

Models of Overt Attention

Wilson S. Geisler and Lawrence Cormack
University of Texas at Austin

Abstract

Formal models of overt attention have played an important role in motivating and interpreting studies of visual search and other tasks. This chapter briefly summarizes some general principles of attention, some general distinctions between models of overt attention, and some examples of existing models of overt attention in visual search, reading, free viewing, and interactive behaviors. Although many of these models have provided new insight into the factors that contribute to task performance, many are still in a formative stage. In the future, the most useful models will likely be those that take images as input, produce eye movements and decisions as output, apply to well-defined tasks, and explicitly represent the variation in early visual processing across the visual field.

Key words: visual search, active vision, peripheral vision, ideal observer, eye-movement statistics

Introduction

Humans and most other animals perform a wide array of tasks including navigating through the environment and manipulating objects. These, in turn, involve sub-tasks such as determining shapes and distances of surfaces, recognizing objects, etc. Each specific task depends on particular kinds of information from the immediate environment and from memory, on particular kinds of neural computations, and on particular kinds of motor control signals. Attention research is concerned with understanding the brain mechanisms that actively select the specific task-relevant information, neural computations, and motor control signals from among the myriad possibilities.

Attention research is enormously diverse and hence there is no entirely satisfactory definition of what would constitute an attention mechanism. The requirement of selectivity alone is not sufficient because every part of the nervous system is selective (e.g., the photoreceptors are selective to light and the hair cells to sound). What is usually meant by an attention mechanism

is a neural mechanism that can perform (or modulate) selection dynamically, in a task dependent way, often under the control of learned cues or instructions. In other words, attention mechanisms are conceptualized as mechanisms that can flexibly control the flow of information from the environment to the organism and through the organism's various stages of neural processing.

This chapter concerns models of *overt attention*. An overt attention mechanism selects specific kinds of neural processing by physically moving the sensory organs. For example, the visual system has finer spatial neural processing at the center of gaze (fovea) and the auditory system has finer spatial neural processing along the midline between the two ears. Thus, the brain has attention mechanisms that can dynamically select where to direct the high resolution circuits of the visual and auditory systems. *Covert attention* concerns attention mechanisms that do not involve explicit movement of the sensory organs. Thus, an obvious advantage of modeling overt attention is that it is possible to test the models by measuring where the sensory organs are directed and hence where the organism has decided to apply (or not apply) its special processing. Currently, this is most commonly done by tracking the orientation of the eyes while a subject, whose head is held still, performs a visual search task on a computer monitor.

Most models of attention can be subsumed under a general framework such as the one in Figure 1. A given task activates attention mechanisms as well as certain processing mechanisms and memory systems. The attention mechanisms select input signals or select specific processing to be applied to the input signals. Conversely, the attention mechanisms may be affected by the input signals (bottom-up effects) or by the output of the selected processing and memory systems (top-down effects). Overt attention is implemented by modulating motor-control signals (e.g., eye movement or head movement signals). Of course, there are mechanisms, such as vestibulo-ocular reflexes, that modulate the motor control signals but would generally not be considered attention mechanisms.

Many specific models of overt attention have been proposed in the literature, and thus it is not possible here to mention more than a representative subset. This chapter emphasizes quantitative models of performance in visual tasks involving eye movements. However, before discussing

specific models we briefly list a few general principles of selective attention that are important to keep in mind, briefly describe a few general distinctions between overt attention models, and briefly summarize the neuroanatomy and neurophysiology of overt attention.

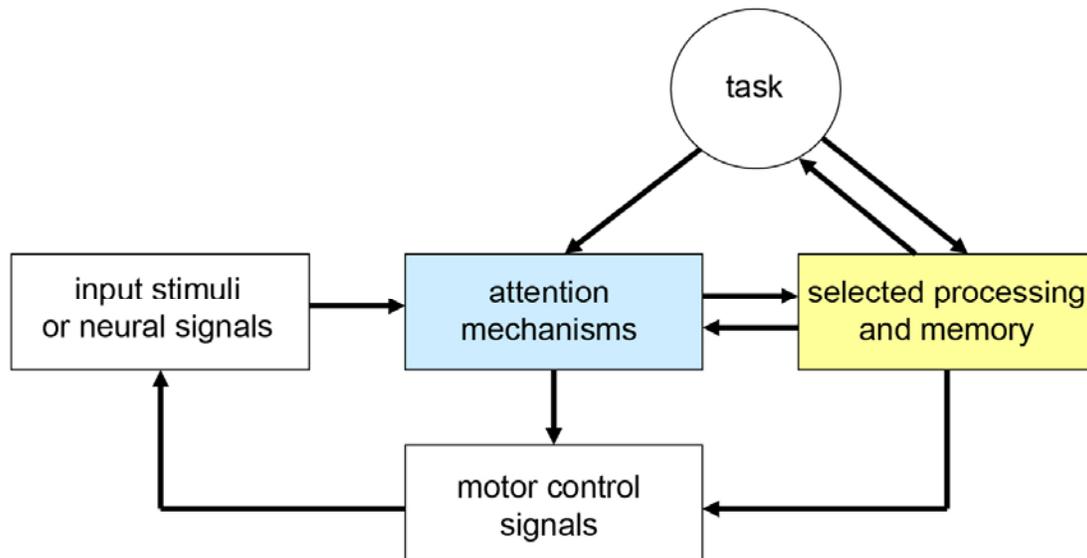


Fig. 1 General framework for models of overt attention.

Some General Principles of Attention

Attention is required for efficient performance in almost all tasks, independent of any capacity limitations in the selected neural processing.

The point here is simply that essentially all tasks involve active selection of neural processing. It is often suggested or implied in the attention literature that attention mechanisms exist because of limited neural processing resources, and thus that the purpose of attention is to allocate these precious resources in the best way possible. While this is sometimes true (e.g., directing the fovea to regions of interest), it is far from given. For example, consider the task of finding a target that may be at one of a few possible locations in a complex background. Efficient performance requires an attention mechanism that selects features from the possible locations and suppresses features from irrelevant locations, because features from irrelevant locations can lead to false detections of the target and potentially other pattern recognition problems. Such an attention mechanism is critical whether or not there is some capacity limitation in how many selected clusters of features can be simultaneously matched against a stored representation of the

target.

The observation of less than perfect performance in a task does not automatically imply limitations in the attention mechanisms or in the capacity of the selected processing.

Often, especially in natural tasks, the requirements of the task and the randomness and complexity of stimuli (or their representations in the early sensory pathways) make perfect performance impossible. Sometimes computational analysis, such as derivation of Bayesian optimal estimators or decision rules (i.e., ideal observers), can help determine the baseline performance that would be attainable with perfect attention and selected processing. This can help identify what aspects of performance cannot be explained by the task requirements and the properties of the input signals.

The flexibility, precision, and speed of an attention mechanism may each be limited.

In addition to the effects of the task requirements and stimuli, performance may be limited by the properties of the relevant attention mechanisms. For example, an attention mechanism may not be able to select or suppress information from arbitrary spatial regions or along arbitrary feature dimensions. For overt attention, the relevant motor systems themselves will place upper bounds on all three of these aspects.

The amount or complexity of selected inputs may exceed the capacity of selected neural processing.

Even if an attention mechanism is able to isolate information from appropriate regions and along appropriate feature dimensions, the selected neural processing may not be able to optimally process that information. Returning to the example above, an attention mechanism may be able to simultaneously select features from all relevant potential target regions and suppress all irrelevant features, but not be able to process in parallel (for target recognition) the features from all the selected locations.

Thus, in general, task performance is limited by task requirements and stimuli, and it may be limited by the flexibility, precision, or speed of the selective attention mechanisms, by the capacity or precision of the selected neural processing, or some combination thereof. Which of

these factors dominates depends on the particular task and stimuli. Quantitative models of attention can play an important role in determining which factors are dominant in a given case and in determining the details of specific attention and selected-processing mechanisms.

Some General Distinctions between Models of Overt Attention

Strongly-specified vs. weakly-specified tasks

Most models of overt attention are proposed for a specific kind of task. Some models are proposed for tasks with well-defined goals such as visual search for a specific target whereas other models are proposed for tasks with poorly-defined goals such as free viewing of natural images. Both kinds of task arise under natural conditions and are worthy of study. An advantage of a strongly-specified task is that it is possible to determine objectively how well the organism is performing and hence potentially evaluate the efficiency of the attention mechanisms. In a weakly-specified task, an organism probably has one or more specific goals, but exactly what they are is more hidden from the experimenter and is likely to be more variable, both within and across experiments.

Explicit vs. implicit representation of the variable-resolution sensory system

In most natural tasks, the primary reason for moving the sensory organs (e.g., the eyes) is to bring the high resolution circuits onto stimuli of interest. For example, if the pinnae (outer ears) of a dog were structured to provide uniform sensitivity in all directions, there would be no need for the dog to be able to move its pinnae and, hence, no overt attention. Obviously, the same argument applies to the foveated nature of the human visual system; nonetheless, many models of overt visual attention do not explicitly represent the variable resolution of the sensory system. Explicit representation is not essential for predicting performance in some circumstances but, obviously, any model of overt attention must correctly specify the input and embody the reason for overt attention in the first place, both of which are accomplished by incorporating an appropriate foveated model of early visual processing (and a moment's reflection should confirm that an appropriate foveated model of early visual processing is also essential for models of covert attention).

Optimal vs. sub-optimal mechanisms

Traditionally, models of attention have been directed at predicting the organism's performance in specific tasks. More recently, there have also been efforts to develop normative (ideal observer) models. The goal of these models is to determine how an optimal attention mechanism should work and how well it would perform in the task. These models play an important role because they rigorously reveal the computational requirements of the task, they provide an appropriate baseline to compare with real performance and with measured neural responses, and they provide principled hypotheses for the underlying attention and selected-processing mechanisms.

Motor movements vs. task performance

Models of overt attention also vary in what aspects of performance they predict. Some are designed to predict only the overt motor movements in a task (e.g., the patterns of eye movements), some only the performance speed (time to task completion), some only the performance accuracy, and some all three aspects (which is the ultimate goal of course). These differences can make it difficult to compare and test models.

Information-processing vs. neurophysiological models

In recent years there has been much progress in understanding the neurophysiology and functional anatomy of attention mechanisms in the mammalian nervous system. The overall picture that emerges is of a heavily interconnected cortical network that relies on V1 for its primary (but not only) input, and the superior colliculus (SC) via the frontal eye fields (FEF) for its primary (but not only) output. Subsequent to V1, the information forks into two streams. One is the ventral stream that is retinotopic but largely concerned with the identification of objects potentially relevant for a given task, and the other is the dorsal stream concerned with movement and object location, though there is also some sensitivity to form. The apexes of these pathways (that is, the anteriormost unarguably "visual" areas of the two forks) are heavily interconnected, presumably to bind objects of interest with potential movement commands via their shared retinotopic representations. These areas also project to (and are heavily innervated by) the FEF, which represents the first stage in the final common pathway required for normal overt attention to occur. In addition to serving as an output stage, however, the FEF is heavily interconnected with both the ventral and dorsal visual processing streams, making it likely that it

is a very dynamic component of larger cortical network that actually embodies the purported “saliency map”. Additional interconnections with other cortical areas are probably responsible for learning, memory, and switching between different tasks requiring overt attention, as well as the more visceral and emotional factors that contribute in certain circumstances. These studies have led to quantitative models of how attention affects neural activity in various brain areas (for a review see Reynolds & Chelazzi 2004), and informed the development of quantitative information-processing models of task performance, which is our focus here.

Tasks

Overt attention mechanisms are engaged in most natural and laboratory tasks, and thus there is no obvious taxonomy of overt-attention tasks, and no limit to the range of tasks that could be modeled in the context of studying the mechanisms of overt attention. Here, we focus on some well-known classes of overt attention task for which there has been some attempt to develop formal quantitative models. These include visual search tasks, visual recognition tasks (reading and shape recognition), free viewing tasks, and visual-motor tasks.

Visual Search

Using the eyes to actively search the environment for specific objects or classes of object is a central subtask in most natural tasks performed by humans and nonhuman primates. Thus, not surprisingly, visual search is the most studied and modeled overt-attention task. In a common form of visual search task, a single target (known to the subject) is randomly located within a display having a well-defined search region that contains some sort of complex background (e.g., a texture of distracter objects, a texture of filtered or unfiltered noise, or a natural image). The subject’s task is to locate the target as rapidly as possible. (Of course, real world search tasks can be much more complicated, e.g., search for “human-made objects” or “anything unusual”.)

A number of overt visual search models are based on the concept of a “conspicuity area”, which is defined to be the spatial region around the center of gaze (assumed to be the center of the fovea) where the target can be detected or identified in the background, within a single “glimpse”

(Engel 1971; 1977; Bloomfield 1972; Geisler & Chou 1995; Toet et al. 1998; 2000).¹ The conspicuity area can vary dramatically depending on the specific target and background, from a fraction of a degree to many degrees of visual angle across. An example conspicuity area measured by Engel (1971) is shown in Fig. 2a.

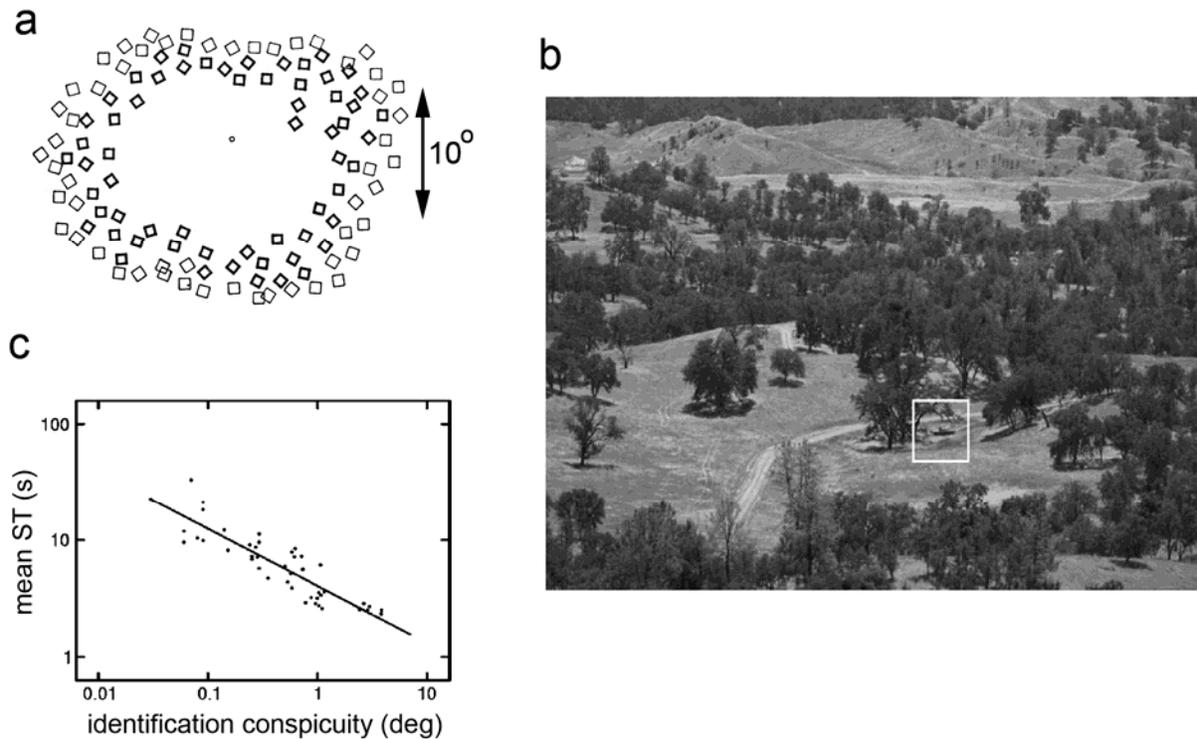


Figure 2. Conspicuity-area models of visual search. **a.** Conspicuity area for a box target in random-line backgrounds. The bold boxes represent locations where the box could be detected in a random-line background, the thin boxes where it could not, with fixation at the small circle. The conspicuity area is the boundary between bold and thin boxes. (Adapted from Engel 1971). **b.** Natural image containing vehicle target (in white box). **c.** Mean search time (for 60 subjects) as function of the radius of the conspicuity area (the average for two subjects). (Adapted from Toet et al. 1998.)

Intuitively, the smaller the conspicuity area, the longer it should take for the subject to find the target, and indeed there is a strong correlation between search time and conspicuity area for simple targets in texture backgrounds (Geisler & Chou, 1995) and for real targets in natural

¹ Although there has not been a consistent definition of what constitutes a glimpse, a sensible one is a period of retinal stimulation, with stable fixation, that is somewhat less than the average fixation duration in the search task (e.g., 250 ms).

images (Toet et al., 1998; Fig. 2b,c). This implies that the variable resolution of the visual system has a profound affect on search performance and that accurate modeling of target visibility as a function of retinal location is essential for developing general models of visual search in laboratory and natural conditions.

The simplest models of visual search based on the concept of conspicuity area postulate a strong top-down attention mechanism that selects fixation locations primarily on the basis of covering the search area rather than being driven by features in the periphery (Bloomfield 1972; Engel 1977; Geisler & Chou 1995). In these models, feature detection, feature interaction (e.g. masking and crowding), and covert attention can affect the conspicuity area, but do not guide the fixation selection *per se*. Rather fixation selection is modeled as a random draw of a 2D location either with or without replacement. If the target happens to fall into the conspicuity area it is detected and the search ends; otherwise the search continues. These random-fixation models can (with certain parameter values) predict approximately the relationship between conspicuity area and search time. However, models in this family that attempt to predict search times directly from images using a model of the conspicuity area are not yet very accurate on natural images (e.g., see Toet et al. 2000). Also, these models are not designed to (and do not adequately) predict eye movement statistics in visual search. There appear to be three major factors contributing to the shortcomings of these models: (1) inadequate specification of the conspicuity area (or more generally of target detectability as a function of retinal position and background context), (2) no role for peripheral features in guiding fixation selection, and (3) no role for prior knowledge in guiding fixation selection (other than the confining of fixations to image regions where the target could be located or to select fixation locations without replacement).

In another general class of search model, the search process is driven largely by feature properties detected during the course of the search (Triesman & Gelade 1980; Wolfe 1994; 2007; Itti & Koch 2000; Rao et al. 2002; Pomplun et al. 2003; Zelinsky 2008). Strictly speaking, some of these models (e.g., Triesman & Gelade 1980; Wolfe 1994; 2007) were developed for covert search. They are often applied, however, to search tasks where reaction times are on the order of 500-2000 ms, which is sufficient time for saccadic eye movements, and thus they can and should be compared (in these tasks) with overt search models. There is a great deal of evidence

demonstrating that eye movements in visual search tasks tend to be directed toward image locations containing features similar to those of the search target (Findlay 1997; Motter & Belky 1998), supporting the general notion of top-down, feature-based guidance in visual search (Wolfe 1994; 2007). For example, Fig. 3a shows the sequence of fixations of a monkey who is searching for a solid black bar that is tilted to the left (Motter & Belky 1998). The fixations tend to occur at locations where the color (black) matches that of the target. Similar results are obtained for search in more complex naturalistic backgrounds of Gaussian noise having the amplitude spectrum of natural images (i.e., $1/f$ noise). The maps at the bottom of Fig. 3b show the average features in the naturalistic noise backgrounds that tend to attract fixations in search tasks where the targets are low contrast versions of the patterns shown above the maps (Rajashekar et al. 2006). These maps (classification images) were obtained by averaging the background noise across all fixations that were not on the target location. If fixations were not being directed at task-relevant features on a substantial proportion of saccades, then these maps would be unstructured. There is also evidence that local features that differ strongly from the features that surround them can attract fixations in free-viewing tasks, suggesting that under some circumstances bottom-up (i.e., non-task-specific), feature-based mechanisms may contribute to fixation selection in visual search (Theeuwes et al. 1998; Ludwig & Gilchrist 2002).

The above findings are clearly not predicted by the random-fixation models that emphasize the role of a conspicuity area. Of course, it is possible that a random-fixation model that accurately models the conspicuity area could still be of practical value in predicting search time and accuracy, even if does not predict the correct patterns of eye movements.

Most models that emphasize the role of feature properties in guiding eye movements during visual search are structured around the intuitive concept of a “saliency map” (Koch & Ullman 1985), which represents the instantaneous attractiveness of each possible fixation location. The saliency map is updated over time as information is collected. If the saliency becomes sufficiently high at some location relative to all other locations, then the eye is directed to that location. Models differ in what kinds of information are hypothesized to contribute to the map and in how the map is updated over time.

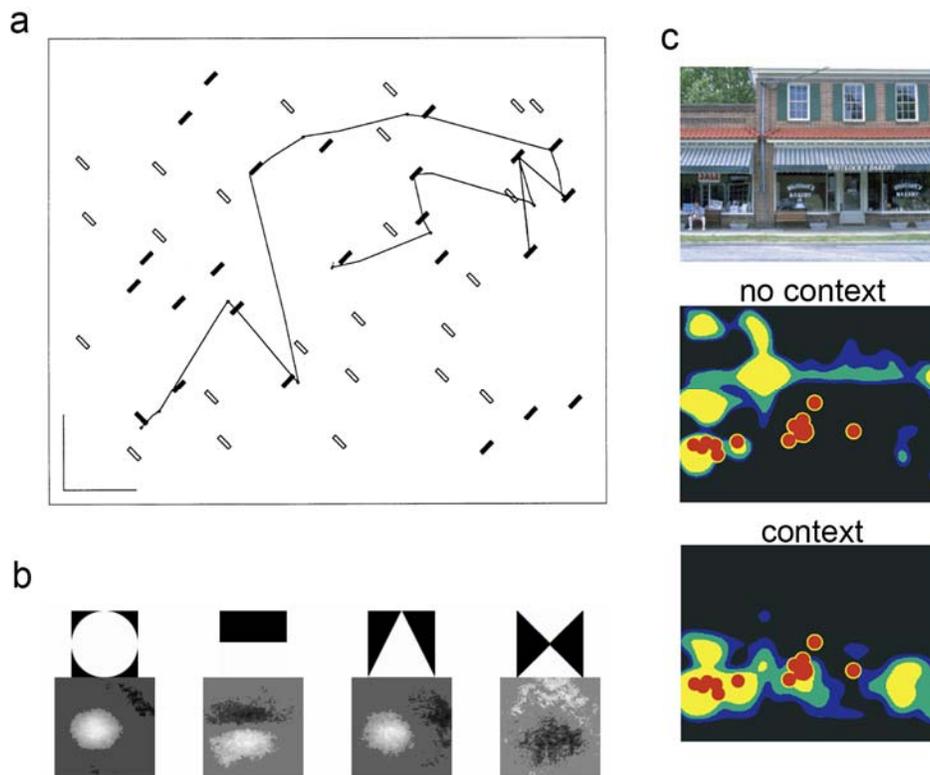


Figure 3. Effect of feature properties and context on fixation selection during visual search. **a.** Sequence of fixations by a monkey searching for a solid black bar that is tilted to the left (adapted from Motter & Belky 1998). **b.** Classification images (below) obtained by averaging the background noise at all the fixation locations of a human searching for targets (above) in a Gaussian noise background having an amplitude spectrum that falls inversely with spatial frequency (adapted from Rajashekar et al. 2006). **c.** Comparison of saliency maps (false-colored regions) that do and do not take scene context into account. Red dots are human fixations recorded when the task is to search for people in the street scene at the top. (Adapted from Torralba et al. 2006).

In one type of model, local feature maps are encoded in parallel from the image along certain stimulus dimensions (e.g., orientation, spatial frequency, color) and then combined in such a way that locations where the features differ more strongly from the surrounding features are given stronger saliency (Itti and Koch 2000). Following fixation at a location, the saliency at that location is suppressed to reduce the chance of fixating that location again. In this kind of model, fixation selection is entirely bottom-up (except for the requirement that fixations stay within the search display). Target recognition is assumed to require direct fixation (at least in the overt-

attention versions of these models), and thus the search ends when and only when a fixation lands on the target (within some predefined error range). Interestingly, such purely bottom-up models can predict some qualitative aspects of visual search performance, illustrating the potential role of simple feature contrasts in determining fixation selection in visual search tasks. However, such models cannot predict results like those in Fig. 3, which show that fixations are often directed at locations with features similar to those of the target. Purely bottom-up models may be more appropriate for free-viewing tasks, where there is no particular target (see later).

In other models, the salience map is determined by a combination of top-down and bottom-up inputs. In the “guided search” models (Wolfe 1994; 2007; Pomplun et al. 2003) feature distinctiveness and feature similarity to the target are combined in determining a salience/activation map. For example, if a subset of “display items” is sufficiently similar to the target and sufficiently distinct from the other display items, then the salience is boosted for the target-similar subset and suppressed for the other items. Fixations are probabilistically selected (“guided”) based on the peaks of the salience map, and on a post-fixation suppression mechanism that reduces the probability of returning to a location after it has been fixated. Target recognition occurs (in the overt versions of these models) only if the fixation lands on the target or nearer to the target than any other item (Pomplun et al. 2003). Importantly, for determining the saliency map, the feature similarity is only computed for a very limited set of simple encoding dimensions such as orientation and “color” (Triesman & Gelade 1980). With appropriate parameters, these models can qualitatively account for a range of search results obtained with simple displays (e.g., search patterns like those in Fig. 3a). However, the current versions of the models are limited because display items and feature dimensions are defined in a relatively simplistic way that limits generalization across experiments and prevents application to natural or naturalistic images. Crucially, they are not “pixels-in / behavior-out”, so their predictive power is very limited. For example, these models are not designed to predict search in naturalistic noise backgrounds and hence cannot generate predictions that could be compared with the classification images in Fig. 3b.

More complete feature-based models of visual search take images as input and generate predicted fixation sequences (Rao et al. 2002; Zelinsky 2008). Like Itti & Koch (2000), they

generate a salience map from feature maps obtained by filtering the input image along various dimensions at various scales; however, the salience at an image location is determined by the correlation between the feature values at that location and those defining the search target. Unlike the guided-search models, the salience map is not restricted to a small set of encoding dimensions. An important property of these models is that fixations are selected on the basis of a weighted average of the salience map (see also, Pomplun et al. 2003). This property is based on evidence that humans often fixate locations that are at the approximate “center-of-gravity” of possible target locations rather than directly at a specific target location (Findlay 1987; He & Kowler 1989; Zelinsky et al. 1997). Recent mathematical analyses show that similar fixation selection strategies are consistent with optimal search performance (Najemnik & Geisler 2005; 2008; 2009). This behavior may seem counterintuitive, but if two targets are similar and separated by some amount, less information might be gained by putting the central fovea on one target and relegating the other to the periphery than if the fovea is placed between them, thus affording at least an intermediate resolution of processing to both targets. The early model (Rao et al. 2002) has only been explored for a limited range of search tasks and it includes an unrealistic assumption about the role of course-to-fine spatial processing on fixation selection. The more recent version (Zelinsky 2008) is more elaborate and realistic, and has been tested against a wider range of search tasks.

The models of visual search described so far have evolved in an effort to predict behavioral data in visual search tasks and to be roughly consistent with certain neurophysiological properties of the primate visual system. A conceptually different approach is to first ask: What kind of overt attention mechanism would a rational (optimal) organism—an ideal observer—use in search tasks, given the properties of natural stimuli and the properties of its sensory and motor systems? The answer to this question (if it can be obtained) can then provide a rigorous basis for formulating and testing more principled models of overt attention. Bayesian statistical decision theory provides a proper theoretical framework for addressing this question, and hence there have been recent attempts to develop Bayesian models of visual search (Torralba 2003; Torralba et al. 2006; Najemnik & Geisler 2005; 2008; 2009; Vincent et al. 2009). These models have their roots in signal detection theory (Green & Swets 1968) and in applications of signal

detection theory to covert visual search (e.g., Palmer et al. 2000; Eckstein 1998; Eckstein et al. 2001; Vincent et al. 2009).

Torralba and colleagues (Torralba 2003; Torralba et al. 2006) developed a Bayesian model in order to understand how a rational observer would combine natural scene statistics and scene context information when searching for objects. The structure of the model is represented by the following formula, which gives the posterior probability that the target object is present at image location x , given the feature values $F(x)$ at that location, and the global features values of the scene G (i.e., the context):

$$p(X = x|F(x), G) = \frac{1}{p(F(x)|G)} p(X = x|G) p(F(x)|X = x, G) \quad (1)$$

where X is the (random) location of a target object. This formula follows directly from Bayes' rule and other rules of conditional probability. (This version of the formula is equivalent to the one in Torralba et al. if we let $X = \emptyset$ represent the event of no target object in the scene.) The key assumption of the Torralba et al. model is that global scene features and local features are encoded rapidly in parallel when a scene is presented, and that together they define a salience map defined by the first two terms on the right side of equation (1), which are estimated from a statistical analysis of natural scenes:

$$s(x) = \frac{1}{p(F(x)|G)} p(X = x|G) \quad (2)$$

The global scene features G determine what kind of scene is being viewed (e.g., a forest scene, a city street scene, an office scene). Thus, the first term in equation (2) asserts that the more unlikely the encoded features at a location, given the type of scene, then the greater the salience. This is similar to the bottom-up definition of salience described earlier. The middle panel of Figure 3c shows a salience map obtained with this component alone for the street scene in the upper panel. The second term asserts that the more likely a target location, given the type of scene, the greater the salience. This term represents the main effect of scene context. The lower

panel in Figure 3c shows the salience map obtained with both terms, for the case where the global context is a street scene and people are the targets. In agreement with intuition, the salient locations are now restricted to the sidewalk/street level where people are most likely to be located. (For simplicity we do not show here how salience is updated across fixations.) The third term in equation (1) is the likelihood of the features at a location given that the target object is at that location in the scene with global features G . This is the object recognition component of a rational observer’s search strategy, and Torralba et al. assume this component requires fixation on the location of interest. Thus, in the model, fixation locations are based on the salience map, and local recognition processing occurs during each fixation.

To implement this model, Torralba et al. start with feature maps similar to those of Itti and Koch (2000). These maps are used both for object recognition and for estimating the kind of scene being viewed. To estimate the kind of scene, the magnitudes of the feature map values are pooled into a small vector of global feature values. A classifier, trained on a large set of natural images, then takes these values as input and returns an estimate of the kind of scene.

The red dots in Figure 3c show human fixation locations when searching for people in the street scene. This example and many others demonstrate the importance of global context information for optimizing visual search performance and they demonstrate that humans make use of the context information.

While conceptually very similar (in that they are captured by equation (1)), the Bayesian models of visual search explored by Najemnik & Geisler (2005; 2008; 2009) differ from that of Torralba et al. in several important ways. The first major difference is that they consider the performance of a full Bayesian ideal observer. In other words, they consider a model where all three terms in equation (1), including the recognition process, are computed in parallel during a fixation. This means, in effect, that the “salience map” is actually a surface giving the posterior probability (following each fixation) that the target is at each scene location:

$$s(x) = p(X = x | \mathbf{F}, G) \quad (3)$$

In equation (3), $F(x)$ has been replaced by $\mathbf{F} = \langle F(1), F(2), \dots \rangle$, where the integers are simply indexes to all possible locations (i.e. pixels), because in some search tasks the posterior probability at a given location depends on the features encoded at multiple locations.

The second major difference is that the variation in visual processing with retinal location is represented explicitly. In terms of equation (3), this means that the features encoded from the scene depend on both the content of the scene and the current fixation location x_0 ; in other words, \mathbf{F} in equation (3) is replaced by $\mathbf{F}(x_0)$, which just means that now the set of feature values at each location are computed after foveation, that is, accounting for the resolution of the image at that location given fixation at x_0 . This adds a great deal of computational complexity to the model, but is essential because the variation in visual processing with retinal location is the very reason eye movements are made (and thus the reason that a chapter on overt attention exists), and because the variation has a huge effect on search performance in many tasks (e.g., see Fig. 2c and Figs. 4a,b). None of the models described above attempt to explicitly represent the effects of retinal location, except for the conspicuity-area models and the models of Zelinsky (2008) and Rajashekar et al. (2008), nor do they have a sufficient representation for determining what would be optimal search strategies and optimal search performance.

The local image features extracted by the visual system limit how detectable (salient) a target is in its background, and thus ultimately, a general model of visual search should include a model for the feature information extracted at each retinal location. Unfortunately, there remains much uncertainty about low-level feature encoding in the visual system as a function of retinal location. Najemnik & Geisler avoided specifying this component of the model by directly measuring the detectability of the target in the search backgrounds (in their case $1/f$ noise) at various retinal locations, in a two-alternative forced choice task, where the target location was cued on each trial. This allowed them to focus on the selective attention mechanisms and eliminate all free parameters. Figure 4a shows the measured detectability as function of retinal eccentricity (averaged across direction from the fovea) for low and high contrast backgrounds, with the target contrast set so the detectability was the same in the center of the fovea ($d' = 3$). The widths of these functions at some criterion height would correspond to the classic definition

of a conspicuity area (also note, perhaps surprisingly, that these would be larger at higher noise contrasts). However, it turns out that the shape of the whole distribution, especially the tails of the distribution, are important for determining optimal search performance and for understanding human search performance.

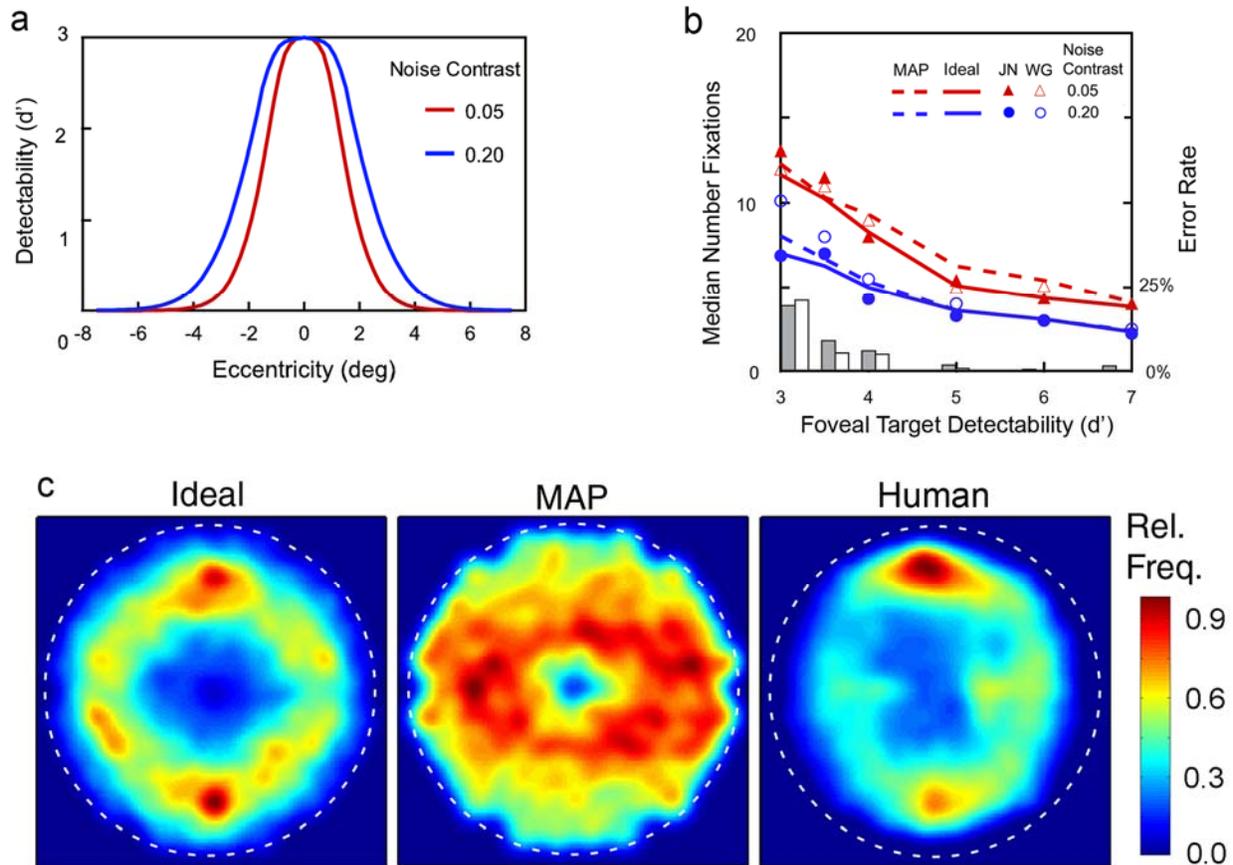


Figure 4. Ideal and human visual search in naturalistic ($1/f$) noise backgrounds. **a.** Target detectability as a function of retinal eccentricity for two contrast levels of the noise background (target contrast was set to produce the same detectability at the point of gaze). **b.** Colored data and curves show the median number of fixations (left axis) to find a target randomly located in noise backgrounds of two different contrasts, as a function of the target detectability in the fovea. Gray bars show error rates (right axis) for model searchers and white bars for the human searchers. **c.** Distribution of fixation locations across all search trials for model and human searchers, with the first fixation at the center of the display excluded. (Adapted from Najemnik & Geisler 2005; 2008.)

The third major difference is that they considered various non-optimal and optimal strategies for fixation selection. As we have seen, a common fixation selection strategy proposed by many

visual search models is to fixate the location with the highest value in the salience map, which is not what the ideal observer would do (although some authors have assumed otherwise). This non-optimal strategy corresponds to fixating the location with the maximum *a posteriori* (MAP) probability after the previous fixation (the location x that maximizes the right side of equation 3). The dashed curves in Figure 4b show the predicted number of fixations to find a target (randomly located in a background of $1/f$ noise), as a function of the detectability of the target in the center of the fovea, for two levels of background noise contrast.

The optimal strategy, however, is to fixate the location where the probability of identifying the target location will be greatest *after* the eye movement is made. In other words, the optimal eye movement is the one that will produce the biggest gain in information about where the target is located, which is not necessarily the single most likely target location. The solid curves show the performance of the ideal fixation selection strategy, and it is only slightly better than the MAP strategy. The symbols show the performance of two practiced observers. Human performance is similar to the parameter free predictions of the ideal and MAP searchers. (The random fixation selection strategies of the conspicuity area models are completely inadequate because they are not capable of reaching human performance levels.) Notice the importance of the detectability functions--the modest differences in the functions shown in Fig. 4a produce big differences in the search performance of the ideal, MAP, and human searchers.

Although the ideal and MAP fixation selection strategies yield similar performance they predict very different fixation patterns. The ideal strategy makes “center of gravity fixations” when appropriate; the MAP strategy never does (by definition). The ideal strategy predicts a distribution of fixations across the search display that has a doughnut shape with higher density at the top and bottom of the display, whereas the MAP strategy predicts a more uniform distribution that is elongated horizontally (Figure 4c). In the models, the asymmetries in these fixation distributions are due to the fact that the detectability functions are elongated along the horizontal axis (Najemnik & Geisler 2008; see also Fig. 2a). These results show that practiced humans are very efficient searchers who have presumably developed heuristics that allow them to closely approximate a Bayesian ideal searcher. These results also underscore the importance

of modeling both performance and the actual patterns of fixations to fully characterize the system under study.

Recognition and Reading

Formal overt attention models have been proposed for other tasks involving object recognition. One important case is the task of reading, for which there are a large number of formal models (for review see Reichle et al. 2003). As in the case of visual search, most of these models have been designed in the process of trying to predict reading performance and eye movement data (Reichle et al. 2003; Engbert et al. 2002; Reilly & Radach 2003), and only one or two have taken the conceptually alternate approach of first trying to derive a ideal observer for the task (Legge et al., 2002). Unlike visual search models, essentially all formal models of reading include an explicit representation of the falloff in resolution of the visual system, perhaps because it is so obvious in typical printed text that letters and words cannot be read in the periphery. An important dimension along which the various models differ is the degree to which eye fixations are driven by the output of linguistic processing versus visual/oculomotor factors, such as word length and font, peripheral acuity, and noise in saccadic eye movement control. Many recent models include both kinds of factors to some extent.

There is not space to review the reading models in any detail, but we briefly describe the ideal reader model of Legge et al. (2002), for comparison with the ideal searcher models described above. We feel that this is important because, while the reading literature on the one hand and the visual search and attention literature on the other are largely separate and distinct, the way in which these particular models are formulated provides an opportunity examine the two domains from a common perspective. The Legge et al. ideal reader (a) knows its own visual acuity along a line of text (a center region with perfect letter identification and a surrounding region where it can discriminate between character and blank space), (b) has a large lexicon of words and knows the marginal probability that each word will appear in the text (based on word frequency counts), and (c) knows the variability of its own saccadic landing points as a function saccade length. With these constraints it then picks fixations that minimize uncertainty (the expected entropy) about each word in sequence, with the goal of maximizing reading speed while correctly identifying each word. Legge et al. (2002) find that the ideal reader quantitatively predicts a

number of statistical properties of human eye movement patterns in reading (e.g., frequency of words skipped), qualitatively predicts a number of other properties (e.g., saccade position with words), but does not predict some properties very well (e.g., percentages of refixations).

Nonetheless, in all cases, the model provides real insight because all the parameters are specified by independently known facts about the lexicon and the visual and oculomotor systems. One way that the ideal reader model differs from the ideal searcher models is that it picks fixations to minimize expected entropy rather than maximize accuracy. However, these measures are closely related, and in fact, Najemnik & Geisler (2009) show that they yield similar performance and eye movement statistics.

Entropy minimization models have also been proposed for shape recognition tasks (Arbel & Ferrie 2001; Renninger et al. 2005; 2007) and scene encoding/recognition tasks (Raj et al. 2005). Renninger et al. (2007) model a task in which observers are given a fixed amount of time to inspect the silhouette of a complex shape before being asked to pick that shape from a test pair of silhouettes placed side by side. Using vernier acuity measurements as a function of eccentricity to estimate the falloff in shape boundary detectability, they determined the fixation pattern that maximally reduces the global shape uncertainty (global entropy). They found that these entropy-minimization fixations better capture human fixation locations than do bottom-up salience models (Itti & Koch 2000), but that a local entropy minimization strategy (like the MAP strategy of visual search) also does well when coupled with a local center-of-gravity rule. Because of the computational complexity of the shape recognition problem, the entropy estimates used in these models are not yet as rigorous as in the reading and visual search models, but they nonetheless demonstrate the useful insights that can be gained with the ideal observer approach.

Free Viewing

There are many situations (e.g., gazing out the window between writing sentences, inspecting a friend's vacation photos or videos) where humans actively direct their eyes to various locations in a visual scene without having an obvious goal or task. A common hypothesis is that under such circumstances attention and hence gaze is attracted by image features that stand out perceptually in some way from the other image features. A number of such models have been proposed for predicting eye movements in free viewing tasks. As mentioned earlier, Itti & Koch (2000)

propose that fixations are directed at “salient” image locations where the local feature values (along certain assumed dimensions) differ markedly from those of the surrounding features. Rajashekar et al. (2008) further incorporated a realistic model of foveation, and found marked improvement over uniform resolution predictions.

Several more recent models are similar in concept, but based on principles from information theory (Itti & Baldi 2006; Bruce & Tsotsos 2006; 2009; Zhang et al. 2008; Gao & Vasconcelos 2009). The central idea is that the current image (or perhaps a larger set of images) is used by the brain to estimate a statistical model of local image features, and that salience is based on how unusual the local image features are given the statistical model. Bruce & Tsotsos (2006; 2009) and Zhang et al. (2008) define salience as “self-information”—essentially the first term in equation (2) above. Gao & Vasconcelos (2009) define salience as the discriminability of a location from its surroundings. Itti & Baldi (2005) are interested in predicting fixations during the free viewing of video, and they define the salience (“surprise”) as the relative entropy (Kullback-Leibler divergence) of the posterior probability distributions over the space of possible image-feature models before and after the current video frame. The critical step in specifying these models of free viewing is specifying either how the statistical models of the image features might be estimated by a rational (optimal) system, or how they might be estimated by the brain.

Although many of recent models of free viewing are computationally interesting and can be applied to arbitrary images (or videos) they are still in a formative state and are difficult to test. One limitation is that the models do not explicitly represent the variation in feature detectability with retinal location; again, this is critically important—features not detected in the periphery cannot attract gaze. A second limitation is that the models make the strong implicit assumption that the task is to fixate (focus neural resources on) statistically unusual locations. Perhaps this is a useful low-level task performed on occasion, but there are other plausible low-level tasks that might drive fixations such as maximally reducing uncertainty about the features in the image (Raj et al. 2005). Furthermore, free-viewing probably involves a wide range of high-level tasks generated internally that vary across individuals and over time within an individual: Are there

any birds out there? What kinds of trees are those? Where was that picture taken? What are the people in this video doing and what are they thinking?

Models of free viewing have been evaluated by comparing the average distribution of human fixations with the models' fixations on the same images or videos. These comparisons go some way toward discriminating between models, but the variability within and across individuals resulting from the unconstrained nature of the task is likely to limit what can be learned about the mechanisms of overt attention. Nonetheless, such models might be of practical value in predicting where people tend to fixate in still images and video.

Interactive Behaviors

Overt attention mechanisms play an important role in every-day tasks where the organism is physically interacting with the environment. There is strong evidence that the eyes are directed toward critical locations in the environment just before the next step in an action sequence is executed (for reviews see Land & Hayhoe 2001, Findlay & Gilchrist 2003, Hayhoe & Ballard 2005). For example, in making a peanut butter and jelly sandwich, humans first make an overt visual search of the scene to find relevant objects (e.g., knife, bread, peanut butter jar, etc.), then they sequentially fixate the objects (or locations within the objects) that are crucial for the each step in the action sequence, often moving the eyes to the next step while the hands finish executing the current step. Developing and testing formal models of such visual-motor tasks is hard because of the difficulty in obtaining the necessary data (recording both eye movements and the other motor movements under controlled conditions), and because of the complexity of the tasks (e.g., visual search tasks are already complex and they are just one sub-task of such interactive behaviors). Not surprisingly there are few formal models of eye movements in such complicated tasks. One potentially promising approach has been explored by Sprague and Ballard (2003). They consider a task where a walker (a model agent) is required to stay on a sidewalk, avoid obstacles, and pick up trash. The key assumptions of the model are that (a) obtaining information about the edge of the sidewalk, the obstacles and the items of trash requires direct fixation, (b) internal noise (uncertainty) about the locations of the observer and relevant scene objects grows over time until there is an appropriate fixation, and (c) the walker uses an optimal fixation selection strategy (based on Kalman filtering) to reduce uncertainty

during the task. An interesting aspect of this model is that it makes predictions for the variation in fixation duration, a dimension of behavior not considered by most other models of overt attention. More mature versions of this type of model may produce eye movement predictions that can be usefully compared against human eye movements in every-day tasks.

Covert and Overt Attention

In most natural tasks, overt and covert attention processes operate in a highly intertwined fashion, and indeed the same cortical areas are consistently implicated in both processes, both at the cellular and systems level (e.g., Corbetta 1998). A key functional difference between covert and overt attention is that the actual input information is changed drastically after each shift in overt attention. The importance of feedback and updating is therefore greatly amplified, and it is thus likely that the attentional circuit(s) of overt attention, to the degree that they are distinct, reflect this fundamental difference. The close connection between covert and overt attention means that realistic models of overt attention must include a model of covert attention. In fact, overt visual attention (i.e., the use of volitional eye movements) probably has its origins in covert attention for the following reasons. As we pointed out earlier, attention mechanisms are necessary for optimal performance regardless of capacity limitations or a non-uniform sensor (such as a foveated retina). Foveation is not universal among vertebrates (e.g. Walls 1963), and thus many species must have covert but not overt attention mechanisms. A contemporary example is the frog, which has a well-developed oculomotor system (following the standard six-muscle vertebrate plan) used for reflexive eye movements, but makes no spontaneous eye movements and thus has no overt attention (Lettvin et al. 1959; Walls 1963). The frog does, however, have experimentally observable covert attention (Ingle 1975). Assuming a similar situation existed at some point in our evolutionary ancestry, it is plausible that a mutation resulting in a non-uniform retina survived and thrived because the existing mechanisms of covert attention and the oculomotor system were already there to be exploited - both (reflexive) eye movements and covert attention are valuable in and of themselves, but a fovea is not very useful unless it can be moved voluntarily.

Conclusion

In recent years there has been much progress in developing formal models of overt attention. Many of these models have provided new insight into the factors that limit or contribute to task performance and into the underlying neural mechanisms. Some models also hold promise of practical application in predicting performance in real-world tasks. Nonetheless, because of the complexity of many overt attention tasks, the models are still largely in the formative stage. A major deficit in many models is lack of an explicit representation (and understanding) of the variation in early visual processing with location in the visual field, and thus there is great need for better general models of early vision. Progress in modeling and understanding overt attention is likely to be most rapid for tasks with well-defined goals, where it is possible to determine optimal performance, or at least determine the general principles of optimal performance.

Acknowledgements

Supported by NIH grant EY02688 to W.S.G. and NSF-IIS-0917175 to L.K.C

References

- Aivar, M. P., Hayhoe, M. M., Chizk, C. L., & Mruczek, R. E. B. (2005). Spatial memory and saccadic targeting in a natural task. *Journal of Vision*, 5, 177-193.
- Arbel T. & Ferrie F.P. (2001). Entropy based gaze planning. *Image and vision computing*, 19, 779-786.
- Beutter, B.R., Eckstein, M.P., Stone, L.S., (2003). Saccadic and perceptual performance in visual search tasks. I. Contrast detection and discrimination. *J Opt Soc Am A Opt Image Sci Vis*, 20, 1341-55
- Bloomfield, J.R., (1972). Visual search in complex fields: size differences between target disc and surrounding discs. *Human Factors* 14, 139–148.
- Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *Advances in neural information processing systems* 18 (pp. 155–162). Cambridge, MA: MIT Press.
- Bruce, N.D.B. & Tsotsos, J.K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3), 1-24.
- Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science*, 9, 111-118.
- Engbert, R., Longtin, A. & Kliegl, R. (2002). A dynamic model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, 42(5), 621–36.
- Engel, F. (1971). Visual conspicuity, directed attention and retinal locus. *Vision Research*, 11, 563–576.
- Engel, F. (1977). Visual conspicuity, visual search and fixation tendencies of the eye. *Vision Research*, 17, 95–108.

- Findlay, J. M. (1997). Saccade target selection in visual search. *Vision Research*, 37, 617–631.
- Findley, J. M. & Gilchrist, I. D. (2003). *Active Vision*. New York: Oxford University Press.
- Gao, D. & Vasconcelos, N. (2009). Decision-theoretic saliency: computational principles, biological plausibility, and implications for neurophysiology and psychophysics. *Neural Computation*, 21(1), 239-271.
- Geisler, W. S. & Chou, Kee-Lee. (1995). Separation of Low-Level and High-Level Factors in Complex Tasks: Visual Search. *Psychological Review*, 102(2), 356-1378.
- Hayhoe, M. and D. Ballard (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4), 188-94.
- Ingle, D. (1975). Focal attention in the frog: behavioral and physiological correlates. *Science*, 188(4192), 1033-1035.
- Itti L. & Baldi P. (2006). Bayesian surprise attracts human attention. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *Advances in neural information processing systems* 18 (pp. 1–8). Cambridge, MA: MIT Press.
- Itti L & Koch C (2000) A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40, 1489-1506.
- Land, M. & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision Research*, 41, 3559-3566.
- Legge, Hooven, Klitz, Mansfield & Tjan (2002). Mr.Chips 2002: new insights from an ideal-observer model of reading. *Vision Research*, 42, 2219-2234.
- Lettvin, J., H. Maturana, et al. (1959). What the frog's eye tells the frog's brain. *Proceedings of the IRE*, 47(11), 1940-1951.
- Ludwig C.J.H & Gilchrist I.D. (2002). Stimulus-drive and goal-driven control over visual selection. *Journal of Experimental Psychology: Human Perception and Performance*, 28, 902-912.
- Motter, B. C., & Belky, E. J. (1998). The guidance of eye movements during active visual search. *Vision Research*, 38, 1805–1815.
- Najemnik, J. & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, 434, 387-391.
- Najemnik, J. & Geisler, W. S. (2008). Eye movement statistics in humans are consistent with an optimal strategy. *Journal of Vision*, 8, 1-14.
- Najemnik, J. & Geisler W.S. (2009) Simple summation rule for optimal fixation selection in visual search. *Vision Research*. 49, 1286-1294.
- Palmer, J., Verghese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research*, 40, 1227–1268.
- Pomplun, M., Reingold, E. M., & Shen, J. (2003). Area activation: A computational model of saccadic selectivity in visual search. *Cognitive Science*, 27, 299-312.
- Raj, R., Geisler, W.S., Frazor, R.A. & Bovik, A.C. (2005). Contrast statistics for foveated visual systems: Fixation selection by minimizing contrast entropy. *Journal of the Optical Society of America A.*, 22 (10), 2039-2049.
- Rajashekar, U., Bovik, A. C. & Cormack, L. K. (2006). Visual Search in Noise: Revealing the Influence of Structural Cues by Gaze-contingent Classification Image Analysis. *Journal of Vision*, 6, 379-386.
- Rajashekar, U., I. Van der Linde, et al. (2008). GAFFE: A gaze-attentive fixation finding engine. *IEEE Transactions on Image Processing*, 17(4), 564-573.

- Rao, R. P. N., Zelinsky, G. J., Hayhoe, M. M. & Ballard, D. H. (2002). Eye movements in iconic visual search. *Vision Research*, 42, 1447-1463.
- Reichle, E. D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26, 445–476.
- Reilly, R. & Radach, R. (2003). Foundations of an interactive activation model of eye movement control in reading. In: *The mind's eye: Cognitive and applied aspects of eye movements*, ed. J. Hyona, R. Radach & H. Deubel. Elsevier.
- Renninger, L. W., Coughlan, J., Verghese, P., & Malik, J. (2005). An information maximization model of eye movements. *Advances in Neural Information Processing Systems*, 17, 1121–1128.
- Renninger, L. W., Verghese, P., & Coughlan, J. (2007). Where to look next? Eye movements reduce local uncertainty. *Journal of Vision*, 7(3), 1-17.
- Reynolds JH & Chelazzi L (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, 27, 611-647.
- Sprague, N. and Ballard, D. (2003). Eye movements for reward maximization. In *Advances in Neural Information Processing Systems* (Vol. 16), Boston: MIT Press
- Theeuwes J., Kramer A.F., Hahn S. & Irwin, D.E. (1998). Our eyes do not always go where we want them to go: capture of the eyes by new objects. *Psychological Science*, 9, 379-385.
- Toet A, Kooi FL, Bijl P & Valeton JM (1998). Visual conspicuity determines human target acquisition performance. *Optical Engineering*, 37, 1969-1975.
- Toet A, Bijl P & Valeton JM (2000). Tests for three visual search and detection models. *Optical Engineering*, 39, 1344-1353.
- Torralba, A., Oliva, A., Castelhana, M. & Henderson, J. M. (2006). Contextual Guidance of Attention in Natural scenes: The role of Global features on object search. *Psychological Review*, 113(4), 766-786.
- Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12, 97–136.
- Vincent, B. T., Baddeley T. J., Troscianko T. & Gilchrist I. D. (2009). Optimal feature integration in visual search. *Journal of Vision*, 9, 1-11.
- Walls, G. L. (1963). *The vertebrate eye and its adaptive radiation*. New York, Hafner.
- Wolfe, J. M. (1994). Guided Search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, 1, 202–238.
- Wolfe, J. M. (2007). Guided Search 4.0: Current Progress with a model of visual search. In W. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 99-119). New York: Oxford.
- Zhang L, Tong MH, Marks TK, Shan H, & Cottrell GW (2008) SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8, 1-20.
- Zelinsky GJ (2008) A theory of eye movements during target acquisition. *Psychological Review*, 115, 787–835.