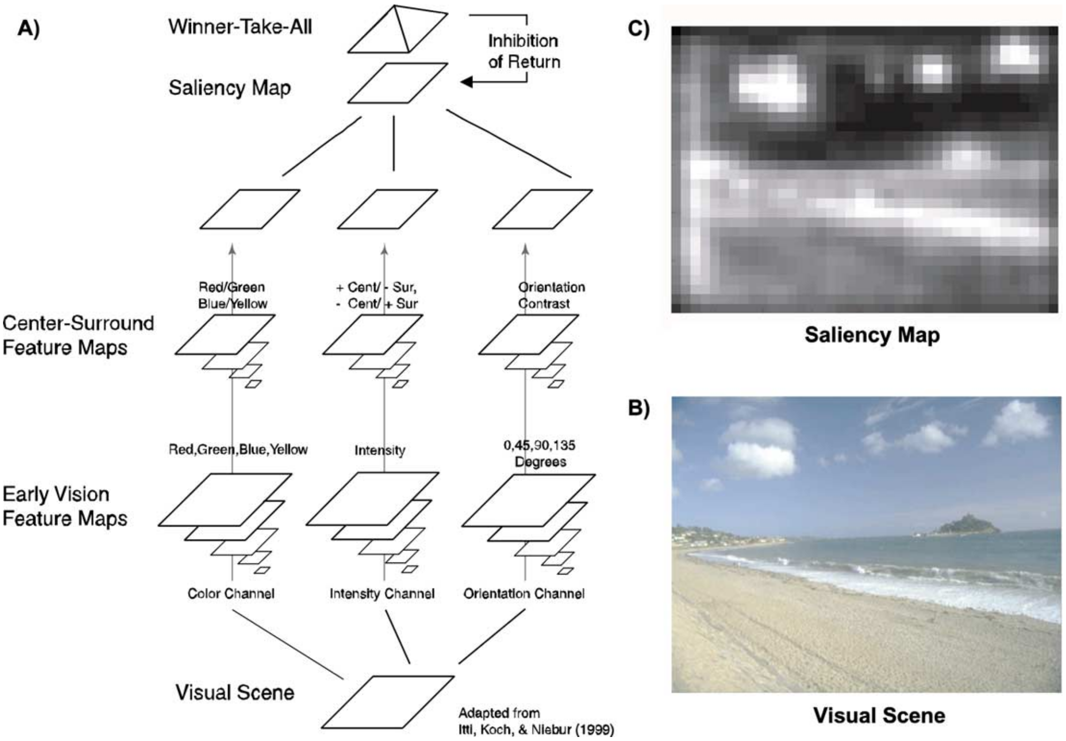

Fig. 1. (A) Schematic view of the model. The input image is separated into three parallel feature channels (color, intensity, and orientation) and sampled at a series of spatial scales. Feature activity is propagated to the next level and reorganized into a center-surround arrangement. Finally, activity is normalized within each feature channel and linearly summed to form the salience map. Dynamic allocation of the focus of attention is determined through a winner-take-all process in the saliency map. Once the focus of attention has been shifted to a location, inhibition of return lowers the salience at that location allowing the focus of attention to shift to a new location. (B) An example image used in the experiment. (C) The saliency map generated from B with regions of high salience shown in white.

In analyzing the correlation between eye movements and salience, we initially focused on the locations of the first fixation after each trial started. Next, we examined the following fixations. Given the slower onset of top-down attentional effects (for a detailed review of the time course of top-down attention see Egeth & Yantis, 1997) and the necessity to acquire at least some information from the visual input before top-down influences can be exerted on a given image, we suspected a stronger correlation with bottom-up influences for early fixations than for later fixations. Subsequently, to further investigate the dependence of attentional allocation on stimulus properties, the ability of each of the three features channels to guide attention was examined independently. Finally, the role of the decline in visual sensitivity as a function of eccentricity in determining attentional allocation was examined in the context of the model.

## 2. Model

Ever since Broadbent (1958) first proposed the filter theory of selective attention, the two-stage framework has pervaded conceptual and computational modeling efforts (Cave & Wolfe, 1990; Findlay & Walker, 1999; Neisser, 1967; Theeuwes, 1993; Treisman & Gelade, 1980; Wolfe, 1994). Within this framework, the first stage preattentively processes all incoming visual information equally and in a parallel fashion. The degree of processing is rudimentary, consisting only of a simple feature-based decomposition (e.g. color and orientation). At the interface between the first and second stages is a *filter* or *bottleneck* that functions as a gate allowing only part of the visual information to proceed to the second stage. The processing of the second stage differs from that of the first in capacity and level of detail. It has a limited capacity, being only able to process one or possibly a few objects simultaneously, and it processes the visual information to a much higher level of detail (e.g. object-based representations). This model framework has served as a cornerstone to interpreting most of the results in the visual attention literature. Indeed, converging evidence from neurophysiological and neuroanatomical studies suggests a plausible neural implementation of the two-stage model in the primate visual cortex. Early cortical areas show cellular response properties similar to some of the

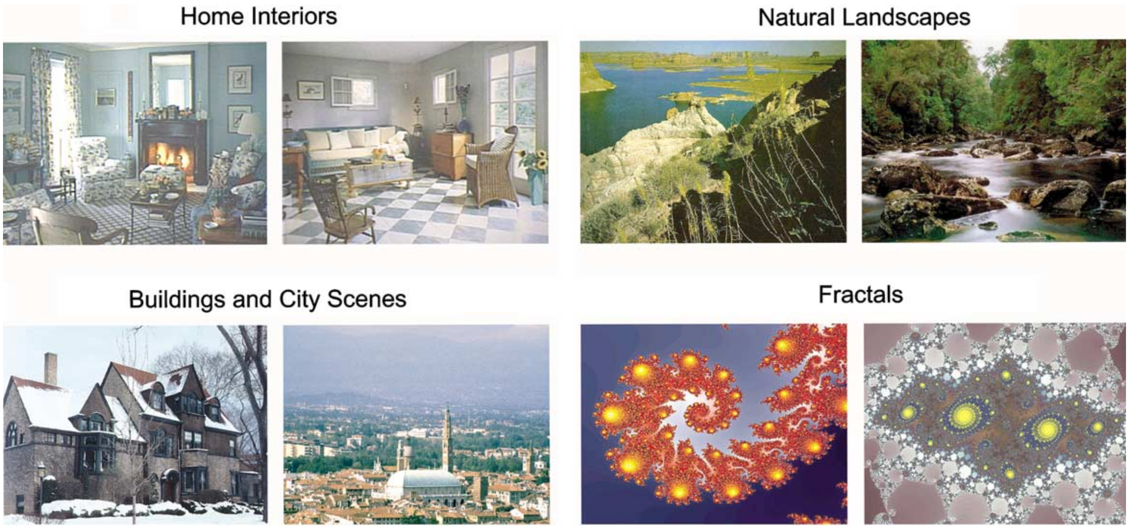

Fig. 2. Examples of the four classes of images used in the experiment.

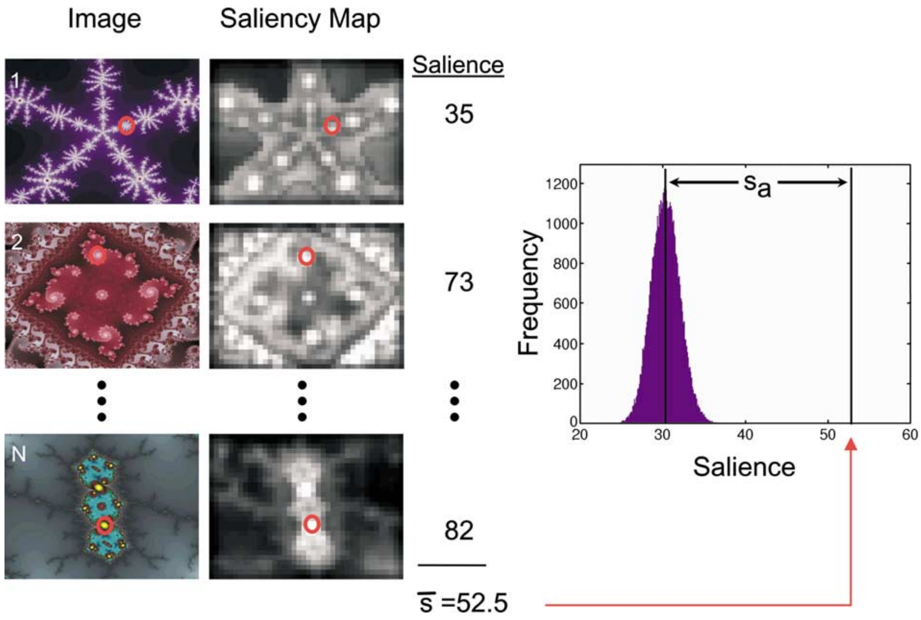

Fig. 3. The method for quantifying the correlation between stimulus salience and fixation locations is illustrated for one image database. The location of the first fixation after stimulus onset is extracted from the eye movement record and indicated by a red circle on each image (left). A saliency map is generated for each image in the database and the saliency at the first fixation location is extracted (center). The mean of the extracted salience values ($\bar{s}$) is calculated across images and compared to the distribution of $\bar{s}$ expected by chance (right). The distance between the $\bar{s}$ obtained as a fixation location and the mean $\bar{s}$ expected by chance alone is referred to as the chance-adjusted salience $s_{\mathrm{a}}$.

stimulus features that support pop-out.[1] These cells typically have small receptive fields scattered throughout the visual field, suggesting parallel processing. Later areas show much more complex response properties and have receptive fields that typically cover a large portion of the visual field, suggesting serial processing.

Described in this section is a biologically plausible model of bottom-up visual selective attention. As can be seen in Fig. 1, visual input is provided to the model and segregated into three separate parallel feature channels, one each for color, intensity, and orientation. Although these feature dimensions do not represent an exhaustive set of those important in determining salience, modeling these features allows the model to account for a large

[1] Pop-out is a perceptual phenomenon where one unique feature will stand out from a field of surrounding features. Pop-out typically occurs for simple features and is attributed to the fast parallel processing that occurs in early vision.

number of psychophysical results. For example, the model can account for many preattentive pop-out effects in visual search (Bravo & Nakayama, 1992; Julesz, 1984; Treisman & Gelade, 1980; Treisman & Gormican, 1988), but as it is exclusively a model of bottom-up attentional allocation, it cannot account for many other results (see Section 7). Within each feature channel, visual input is first sampled at a series of spatial scales to form a set of topographic feature maps. Then at the next level of processing, these features are reorganized into a center-surround arrangement that is characteristic of receptive field organization seen throughout the primate visual system (Wandell, 1995). The center-surround feature maps are then combined across scale and normalized within each channel. Next, the resulting maps are linearly summed across feature channels to form the saliency map. Following Koch and Ullman (1985), the fundamental hypothesis in our approach is that the saliency map indicates which locations in the visual input are most salient, or are most visually important. As the model only takes bottom-up information into account, in this framework the focus of attention is solely determined by a winner-take-all process in the saliency map (ibid). In other words the focus of attention is directed to the location of peak salience. Once attention is shifted to that location, inhibition of return, a bias that inhibits attentional selection of previously selected locations (Posner & Cohen, 1984) reduces salience at the current focus of attention. Consequently, a new salience peak will dominate and cause attention to shift to another location. Although the dynamics of attentional shifts are clearly an important aspect of visual selection, this study focuses primarily on the ability of the saliency map to predict fixation locations. Other work has focused on the dynamic aspects of attentional allocation in the model (Niebur & Koch, 1996; Itti & Koch, 1999, 2000).

A rudimentary description of the model is presented in the remainder of this section. For a more comprehensive treatment see Niebur and Koch (1996) and Itti, Koch, and Niebur (1998). The model takes digitized images as input (see Fig. 1B). This input is then broken down and processed in parallel in the aforementioned three feature pathways. At the first stage of each of the pathways, the input image is sampled at nine spatial scales in a Gaussian pyramid scheme (Burt & Adelson, 1983). The spatial scales of the pyramid range from no reduction (1:1) to maximal reduction (1:256) in powers of 2. For the first feature channel, four color pyramids (red, green, blue and yellow) are constructed from the RGB formatted input image. For the second feature channel, one intensity pyramid is constructed by taking the average luminance across all the RGB image color components. The third feature channel consists of four orientation pyramids (0°, 45°, 90°, 135°) constructed by convolving the input image with an oriented Gabor function of the appropriate orientation and scale. Gabor convolution was used because it approximates the receptive field structure of orientation selective neurons commonly found in the primary visual cortex (Hubel & Wiesel, 1968, 1977).

At the second stage of processing, the features in each channel and each scale are reorganized into a center-surround arrangement. This arrangement, which is ubiquitous throughout the visual system, functions to maximize local differences and increase contrast. In the model, center-surround reorganization is accomplished by taking the difference of simple features at different spatial scales within each pyramid. For the color channel, two center-surround pyramids are created to model the double-opponent color system of early vision. One pyramid is sensitive to center-surround differences in red and green and the other to differences in blue and yellow. For the intensity channel, a single center-surround pyramid is constructed by computing differences across scales within the intensity feature map and effectively models the color-blind magnocellular pathway (Hubel & Livingstone, 1990). Similarly for the orientation channel, each of four center-surround pyramids is constructed by computing across-scale differences within the corresponding orientation pyramid. This representation of orientation contrast models the non-classical receptive field influences typically seen in primary visual cortex (Allman, Miezin, & McGuinness, 1985) and implements the psychophysical orientation contrast that drives texture segmentation (Leonards & Singer, 1998; Nothdurft, 1991, 1993, 2000).

Within each center-surround feature map, a normalization procedure is then applied. The normalization scales the activations of a map to the squared difference between its global maximum and the average of the remaining (local) maxima in the map (such a normalization has been used successfully to explain single-cell behavior in primary visual cortex; Carandini & Heeger, 1994). Functionally, this operation is akin to a lateral inhibition mechanism between neighbors with similar activation values (a widespread neural mechanism). Furthermore the normalization remaps the activations from dynamic ranges specific to a particular feature or particular spatial scale to a range that depends on how much a particular feature stands out. This remapping across different features and spatial scales is required for the subsequent development of a single saliency map which is common to all feature types and scales. Although more realistic normalization procedures have been implemented in the context of this model (for example see Itti & Koch, 2000), the results obtained using the current implementation are qualitatively similar and preferred for reasons of computational and conceptual simplicity.

The final processing stage involves the construction of the saliency map. First, for the color and orientation channels, the individual pyramids are linearly summed within each channel to obtain one pyramid per channel. Then a unimodal saliency map is created for each feature channel through cross-scale addition of the feature maps. The cross-scale addition reduces all the center-surround feature maps to the spatial scale of the saliency map (1:16 relative to the original images). This scale was chosen because it leads to a resolution of 40 × 30, similar to what has been proposed for the resolution of selective attention in human observers (Verghese & Pelli, 1992). The resulting unimodal saliency maps are then normalized again before summing them to form the final, common saliency map. The last normalization serves to enhance within-feature competition across spatial scales and to remap all the unimodal saliency maps to equivalent dynamic ranges. Finally, it is convenient to scale the saliency map to a range from 0 (global minimum) to 100 (global maximum).

## 3. Experimental methods

Four Johns Hopkins students (two female) were paid for participation in the experiment. Each participated in three half-hour sessions, with a break every five minutes. All participants had normal or corrected-to-normal vision and all were naive with respect to the purpose of the study.

Over the course of the experiment, participants were presented images from four different databases. For three of the databases (home interiors, natural landscapes, and buildings and city scenes) the images were digitized from photographs whereas the remaining database (fractals) was computer generated. All images were displayed fullscreen (30° × 22.4°) at a resolution of 640 × 480 pixels in 16-bit color mode. Representative images are shown in Fig. 2.

Prior to beginning the experiment, participants were told to free view the images and that the only requirement was that they "look around at the images". At the beginning of each trial only a fixation cross was presented at the center of the screen and the participants were required to fixate and to press a mouse button to commence the trial. At this time, an image was presented for a period of five seconds. Subsequently, the display was blanked and the fixation cross for the next trial was displayed. Participants viewed each image once until the database was exhausted. Image type was blocked such that all images from one database were viewed before proceeding to another database.

At the beginning of each section of 25 trials, the eye tracker was calibrated. The calibration phase consisted of a series of nine fixation crosses that the participants were required to sequentially fixate. At the end of each section, an eye tracking error measurement was taken by having the participants fixate 10 randomly positioned crosses. The mean distance between actual and obtained positions was used to estimate the quality of eye tracking. The participants were given the opportunity to take a short break before the next section of trials.

Participants were seated comfortably at a normal viewing distance (58 cm) in front of a standard 17 in. computer screen (cathode ray tube) that was used for stimulus presentation. All stimuli were presented full-screen and subtended 30.0° of visual angle horizontally and 22.4° vertically. A custom-made head rest provided support for chin and forehead in order to minimize the effects of head movements.

An ISCAN model RK-416 eye tracker was used to monitor eye position. This model is a real time digital image processor that tracks the center of the participant's pupil and measures its size from an infrared video image of the participant's eye. The unit automatically computes the position of the pupil over the two-dimensional matrix of the eye imaging camera. Pupil coordinates and diameter are computed at a rate of 60 Hz. A bi-cubic non-linear interpolation (cubic in both horizontal and vertical dimensions) between a grid of nine calibration points was used to calibrate the eye tracker (Stampe, 1993). This procedure helped to minimize errors from non-linearities due to infrared source reflections. Additionally, the calibration was adjusted using a procedure where an eye sample from the fixation point at the beginning of each trial was used to re-align the original nine point interpolation. Full recalibration and adjustment of the eye tracker was intermittently required during a block of trials in the case of excessive head movements. Fixation locations and durations were extracted from the raw eye tracking data using velocity (saccadic velocity: greater than 30°/s) and duration (fixation duration: greater than 100 ms) criteria (Stampe, 1993).

## 4. Stimulus dependence of attentional guidance

### 4.1. Data analysis

In order to quantify the correlation between stimulus salience and eye movements the following procedure (as diagrammed in Fig. 3) was conducted for each image database on a participant by participant basis. To begin, the coordinates $(f_x^k, f_y^k)$ of the $k$th fixation location following stimulus onset were extracted from the raw eye tracking data for a given image. The model was presented with the same image and allowed to generate a saliency map ($S$) as described in Section 2. Next, the salience at the fixation locations was extracted from the corresponding saliency map and the average salience ($\bar{s}^k$) was computed across all salience values obtained for the images in a given database:

$$\bar{s}^k = \frac{1}{n}\sum_{i=1}^{n} S_i(f_x^k, f_y^k) \qquad (1)$$

where $i$ is the image number and $n$ is the number of images in the database.

Given that salience is scaled to range from 0 to 100, a value of $\bar{s} = 100$ would indicate a perfect correlation between the location of highest salience in the salience map and the observed fixation locations. On the other hand, if $\bar{s} = 0$ were found, consistently low salience values would be associated with fixation locations of high probability. To quantify the correlation between stimulus salience and fixation locations, we compare the values of $\bar{s}$ obtained at the observed locations of fixation with the salience value expected by chance. We refer to the difference between the mean salience obtained at the observed fixation locations and the mean salience expected by chance as the chance-adjusted salience ($s_a$). The chance-adjusted salience is the preferred metric for comparison across conditions because the salience expected by chance alone varies with image database (see Fig. 4). Note that chance-adjusted salience ($s_a$) can be used to estimate the relative probability of fixating regions of high salience or the inverse, the probability of fixating regions of low salience. If $s_a$ is positive, regions of high salience would be fixated with a greater probability than other regions and if $s_a$ is negative, regions of low salience would be fixated with a greater probability.

One way to estimate the $\bar{s}$ expected by chance alone is by recalculating $\bar{s}$ using randomly chosen locations in the saliency map instead of the observed fixation locations. If the $\bar{s}^k$ obtained using the observed fixation locations is similar to the $\bar{s}$ obtained at random locations, then this would indicate that attention is not influenced by stimulus properties. If on the other hand, the $\bar{s}^k$ obtained using the observed fixation locations is greater (or possibly smaller) than the $\bar{s}$ obtained at random locations, this would indicate that attention is guided by stimulus properties.

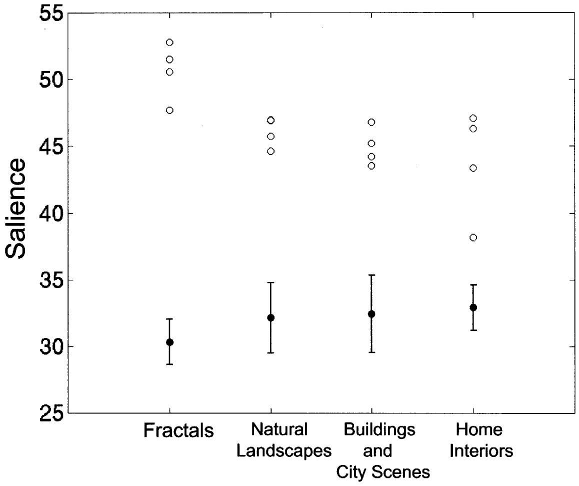

Fig. 4. The mean salience at the first fixation location is shown as an open circle for each participant within each database. The mean salience expected by chance for each database is shown as a closed circle with errorbars indicating plus/minus one standard error of the mean. Each observation significantly differs from chance. Stimulus dependence for the fractal images was the highest.

By recalculating $\bar{s}$ many times using randomly chosen locations in the saliency map a sampling distribution of $\bar{s}$ due to chance factors alone can be generated. The distribution is presented as a histogram for the fractals image database in Fig. 3. The mean of this sampling distribution gives the average salience that would be expected by chance alone, and the standard deviation gives the standard error of the mean (Efron & Tibshirani, 1993). It is clear that $\bar{s}^1$ (vertical line in Fig. 3) significantly differs from chance as it is many standard errors away from the mean $\bar{s}$ expected by chance alone. Although an estimate of the mean salience expected by chance alone and standard error of that mean can be calculated using a bootstrap, an exact analytic solution is available in this case (Efron & Tibshirani, 1993). Therefore, the $\bar{s}$ expected by chance and its standard error are analytically computed rather than using the bootstrap which was described to provide an intuitive understanding of chance performance. To evaluate the significance of an observed $\bar{s}$, a $z$-score is calculated and the implied $p$-value is derived from the normal distribution.

### 4.2. Results

The mean salience at the first fixation location of each participant for each image database are shown in Fig. 4. The salience for each participant is plotted as an open circle and the mean salience expected by chance, as computed in Section 4.1, is plotted as a closed circle, with the vertical bars indicating plus/minus one standard error of the mean. Significance was evaluated for each participant within each image database as described in Section 4.1. In every case, a significant result was obtained (always $\underline{p} < 0.001$, although usually with much lower $\underline{p}$-values). In addition, a one-way repeated-measures ANOVA was conducted with image type (fractals, natural landscapes, buildings and city scenes, and home interiors) as the relevant factor. A significant main effect of image type was observed ($\underline{F}(3,9) = 5.12$, $\underline{p} < 0.05$). Post hoc comparisons between the means using the Newman–Keuls procedure ($\alpha = 0.05$) indicated a stronger correlation between salience an fixation locations for the fractals ($\underline{M} = 50.65$) than for the natural landscapes ($\underline{M} = 46.08$), buildings and city scenes ($\underline{M} = 44.95$), or for the home interiors ($\underline{M} = 43.74$). No other effects were significant.

The chance-adjusted salience averaged across image databases and participants is shown in Fig. 5 as a function of fixation number. The error bars represent plus/minus one standard error of the mean across participants. A two-way repeated-measures ANOVA was conducted with image type and fixation number as the relevant factors. A significant main effect of fixation

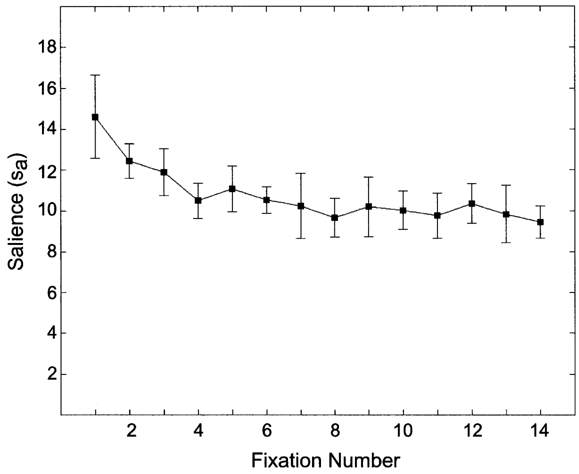

Fig. 5. The mean chance-adjusted salience for all databases is shown averaged across participants as a square where the errorbars represent plus or minus one standard error of the mean. Stimulus dependence is greatest for early fixations, but remains highly above chance levels throughout the trial.

number was observed ($\underline{F}(13,39) = 6.96$, $\underline{p} < 0.001$) where early fixations showed a higher chance-adjusted salience than later fixations. Comparing the first fixation to the remaining fixations, an approximately 40% stronger correlation is observed. A significant main effect of image type was also observed ($\underline{F}(3,9) = 20.79$, $\underline{p} < 0.001$). Post hoc comparisons between the image types using the Newman–Keuls procedure ($\alpha = 0.05$) indicated that the correlation between salience an fixation locations for the home interiors ($\underline{M} = 7.76$) was weaker than the correlation for the fractals, natural landscapes, or buildings and city scenes ($\underline{Ms} = 12.92$, 10.59 and 11.41). In addition, fractals showed a stronger correlation than the natural landscapes. No other effects were significant.

### 4.3. Discussion

The goal of this experiment was to examine the stimulus dependence of eye movements under normal viewing conditions using a bottom-up computational model of attention. If attention is stimulus-driven, there should be a positive correlation between the locations of fixation and the salience of the stimulus at those locations. The first fixation location analysis results, shown in Fig. 4, clearly indicate a significant correlation between stimulus salience and fixation locations. For each participant within each image database the mean salience at the first fixation locations was many standard errors away from the mean salience expected by chance alone. This significant correlation indicates that the selection of the first fixation location is indeed guided by stimulus properties.

The stimulus dependence of later fixations was also examined and plotted in Fig. 5. Stimulus dependence was highest for fixations that immediately followed stimulus onset. This is consistent with the assumption of a slow onset of top-down attentional effects. In all cases, however, stimulus dependence reached an asymptotic level for later fixations rather than dropping to chance levels or below. Both the fact that the time course to asymptote is over many fixations (many hundreds of milliseconds), and the fact that stimulus dependence remained significantly greater than expected by chance throughout the trial indicates that stimulus properties are more influential in attentional guidance than previously thought. Overall, these results indicate that eye movements are indeed stimulus-driven under normal viewing conditions.

## 5. Relative strength of each feature dimension

To further probe the nature of the stimulus dependence of attentional allocation, the relative contributions of different feature dimensions were examined. As illustrated in Fig. 6, visual input to the model is processed in three independent feature channels representing color, intensity, and orientation. The culmination of processing in each channel is a map that indicates the salient locations in the image with respect to only one feature dimension. The final saliency map is a linear sum of these three submodality saliency maps. The following analysis aims to determine the relative contribution of each feature dimension by examining the correlations between the submodality saliency maps and the observed fixation locations.

### 5.1. Data analysis

In order to quantify the relative strength of each feature dimension, the procedure described in Section 4.1 (as diagrammed in Fig. 3) was repeated using each submodality salience map instead of the combined salience map. The relative strength of each feature channel was calculated as a ratio of the chance-adjusted salience obtained using each submodality map relative to the chance-adjusted salience obtained using the combined salience map. A relative strength of one would mean that fixation locations correlated with regions of high salience in the submodality salience map as well as they did in the combined salience map. This would indicate that the relevant stimulus property strongly underlies the observed stimulus dependence. On the other hand, relative strength zero would mean that fixation locations did not correlate at all with regions of salience in the submodality map, and therefore would indicate that the stimulus property has little influence on the guidance of attention. A measure of relative strength was calculated

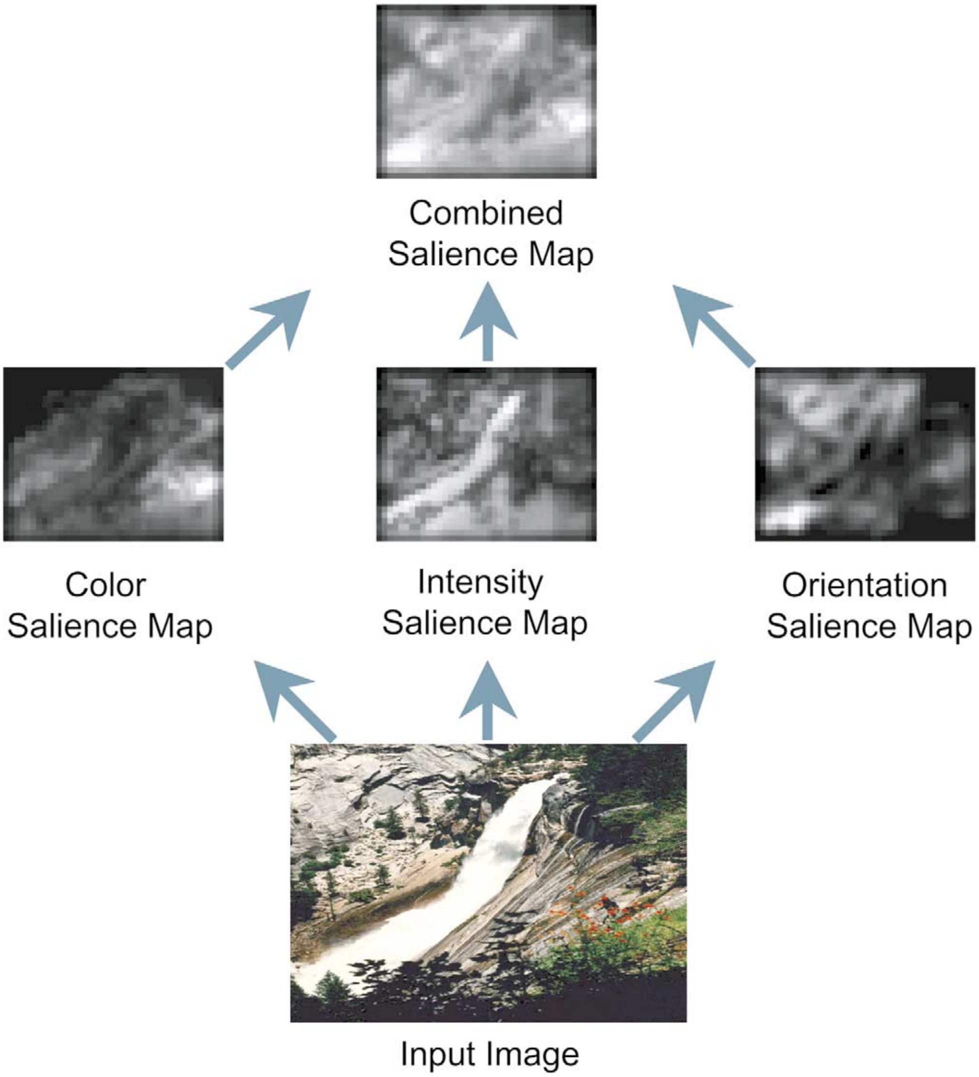

Fig. 6. The generation of the combined salience map is illustrated. The input image is separated into three independent feature channels and processed to create submodality salience maps. The salience maps indicate locations in the image which are highly salient based on only one stimulus property. For example, in the color salience map, the area corresponding to the greenery and red flowers is salient whereas in the intensity salience map, the location corresponding to the bright reflection from the waterfall is salient. The combined salience map is generated by summing together the three submodality salience maps. It indicates the locations in the image which are highly salient based on all the computed features.

independently for each feature channel, each image database, and each participant by averaging over the results obtained from the analysis using each of the first 14 fixation locations.

### 5.2. Results

Shown in Fig. 7 are the relative strengths obtained for each image database and each feature channel averaged across participants where the error bars represent plus/minus one standard error of the mean across participants. A two-way repeated-measures ANOVA was conducted with image type and feature channel (color, intensity and orientation) as the relevant factors. A significant main effect of feature channel was observed ($\underline{F}(2,6) = 15.33$, $\underline{p} < 0.005$). Post hoc comparisons between the feature channels using the Newman–Keuls procedure ($\alpha = 0.05$) indicated that the correlation between salience and fixation locations in the color and intensity salience maps ($\underline{M}s = 0.81$, 0.81) was stronger than the correlation for the orientation salience map ($\underline{M} = 0.64$). A small but significant main effect of image type was observed ($\underline{F}(3,9) = 32.80$, $\underline{p} < 0.001$) as was an interaction between image type and feature channel ($\underline{F}(6,18) = 5.88$, $\underline{p} < 0.005$). No other effects were significant.

### 5.3. Discussion

The goal of this analysis was to examine the role that different feature dimensions play in attentional allocation. One might have suspected to find a clear rank

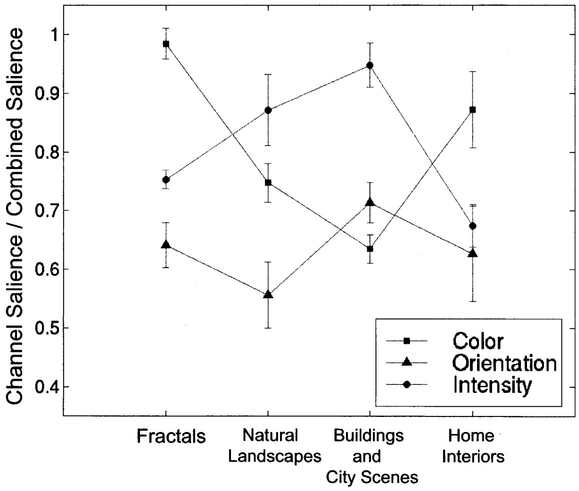

Fig. 7. The mean ratio of chance-adjusted salience for a single feature channel relative to the chance-adjusted salience for all channels is shown for each image database and each feature channel. The ratios represent the mean across participants and are shown as circles plus or minus one standard error of that mean.

ordering of feature dimension strengths based on their behavioral importance. Instead, we found that the relative strength of each feature dimension depends strongly on the image type. Although the range of image types used was small, the rank ordering of importance, shown in Fig. 7, varies dramatically. For instance, in the fractals and home interior image databases, color played a dominant role in the guidance of attention whereas intensity dominated in the natural landscapes and buildings and city scenes databases. In general, the color and intensity feature channels (both with equal contributions of 0.81 when averaged across image databases) contributed more than the orientation feature channel (0.64). A notable exception to this pattern is for the buildings and city scenes where orientation contributes more than color. This is most likely due to the strong lines present in the architecture and the low saturation typical of the colors in these images. The results of this analysis clearly indicate the importance of integrating a range of features to calculate stimulus salience when modeling attentional allocation.

It is interesting to note that relative strengths obtained for each feature dimension do not sum to 1, but rather, sum to approximately 2.3. This indicates the presence of redundancy between channels since summing to 1 would indicate independence of the feature channels. This is not surprising given that the submodality salience maps show an average correlation of 0.34 with each other. There are several potential explanations for this redundancy. First, although each feature channel operates independently, the extraction of information for different channels may partially rely on the same information. For example, two adjacent areas with different contrast usually generate an oriented edge. Additionally, multiple feature properties in the scene may be spatially co-located, for example when both color and intensity contrasts are present at the same location. Interestingly, recent psychophysical results suggest that salience, even when signaled independently by more than one dimension, fails to sum linearly (Nothdurft, 2000), which is in agreement with the present results.

## 6. The role of visual sensitivity in attentional selection

Although the results of the previous analyses clearly indicate that stimulus properties play an important role in guiding attention and eye movements, qualitative comparison between the model predictions and the observed eye movements indicate an interesting discrepancy which is shown in Fig. 8. The location of the first fixation following stimulus onset is plotted as points for the participants and as boxes for the model. The participants' first fixation locations tend to be clustered around the center of the image, site of the fixation point before the test images are presented, whereas the locations predicted by the model are more uniformly distributed across the entire image. Shown in Fig. 9A and B are histograms of the first saccade distances (i.e. the distance between the fixation point and the first fixation location) for the participants and the model, respectively. The participant distribution is positively skewed with a mean of about 5° whereas the model distribution is less skewed and has a much larger mean.

Both Figs. 8 and 9 show that participants are biased to preferentially make saccades to targets that are positioned close to the current fixation location. A bias towards central targets is commonly observed in studies

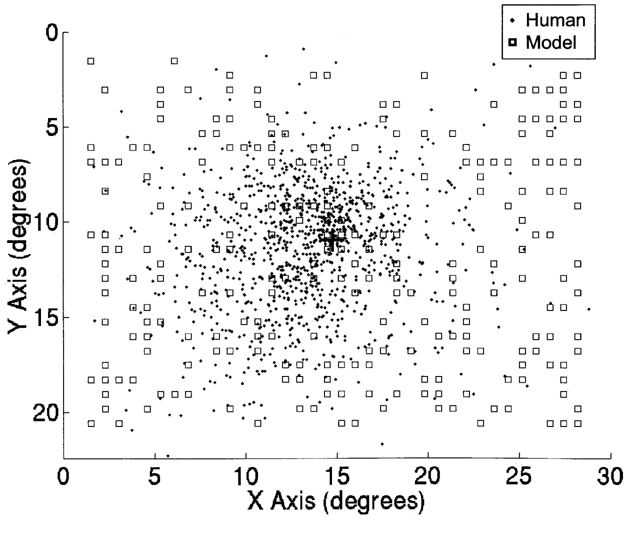

Fig. 8. First fixation locations for the participants (shown as points) and the model (shown as boxes). Fixations for the participants are noticeably biased towards the center, while those predicted by the model are more uniformly distributed.

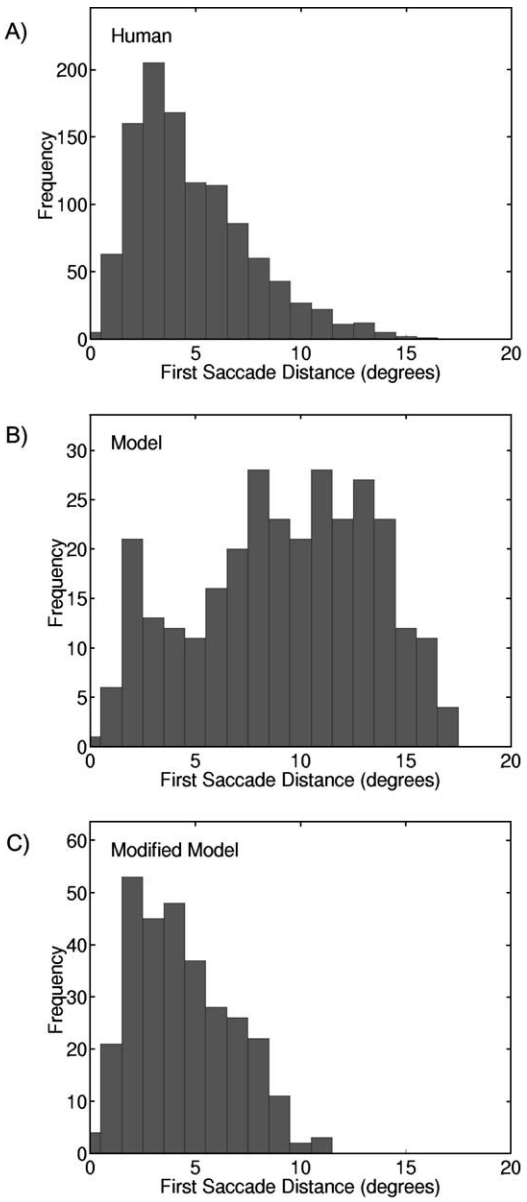

Fig. 9. (A) and (B) Histogram of saccade distances between the fixation point and the first fixation location for the participants and the model, respectively, across all databases. Note that the participant data are positively skewed with a mean of about 5° and a mode of 3° whereas the modeled data is less skewed and has a much larger mean. (C) Histogram of saccade distances between the fixation point and the first fixation location for the fixation model with ($\sigma = 9.5°$), see text. The distribution of the fixation model resembles that of the participants seen in (A). Each histogram was calculated using the data collected from all four image databases.

of scene viewing (Mannan et al., 1996, 1997) as well as visual search (Carrasco, Evert, Chang, & Katz, 1995; Engel, 1971, 1977; Wolfe, O'Niel, & Bennett, 1998). A potential explanation for this pattern of results is the drop of visual system sensitivity to high spatial frequencies in the periphery. It is known that spatial frequency sensitivity in the central region of the visual field exceeds that in peripheral regions by an order of magnitude (Wandell, 1995). One reason for this difference is that photoreceptor density varies across the surface of the retina. It is greatest in the fovea, the part of the retina to which the central visual field projects, and lowest in the surrounding regions. This inhomogeneous representation of the visual field continues at subsequent stages of the visual pathway. For instance, the neural projections from the retina (through the lateral geniculate nucleus of the thalamus) to the visual cortex follow a non-linear mapping with more cortical processing power being dedicated to the central visual field (Schwartz, 1977). Moreover, as one considers more downstream visual cortical areas, visual receptive fields of most neurons grow and eventually most come to encompass the central visual field (Boussaoud, Desimone, & Ungerleider, 1991; Gattass, Sousa, & Gross, 1988).

This drop of visual sensitivity is not implemented in our model; resolution is kept constant across the feature maps. To evaluate the effects of decreased visual sensitivity in the periphery, a set of images from the home interior database was filtered such that the spatial frequency content progressively fell as the distance from the central fixation point increased. A seventh-order Butterworth filter was used to achieve a spatial frequency reduction that correlated with the maximum detectable spatial frequency of a sinusoidal grating presented at the relevant eccentricity (Virsu & Rovamo, 1979). These images, which simulate the drop in sensitivity of the visual system at peripheral locations, were used as input to the model. Qualitatively, the resulting saliency maps showed a reduction of salience in the periphery that was proportional to the decrease in spatial frequency content. This result suggests that the decrease in sensitivity as a function of eccentricity functions to reduce the salience of objects located in the periphery.

To quantitatively examine the role that the decline of visual sensitivity in the periphery has in determining attentional allocation, the model was modified and the results re-analyzed using the methods of Section 4.1. The model was modified such that when analyzing the salience at a particular fixation, the salience peripheral to the previous fixation location was reduced. This simulates the reduction in peripheral salience that would occur due to the drop in visual sensitivity in the periphery. If this drop plays a role in attentional allocation, then there should be an increase in the correlation between stimulus salience and fixation locations. The next section describes in detail how this prediction was tested.

### 6.1. *Modified model*

In order to take into account the drop of visual sensitivity in the periphery, and the resulting drop in peripheral salience, the model was modified in the following manner. A first-order approximation to the reduction in peripheral salience was used where salience is scaled by a weighted function of distance from the current fixation location in the image. This reduction was modeled by weighting the saliency map, generated as described in Section 2, by a two-dimensional Gaussian filter. The resulting saliency map ($S'$) is given by

$$S'(x,y) = S(x,y)\exp\left(-\frac{(x-f_x^0)^2+(y-f_y^0)^2}{2\sigma^2}\right) \qquad (2)$$

where $\sigma$ determines the width of the Gaussian and consequently controls to what degree salience is reduced in the periphery. When creating the saliency map relevant for analyzing the first fixation location after stimulus onset, the Gaussian reduction must be centered on the fixation location prior to the first fixation location. In other words, it must be centered on the location where the participant was fixating prior to stimulus onset; the central fixation point whose coordinates we will call $(f_x^0, f_y^0)$. More generally, to generate the relevant saliency map for analyzing a particular fixation $(f^n)$, the Gaussian reduction must be centered on the previous fixation $(f^{n-1})$. We refer to this model as the modified fixation model.

Since all trials begin by the participants fixating the center of the screen (the fixation cross), reducing peripheral salience in the model has the effect of reducing the number of first fixation locations seen in the periphery and increasing the number near the center. This redistribution can be seen in the histogram of simulated saccade distances for the modified model plotted in Fig. 9C. Compared to the original model distribution shown in Fig. 9B, the modified model distribution has a smaller mean and is more positively skewed resembling that observed for the participants shown in Fig. 9A.

### 6.2. Data analysis

Given that modeling the reduction in peripheral salience improves the correspondence between the predicted and observed distributions of first fixation locations, the correlation between salience and fixation locations, as described in Section 4.1, was also examined. For the modified model, the salience map is weighted by a Gaussian with a standard deviation ($\sigma$) chosen individually to optimize the correspondence for each image database of each participant. Plotted in Fig. 10 is the correlation between stimulus salience and the first fixation location across a range of $\sigma$ for one participant who viewed the fractal image database. The correlation is plotted in terms of the chance-adjusted salience ($s_a$). This representation is useful in the comparison across different values of $\sigma$ because as salience is reduced in the periphery, the mean salience expected by chance is also reduced. For comparison purposes, the chance-adjusted salience obtained from the uniform model is also plotted (diamond at far right of Fig. 10). As expected, for large $\sigma$ with correspondingly small reduction in peripheral salience, the chance-adjusted salience approaches the value seen with the uniform model. On the other hand, for small $\sigma$, corresponding to a very strong reduction in peripheral salience, the chance-adjusted salience approaches zero. This is again expected since the only salient locations remaining are those in direct proximity to the fixation point. In between these two extremes lies an optimal $\sigma$ (in this example, $\sigma \approx 5°$) that maximizes the chance-adjusted salience by reducing the peripheral salience of distant locations, yet not so much as to overly restrict the range of potential fixation locations.

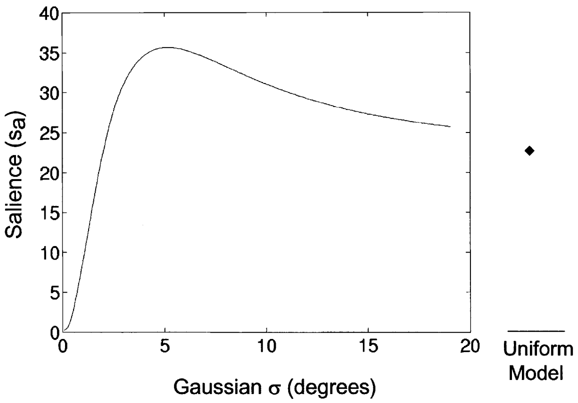

Fig. 10. Chance-adjusted salience ($s_a$) as a function of the standard deviation ($\sigma$) of the Gaussian function used to weight the saliency map. Optimal performance is obtained at an intermediate $\sigma$. The chance-adjusted salience for the uniform model is plotted to the right for comparison (diamond, same scale).

### 6.3. Results

The chance-adjusted salience observed for the first fixation locations for each participant and each image database is illustrated in Fig. 11. Open circles and triangles represent the uniform model and the modified fixation model respectively. The means across participants for each model and image database are plotted as closed circles plus/minus one standard error of that mean. A two-way repeated measures ANOVA was conducted with model type (uniform/modified) and image type as factors. A significant main effect of model type was observed ($\underline{F}(1,3) = 45.66$, $\underline{p} < 0.01$) where the modified fixation model ($\underline{M} = 23.37$) showed a stronger correlation than the uniform model ($\underline{M} = 14.38$). A significant main effect of image type was also observed ($\underline{F}(3,9) = 6.51$, $\underline{p} < 0.05$). Post hoc comparisons between the means using the Newman–Keuls procedure ($\alpha = 0.05$) indicated a stronger correlation for the fractals

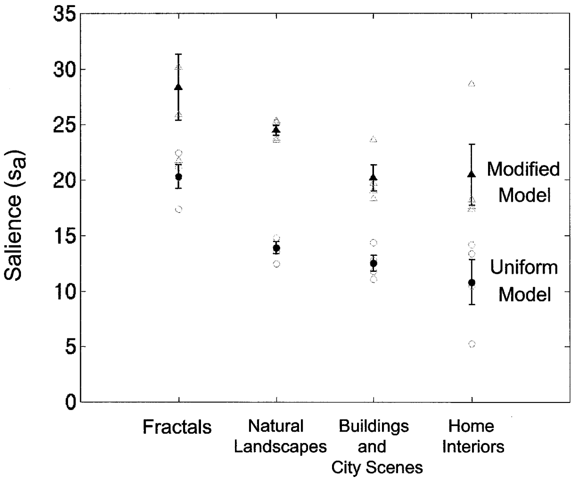

Fig. 11. Chance-adjusted salience ($s_a$) for the uniform model and the fixation-centered model are shown as open circles and triangles, respectively. The mean performance across participants for the models are shown as a closed circle or triangle plus or minus one standard error of the mean. A main effect of model type is seen with performance being greatest for the modified model.

($M = 24.34$) than for the other databases ($M = 19.20$, 16.34, and 15.64). No other effects were significant.

The mean chance-adjusted salience observed for the first fourteen fixations locations following stimulus onset is plotted in Fig. 12 where the means across participants are plotted as closed squares and circles representing the uniform and modified fixation model respectively. The error bars represent plus/minus one standard error of that mean. Shown also is the same quantity for a central model in which Gaussian reduction is always around the center of the screen, rather than on the current fixation location (see Section 6.4). A two-way repeated measures ANOVA was conducted with model type (uniform/modified) and fixation number as factors. A significant main effect of model type was observed ($F(1,3) = 1310.77$, $p < 0.001$) where the modified fixation model ($M = 16.02$) showed a stronger correlation than the uniform model ($M = 10.56$). A significant main effect of fixation number was also observed ($F(13,39) = 6.37$, $p < 0.001$) where early fixations showed a higher chance-adjusted salience than later fixations. With the modified fixation model, comparing the first fixation to the remaining fixations, an approximately 55% stronger correlation is observed. A significant interaction between model type and fixation number was observed ($F(13,39) = 2.85$, $p < 0.01$) caused by the larger difference between earlier and later saccades for the modified fixation model than for the uniform model.

### 6.4. Discussion

The goal of this analysis was to study the effects of the decreasing visual sensitivity in the periphery on the correlation between stimulus salience and fixation locations. As discussed earlier, the earliest selection mechanism in primate vision is based on the differential resolution of the retina (high in the center of the visual field, progressively lower towards the periphery). Though various types of eye movements (mainly saccades but also smooth pursuit), different location in the visual field are sequentially sampled. Because the amount of information available already at the retinal level (and, consequently, at all higher processing levels) varies with eccentricity, this differential resolution can be expected to lead to different weightings in the saliency map (but see Wolfe et al. (1998) discussed below). Emphasis of the central part of the visual field continues in higher stages of primate vision, as evidenced by the fact that the receptive fields of many if not most cells in inferotemporal cortex include the fovea. It stands to reason that this emphasis of stimuli in central vision gives rise to preferential treatment during the visual search task studied. This is exactly what we observed for the first fixation (see Fig. 11) as well as subsequent fixations (see Fig. 12). The correlation between stimulus salience and fixation locations increased substantially (independent of image databases) when the modified version of the saliency map, as described in Section 6.1, was used instead of the original, uniform saliency map.

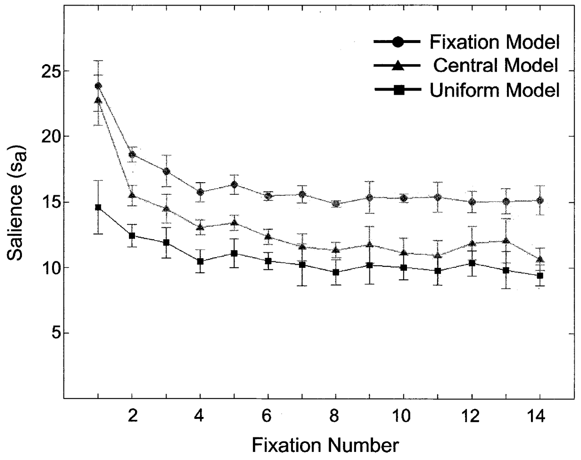

Fig. 12. The mean chance-adjusted salience for all databases is shown averaged across participants where the errorbars represent plus or minus one standard error of the mean. The results from the fixation model are shown as circles and represent the case where the Gaussian reduction is centered on the previous fixation location in order to simulate the falloff of visual system sensitivity in the periphery. The results from the central model are shown as triangles and represent the case where the Gaussian reduction is always centered in the middle of the image. The results of the uniform model are shown as squares. Although the mean salience for both the central model and the fixation model are higher than that of the uniform model, only for the fixation model is this difference significant for all fixations. Performance of the central model is high for the first few fixations but declines later.

An alternative account of the tendency for fixation locations to be clustered around the center of the image is that observers may have a general bias or top-down strategy to select central locations. Many visual media sources (e.g. television) present important information centrally, and consequently, attending to central regions represents an efficient information selection strategy. Or possibly, the presentation, by media producers, of important information in a central location is motivated by knowledge of a general bias to attend centrally. Either way, a central bias account differs from the visual sensitivity account only by the fact that the reduction of peripheral salience is always centered on the screen rather than the current fixation location. Therefore to evaluate this account, a central bias model was implemented in a similar fashion to the modified fixation model with the exception that the Gaussian reduction in salience was restricted to the center of the image rather than following the current fixation location. The data analysis of Section 6.2 was repeated and the mean results across participants are shown in Fig. 12 as triangles where the errorbars represent plus/minus one standard error of the mean. For early fixations there is a clear benefit for the central model over the uniform model, which is to be expected because most early fixations also tend to be centrally located. Even so, the central model does not perform quite as well as the fixation model on the first fixation because there is a some variability in participants' ability to fixate centrally before beginning each trial which is taken into account by the fixation model. For later fixations the benefit of the central model over the uniform model almost disappears. On the other hand, the results for the fixation model show a clear benefit over the uniform model that only slightly decreases for later fixations. This pattern of results suggests that the decrease of visual sensitivity in the periphery rather than a bias to select central locations accounts for the observed pattern of eye movements.

The implementation of the reduction of visual sensitivity in the periphery that we used is quite crude. In the spirit of keeping the model as simple as possible, we applied a non-uniform weighting only at the final processing stage, the saliency map itself, rather than at all of the earlier stages. We found that this simple weighting scheme, a convolution of the saliency map with a Gaussian, dramatically improves the correlation between salience and fixation locations. In addition to simplifying the computations, our implementation has the advantage that no assumptions are made about the source of the non-homogeneity, whether it is the lower resolution of the optical system/retina in the periphery, the influence of the cortical magnification factor (Carrasco et al., 1995; Geisler & Chou, 1995), or rather attentional effects that favor the vicinity of the fixation point over the surround (Wolfe et al., 1998). More realistic implementations will probably improve performance but at the cost of increasing model complexity. It may also be advantageous to take into account other dimensions of retinal inhomogeneity, for example the dependence of color receptor distributions on eccentricity (Curcio et al., 1991). In either case, we suggest that the reduction in peripheral salience is an important factor to consider in the control of visual selective attention and eye movements.

## 7. General discussion

The goal of this study was to elucidate the extent to which stimulus-driven factors influence the allocation of attention by examining the correlation of stimulus salience, as determined by a biologically plausible computational model of bottom-up visual selective attention, and eye movements while observers viewed complex scenes. The results of the primary analysis indicated that attention is stimulus-driven throughout the trial, and furthermore that attention was most stimulus-driven just after stimulus onset when top-down influences are presumably weakest. The second analysis examined the relative strength of individual feature channels and indicated a strong dependence of the relative strengths on image type. The final analysis examined the effect of the reduced visual sensitivity in the periphery on the development of stimulus salience and indicated that the reduction in visual sensitivity correspondingly reduces stimulus salience in the periphery.

A side result of interest is that the correlation between stimulus salience and fixation locations was greatest for the fractal image database. Two possible explanations for this result include the influence of top-down attentional biases as well as differences in bottom-up stimulus characteristics across image databases. First, eye movements may have been influenced by top-down attentional biases stemming from internal models (Noton & Stark, 1971; Yarbus, 1967). For example, we observed in the home interiors database that participants often examined objects on table tops independent of their salience. Searching table tops is a reasonable strategy for finding the position of interesting objects in home interiors whereas such strategies are less likely to be established for fractals. Given that such strategies are at least partially idiosyncratic, one might expect to observe greater interparticipant variability of fixation locations when top-down strategies are influencing attention than when attention is controlled predominantly by bottom-up stimulus properties. This is exactly what is observed; interparticipant variability of fixation locations is much lower for fractals than for the other image types. This reduced variability is consistent with our results that stimulus properties have a stronger influence on attentional guidance in these images.

On the other hand, differences in bottom-up stimulus characteristics across image types could also explain the pattern of model performance. Qualitative examination of the saliency maps indicated that there were often fewer areas of high salience (or, in other words, a greater separation between the salience of the peaks and the average background level) for the fractals than for other image types. This can be seen in Fig. 4 as the lower salience expected by chance for fractals as compared to the other databases. Possibly, the fact that certain features in fractals pop-out due to their contrast to the background acts to increase the stimulus dependence when viewing fractals.

Given that this study examined the attentional allocation of observers viewing static images, we were unable to evaluate the ability of motion or temporal change to guide attention. Clearly, dynamic stimuli are an important aspect of natural viewing conditions and have been shown to guide or even capture attention in experimental settings (Dick, Ullman, & Sagi, 1987; Folk, Remington, & Wright, 1994; Yantis & Jonides, 1984). It may be possible to evaluate the degree of stimulus dependence of eye movements made by observers viewing dynamic stimuli using a similar analysis with the addition of a motion feature channel to the model (for efforts to model motion in the context of this model see Niebur & Koch, 1996).

The overall approach we have taken to understand and quantitatively measure the stimulus dependence of attention has been strongly influenced by both neuroscience and psychology. In designing a computational model of visual selective attention, an effort was made to functionally implement those neural mechanisms which are thought to be important in early visual processing. Experimental and theoretical results from psychology also served to constrain the model's implementation. For the sake of simplicity in these early stages of development and testing of this model, we do not attempt to implement the biological mechanisms with a high degree of detail. Rather, we feel that capturing the functional aspects of these mechanisms is most important and that implementational details can be incorporated at a later time.

It is important to note that although the model can account for many psychophysical results relevant to natural scene viewing and visual search, there are many results for which the model cannot account. One reason for this is that we may have not yet fully explored or implemented all the mechanisms important in determining visual salience. For example, some results suggested that visual salience may be affected by stimulus repetition (McPeek et al., 1999) and familiarity (Suzuki & Cavanagh, 1995; Wang, Cavanagh, & Green, 1994). Although we may be able to account for these factors by altering low-level calculations of salience, it is still unclear whether these effects are bottom-up or top-down let alone their neural implementation. Another example is the fact that we have not fully explored the parameter space of the model. Throughout this study, all feature maps were treated equally (i.e. linearly summed with constant and equal weighting after normalization). A potential modification to our approach would be to tailor the weights of each feature map for each participant. Certain participants may have a bias for or against a particular feature map or spatial scale. Given that the documented variability of eye movements within as well as between participants is quite high (Mannan, Ruddock, & Wooding, 1995; Mannan et al., 1996, 1997), tailoring the weighting of feature maps to individual participants may lead to substantive gains in model performance.

Furthermore, our modeling approach has focused on bottom-up attentional allocation and therefore by its very nature cannot account for top-down effects. There are many potential ways to extend this model. One way might be to adjust the feature map weights with the purpose of simulating top-down influences, assuming that those influences are implemented in the form of biased neural processing. For instance, in the context of a search task, those neurons which represent features of the item to be searched would be preferentially activated, as has been suggested, for instance, for the guided search model of attentional selection (Wolfe, 1994; Wolfe, Cave, & Franzel, 1989). In the framework of the model used in this report, this approach was demonstrated by using a supervised learning scheme to determine weights to optimize performance in a visual search task (Itti, Niebur, Braun, & Koch, 1996).

In conclusion, we used a purely bottom-up model of selective visual attention based on the architecture and neural mechanisms of the primate visual cortex to examine the degree to which eye movements are determined by stimulus properties alone. It was found that stimulus salience correlated with fixation locations much better than expected by chance alone. The best correlation was observed just after stimulus onset, but even later in the trial, eye movements were still influenced by stimulus properties. Overall, our results indicate that attention is indeed guided by stimulus-driven, bottom-up mechanisms under natural viewing conditions even when top-down mechanisms are presumably operating.

## Acknowledgements

This research was supported by the National Science Foundation (CAREER and LIS grants) and by an Alfred P. Sloan Fellowship to EN as well as a NIH-NEI visual neuroscience training fellowship to DP.

## References

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

Andrews, T. J., & Coppola, D. M. (1999). Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision Research*, *39*, 2947–2953.

Antes, J. R. (1976). The time course of picture viewing. *The Journal of Experimental Psychology*, *103*, 62–70.

Bacon, W. F., & Egeth, H. E. (1994). Overriding stimulus-driven attentional capture. *Perception and Psychophysics*, *55*, 485–496.

Boussaoud, D., Desimone, R., & Ungerleider, L. G. (1991). Visual topography of area TEO in the macaque. *Journal of Comparative Neurology*, *36*, 554–575.

Brant, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *The Journal of Cognitive Neuroscience*, *9*(1), 27–38.

Bravo, M. J., & Nakayama, K. (1992). The role of attention in different visual search tasks. *Perception and Psychophysics*, *51*(5), 465–472.

Broadbent, D. E. (1958). *Perception and communication*. London: Pergamon.

Burt, P. J., & Adelson, E. H. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, *31*, 532–540.

Buswell, G. T. (1935). *How people look at pictures: a study of the psychology of perception in art*. Chicago: University of Chicago Press.

Carandini, M., & Heeger, D. (1994). Summation and division by neurons in primate visual cortex. *Science*, *264*, 1333–1336.

Carrasco, M., Evert, D. L., Chang, I., & Katz, S. M. (1995). The eccentricity effect: target eccentricity aspects performance on conjunction searches. *Perception and Psychophysics*, *57*, 1241–1261.

Cave, K., & Wolfe, J. (1990). Modeling the role of parallel processing in visual search. *Cognitive Psychology*, *22*, 225–271.

Curcio, C. A., Allen, K. A., Sloan, K. R., Lerea, C. L., Hurley, J. B., Klock, I. B., & Milam, A. H. (1991). Distribution and morphology of human cone photoreceptors stained with anti-blue opsin. *Journal of Comparative Neurology*, *312*, 610–624.

DeGraef, P., Christiaens, D., & d'Ydewalle, G. (1990). Perceptual effects of scene context on object recognition. *Psychological Research*, *52*, 317–329.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837.

Dick, M., Ullman, S., & Sagi, D. (1987). Parallel and serial processes in motion detection. *Science*, *237*, 400–402.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*, 501–517.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall.

Egeth, H. E., & Yantis, S. (1997). Visual attention: control representation and time course. *Annual Review of Psychology*, *48*, 269–297.

Ellis, S. R. (1986). Statistical dependency in visual scanning. *Human Factors*, *28*, 421–438.

Engel, F. L. (1971). Visual conspicuity, directed attention and retinal locus. *Vision Research*, *11*, 563–576.

Engel, F. L. (1977). Visual conspicuity, visual search and fixation tendencies of the eye. *Vision Research*, *17*, 95–108.

Findlay, J. M., & Walker, R. (1999). A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences*, *22*(4), 661–721.

Folk, C. L., Remington, R., & Wright, J. H. (1994). The structure of attentional control: contingent attentional capture by apparent motion, abrupt onset, and color. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 317–329.

Gattass, R., Sousa, A., & Gross, C. (1988). Visuotopic organization and extent of V3 and V4 of the macaque. *Journal of Neuroscience*, *8*, 1831–1845.

Geisler, W. S., & Chou, K. L. (1995). Separation of low-level and high-level factors in complex tasks: visual search. *Psychological Review*, *102*(2), 356–378.

Henderson, J. H., Weeks, P. A., & Hollingsworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *The Journal of Experimental Psychology: Human Perception and Performance*, *25*(1), 210–228.

Henderson, J. M., & Hollingsworth, A. (1998). Eye movements during scene viewing: an overview. In G. Underwood (Ed.), *Eye guidance while eading and while watching dynamic scenes* (pp. 269–293). Amsterdam: Elsevier.

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Vision Research*, *57*(6), 787–795.

Hubel, D. H., & Livingstone, M. S. (1990). Color and contrast sensitivity in the lateral geniculate body and primary visual cortex of the macaque monkey. *Journal of Neuroscience*, *1*, 2223–2237.

Hubel, D., & Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, *195*, 215–243.

Hubel, D., & Wiesel, T. (1977). Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London Series B*, *198*, 1–59.

Itti, L., & Koch, C. (1999). Target detection using saliency-based attention. *RTO/SCI-12 Workshop on Search and Target Acquisition* (Vol. 20(11)). Utrecht, The Netherlands: NATO.

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

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

Itti, L., Niebur, E., Braun, J., & Koch, C. (1996). A trainable model of saliency-based visual attention. In *Society for Neuroscience Abstracts, vol. 22* (p. 270). Washington, DC: Society for Neuroscience.

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

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

Kowler, E., Anderson, E., Dosher, B., & Blaser, E. (1995). The role of attention in the programming of saccades. *Vision Research*, *35*(13), 1897–1916.

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*, *13*, 201–214.

Leonards, U., & Singer, W. (1998). Two segmentation mechanisms with differential sensitivity for color and luminance contrast. *Vision Research*, *38*(1), 101–109.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *The Journal of Experimental Psychology: Human Perception and Performance*, *4*, 565–572.

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception and Psychophysics*, *2*, 547–552.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1995). Automatic control of saccadic eye movements made in visual inspection of briely presented 2-D images. *Spatial Vision*, *9*(3), 363–386.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those oxations made during visual examination of briefly presented images. *Spatial Vision*, *10*(3), 165–188.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1997). Fixation sequences made during visual examination of briefly presented 2D images. *Spatial Vision*, *11*(2), 157–178.

McPeek, R. M., Maljkovic, V., & Nakayama, K. (1999). Saccades require focal attention and are facilitated by a short-term memory. *Vision Research*, *39*, 1555–1566.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Niebur, E., & Koch, C. (1996). Control of selective visual attention: modeling the "where" pathway. In D. S. Touretzky, M. C. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems* (vol. 8) (pp. 802–808). Cambridge, MA: MIT Press.

Niebur, E., & Koch, C. (1998). Computational architectures for attention. In R. Parasuraman (Ed.), *The attentive brain* (pp. 163–186). Cambridge, MA: MIT Press.

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

Nothdurft, H.-C. (1993). Saliency effects across dimensions in visual search. *Vision Research*, *33*.

Nothdurft, H.-C. (2000). Salience from feature contrast: additivity across dimensions. *Vision Research*, *40*, 1183–1201.

Noton, D., & Stark, L. (1971). Scanpaths in eye movements. *Science*, *171*, 308–311.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*, 3–25.

Posner, M.I., & Cohen, Y. (1984). Components of visual orienting. In H. Bouma, & D. G. Bouwhuis (Eds.), Attention and performance X (pp. 531–556). Hilldale, NJ: Erlbaum.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the center of gaze. *Network: Computation in Neural Systems*, *1*, 341–350.

Rock, I., & Gutman, D. (1981). The effect of inattention on form perception. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(2), 275–285.

Schneider, W. X., & Deubel, H. (1995). Visual attention and saccadic eye move ments: evidence for obligatory and selective spatial coupling. In J. M. Findlay, R. Kentridge, & R. Walker (Eds.), *Eye movement research: mechanisms, processes and applications* (pp. 317–324). New York: Elsevier.

Schwartz, E. (1977). Spatial mapping in the primate sensory projection: analytic structure and relevance to perception. *Biological Cybernetics*, *25*, 181–194.

Shepherd, M., Findlay, J. M., & Hockey, R. J. (1986). The relationship between eye movements and attention. *Quarterly Journal of Experimental Psychology*, *38A*, 475–491.

Stampe, D. M. (1993). Heuristic filtering and reliable calibration methods for video based pupil tracking systems. *Behavior Research Methods, Instruments, and Computers*, *25*(2), 137–142.

Stark, L. W., & Ellis, S. R. (1981). Scan paths revisited: cognitive models direct active looking. In D. F. Fisher, R. A. Monty, & J. W. Senders (Eds.), *Eye movements: cognition and visual perception* (pp. 193–226). Hillside, NJ: Lawrence Erlbaum Associates.

Suzuki, S., & Cavanagh, P. (1995). Facial organization blocks access to low-level feature: An object inferiority effect. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(4), 901–913.

Theeuwes, J. (1993). Visual selective attention: a theoretical analysis. *Acta Psychologica*, *83*, 93–154.

Tipper, S. P., Weaver, B., Jerreat, L. M., & Burak, A. L. (1994). Object- and environment-based inhibition of return of visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 478–499.

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

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

Verghese, P., & Pelli, D. (1992). The information capacity of visual attention. *Vision Research*, *32*(5), 983–995.

Virsu, V., & Rovamo, J. (1979). Visual resolution contrast sensitivity and the cortical magnification factor. *Experimental Brain Research*, *37*(3), 475–494.

Wandell, B. A. (1995). *Foundations of Vision*. Sunderland, MA: Sinauer Associates.

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

Wolfe, J. (1994). Guided Search 2.0—a revised model of visual search. *Psychonomics Bulletin and Review*, *1*(2), 202–238.

Wolfe, J., Cave, K., & Franzel, S. (1989). Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology*, *15*, 419–433.

Wolfe, J. M., O'Niel, P., & Bennett, S. A. (1998). Why are there eccentricity effects in visual search? Visual and attentional hypotheses. *Perception and Psychophysics*, *60*, 140–156.

Yantis, S., & Hillstrom, A. P. (1994). Stimulus-driven attentional capture: evidence equiluminant visual objects. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 95–107.

Yantis, S., & Jonides, J. (1984). Abrupt visual onsets and selective attention: evidence from visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 601–621.

Yarbus, A. (1967). *Eye Movements and Vision*. New York: Plenum Press.