VR 3344
9 March 2002 Disk used
ARTICLE IN PRESS
No. of Pages 17, DTD = 4.3.1
SPS, Chennai

PERGAMON

Vision Research xxx (2002) xxx–xxx

Vision Research

www.elsevier.com/locate/visres

Eye movements in iconic visual search

Rajesh P.N. Rao [a], Gregory J. Zelinsky [b], Mary M. Hayhoe [c], Dana H. Ballard [d,*]

[a] Department of Computer Science, University of Rochester, Rochester, NY 14627, USA
[b] Beckman Institute, University of Illinois, Urbana, IL 61801, USA
[c] Center for Visual Science, University of Rochester, Rochester, NY 14627, USA
[d] Department of Computer Science, University of Rochester, Rochester, NY 14627, USA

Received 15 January 2001; received in revised form 17 July 2001

Abstract

Visual cognition depends critically on the moment-to-moment orientation of gaze. To change the gaze to a new location in space, that location must be computed and used by the oculomotor system. One of the most common sources of information for this computation is the visual appearance of an object. A crucial question is: How is the appearance information contained in the photometric array is converted into a target position? This paper proposes a such a model that accomplishes this calculation. The model uses iconic scene representations derived from oriented spatiochromatic filters at multiple scales. Visual search for a target object proceeds in a coarse-to-fine fashion with the target's largest scale filter responses being compared first. Task-relevant target locations are represented as saliency maps which are used to program eye movements. A central feature of the model is that it separates the targeting process, which changes gaze, from the decision process, which extracts information at or near the new gaze point to guide behavior. The model provides a detailed explanation for center-of-gravity saccades that have been observed in many previous experiments. In addition, the model's targeting performance has been compared with the eye movements of human subjects under identical conditions in natural visual search tasks. The results show good agreement both quantitatively (the search paths are strikingly similar) and qualitatively (the fixations of false targets are comparable).

Keywords: Saccades; Computation; Attention; Visuomotor control

1. Introduction

Human vision relies extensively on the ability to make saccadic eye movements to orient the high-acuity foveal region of the eye over targets of interest in a visual scene. However, resolution per se is not the only determinant of gaze location. Starting from Yarbus' classical work (Yarbus, 1967), many studies have suggested that gaze changes are directed according to the ongoing cognitive demands of the task at hand. The task-specific use of gaze is best understood for reading text (O'Regan, 1990) where the eyes fixate almost every word, sometimes skipping over small function words. In addition, saccade size during reading is modulated according to the specific nature of the pattern recognition task at hand (Kowler & Anton, 1987). Tasks requiring comparison of complex patterns also elicit characteristic saccades back and forth between the patterns (Just & Carpenter, 1976). In copying of a model block pattern on a board, subjects have been shown to employ fixations for accessing crucial information during different stages of the task (Ballard, Hayhoe, & Pelz, 1995; Ballard, Hayhoe, Pook, & Rao, 1997). In natural language processing, fixations can reflect the instantaneous parsing of a spoken sentence in the current visual context (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). The role of gaze has been studied by in a variety of natural visuomotor tasks such as driving, music reading and playing ping-pong (Land & Furneaux, 1997). In each case, gaze was found to play a central *functional* role, closely linked to the immediate task demands. All these tasks have very different kinds of fixation targets, sometimes only defined in terms of functional needs. For example, in driving around a bend, subjects fixate the tangent point of the curve to control steering angle, and in ping-pong, subjects fixate the bounce point in advance, in order to estimate the ball's trajectory.

* Corresponding author. Tel.: +1-716-275-3772; fax: +1-716-461-2018.
E-mail address: dana@cs.rochester.edu (D.H. Ballard).

0042-6989/02/$ - see front matter
PII: S0042-6989(02)00040-8

The general utility of saccadic eye movements has spurred an extensive effort to characterize their properties. A variety of studies have revealed the importance of task, acuity, and visual features in determining the stimulus for target selection together with accompanying metrics of accuracy and fixation duration (e.g. Findlay, 1997; Hooge & Erkelens, 1998; Motter & Belky, 1998; Viviani, 1990; Zelinsky & Sheinberg, 1997). However, much less is known about the underlying computational operations that determine these properties, although some ground-breaking work has been done. Itti and Koch (2000) use the coincidental alignment of visual features to define a saliency map of possible targets. Moving the gaze to these points successively has some resemblance to human visual search but there is no model of how specific targets are selected. Tsotsos et al. (1995) use an hierarchical attractor network to define interesting targets. Unlike Itti and Koch, Tsotsos's network can be driven by selected target features, however the representation cannot define completely general image targets. There also has been no attempt in either of these models to compare their performance with human visual search.

This paper describes a general model for fixating and remembering appearance-based encodings of targets in natural scenes. The model uses iconic (appearance-based) target representations to search arbitrary visual scenes. Iconic representations are specified by the responses of oriented spatiochromatic filters at multiple scales. This has been demonstrated to be a very robust computational mechanism for target selection in natural scenes (Rao & Ballard, 1997). The computation of target coordinates for a saccade reduces to correlation between a "top-down" iconic target representation and the "bottom-up" iconic scene representations. The model provides a good fit to visual search data where the target is defined predominantly from its appearance. An key feature of the model is that it separates the targeting process, which changes gaze, from the decision process, which uses the information at the new gaze point. The virtue of this separation is that decision-making about the target can be separated from the process of fixating it. Thus there is no additional control structure to make the gaze change contingent on the decision process. If the decision process is slow with respect to the time needed for target selection, then gaze can be moved to the target more accurately. If the decision process is fast, then gaze does not have to be changed at all, as is observed in a huge number of studies of attention.

2. General purpose iconic representations

In many experiments that study saccades, the targets themselves are simple colored shapes that are presented on a blank background. While extensive useful data has been collected using this paradigm, this setup does not address issues of target selection in natural viewing. In natural scenes, the saccadic target may be composed of complex photometric intensity patterns, produced by cluttered scenes. In order to move the eyes in this case, there must be a mechanism that translates the intensity image on the retina into a representation that can be used by the oculomotor system. Such a mechanism must meet at least the following three criteria:

1. *Generality:* Any proposed mechanism for targeting parts of an image must have broad generality since saccadic targets can vary greatly according to the requirements of the current task.
2. *Speed:* Targets must be computed quickly in order to model observed human performance. Using millisecond neural circuitry, the targets for the next fixation need to be computed in approximately 80–100 ms, allowing barely one pass through the cortex (Oram & Perrett, 1992; Thorpe & Imbert, 1989).
3. *Resolution:* The computation of the target must use spatial scales that are available extrafoveally, since it is unlikely that the target is already at the gaze point.

One representation that meets these criteria employs low resolution iconic representations of targets and scenes that can be extracted directly from the optic array. This allows general portions of a scene to be represented in a precategorical format without requiring any elaborate segmentation. This is an essential property, since the information required for such complex operations is frequently the goal of the eye movement itself. The computation of saccadic target coordinates is accomplished by correlating the iconic memory of the target with the iconic representation of the current optic array. A correlation peak indicates the most likely location of the target in the current image, allowing a saccade to be executed to that location. We regard the notion of "icon" as completely general. The idea is that any criterion for a gaze point can be transformed into an appearance model which captures how that criterion should appear in the scene. Then the resultant appearance image, or icon, is used as a correlation template.

It would be prohibitively expensive to encode icons literally as gray-level images, since the memory needed would then scale with the size and number of icons. A more efficient alternative is to encode the icons as their responses to a set of spatiochromatic basis functions, or spatial filters (Itti & Koch, 2000; Poetzsch, Krueger, & Von der Malsburg, 1996; Weber & Malik, 1995). One motivation for this is that it approximates the transformations imposed by the receptive fields of striate cortical cells. Another motivation is the psychophysical evidence of suggesting that the human visual system uses such channels (Graham, 1989; Wilson & Wikinson, 1997). The particular filters we use are the steerable filters, so-called because the responses of these filters at any given orientation can be used to produce the responses at any other location by interpolation formulae. A local image patch can be characterized using a zeroth order Gaussian G_0 and nine of its oriented derivatives (Fig. 1) as follows (Freeman & Adelson, 1991):

$$G_n^{\theta_n}, \quad n = 1, 2, 3, \quad \theta_n = 0, \ldots, m\pi/(n+1), \quad m = 1, \ldots, n \tag{1}$$

where n denotes the order of the filter and θ_n refers to the preferred orientation of the filter. The response of an image patch I centered at (x_0, y_0) to a particular basis filter $G_i^{\theta_j}$ can be obtained by convolving the image patch with the filter:

$$r_{i,j}(x_0, y_0) = \int\int G_i^{\theta_j}(x_0 - x, y_0 - y) I(x, y)\,\mathrm{d}x\,\mathrm{d}y \tag{2}$$

The iconic representation for the local image patch centered at (x_0, y_0) is formed by combining into a high-dimensional vector the responses from the 10 basis filters above at different scales

$$\mathbf{r}(x_0, y_0) = [r_{i,j,s}(x_0, y_0)] \tag{3}$$

where $i = 0, 1, 2, 3$ denotes the order of the filter, $j = 1, \ldots, i+1$ denotes the different filters per order, and $s = s_{\min}, \ldots, s_{\max}$ denotes the different scales of the filters. For computational efficiency, a Gaussian pyramid representation of the image can also be used to generate multi-scale responses from a set of basis filter kernels at a fixed scale. This strategy was used in the visual search simulations. As an example, Fig. 2 shows the filter-based responses at a given location in a cluttered scene for filters G_1 and G_2 and five spatial scales. The filter response vector at every image location in general provides an almost unique representation of the local image region surrounding that location (Rao & Ballard, 1996).

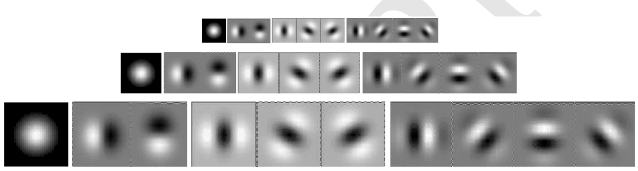

Fig. 1. Spatiochromatic basis functions. Motivation for these basis functions comes from statistical characterizations of natural image stimuli (Bell & Sejnowski, 1997; Derrico & Buchsbaum, 1991; Hancock, Baddeley, & Smith, 1992; Olshausen & Field, 1996; Rao & Ballard, 1997). The nine oriented spatial filters at three octave-separated scales for each of the three channels in (a) (bright regions denote positive magnitude while darker regions denote negative magnitude). At each scale, these nine filters are comprised of two first-order derivatives (G_1) of a 2D photometric Gaussian, three second-order derivatives (G_2), and four third-order derivatives (G_3). Thus, there are three scales per channel, and nine spatial filters per scale, for a total of 27 filter responses characterizing each location in the image. These 27 spatiochromatic measurements at a given image location can be regarded as a photometric signature of the local image region centered at that location.

The model search simulations used gray scale stimuli, with three spatial scales and nine filters per scale for a total of 27 measurements per image location. The scales used in our tests range from approximately 1–6 cycles per degree, well within the limits of human spatial resolution at the eccentricities involved in the experiments described here. The basis functions described above were picked a priori, but very similar functions can be learned from samples of natural images (Ballard et al., 1997; Barrow, 1987; Bell & Sejnowski, 1997; Hancock et al., 1992; Olshausen & Field, 1996).

The use of multiple scales is crucial to the visual search model. In particular, the larger the number of scales, the greater the perspicuity of the representation as depicted in Fig. 3, which shows the frequency distribution of correlations between all points in the dining table image (Fig. 8(d)) and a fixed target point in the same image. The distribution on the left shows how using filter responses at a single scale causes ambiguity in the iconic scene representations, with as many as 936 points in the scene having correlations greater than 0.94 with respect to a fixed target. However, when five scales are used, the ambiguity is resolved, and only a single point (the target point) correlates greater than 0.94 (indicated by the arrow for both histograms). The greater perspicuity results partly due to the inclusion of information from additional scales and partly due to the high-dimensionality of the multi-scale vectors. The high-dimensionality of the vectors makes them remarkably robust to noise due to the *orthogonality* inherent in high-dimensional spaces: given any vector, almost all of the other vectors in the space tend to be relatively uncorrelated with the given vector (Kanerva, 1988; Rao & Ballard, 1995a), and almost none are identical with respect to each other. The result is that the filter response vector for a given point is unique for all practical purposes and can therefore be used to define search targets. This property also makes the filter template robust to partial occlusions, which commonly occur in natural viewing (see Rao & Ballard (1995a) for some examples).

The representation works best when the gross viewpoint of the scene does not change drastically from moment-to-moment. The filter responses are dominated by a cosine envelope, so that there is a useful range of rotations for which the responses will be effectively invariant. Drastic rotations are handled by storing feature vectors from different views (Bulthoff & Edelman, 1992). This is consistent with psychophysical evidence that shows that subjects represent objects using a small number of separate viewpoints. The multi-scale representation also allows interpolation strategies for scale invariance (Rao & Ballard, 1995a).

VR 3344
9 March 2002 Disk used
ARTICLE IN PRESS
No. of Pages 17, DTD = 4.3.1
SPS, Chennai

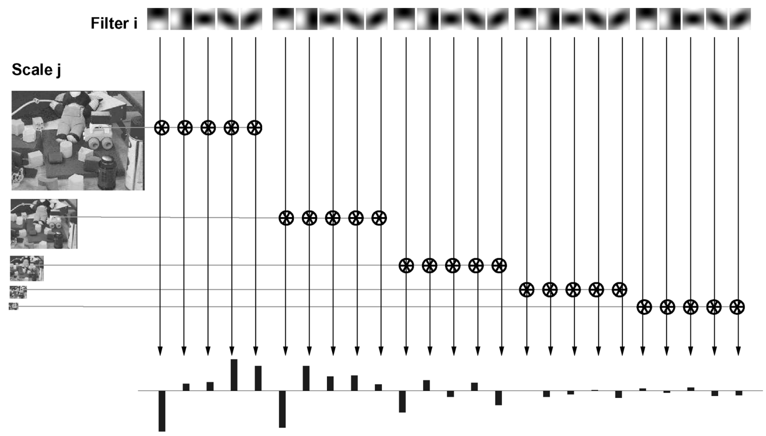

Fig. 2. Using spatiochromatic filters to extract task-dependent properties. A portion of a cluttered image. The scales at which the filters of Fig. 1 were applied to the image are shown on the left. Each individual filter, when convolved with the local image intensities near the given image location, results in one measurement. This example uses the first two filters and five spatial scales for a total of 25 measurements per point. Positive responses in the vector are represented as an upward bar above the horizontal, negative responses as a downward bar below the horizontal. For reasons of economy, large scale filters are modeled by using the standard size filter and shrinking the image.

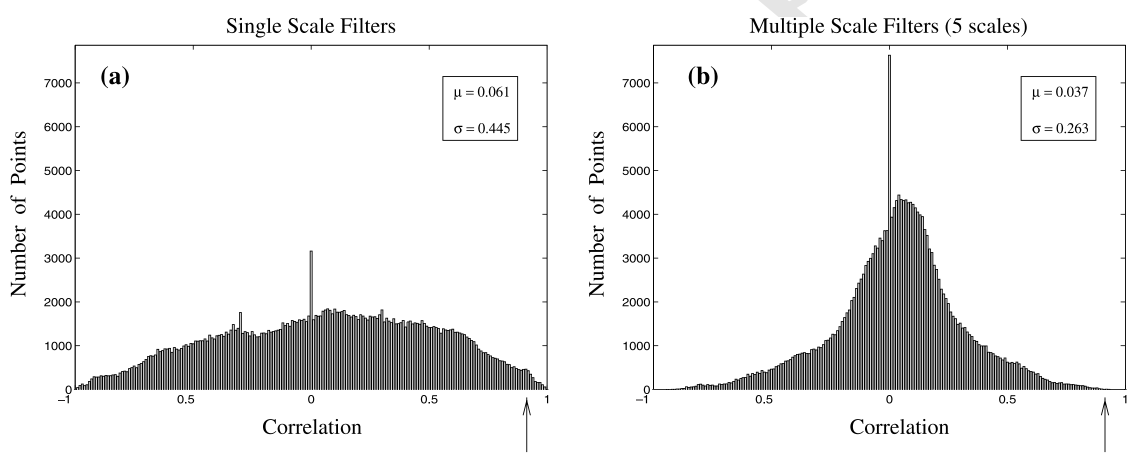

Fig. 3. The effect of scale. The distribution of distances (in terms of correlations) between the filter response vector for a selected target point in the dining table scene (Fig. 7(a)) and all other points in the scene is shown for single scale response vectors (a) and multiple scale vectors (b). Using responses from multiple scales (five in this case) results in greater perspicuity and a sharper peak near 0.0. The most important feature of these plots appears at the extreme right hand side. Only one point (the target point) has a correlation greater than 0.94 (demarcated by an arrow) in the multiple scale case (b) whereas 936 candidate points fall in this category in the single scale case (a).

To summarize, the representation meets the criterion of generality since any gaze target can be translated internally into a local appearance, which in turn can be expressed in terms of filter responses. The representation can be used quickly since targeting reduces to filter correlations, which we assume can be done in parallel without penalty over the retinal array. Finally the use of multiple scales means that the range of resolutions used can be adjusted to trade-off speed with accuracy as suggested by Geisler and Chou (1995).

3. Modeling visual search

Early models of visual search suggested that the search process proceeds item-by-item (Treisman, 1988) but data showing fast search times for some multiple conjunctions were hard to model. More recent models, guided by Palmer, Vergese, and Pavel (2000) assume that search is area-based, aimed at detecting targets within a window centered around the center of gaze (Eckstein, 1998; Geisler & Chou, 1995). The size of the window is a function of the speed and accuracy required of the task, and reflects the signal-to-noise characteristics of the display (Motter & Belky, 1998). In the latter

VR 3344
9 March 2002 Disk used
ARTICLE IN PRESS
No. of Pages 17, DTD = 4.3.1
SPS, Chennai

case, the search task can be seen as one of covering the scene while prioritizing likely locations. As a consequence the gaze point need not search item-by-item but can delimit large areas.

Fig. 4 motivates the model's use of area-based search in terms of the resolution of the retinal image as reported by Hess. For each search task, a resolution needs to be chosen based on signal-to-noise conditions of the display and the spatial properties of the target. The resolution chosen for the search process defines a search window width. Higher signal-to-noise means that the object can be recognized at a lower resolution and hence a bigger search window can be used. A consequence of this choice is that the same resolution is used throughout the search window, even though higher resolution is available. The use of a set resolution in this manner by our model is counterintuitive, as it is more natural to assume that all the available resolution is continuously available. However, the use of resolution as a search parameter is motivated by search experiments that show that other search parameters are set and changed with temporal cost. For example, Sperling (Sperling & Dosher, 1986) showed that searching displays of two different font sizes incurred a cost that suggested the scale had to be set for each size.

The visual search model is composed of three separate procedures that each operate largely independent of each other, while at the same time cooperating to solve the current visual search task:

1. A *targeting process* (or "where" process) that computes the next location to be fixated.
2. A *decision process* (or "what" process) that matches a stored iconic object representation to the current foveated image region.
3. An *oculomotor process* that accepts retinotopic target locations from the "where" process and executes a saccade to the target location (a method for learning this sensorimotor mapping is given in (Rao & Ballard, 1995b)).

The model assumes that these processes are running concurrently, but that they do not have to be precisely coordinated in time. The oculomotor process will continue to execute eye movements as long as the decision process has not terminated. The current best guess of target location is updated as fixations increase the available resolution. Although we do not model the decision process, a key point is that the decision process needs to choose a resolution and window in the same way as the search process, but the two resolutions need not be the same, since getting the gaze to the target and analyzing a property of the target are different computations.

All three processes use a *saliency map* (Koch & Ullman, 1985) whose value at a given location represents the weight determined by multi-scale filter-based correlation. This weight map has a dual purpose: (1) it allows the oculomotor process to fixate target locations with high correlations, and (2) its maximum value is used by

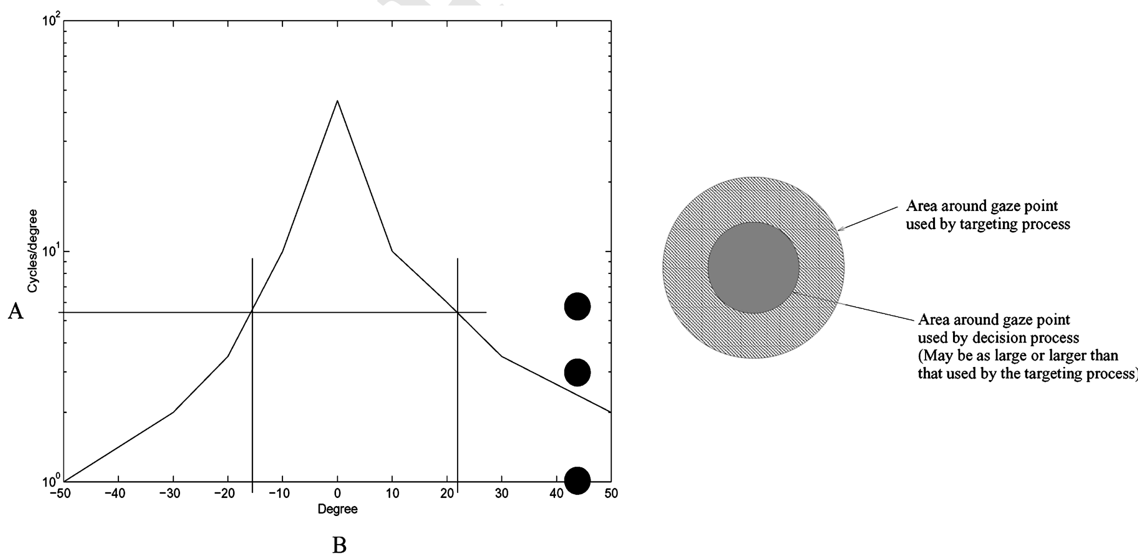

Fig. 4. How the model chooses resolutions. Left: Resolution as a function of retinal eccentricity, with a hypothetical search window. Data are replotted from (Anderson, Mullen, & Hess, 1991). For a given search task our model assumes that the subject chooses a signal-to-noise ratio. That defines a maximum resolution to be used in the search (A). Given this resolution value, the resolution available on the retina defines a search width (B). The three frequency scales used by the model are shown at right as filled circles. Right: Separate search windows are used for targeting, which changes gaze, and decisions, which extract information needed for behavior.

the decision process to judge the presence or absence of the target. The decision process need only use a signal-to-noise criterion to decide whether the correlation peak in the saliency map is high enough so that the target can be assumed to be present. It does not need information on where that measurement came from.

The computation of such a saliency map usefully can be described in an oversimplified form as follows. Objects of interest to the current search task are assumed to be represented by a set of memorized filter response vectors $\mathbf{r}_s^m$ where s denotes the scale of the filters and m denotes a particular target object in memory. Given a new input image, the targeting process computes the most likely location of the target as follows:

1. Compute the saliency map S across all locations (x, y) as

$$S(x,y) = \sum_{s=1}^{\max} ||\mathbf{r}_s(x,y) - \mathbf{r}_s^m||^2 \tag{4}$$

   where $||\mathbf{x}||$ denotes the Euclidean norm of the vector $\mathbf{x}$. In other words, the saliency value at location (x, y) is simply the sum of squared differences between the corresponding components of the filter response vector $\mathbf{r}_s$ at that image location and the memorized target object vector $\mathbf{r}_s^m$, across all filter scales $s = 1, \ldots, \max$.
2. The location for saccadic targeting is the one that is most similar to the target, where similarity is given by Euclidean distance

$$(\hat{x}, \hat{y}) = \arg\min S(x,y) \tag{5}$$

In this targeting process, a single saliency map is calculated across all filter scales for a given image, and the location $(\hat{x}, \hat{y})$ to be foveated is chosen to be the one with the highest correlation value with respect to the memorized target i.e. the one with the least $S(x, y)$. These computations have been implemented using the Datacube MV200 image processor and the Rochester dual-camera robot head to perform targeting movements in real time in natural scenes. The virtue of this system is that the Datacube MV200 can compute convolutions at frame rates (30 s^{-1}) and this allows for extensive experimentation. Details of the hardware implementation are given in (Rao & Ballard, 1995a). Figs. 5 and 6 illustrates the utility of this simple algorithm in a search task. Gaze, as denoted by the cross-hairs, is first directed to a given scene location as shown in (a). At that point the filter responses are memorized. Next, at some point in the course of the rest of the behavior, it may be desirous to return to the original location from a distal point. The targeting algorithm is used to correlate the memorized features with the current retinotopic image, resulting in a saliency map as shown in (c). Note that the coordinate system of the saliency map can also be interpreted in terms of a motor error signal. Thus, the saliency peak can be used to drive the oculomotor command for returning the eyes to the original target without involving complex object properties.

4. Human fixation patterns in appearance-based visual search

Human fixation patterns are more complicated than those predicted by the simple search model. In order to compare the model's performance with human search and targeting behavior we used the data from eye movements in a visual search task described in (Zelin-

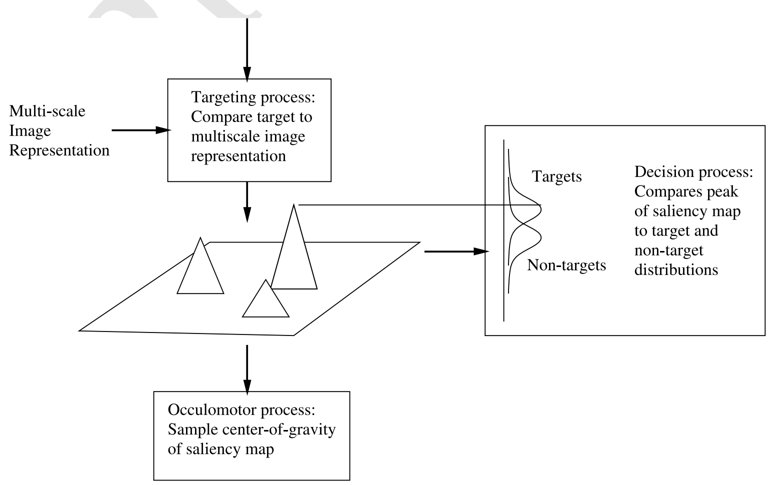

Fig. 5. Visual search using spatial filter responses. The simplest form of the visual search model is based on winner-take-all correlation matching. (a) At a given location, the filter responses are remembered. (b) Next, gaze is transferred to another point. The search problem is to find the original location in this new view. (c) The saliency map, showing the highest correlation value (brightest point) at the original location. (d) Gaze is transfered back to that location.

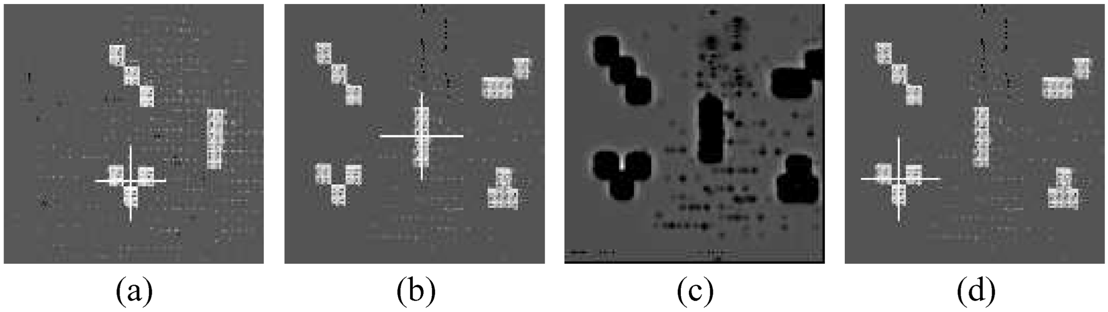

Fig. 6. Visual search using spatial filter responses. The simplest form of the visual search model is based on winner-take-all correlation matching. (a) At a given location, the filter responses are remembered. (b) Next, gaze is transfered to another point. The search problem is to find the original location in this new view. (c) The saliency map, showing the highest correlation value (brightest point) at the original location. (d) Gaze is transfered back to that location.

sky, Rao, Hayhoe, & Ballard, 1997). In this experiment, fixation patterns were observed in a simple search paradigm using natural images of three different scenes: a crib, a workbench and a dining table. Subjects were asked to fixate a point near the bottom of a $12° \times 16°$ display. They were given a one second presentation of an image containing a single object (e.g. a tool) at the fixation point, defining the search target, on a realistic background (e.g. the workbench). This was followed approximately one second later by a scene that filled the display and contained one, three, or five objects (e.g. various tools) on the same background. Images of the objects were placed on the background on-line at one to five of the six possible equi-eccentric locations (22.5°, 45°, 67.5°, 112°, 135°, and 157.5°, each located at an eccentricity of 7°) along an arc centered on the subject's initial fixation point (see Fig. 7(a)). The objects themselves subtended about 2° of visual angle. The subjects were asked to indicate (by pressing a button), as quickly and accurately as possible, whether the previewed object was among the group of one to five objects in the subsequent view. Note that the configuration of the objects in the experiment was like that shown in the following figure (see Fig. 7(a)). For each subject, each of the search trials tested a unique configuration of objects and positions. The trials were evenly divided into randomly interleaved target-present and target-absent conditions for set sizes of one, three, and five objects. The background objects were always present. Eye movements were recorded when the subjects performed this visual search task for both color and gray scale images of the targets and scenes. The subject's eye was tracked using a Generation-V Dual Purkinje image eye tracker. Note that although eye movements were recorded, the subject was given no instructions about eye movements except to hold fixation before the stimulus presentation. The task was described simply to respond whether the target was present or absent.

The typical eye movements elicited in this particular task are shown in Fig. 7(a). The surprising result was

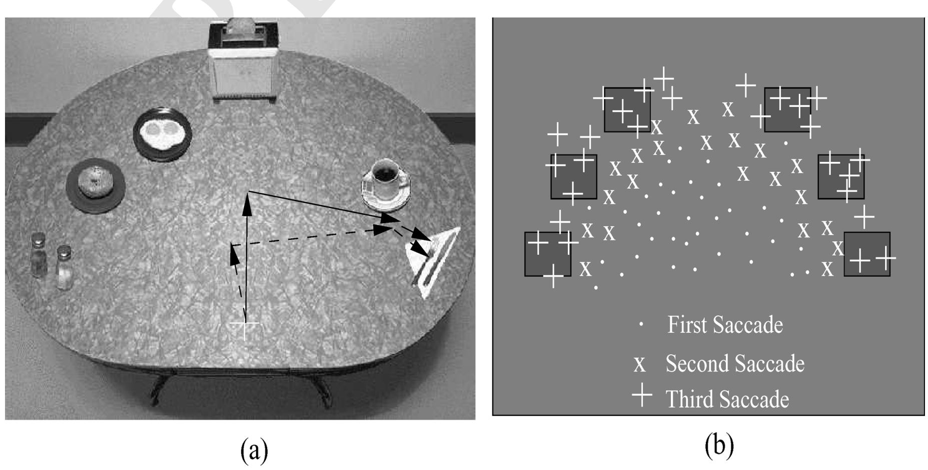

Fig. 7. Eye movements in the visual search task. Measurements from actual human data show marked differences from the simple winner-take-all model: (a) shows the typical pattern of multiple saccades (shown here for two different subjects) elicited during the course of searching for the object composed of the fork and knife. The initial fixation point is denoted by "+"; (b) depicts a summary of such movements over many target-present search trials as a function of the six possible locations of a target object on the table.

that rather than a single movement to the location of the memorized target, several saccades are typical, with each successive saccade moving closer to the target location (Fig. 7(b)). This "skipping" of the saccades in this search paradigm proved to be an extraordinarily robust finding, occurring in almost all 480 trials across all four subjects (Zelinsky et al., 1997).

5. Appearance-based search model

The simple model described in Section 3 cannot account for the experimentally observed multiple fixations, since its winner-take-all strategy means that only a single saccade is computed. However, multiple fixations can be fairly easily modeled if the computation of the saliency map is modified in the following three ways:

(1) The saliency map computation is made to be slower than the time needed to make an eye movement. This would imply that eye movements are made to target locations as determined by the *current* state of the saliency map, rather than waiting until the final state has been computed.

(2) The saliency map is computed using the larger spatial scale filters first, adding saliency information from successively finer scales as the search process evolves over time. Motivation comes both from the data and several studies that show that lower spatial frequencies influence the decision process earlier than higher spatial frequencies (Bichot & Schall, 1999; Gilchrist & Heywood, 1999; McPeek & Keller, 2001; Schyns & Oliva, 1994).

(3) The most likely target location is computed using a weighted averaging scheme rather than a pure winner-take-all mechanism. In conjunction with (1) and (2) above, this would imply that early eye movements are directed to "center-of-gravity" locations since only coarse scale information regarding the objects and the background is available at the early stages of the search, thereby biasing the weighted averaging model towards the center of the scene. The motivations for doing this is that it is known that in some circumstances saccades display a "center-of-gravity" property and fall midway between potential targets (Coren & Hoenig, 1972; Findlay, 1982, 1987; He & Kowler, 1989). The movement of the first saccade to the center of the image is likely to be a center-of-gravity effect, caused by the presence of many potential targets in the scene.

To implement these modifications, the simple winner-take-all model of Section 3 was changed to the following:

1. Set the initial scale of analysis k to the largest scale i.e. $k = \max$; set $S(x,y) = 0$ for all (x,y).
2. Compute the current saliency map across all locations (x,y) based on filter responses from the current scale k up to the maximum scale
$$S(x,y) = \sum_{s=k}^{\max} ||\mathbf{r}_s(x,y) - \mathbf{r}_s^m||^2 \qquad (6)$$
As before, $S(x,y)$ is the square of the Euclidean distance between the filter response vector $\mathbf{r}_s$ for image location (x,y) and the memorized target response vector $\mathbf{r}_s^m$, summed over the scales $s = k, \dots, \max$.
3. Find the location for saccadic targeting using the following *weighted population averaging scheme*:
$$(\hat{x},\hat{y}) = \sum_{(x,y)} F(S(x,y))(x,y) \qquad (7)$$
where F is an interpolation function. For the experiments, we used
$$F(S(x,y)) = \frac{\exp(-S(x,y)/\lambda(k))}{\sum_{(x,y)} \exp(-S(x,y)/\lambda(k))} \qquad (8)$$
This choice is attractive since it allows an interpretation of the search algorithm as computing *maximum likelihood estimates* (cf. Nowlan, 1990) of target locations. In the above, $\lambda(k)$ is a "temperature" parameter that is decreased with k. Decreasing $\lambda(k)$ allows the search to evolve from an initial state where all target locations compete equally for a saccade to a final state where only a few most likely target locations remain.
4. Move the eye to the location found by step (3). Although in our simulations we can get away with not actually implementing this step, as explained below.
5. Repeat steps (2), (3) and (4) above with $k = \max - 1, \max - 2, \dots$ until either the target object has been foveated or the number of scales has been exhausted. In the former case, the decision process signals the termination of the search process. In the latter case, subsequent eye movements are made using saliency maps based on all the scales.

The model has only one parameter, the initial value of $\lambda(1)$. The function of $\lambda(k)$ is to sharpen the peaks in the saliency map. The specific initial value of $\lambda(1)$ is dependent on the values in the filter kernels. With each target computation, $\lambda(k)$ was decreased by a factor of two, thereby allowing the search to evolve from an initial coarse resolution state where many target correlations contribute to a saccade, to a final state where only a single most likely target location contributes. The values for $\lambda(k)$ used were 4, 2 and 1 for $k = 1$, 2 and 3 respectively. The exact values are not crucial; the data can be fit qualitatively with values of $\lambda(1)$ ranging from 1 to 20. The same values of $\lambda(k)$ are used for all scenes and target locations within a scene.

The modified targeting model was implemented on our pipeline image processor. Fig. 8 shows the saliency

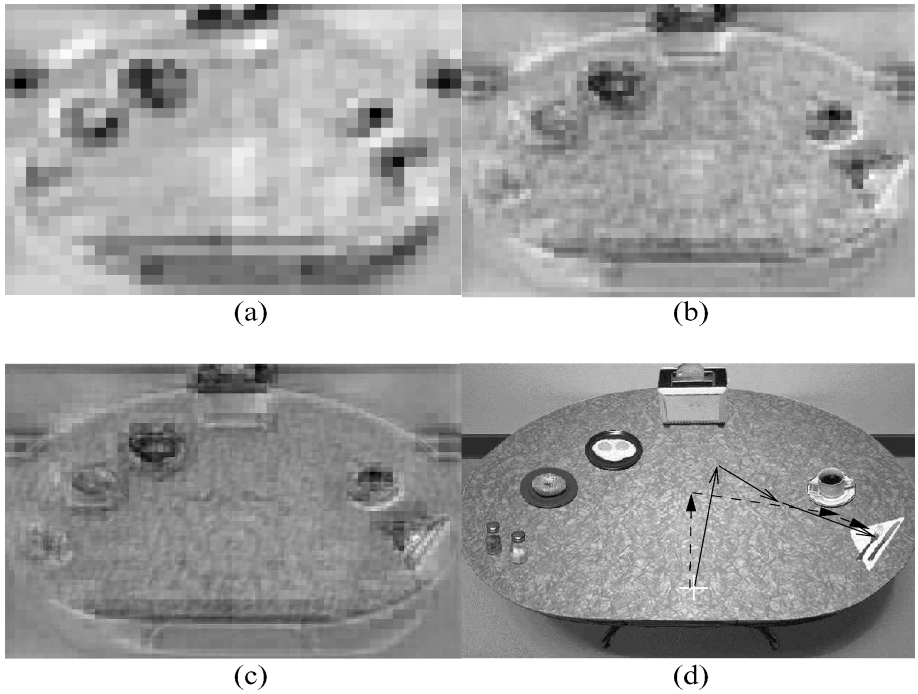

Fig. 8. Illustration of coarse-to-fine saccadic targeting. The saliency map $S(x, y)$ after the inclusion of the largest (a), intermediate (b), and smallest scale (c) as given by filter response distances to the prototype (the fork and knife); the brightest points are the closest matches; (d) shows the predicted eye movements as determined by the weighted population averaging scheme. For comparison, saccades from a human subject are given by the dotted arrows.

maps for this image after each of three iterations, with the middle and highest frequencies included in (b) and (c) respectively. Part (d) of the figure shows the sequence of fixations generated by the model for this image, together with those from a human subject. The target (composed of the fork and the knife) was the same in both cases. Thus the coarse-to-fine analysis, together with center-of-gravity effects, can produce the kind of fixation patterns that human subjects generate with this image.

In Fig. 8 the saliency map should of course be shifted with gaze. The reason we do not do this is simple expediency. Since we assume the resolution is chosen at the outset of the search, this implies that it is not changed during the target selection, therefore the saliency map cannot take advantage of the resolution of the fovea during the targeting period. The reason for this may not be obvious: if the target is being decided upon by some kind of correlation, the correlation function for foveated targets and non-foveated targets must be adjusted in a way that depends on the eccentricity and target. Otherwise a false target near the fovea might appear better than an eccentric true target. This is avoided in the model by selecting a resolution based on the signal-to-noise properties of the display and using that resolution cutoff everywhere in the resultant search window. As a consequence the saliency map is, to a first approximation, just shifted by saccades. We do not shift it in our figures in order to more easily compare visually the temporal effect of sequentially applying the multiple-scale filters.

6. Model–data comparison

The model's performance was compared to human data taken with 480 search trials pooled over four subjects. Owing to the nature of the different distractors and targets, there is substantial intersubject variability for each configuration, nonetheless, on the average, the model is remarkably good at approximating the actual gaze changes that subjects make. To show this we did the following analyses. The first step was to separate the sequences that ended up on the target with those that went to neighboring targets. Over the 480 trials, many records showed eye movements to nearby targets. This data is consistent with observations of both Kowler and Findlay who showed, particularly in the case when eye movements are made immediately upon the onset of the display, that a percentage of the movements were to false targets. Interestingly, the model also makes eye movements to false targets, but generally not to the same ones made by the subjects. Thus to compare the two sets of data we did the following:

(1) We generated an *average observer's* path to each of the six locations by averaging the fixations over subjects and target images. The coordinates were weighted by the variance between subjects. This meant that if a subject's movements were dissimilar to the group, they counted less in the sum. In the small number of cases where there were more than three saccades, only the first three were counted, as by the third saccade the eyes were always very close to one of the targets.

(2) The model data was averaged over the different targets for each location. In addition, trials where the final saccade was closer to a false target were excluded from the data and scored as errors. This resulted in 27 false targets in 120 model trials. In comparison, if we count human subject trials that had a standard deviation of the subjects' final gaze points of more than 75% of the intertarget separation difference as errors, then 29 of the records averaged over subjects are counted as false targets.

After these steps the results are shown in Fig. 9. The box in each sub-figure represents a 1° region centered on each target location. As is evident there is very good agreement between the model and human data for each location. Furthermore the number of errors made by the model is in very close agreement with the number of errors made by human subjects. It would be perhaps desirable to have the model represent an average or prototypical subject, but we cannot do this as the filters used by the model are probably slightly different than those used by the subjects, as described subsequently. However, we can ask whether the model is representative of an individual subject, and there the evidence is very encouraging. The average standard deviation for the subjects, averaged overall fixations is 1.5° whereas the average difference between model and average subject fixations is 0.7°. Thus the model behavior is well within the profile expected of an individual subject.

We also examined the saccades to false targets to see if there was any systematic bias in terms of location, target or scene type. One might well ask why there should be *any* false targets, since the decisions made by the subjects as to target presence are 100% accurate. We believe that the model provides an answer: (a) the decision process is separate from the targeting process and thus can still function when the ultimate target is eccentric, and (b) gaze can be mislocated since the template is defined on a neutral background and the background of the display bleeds into the larger filters, disturbing the correlation computation.

Table 1 shows this data for target location. The table shows the principal difference between the human and model data. The model had no difficulty with the crib scene, where targets were arrayed on a high contrast background, but the human subjects spread their errors around all three scenes uniformly. We interpret this to mean that the filter model is not identical to that used by the human subjects in that the filters are too sensitive to contrast and not sensitive enough to the fine structure in the targets. Nonetheless, given this caveat, the overall pattern of errors among locations is fairly uniform in both data sets.

Additional evidence for the correlation model comes from a control experiment that we performed, in which

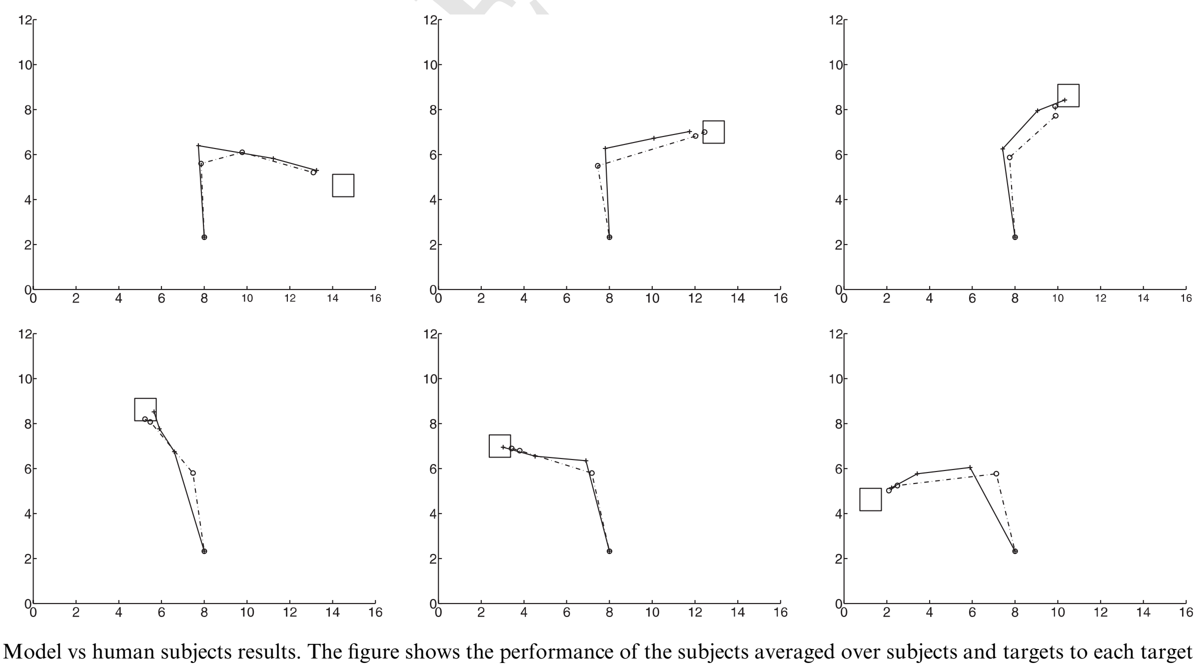

Fig. 9. Model vs human subjects results. The figure shows the performance of the subjects averaged over subjects and targets to each target location (see text). The scale is in degrees and the box shows a 1° region centered around each target. Circles and plus symbols mark the fixation points for human and model data respectively.

Table 1
Number of false targets for the model and human subjects, broken down by target location and scene

Location	Crib	Dine	Work
Model			
1		3	3
2		3	1
3			3
4		2	
5		2	3
6		4	3
Subjects			
1	3	4	3
2	2	3	1
3		1	
4	3	2	2
5	1	1	
6		1	2

the contextual background (e.g. the workbench and other objects) was removed, and the search objects were presented on a uniform background. Table 2 shows the results in the form of initial endpoint error after the first saccade. A striking point of comparison is the difference in error for search scenes containing a single object in the case of a uniform color background (c) vs a non-uniform realistic background (a) and (b). In the former case, the error is reduced by a factor of two for color images and slightly more than that for the gray scale images. This result implies an interference due to the background in the targeting process, as assumed by the model. As one might expect, the effect of the background is less as the number of target objects increases. This experiment is described in more detail in (Zelinsky et al., 1997). It is also of interest to compare the endpoint error for color and gray scale images. A small difference is evident after the first saccade. After the second saccade, the endpoint error was a full 1° less in the case of color images, strongly suggesting that color information is being used in the targeting computation. Although the simulation results described in this section modeled human eye movement data from gray scale images, the model can be readily extended for saccadic targeting based on color information.

Table 2
The effect of background on saccade accuracy. Mean endpoint error (in degrees) across all four subjects after the first saccade as a function of three different display conditions: (a) color images with a realistic background, (b) gray scale images with a realistic background, and (c) color images with a uniform background

Condition	Set size		
	1	3	5
(a) Color	3.2	4.8	5.1
(b) Gray	3.8	5.0	5.2
(c) Uniform background	1.6	4.8	5.1

The errors are shown for set sizes of one, three, and five objects in the search scene. Note that a uniform background for one target causes initial saccade accuracy to increase by a factor of two, implying that the background and other targets are deviating the saccade trajectory.

7. Appearance-based search vs spatial memory search

In both the model and experiment there is no prior knowledge of the specific location of the target before the presentation of the search array. Thus the only information available in both cases is *what* the target looks like, not *where* it is, and the search strategy is based primarily on the object's appearance. However, it seems intuitively likely that information about an object's location based on previous fixations in a continuously present scene, would contribute to the search process. Both physiological and psychophysical evidence reveal the ability to make saccades purely on the basis of information about spatial location (Colby & Goldberg, 1999). Precuing a location also reduces saccade latencies to that location. However, it is not clear what role spatial information plays when the stimulus is present on the retina and can be chosen on the basis of appearance, as is ordinarily the case in natural viewing, where subjects have usually made multiple fixations in a scene. Evidence from natural tasks such as tapping (Epelboim, Steiman, Kowler, & Pizlo, 1997; Land, Mennie, & Rusted, 1999) suggest that spatial information does ordinarily play a role in the targeting process. Thus adding spatial information to the task should affect the targeting strategy.

To test whether spatial information in addition to appearance factors would change the search pattern, a modification of the visual search task described above was run, where subjects were allowed to briefly *preview* the search scene (without knowing the search target) in a separate interval just before the search target was presented. Subjects were given a one second opportunity to preview the search scene prior to the presentation of the target. In this period, they were allowed to move their gaze freely, allowing them to fixate individual targets. The rest of the experiment remained the same as before (Zelinsky & Sheinberg, 1997). The subjects held fixation on a fixation cross, an icon of the target was then presented at the fixation point, followed by the search scene. An analysis of the eye movement data revealed that single saccades were by far the most common, as summarized in Fig. 10. The histograms show the initial endpoint error after the first saccade for the original search paradigm and the same for the case where subjects had a one second preview of the scene containing the potential targets. For most but not all of the preview cases, the initial endpoint error is 1° or less, strongly suggesting that subjects use the spatial location of the targets as an integral part of the search process. In addition, the reaction time for the decision was about 100 ms faster when the preview was presented, suggesting

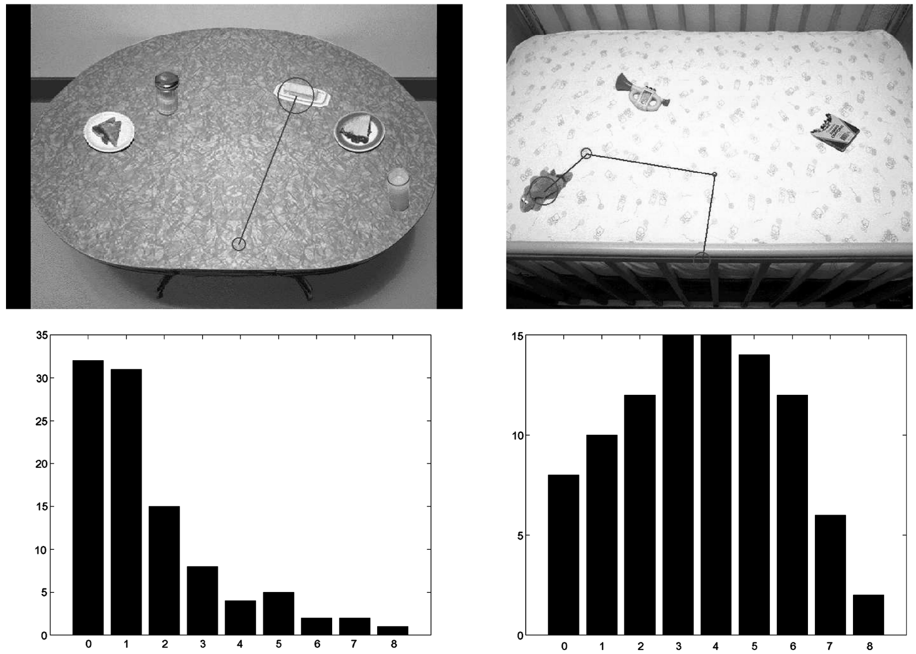

Fig. 10. Comparing preview vs no preview. The graph shows histograms of the endpoint error after the first saccade for the original search paradigm and the case where subjects had a one second preview of the potential targets. For most but not all of the preview cases the endpoint error is 1° or less, implying that subjects were able to remember and use the spatial location of the targets. Histograms: vertical axis = frequency of occurrences, horizontal axis = degrees.

that the location information facilitated the search process (Zelinsky & Sheinberg, 1997). This might occur if subjects were able to associate locations in the saliency map with the filter response vectors for objects, so that seeing one of these objects would now "prime" the corresponding location in the saliency map. This priming would in turn allow more accurate saccadic targeting in the cases where the target location happened to be inventoried during the preview period. It is important to remember that the subjects' task was simply to respond with a key press whether the target was present or not. No instructions were made about eye movements except that the subject should fixate the cross before the stimulus presentation. Thus it is likely that the observers are integrating the spatial and appearance information as part of a natural search strategy that results in more direct saccades. A way to extend the model to do this is described in (Ballard et al., 1997).

8. Discussion

The current model shares some mechanisms used by Itti and Koch (2000). They also propose a specific computational implementation of stimulus saliency for general scenes. Itti and Koch also propose filtering the image at different spatial scales. However, their model differs in that separate saliency maps are computed for color, intensity, and orientation. These separate maps are linearly combined following iterative lateral competition within each map. The saliency peak is then found using a winner-take-all network. Our model has a single saliency map by using oriented spatiochromatic filters, but the most important difference is that it uses a top-down search template to locate the saliency peak. Itti and Koch have no obvious way of searching for specific targets that are not contained in their bottom-up maps. Furthermore, our model works with unsegmented images, and thus avoids the difficult task of deciding what constitutes a "feature." The other important difference is our evolution of the signal in time with the addition of information at higher spatial frequency which is needed to fit the human data. Itti and Koch also have no direct comparisons with human data.

The model used by Tsotsos (Tsotsos et al., 1995) is more similar to that described here in that it has a top-down target component. However, there is no attempt in the Tsotsos model to model the details of eye movements in a way that could capture the skipping saccades seen in human data.

The model shares some general similarities with the visual search model proposed (but not implemented) by Findlay and Walker (1999), as well as that of Hooge (1996). Their suggestion of a temporal evolution of the saliency map takes specific form here. We differ most from Findlay and Walker in the representation of temporal control. In our model there is no explicit temporal control of saccades other than the assumption that the saliency map takes about 400 ms to evolve. We see this as a distinct biological advantage. By decoupling the dependence of the saliency map dynamics with the targeting system, they can be simpler and work independently.

Although not a computer model, the model used by Motter and Holsapple (2000) is very relevant to our work. Motter's studied monkeys search patterns in looking for small conjunction targets of color and shape and found that the data in different displays could be normalized by dividing by the density of search patterns of the correct color. He terms this an adjusted nearest neighbor distance (ANND). The reason this is relevant to our own model is that although not implemented, we conceptualize the search window as being adjusted based on signal-to-noise characteristics. The ANND concept can be seen as making a similar suggestion as dense target arrays can reduce signal-to-noise as shown by Palmer et al. (2000).

The model also shares some similarities with that of Wolfe, Cave, and Franzel (1989) and might be seen as an extension that fixes important problems with that model. In the Wolfe and Cave model, top-down priming of features in the saliency map computations can direct the search. Important differences arise in how these computations are carried out. To implement these calculations, their model requires that the features be segmented from the background, an unrealistic requirement in general. In contrast, our general correlation-based targeting method can handle arbitrary targets. More importantly, by separating the eye movements from the decision process, as is done in our model, means that gaze does not have to search every item in a multiple-item search task, but can use area-based calculations. The skipping data provides evidence that this can happen as the eyes move to non-target locations en route to making a decision. Motter's ANND data and Zelinsky's data provide further evidence for area-based vs item-by-item search.

Explaining the observed skipping saccades is done using a coarse-to-fine matching mechanism. The main benefit of a coarse-to-fine strategy is that it allows continuous execution of the decision and oculomotor processes, thereby increasing the probability of an early match. Coarse-to-fine strategies have also enjoyed recent popularity in computer vision with the advent of image pyramids for tasks such as motion detection (Burt, 1988). One key question that remains is the source of sequential application of the filters in the human visual system. This will usually result from the variation in resolution of the retina. Since resolution falls off with distance from the fovea, the fine spatial scales could be ineffective during early stages of search simply because the fixation point is distant from the target. However, our model suggests a different explanation. First, the three filters used in the model predictions were centered about 1, 3, and 6 cycles per degree. Even the highest of these should be visible at an eccentricity of 7° (Anderson et al., 1991). To test if the targets were identifiable at this eccentricity, in a control experiment observers were required to identify the targets while maintaining fixation. They were able to do this with negligible errors but used much longer reaction times (Zelinsky & Sheinberg, 1997). In addition, in the experiment where subjects were given a preview, many saccades went directly to the target, suggesting that resolution did not preclude direct targeting. Since the model fits the data well, it suggests that the additional effects on targeting from higher acuity measurements might be small.

An additional explanation for the sequential application of the filters is that the cortical machinery is setup to match the larger scales first, as target information is propagated via cortico-cortical feedback from higher to lower areas in the visual cortical hierarchy. If this were the case, the observed data would result from the fact that the oculomotor system is ready to move before all the scales can be matched, and thus the eyes move to the current best target position. This interpretation of the data is appealing for two reasons. First, it reflects a long history of observations on the priority of large scale channels in vision (Breitmeyer, 1975; Navon, 1977; Parker & Dutch, 1987). A particularly relevant experiment is that of Schyns and Oliva (1994). This shows that in a recognition task with 30 ms exposures, subjects are sensitive to the low frequencies in the image whereas with 150 ms exposures, subjects respond to the high frequency content. Second, in a search experiment similar to ours done by Findlay (1997), when subjects held their gaze before before starting the search, the pattern of saccades was more direct, suggesting that the target location had been refined during the wait. In another experiment using pairs of targets, Findlay (1997) found evidence that the saccade target signal is initially coarsely localized, and becomes more refined with increasing duration. Thus it is not clear whether the coarse-to-fine analysis is instantiated in the hardware or whether it is a de facto consequence of peripheral resolution fall off. Even if peripheral information is not limiting in a particular instance, coarse-to-fine analysis may develop as a naturally efficient strategy, since foveation will invariably lead to additional high frequency information for the current perceptual decision.

An alternative explanation for the initial saccade towards the center of the display is that it is a preplanned saccade to facilitate the search by centering fixation within the search array. The brief latencies before the first saccade support the idea of some kind of preprogramming. However, it is not likely to be entirely strategic (as opposed to a center-of-gravity saccade) because the initial fixation is biased toward the target.

One might suspect that the findings were a product of the experimental setup, which had subjects's heads fixed in a bite-bar. To check this we repeated the tests using a stereoview head mounted display which contained an eye tracker. We did not analyze the results quantitatively, but skipping movements were ubiquitous in the data.

Normally, a saccade is followed by a 200–300 ms fixation period before the next saccade is generated. Under certain circumstances, *express saccades* are also observed (Fischer & Boch, 1983; Fischer & Ramsperger, 1984; Fischer & Weber, 1993). The fixation periods for express saccades are much shorter, in the range 70–100 ms. An analysis of the visual search results (Zelinsky et al., 1997) revealed that the fixation periods of some of the center-of-gravity "skipping" eye movements are much smaller than normal (around 80–130 ms), small enough to qualify them as express saccades. There is a very simple explanation of these short latencies in the context of the proposed model. In a normal fixation, information from that fixation is presumably used in the computation of the next target. This necessitates some setup time for the information to be part of the targeting computation. However, in some cases, the next target may not require information from the current fixation. In such cases, the fixation times can be made much shorter. Such a situation may occur in the case of the "skipping" eye movements, as the targeting is based on a correlation process which is being done sequentially across scales. Of course, the partial correlation results contained in the saliency map have to be "shifted" due to the intermediate eye movements, before being integrated, but the eye movement itself contains the information necessary to perform this shifting. The crucial point is that express saccades may simply reflect a simple relationship between the ongoing computation of the saliency map and the motor command that executes eye movements. When the saliency map computations can be speeded up, the rate of saccades can be made correspondingly faster.

There exists a vast literature on the role of attention in visual cognition (Duncan & Humphreys, 1992; Krose & Julesz, 1989; Posner & Petersen, 1990; Saarinen & Julesz, 1991; Treisman, 1988; Treismann & Gelade, 1980). Attention has been characterized as covert search based on the metaphor of an attentional spotlight. Some of the search results have suggested that targets can be examined at the rate of about 25 ms per item, with the attentional spotlight moving from one location to the next at a speed of about one attentional shift every 30–50 ms (Krose & Julesz, 1989; Saarinen & Julesz, 1991). Models of attention (for example, Niebur & Koch, 1996) have in fact literally modeled this shift of the "focus of attention". The technical advantage of such a strategy is that, since gaze is fixed, retinal coordinates can be used for keeping track of examined locations. However, since signal transmission through visual cortex is on the order of 80–100 ms, performing covert search with an attentional spotlight while simultaneously obeying this stringent time constraint seems a difficult endeavor. An alternate explanation provided by the present model is that covert search occurs whenever the decision process finishes before an eye movement is made. This would occur, for example, in the cases where the presence of the target in a peripheral location can be judged directly from the correlation peaks in the saliency map using a signal-to-noise criterion. Under such circumstances, the eye movement becomes superfluous and a decision as to the presence or absence of the target can be made immediately without the need for an overt saccade. Such an interpretation is especially attractive since it allows a single targeting mechanism to parsimoniously account for both covert and overt search. It is also consistent with a body of evidence suggesting that the "attentional" (decision-making) and saccadic systems are regulated by different but closely related oculomotor control systems (Shepherd, Findlay, & Hockey, 1986; Groner, 1988; Corbetta, 1999; Findlay, 1997; Motter & Belky, 1998; Rizzolatti, 1996). The model has the additional advantage of being simpler than models that use additional machinery to couple the decision and targeting systems (e.g. Findlay, 1997).

9. Conclusion

A large number of computational models pertaining to human visual search and attention have previously been proposed (Chapman, 1991; Niebur & Koch, 1996; Olshausen, Van Essen, & Anderson, 1993; Tsioutsias & Mjolsness, 1996; Tsotsos et al., 1995; Wolfe, 1994). Many of these rely on predominantly bottom-up attentional processes based on various forms of feature maps that are used to facilitate search. Some of these models were motivated primarily by the need to explain classical reaction time results rather than the pattern of eye movements observed during visual tasks. Other models have explored the use of bottom-up saliency maps and have used eye movement scan-paths as sensorimotor memories for recognition (Didday & Arbib, 1975; Giefing, Janßen, & Mallot, 1991; Rimey & Brown, 1991; Rybak, Gusakova, Golovan, Podladchikova, & Shevtsova, submitted for publication; Yamada & Cottrell, 1995). This paper proposes a new model of the gaze targeting process in natural tasks based on observations of (Geisler & Chou, 1995; Motter & Holsapple, 2000; Palmer et al., 2000) that uses both bottom-up scene representations as well as top-down target descriptions for gaze control.

The model has four principal features:

(1) Instead of "features" that are preselected independently of a task, the model uses iconic templates that are task-dependent. As they are expressed in terms of

image filter responses, that are both more general and simpler to use than features. Eye movement models that are based on a fixed library of features cannot explain how arbitrary targets are computed.

(2) The model separates the process of changing gaze from that of deciding on properties of a target. This has the virtue of allowing the timing relationships between these two processes to be a natural consequence of the properties of the scene. This greatly simplifies the control problem of coordinating eye movements and decisions.

(3) The model specifies that the correlation used to select targets proceeds in a coarse-to-fine manner that takes time. If the target is novel and its location must be determined solely on appearance, this time is longer than that needed to generate an eye movement, and consequently effects the gaze trajectory in a predictable way. This result provides a concrete model of a myriad of experimentally observed "center-of-gravity" observations. Since our center of gravity is correlation-based, it is readily tested experimentally.

(4) The most controversial aspect of the model is its use of area-based search. The assumption is that the resolution used to search for the target can be chosen at the beginning of the search based on the signal-to-noise properties of the search area. The motivation for being able to do this is to search large areas at comparable resolution. The assumption that humans would not make continuous sue of all the available resolution in the retinotopic array is counterintuitive. We have argued that it has precedents in search models, and our experiments show (1) that the model fits the data well and (2) foveal resolution is not necessary for target location. However we cannot rule out the use of all the available resolution by human subjects, so that this question needs to be settled by further experiments.

The model is constructive, has a specific computational prescription for target computation, and fits experimental observations. Its most controversial claim is that, for the experimental conditions tested, it can use resolutions much lower than that ultimately available from the scene to guide gaze changes. As a consequence, the effect of additional foveal resolution has minimal effects on the gaze trajectory. We anticipate that situations could be constructed for which foveal effects would be seen, but those effects may prove a refinement on the model presented here.

The main goal of the model was to capture the exogenous effects of the visual stimulus. There has been no attempt to model endogenous target specifications e.g. anti-saccades. However these effects have been modeled by Kopecz and Schoner (1995) and Trappenberg, Dorris, Munoz, and Klein (2001) in a way that is compatible with our model.

Acknowledgements

This work was supported by NSF research grant no. CDA-8822724, NIH/PHS research grants no. 1-R24-RRO6853, EY-05729 and 1-P41-RR09283, and a grant from the Human Science Frontiers Program. The paper greatly benefitted from the reviewers' comments.

References

Anderson, S. J., Mullen, K. T., & Hess, R. F. (1991). Human peripheral resolution for chromatic and achromatic stimuli-limits imposed by optical and retinal factors. *Vision Research*, *442*, 47–64.

Ballard, D. H., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, *7*(1), 66–80.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. N. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*.

Barrow, H. G. (1987). Learning receptive fields. In *Proceedings of the IEEE International Conference on Neural Networks* (pp. 115–121).

Bell, A. J., & Sejnowski, T. J. (1997). The 'independent components' of natural scenes are edge filters. *Vision Research*.

Bichot, M., & Schall, J. (1999). Saccadic target selection in the maccaque during feature and conjunction visual search. *Visual Neuroscience*, *16*, 81–89.

Breitmeyer, B. G. (1975). Simple reaction time as a measure of the temporal response properties of transient and sustained channels. *Vision Research*, *15*, 1411–1412.

Bulthoff, H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Science USA*, *89*, 60–64.

Burt, P. J. (1988). Attention mechanisms for vision in a dynamic world. In *ICPR* (pp. 977–987).

Chapman, D. (1991). *Vision, instruction, and action*. Cambridge, MA: MIT Press.

Colby, C., & Goldberg, M. (1999). *Annual Review of Neuroscience*, *22*, 319–349.

Coren, S., & Hoenig, P. (1972). Effect of non-target stimuli upon length of voluntary saccades. *Perceptual and Motor Skills*, *34*, 499–508.

Derrico, J. B., & Buchsbaum, G. (1991). A computational model of spa-tiochromatic image coding in early vision. *Journal of Visual Communication and Image Representation*, *2*(1), 31–38.

Didday, R. L., & Arbib, M. A. (1975). Eye movements and visual perception: A two visual system model. *International Journal of Man–Machine Studies*, *7*, 547–569.

Duncan, J., & Humphreys, G. W. (1992). Beyond the search surface: visual search and attentional engagement. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 578–588.

Eckstein, MIguel P. (1998). The lower visual search efficiency for conjunctions is due to noise not serial attentional processing. *Psychological Science*, *9*, 111–118.

Epelboim, J., Steiman, R. M., Kowler, E., & Pizlo, Z. (1997). Gaze-shift dynamics in two kinds of sequential looking tasks. *Vision Research*, *37*, 2597.

Findlay, J. (1982). Global visual processing for saccadic eye movements. *Vision Research*, *22*, 1033–1045.

Findlay, J. (1987). Visual computation and saccadic eye movements: A theoretical perspective. *Spatial Vision*, *2*, 175–189.

Findlay, J. (1997). Saccade target selection during visual search. *Vision Research*, *37*, 617–631.

Fischer, B., & Boch, R. (1983). Saccadic eye movements after extremely short reaction times in the monkey. *Brain Research*, *260*, 21–26.

Fischer, B., & Ramsperger, E. (1984). Human express-saccades: extremely short reaction times of goal directed eye movements. *Experimental Brain Research*, *57*, 191–195.

Fischer, B., & Weber, H. (1993). Express saccades and visual attention. *Behavioral and Brain Sciences*, *16*, 553–610.

Freeman, W. T., & Adelson, E. H. (1991). The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *13*(9), 891–906.

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.

Giefing, G.-J., Janßen, H., & Mallot, H. (1991). A saccadic camera movement system for object recognition. In T. Kohonen, K. Makisara, O. Simula, & J. Kangas (Eds.), *Artificial Neural Networks* (vol. 1) (pp. 63–68). Amsterdam: Elsevier.

Gilchrist, I., & Heywood, C. (1999). Saccade selection in visual search: evidence for spatial frequency specific between item interactions. *Vision Research*, *39*, 1373–1393.

Graham, N. (1989). *Visual pattern analyzers*. New York: Oxford University Press.

Groner, R. (1988). Eye movements, attention and visual information processing: some experimental results and methodological considerations. In G. Luer, U. Lass, & J. Shallo-Hoffman (Eds.), *Eye Movement Research: Physiological and Psychological Aspects* (pp. 295–319). Gottingen, Germany: Hogrefe.

Hancock, P. J. B., Baddeley, R. J., & Smith, L. S. (1992). The principal components of natural images. *Network*, *3*, 61–70.

He, P., & Kowler, E. (1989). The role of location probability in the programming of saccades: Implications for "center-of-gravity" tendencies. *Vision Research*, *29*, 1165–1181.

Hooge, I., & Erkelens, C. (1998). Adjustment of fixation duration during visual search. *Vision Research*, *38*, 1295–1302.

Hooge, I. T. C. (1996). *Control of eye movements in visual search. Ph.D. Thesis*. Netherlands: University of Utrecht.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 11–46.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*, 441–480.

Kanerva, P. (1988). *Sparse distributed memory*. Cambridge, MA: Bradford Books.

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

Kopecz, K., & Schoner, G. (1995). Saccadic motor planning by integrating visual information and pre-information on neural dynamic fields. *Biological Cybernetics*, *73*, 49–60.

Kowler, E., & Anton, S. (1987). Reading twisted text: implications for the role of saccades. *Vision Research*, *27*, 45–60.

Krose, B. J. A., & Julesz, B. (1989). The control and speed of shifts of attention. *Vision Research*, *29*(11), 1607–1619.

Land, M. F., & Furneaux, S. (1997). The knowledge base of the oculomotor system. In *Proceedings of the Royal Society Conference on Knowledge-Based Vision*, February.

Land, M., Mennie, M., & Rusted, J. (1999). An activethe roles of vision and eye movements in the control of activities of daily living. *Perception*, *28*, 1311–1328.

McPeek, R., & Keller, E. (2001). Short term priming, concurrent processing and saccade curvature during a target selection task in the monkey. *Vision Research*, *41*, 785–800.

Motter, B., & Belky, E. (1998). The guidance of eye movements during active visual search. *Vision Research*, *38*, 1805–1815.

Motter, B. C., & Holsapple, J. W. (2000). Cortical image density determines the probability of target discovery during active search. *Vision Research*, *40*, 1311.

Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cognitive Psychology*, *9*, 353–383.

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

Nowlan, S. J. (1990). Maximum likelihood competitive learning. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems* (vol. 2) (pp. 574–582). Morgan Kaufmann: San Mateo, CA.

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

Olshausen, B. A., Van Essen, D. C., & Anderson, C. H. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, *13*, 4700–4719.

Oram, M. W., & Perrett, D. I. (1992). Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology*, *68*(1), 70–84.

O'Regan, J. K. (1990). Eye movements and reading. In E. Kowler (Ed.), *Eye Movements and Their Role in Visual and Cognitive Processes* (pp. 455–477). New York: Elsevier.

Palmer, J., Vergese, P., & Pavel, M. (2000). The psychophysics of visual search. *Vision Research*, *40*, 1227.

Parker, D. M., & Dutch, S. (1987). Perceptual latency and spatial frequency. *Vision Research*, *27*, 1279–1283.

Poetzsch, M., Krueger, N., & Von der Malsburg, C. (1996). Improving object recognition by transforming gabor filter responses. *Network*, *11*, 341.

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*, 25–42.

Rao, R. P. N., & Ballard, D. H. (1995a). An active vision architecture based on iconic representations. *Artificial Intelligence (Special Issue on Vision)*, *78*, 461–505.

Rao, R. P. N., & Ballard, D. H. (1995b). Learning saccadic eye movements using multiscale spatial filters. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in Neural Information Processing Systems* (vol. 7) (pp. 893–900). Cambridge, MA: MIT Press.

Rao, R. P. N., & Ballard, D. H. (1996). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, *9*, 721–763.

Rao, R. P. N., & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, *9*(4), 721–763.

Rimey, R. D., & Brown, C. M. (1991). Controlling eye movements with hidden Markov models. *International Journal of Computer Vision*, *7*(1), 47–65.

Rybak, L. A., Gusakova, V. I., Golovan, A. V., Podladchikova, L. N., & Shevtsova, N. A. (submitted for publicat). A model of attention-guided visual perception and recognition. *Vision Research*.

Saarinen, J., & Julesz, B. (1991). The speed of attentional shifts in the visual field. *Proceedings of the National Academy of Science, USA*, *88*, 1812–1814.

Schyns, P. G., & Oliva, A. (1994). From blobs to edges: evidence for time and spatial scale dependent scene recognition. *Psychological Science*, *5*, 195–200.

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

Sperling, G., & Dosher, B. A. (1986). The attention operating characteristic: some examples from visual search. *Science*, *202*, 315–318.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. E. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*, 632–634.

Thorpe, S. J., & Imbert, M. (1989). Biological contraints on connectionist modelling. In R. Pfeifer, Z. Schreter, F. Fogelman-Soulie, &

L. Steels (Eds.), *Connectionism in Perspective* (pp. 63–92). Amsterdam: Elsevier.

Trappenberg, T. P., Dorris, M. C., Munoz, D. P., & Klein, R. M. (2001). A model of saccade initiation based on the competitive integration of exogenous and endogenous signals in the superior colliculus. *Journal of Cognitive Neuroscience*, *13*, 256–271.

Treisman, A. (1988). Features and objects the fourteenth Bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology*, *40*(2), 201–237.

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

Tsioutsias, D. I., & Mjolsness, E. (1996). A multiscale attentional framework for relaxation neural networks. In D. Touretzky, M. Mozer, & M. Hasselmo (Eds.), *Advances in Neural Information Processing Systems* (Vol. 8) (pp. 633–639). Cambridge, MA: MIT Press.

Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y., Davis, N., & Nuflo, F. (1995). Modeling visual attention via selective tuning. *Artificial Intelligence (Special Issue on Vision)*, *78*, 507–545.

Viviani, P. (1990). Eye movements in visual search: cognitive, perceptual and motor control aspects. In *Eye Movements and Their Role in Visual and Cognitive Processes*, New York.

Weber, J., & Malik, J. (1995). Robust computation of optic flow in a multiscale differential framework. *International Journal of Computer Vision*, *14*, 67–81.

Wilson, H. R., & Wikinson, F. (1997). Evolving concepts of spatial channels in vision: from independence to nonlinear interactions. *Perception*, *26*, 939–960.

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

Wolfe, J. (1994). Visual search in continuous naturalistic stimuli. *Vision Research*, *34*, 1187–1195.

Yamada, K., & Cottrell, G. W. (1995). A model of scan paths applied to face recognition. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society* (pp. 55–60).

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

Zelinsky, G., & Sheinberg, D. (1997). Eye movements during parallel-serial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *23*, 244–262.

Zelinsky, G. J., Rao, R. P. N., Hayhoe, M. M., & Ballard, D. H. (1997). Eye movements reveal the spatio-temporal dynamics of visual search. *Psychological Science*.