A Saliency-Based Search Mechanism for Overt and Covert Shifts of Visual Attention

Laurent Itti and Christof Koch*

Computation and Neural Systems Program, California Institute of Technology, Pasadena, CA 91125.

Tel: (626) 395-6855; Fax: (626) 796-8876; Email: {itti,koch}@klab.caltech.edu

Abstract

Most models of visual search, whether involving overt eye movements or covert shifts of attention, are based on the concept of a *saliency map*, that is, an explicit two-dimensional map that encodes the saliency or conspicuity of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the next attended target. Inhibiting this location automatically allows the system to attend to the next most salient location. We describe a detailed computer implementation of such a scheme, focusing on the problem of combining information across modalities, here orientation, intensity and color information, in a purely stimulus-driven manner. The model is applied to common psychophysical stimuli as well as to a very demanding visual search task. Its successful performance is used to address the extent to which the primate visual system carries out visual search via one or more such saliency maps and how this can be tested.

1 Introduction

Most biological vision systems (including *Drosophila*; [1] appear to employ a serial computational strategy when inspecting complex visual scenes. Particular locations in the scene are selected based on their behavioral relevance or on local image cues. In primates, the identification of objects and the analysis of their spatial relationship usually involves either rapid, saccadic eye movements to bring the fovea onto the object, or covert shifts of attention.

It may seem ironic that brains employ serial processing, since one usually thinks of them as paradigmatic "massively parallel" computational structures. However, in any physical computational system, processing resources are limited, which leads to bottlenecks similar to those faced by the *von Neumann* architecture on conventional digital machines. Nowhere is this more evident than in the primate's visual system, where the amount of information coming down the optic nerve—estimated to be on the order of of 10^8 bits per second—far exceeds what the brain is capable of fully processing and assimilating into conscious experience. The strategy nature has devised for dealing with this bottleneck is to select certain portions of the input to be processed preferentially, shifting the processing focus from one location to another in a serial fashion.

Despite the widely shared belief in the general public that "we see everything around us," only a small fraction of the information registered by the visual system at any given time reaches levels of processing that directly influence behavior. This is vividly demonstrated by *change blindness* [2, 3, 4] in which significant image changes remain nearly invisible under natural viewing conditions, although observers demonstrate no difficulty in perceiving these changes once directed to them. Overt and covert attention controls access to these privileged levels and ensures that the selected information is relevant to behavioral priorities and objectives. Operationally, information can be said to be "attended" if it enters short-term memory and remains there long enough to be voluntarily reported. Thus, visual attention is closely linked to *visual awareness* [5].

But how is the selection of one particular spatial location accomplished? Does it involve primarily bottom-up, sensory-driven cues or does expectation of the targets characteristics play a decisive role? A large body of literature has concerned itself with the psychophysics of visual search or orienting for targets in sparse arrays or in natural scenes using either covert or overt shifts of attention (for reviews, see [6] or the survey article [7]).

Much evidence has accumulated in favor of a two-component framework for the control of where in a visual scene attention is deployed [8, 9, 10, 11, 12, 13, 14, 15, 16]: A bottom-up, fast, primitive mechanism that biases the observer towards selecting stimuli based on their *saliency* (most likely encoded in terms of center-surround mechanisms) and a second slower, top-down mechanism with variable selection criteria, which directs the "spotlight of attention" under cognitive, volitional control. Whether visual consciousness can be reached by either saliency-based or top-down attentional selection or by both remains controversial.

Preattentive, parallel levels of processing do not represent all parts of a visual scene equally well, but instead provide a weighted representation with strong responses to a few parts of the scene and poor responses to everything else. Indeed, in an awake monkey freely viewing a natural visual scene, there are not many locations which elicit responses in visual cortex comparable to those observed with isolated, laboratory stimuli [17]. Whether a given part of the scene elicits a strong or a poor response is thought to depend very much on "context", that is, on what stimuli are present in other parts of the visual field. In particular, the recently accumulated evidence for "non-classical" modulation of a cell's response by the presence of stimuli outside of the cell's receptive field provides direct support for the idea that different visual locations compete for activity [18, 19, 20]. Those parts which elicit a strong response are thought to draw visual attention to themselves and to therefore be experienced as "visually salient". Directing attention at any of the other parts is thought to require voluntary "effort".

Both modes of attention can operate at the same time and visual stimuli have two ways of penetrating to higher levels of awareness: Being willfully brought into the focus of attention, or winning the competition for saliency.

Koch and Ullman [21] introduced the idea of a saliency map to accomplish preattentive selection (see also the concept of a "master map" in [11]). This is an explicit two-dimensional map that encodes the saliency of objects in the visual environment. Competition among neurons in this map gives rise to a single winning location that corresponds to the most salient object, which constitutes the next target. If this location is subsequently inhibited, the system automatically shifts to the next most salient location, endowing the search process with internal dynamics (**Fig. 1.a**).

Many computational models of human visual search have embraced the idea of a saliency map under different guises [11, 22, 23, 24, 25]. The appeal of an explicit saliency map is the relatively straightforward manner in which it allows the input from multiple, quasi-independent feature maps to be combined and to give rise to a single output: The next location to be attended. Electrophysiological evidence points to the existence of several neuronal maps, in the pulvinar, the superior colliculus and the intraparietal sulcus, which appear to specifically encode for the saliency of a visual stimulus [26, 27, 28, 29].

However, some researchers reject the idea of a topographic map in the brain whose *raison d'être* is the representation of salient stimuli. In particular, Desimone and Duncan [30] postulate that selective attention is a consequence of interactions among feature maps, each of which encodes in an implicit fashion, the saliency of a stimulus in that particular feature. We know of only a single implementation of this idea in terms of a computer algorithm [31].

We here describe a computer implementation of a preattentive selection mechanism based on the architecture of the primate visual system. We address the thorny problem of how information from different modalities—in the case treated here from 42 maps encoding intensity, orientation and color in a center-surround fashion at a number of spatial scales—can be combined into a single saliency map. Our algorithm qualitatively reproduces human performance on a number of classical search experiments.

Vision algorithms frequently fail when confronted with realistic, cluttered images. We therefore studied the performance of our search algorithm using high-resolution (6144×4096 pixels) photographs containing images of military vehicles in a complex rural background. Our algorithm shows, on average, superior performance compared to human observers searching for the same targets, although our system does not yet include any top-down task-dependent tuning.

Finally, we discuss future computational work that needs to address the physiological evidence for multiple saliency maps, possibly operating in different coordinate systems (e.g. retina versus head coordinates), and the need to integrate information across saccades.

The work presented here is an considerable elaboration upon the model presented in [23] and has not been reported previously.

2 The Model

Our model is limited to the bottom-up control of attention, *i.e.* to the control of selective attention by the properties of the visual stimulus. It does not incorporate any top-down, volitional component. Furthermore, we are here only concerned with the localization of the stimuli to be attended ("Where"), not their identification ("What"). A number of authors [25, 32] have presented models for the neuronal expression of attention along the occipital-temporal pathway once spatial selection has occurred.

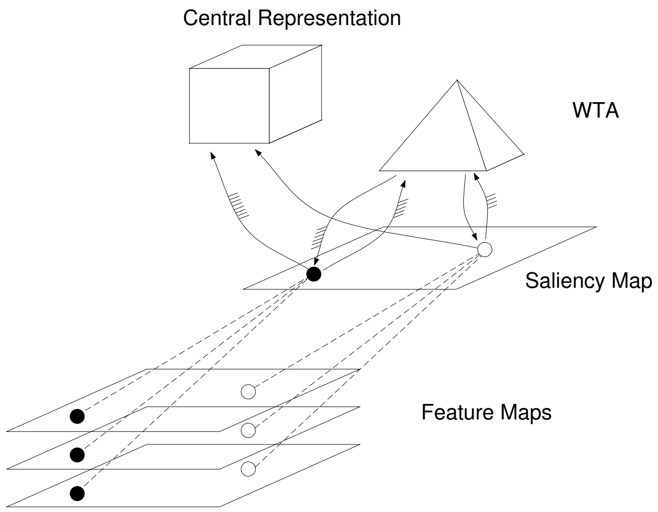
In the present work, we make the following four assumptions: First, visual input is represented, in early visual structures, in the form of iconic (appearance-based) topographic feature maps. Two crucial steps in the construction of these representations consist of center-surround computations in every feature at different spatial scales, and within-feature spatial competition for activity. Second, information from these feature maps is combined into a single map which represents the local "saliency" of any one location with respect to its neighborhood. Third, the maximum of this *saliency map* is, by definition, the most salient location at a given time, and it determines the next location of the attentional searchlight. And fourth, the saliency map is endowed with internal dynamics allowing the perceptive system to scan the visual input such that its different parts are visited by the focus of attention in the order of decreasing saliency.

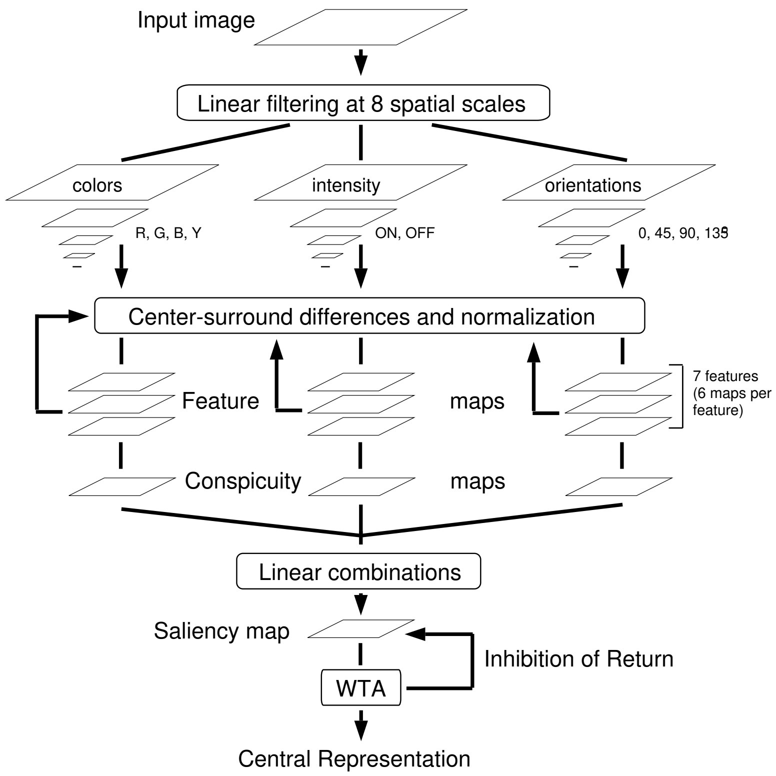
Fig. 1.b shows an overview of our model. Input is provided in the form of digitized images, from a variety of sources including a consumer-electronics NTSC video camera.

2.1 Extraction of Early Visual Features

Low-level vision features (color channels tuned to red, green, blue and yellow hues, orientation, and brightness) are extracted from the original color image at several spatial scales, using linear filtering. The different spatial scales are created using Gaussian pyramids [33], which consist of progressively low-pass filtering and subsampling the input image. In our implementation, pyramids have a depth of 9 scales, providing horizontal and

(a)





(b)

Figure 1: **(a)** Original model of saliency-based visual attention, adapted from Koch and Ullman [21]. Early visual features such as color, intensity or orientation are computed, in a massively parallel manner, in a set of pre-attentive feature maps based on retinal input (not shown). Activity from all feature maps is combined at each location, giving rise to activity in the topographic saliency map. The winner-take-all (WTA) network detects the most salient location and directs attention towards it, such that only features from this location reach a more central representation for further analysis. **(b)** Schematic diagram for the model used in this study. It directly builds on the architecture proposed in **(a)**, but provides a complete implementation of all processing stages. Visual features are computed using linear filtering at eight spatial scales, followed by center-surround differences, which compute local spatial contrast in each feature dimension for a total of 42 maps. An iterative lateral inhibition scheme instantiates competition for salience within each feature map. After competition, feature maps are combined into a single "conspicuity map" for each feature type. The seven conspicuity maps then are summed into the unique topographic saliency map. The saliency map is implemented as a 2-D sheet of Integrate-and-Fire (I&F) neurons. The WTA, also implemented using I&F neurons, detects the most salient location and directs attention towards it. An Inhibition-of-Return mechanism transiently suppresses this location in the saliency map, such that attention is autonomously directed to the next most salient image location. We here do not consider the computations necessary to identify a particular object at the attended location.

vertical image reduction factors ranging from 1:1 (level 0; the original input image) to 1:256 (level 8) in consecutive powers of two.

Each feature is computed in a center-surround structure akin to visual receptive fields. Using this biological paradigm renders the system sensitive to local spatial contrast in a given feature rather than to amplitude in that feature map. Center-surround operations are implemented in the model as differences between a fine and a coarse scale for a given feature: The center of the receptive field corresponds to a pixel at level $c \in \{2, 3, 4\}$ in the pyramid, and the surround to the corresponding pixel at level $s = c+\delta$, with $\delta \in \{3, 4\}$. We hence compute six feature maps for each type of feature (at scales 2-5, 2-6, 3-6, 3-7, 4-7, 4-8). Seven types of features, for which wide evidence exists in mammalian visual systems, are computed in this manner from the low-level pyramids: As detailed below, one feature type encodes for on/off image intensity contrast [34], two encode for red/green and blue/yellow double-opponent channels [35, 36], and four encode for local orientation contrast [37, 38].

The six feature maps for the intensity feature type encode for the modulus of image luminance contrast, i.e., the absolute value of the difference between intensity at the center (one of the three c scales) and intensity in the surround (one of the six $s = c + \delta$ scales). To isolate chromatic information, each of the red, green and blue channels in the input image are first normalized by the intensity channel; a quantity corresponding to the double-opponency cells in primary visual cortex is then computed by center-surround differences across scales. Each of the six red/green feature maps is created by first computing (red-green) at the center, then subtracting (green-red) from the surround, and finally outputting the absolute value. Six blue/yellow feature maps are similarly created. Local orientation is obtained at all scales through the creation of oriented Gabor pyramids from the intensity image [39]. Four orientations are used (0, 45, 90 and 135°) and orientation feature maps are obtained from absolute center-surround differences between these channels. These maps encode, as a group, how different the average local orientation is between the center and surround scales. A more detailed mathematical description of the preattentive feature extraction stage has been presented previously [23].

2.2 Combining Information Across Multiple Maps

Our modeling hypotheses assume the existence of a unique topographic saliency map. At each spatial location, activity from the 42 feature maps consequently needs to be combined into a unique scalar measure of salience. One major difficulty in such combination resides in the fact that the different feature maps arise from different visual modalities, which encode for *a priori* not comparable stimulus dimensions: For example, how should a 10° orientation discontinuity compare to a 5% intensity contrast?

In addition, because of the large number of maps being combined, the system is faced with a severe signal-to-noise ratio problem: A salient object may only elicit a strong peak of activity in one or a few feature maps, tuned to the features of that object, while a larger number of feature maps, for example tuned to the features of distracting objects, may show strong peaks at numerous locations. For instance, a stimulus display containing one vertical bar among many horizontal bars yields an isolated peak of activity in the map tuned to vertical orientation at the scale of the bar; the same stimulus display however also elicits strong peaks of activity, in the intensity channel, at the locations of all bars, simply because each bar has high intensity contrast with the background. When all feature maps are combined into the saliency map, the isolated orientation pop-out hence is likely to be greatly weakened, at best, or even entirely lost, at worst, among the numerous strong intensity responses.

Previously, we have shown that the simplest feature combination scheme—to normalize each feature map to a fixed dynamic range, and then sum all maps—yields very poor detection performance for salient targets in complex natural scenes [40]. One possible way to improve performance is to learn linear map combination weights, by providing the system with examples of targets to be detected. While performance improves greatly, this method presents the disadvantage of yielding different specialized models (that is, sets of synaptic weights), one for each type of target studied [40].

In the present study, we derive a generic model which does not impose any strong bias for any particular feature dimension. To this end, we implemented a simple within-feature spatial competition scheme, directly inspired by physiological and psychological studies of long-range cortico-cortical connections in early visual areas. These connections, which can span up to 6-8mm in striate cortex, are thought to mediate "non-classical" response modulation by stimuli outside the cell's receptive field. In striate cortex, these connections are made by axonal arbors of excitatory (pyramidal) neurons in layers III and V [41, 42, 43, 44]. Non-classical interactions are thought to result from a complex balance of excitation and inhibition between neighboring neurons as shown by electrophysiology [18, 19, 20], optical imaging [45], and human psychophysics [46, 47, 48].

Although much experimental work is being deployed in the characterization of these interactions, a precise quantitative understanding of such interactions still is in the early stages [48]. Rather than attempting to propose a detailed quantitative account of such interactions, our model hence simply reproduces three widely observed features of those interactions: First, interactions between a center location and its non-classical surround appear to be dominated by an inhibitory component from the surround to the center [49], although this effect is dependent on the relative contrast between center and surround [20]. Hence our model focuses on non-classical surround inhibition. Second, inhibition from non-classical surround locations is strongest from neurons which are tuned to the same stimulus properties as the center [18, 50, 51, 43, 52, 53]. As a consequence, our model implements interactions within each individual feature map rather than between maps. Third, inhibition appears strongest at a particular distance from the center [48], and weakens both with shorter and longer distances. These three remarks suggest that the structure of non-classical interactions can be coarsely modeled by a two-dimensional difference-of-Gaussians (DoG) connection pattern **(Fig. 2)**.

The specific implementation of these interactions in our model is as follows: Each feature map is first normalized to a fixed dynamic range (between 0 and 1), in order to eliminate feature-dependent amplitude differences due to different feature extraction mechanisms. Each feature map is then iteratively convolved by a large 2-D DoG filter, the original image is added to the result, and negative results are set to zero after each iteration. The DoG filter, a section of which is shown in **Fig. 2**, yields strong local excitation at each visual location, which is counteracted by broad inhibition from neighboring locations. Specifically, we have:

$$\mathcal{DoG}(x,y) = \frac{c_{ex}^2}{2\pi\sigma_{ex}^2} e^{-\frac{x^2+y^2}{2\sigma_{ex}^2}} - \frac{c_{inh}^2}{2\pi\sigma_{inh}^2} e^{-\frac{x^2+y^2}{2\sigma_{inh}^2}} \qquad (1)$$

In our implementation, $\sigma_{ex} = 2\%$ and $\sigma_{inh} = 25\%$ of the input image width, $c_{ex} = 0.5$ and $c_{inh} = 1.5$ **(Fig. 2)**. At each iteration of the normalization process, a given feature map $\mathcal{M}$ is then subjected to the following transformation:

$$\mathcal{M} \leftarrow |\mathcal{M} + \mathcal{M} * \mathcal{DoG} - C_{inh}|_{\geq 0} \qquad (2)$$

where $\mathcal{DoG}$ is the 2D Difference of Gaussian filter described above, $|.|_{\geq 0}$ discards negative values, and C_{inh} is a constant inhibitory term ($C_{inh} = 0.02$ in our implementation with the map initially scaled between 0 and 1). C_{inh} introduces a small bias towards slowly suppressing areas in which the excitation and inhibition balance almost exactly; such regions typically correspond to extended regions of uniform textures (depending on the DoG parameters), which we would not consider salient.

Each feature map is subjected to 10 iterations of the process described in **Eq. 2**. The choice of the number of iterations is somewhat arbitrary: In the limit of an infinite number of iterations, any non-empty map will converge towards a single peak (except for a few unrealistic, singular configurations), hence constituting only a poor representation of the scene. With few iterations however, spatial competition is weak and inefficient. Two examples of the time evolution of this process are shown in **Fig. 3**, and illustrate that using on the order of 10 iterations yields adequate distinction between the two example images shown. As expected, feature maps with initially numerous peaks of similar amplitude are suppressed by the interactions, while maps with one or a few initially stronger peaks become enhanced. It is interesting to note that this within-feature spatial competition scheme resembles a "winner-take-all" network with localized inhibitory spread, which allows for a sparse distribution of winners across the visual scene (see [54] for a 1-D real-time implementation in Analog-VLSI).

After normalization, the feature maps for intensity, color, and orientation are summed across scales into three separate "conspicuity maps", one for intensity, one for color and one for orientation (Fig. 1b). Each conspicuity map is then subjected to another 10 iterations of **Eq. 2**. The motivation for the creation of three separate channels and their individual normalization is the hypothesis that similar features compete strongly for salience, while different modalities contribute independently to the saliency map. Although we are not aware of any supporting experimental evidence for this hypothesis, this additional step has the computational advantage of further enforcing that only a spatially sparse distribution of strong activity peaks is present within each visual feature type, before combination of all three types into the scalar saliency map.

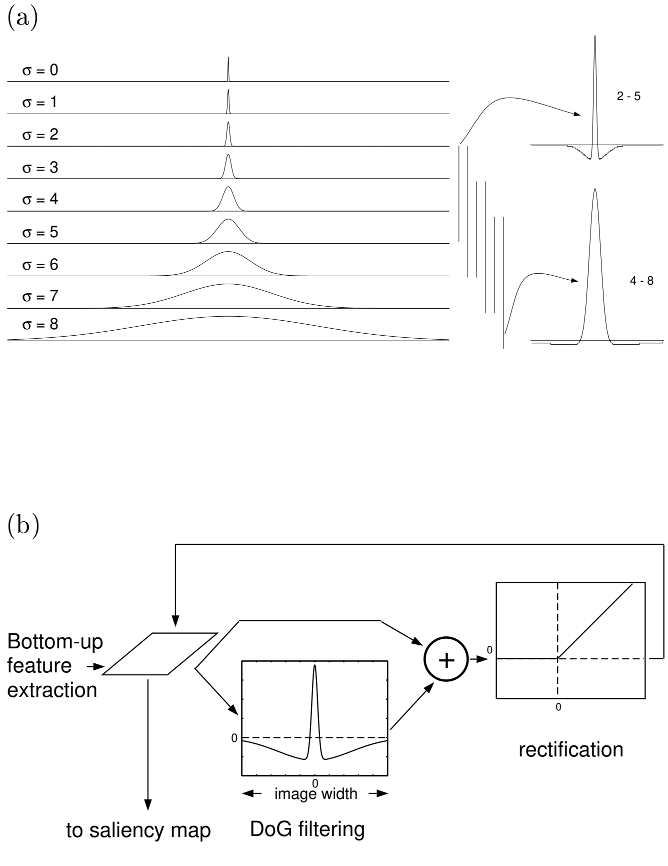


Figure 2: **(a)** Gaussian pixel widths for the 9 scales used in the model. Scale $\sigma = 0$ corresponds to the original image, and each subsequent scale is coarser by a factor 2. Two examples of the six center-surround receptive field types are shown, for scale pairs 2-5 and 4-8. **(b)** Illustration of the spatial competition for salience implemented within each of the 42 feature maps. Each map receives input from the linear filtering and center-surround stages. At each step of the process, the convolution of the map by a large Difference-of-Gaussians (DoG) kernel is added to the current contents of the map. This additional input coarsely models short-range excitatory processes and long-range inhibitory interactions between neighboring visual locations. The map is half-wave rectified, such that negative values are eliminated, hence making the iterative process non-linear. Ten iterations of the process are carried out before the output of each feature mapped is used in building the saliency map.

2.3 The Saliency Map

After the within-feature competitive process has taken place in each conspicuity map, these maps are linearly summed into the unique saliency map, which resides at scale 4 (reduction factor 1:16 compared to the original image). At any given time, the maximum of the saliency map corresponds to the most salient stimulus to which

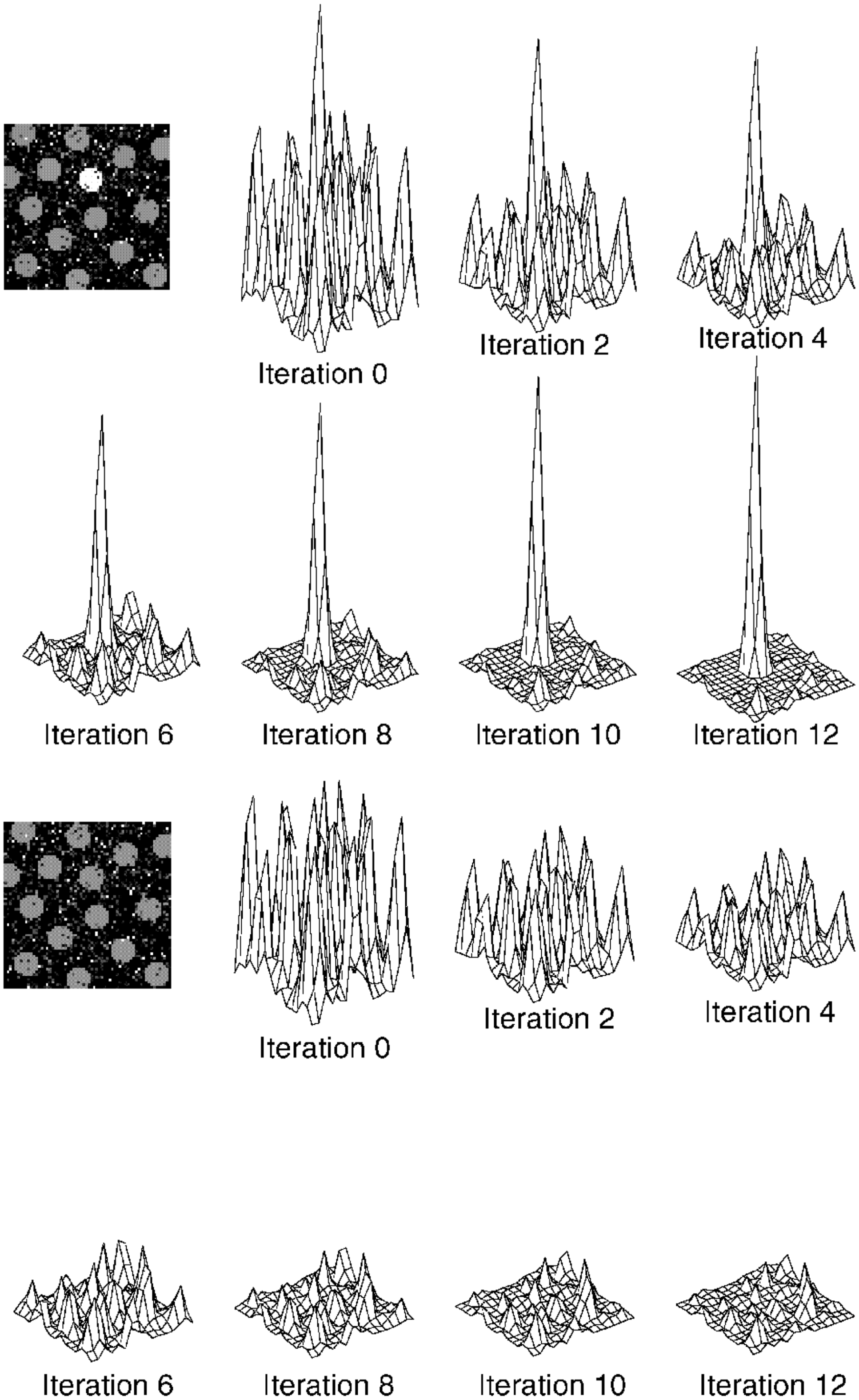


Figure 3: **(a)** Iterative spatial competition for salience in a single feature map with one strongly activated location surrounded by several weaker ones. After a few iterations, the initial maximum has gained further strength while at the same time suppressing weaker activation regions. **(b)** Iterative spatial competition for salience in a single feature map containing numerous strongly activated locations. All peaks inhibit each other more-or-less equally, resulting in the entire map being suppressed.

the focus of attention should be directed next, in order to allow for more detailed inspection by neurons along the occipito-temporal pathway. To find the most salient location, we have to determine the maximum of the saliency map.

This maximum is selected by application of a winner-take-all algorithm. Different mechanisms have been suggested for the implementation of neural winner-take-all networks [21, 55]; in particular see [56] for a multi-scale version of the winner-take-all network. In our model, we

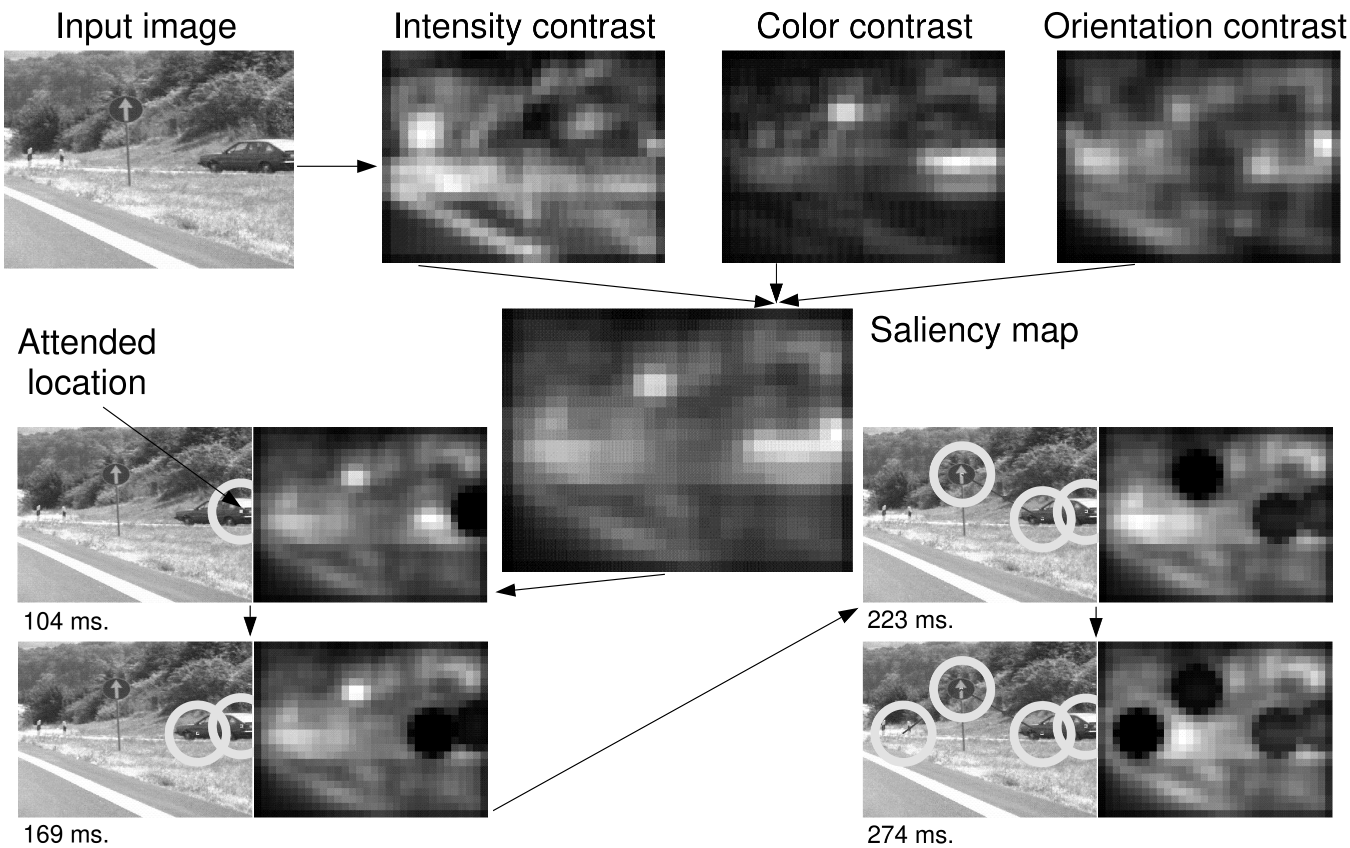


Figure 4: Example of the working of our model with a 512 × 384 pixels color image. Feature maps are extracted from the input image at several spatial scales, and are combined into three separate conspicuity maps (Intensity, Color and Orientation; see Fig. 1b) at scale 4 (32 × 24 pixels). The three conspicuity maps that encode for saliency within these three domains are combined and fed into the single saliency map (also 32 × 24 pixels). A neural winner-take-all network then successively selects, in order of decreasing saliency, the attended locations. Once a location has been attended to for some brief interval, it is transiently suppressed in the saliency map by the inhibition of return mechanism (dark round areas). Note how the inhibited locations recover over time (e.g., the first attended location has regained some activity at 274 ms), due to the integrative properties of the saliency map. The radius of the focus of attention was 64 pixels.

used a two dimensional layer of integrate-and-fire neurons with strong global inhibition in which the inhibitory population is reliably activated by any neuron in the layer (a more realistic implementation would consist of populations of neurons; for simplicity, we model such populations by a single neuron with very strong synapses). When the first of these integrate-and-fire cells fires (winner), it will generate a sequence of action potentials, causing the focus of attention (FOA) to shift to the winning location. These action potentials will also activate the inhibitory population, which in turn inhibits all cells in the layer, hence resetting the network to its initial state.

In the absence of any further control mechanism, the system described so far would direct its focus of attention, in the case of a static scene, constantly to one location, since the same winner would always be selected. To avoid this undesirable behavior, we follow Koch and Ullman [21] and introduce inhibitory feedback from the winner-take-all (WTA) array to the saliency map. When a spike occurs in the WTA network, the integrators in the saliency map transiently receive additional input with the spatial structure of a difference of Gaussians. The inhibitory center (with a standard deviation of half the radius of the FOA) is at the location of the winner; it and its neighbors become inhibited in the saliency map. As a consequence, attention switches to the next-most conspicuous location (Fig. 4). Such an "inhibition of return" has been well demonstrated for covert attentional shifts in humans [57, 58]. There is much less evidence for inhibition-of-return for eye movements in either humans or trained monkeys [59].

The function of the excitatory lobes (half width of four times the radius of the FOA) is to favor locality in the displacements of the focus of attention: If two locations are of nearly equal conspicuity, the one closest to the previous focus of attention will be attended next. This implementation detail directly follows the idea of "proximity preference" proposed by Koch and Ullman [21].

The time constants, conductances, and firing thresholds of the simulated neurons are chosen so that the FOA jumps from one salient location to the next in approximately 30–70 ms (simulated time; [60]), and so that an attended area is inhibited for approximately 500–900 ms (see Fig. 4). These delays vary for different locations with the strength of the saliency map input at those locations. The FOA therefore may eventually return to previously attended locations, as it is observed psychophysically. These simulated time scales are related to the dynamical model of integrate-and-fire neurons used in our model (see `http://www.klab.caltech.edu/~itti/` for the implementation source code, which clearly specifies all parameters of the simulated neurons using SI units).

3 Results

We tested our model on a wide variety of real images, ranging from natural outdoor scenes to artistic paintings. All images were in color, contained significant amounts of noise, strong local variations in illumination, shadows and reflections, large numbers of "objects" often partially occluded, and strong textures. Most of these images can be interactively examined on the World-Wide-Web, at `http://www.klab.caltech.edu/~itti/attention/`. Overall, the results indicate that the system scans the image in an order which makes functional sense in most behavioral situations.

It should be noted however that it is not straightforward to establish objective criteria for the performance of the system with such images. Unfortunately, nearly all quantitative psychophysical data on attentional control are based on synthetic stimuli similar to those discussed in the next section. In addition, although the scan paths of overt attention (eye movements) have been extensively studied [61, 62], it is unclear to what extent the precise trajectories followed by the attentional spotlight are similar to the motion of covert attention. Most probably, the requirements and limitations (e.g., spatial and temporal resolutions) of the two systems are related but not identical [56, 63]. Although our model is mostly concerned with shifts of covert attention, and ignores all of the mechanistic details of eye movements, we attempt below a comparison between human and model target search times in complex natural scenes, using a database of images containing military vehicles hidden in a rural environment.

3.1 Pop-Out and Conjunctive Search

A first comparison of the model with humans can be made using the type of displays used in "Visual Search" tasks [11]. A typical experiment consists of a speeded alternative forced-choice task in which the presence of a certain item in the presented display has to be either confirmed or denied. It is known that stimuli which differ from nearby stimuli in a single feature dimension can be easily found in visual search, typically in a time which is nearly independent of the number of other items ("distractors") in the visual scene. In contrast, search times for targets which differ from distractors by a combination of features (a so-called "conjunctive task") are typically proportional to the number of distractors [9].

We generated three classes of synthetic images to simulate such experiments: (1) one red target (rectangular bar) among green distractors (also rectangular bars) with the same orientation; (2) one red target among red distractors with orthogonal orientation; and (3) one red target among green distractors with the same orientation, and red distractors with orthogonal orientation. In order not to artifactually favor any particular orientation, the orientation of the target was chosen randomly for every image generated. Also, in order not to obtain ceiling performance in the first two tasks, we added strong orientation noise to the stimuli (between -17 and +17 degrees with uniform probability) and strong color speckle noise to the entire image (each pixel in the image had a 15% uniform probability to become a maximally bright color among red, green, blue, cyan, purple, yellow and white). The positioning of the stimuli along a uniform grid was randomized (by up to ± 40% of the spacing between stimuli, in the horizontal and vertical directions), to eliminate any possible influence of our discrete image representations (pixels) on the system. Twenty images were computed for a total number of bars per image varying between 4 and 36, yielding the evaluation of a total of 540 images. In each case, the task of our model was to locate the target, whose coordinates were externally known from the image generation process, at which point the search was terminated. We are here not concerned with the actual object recognition problem within the focus of attention. The diameter of the FOA was fixed to slightly more than the longest dimension of the bars.

Results are presented in **Fig. 5** in terms of the number of false detections before the target was found. Clear pop-out was obtained for the first two tasks (color only and orientation only), independently of the number of distractors in the images. Slightly worse performance is found when the number of distractors is very small, which seems sensible since in these cases the distractors are nearly as salient as the target itself. Evaluation of these types of images without introducing any of the distracting noises described above yielded systematic pop-out (target found as the first attended location) in all images. The conjunctive search task showed that the number of shifts of the focus of attention prior to the detection of the target increased linearly with the number of distractors. Notice that the large error bars in our results indicate that our model usually finds the target either quickly (in most cases) or only after scanning a large number of locations.

3.2 Search Performance in Complex Natural Scenes

We propose a second test in which target detection is evaluated using a database of complex natural images, each containing a military vehicle (the "target"). Contrary to our previous study with a simplified version of the model [23], which used low-resolution image databases with relatively large targets (typically about 1/10th the width of the visual scene), this study uses very-high resolution images (6144×4096 pixels), in which targets appear very small (typically 1/100th the width of the image). In addition, in the present study, search time is compared between the model's predictions and the average measured search times from 62 normal human observers [7].

The 44 original photographs were taken during a DISSTAF (Distributed Interactive Simulation, Search and Target Acquisition Fidelity) field test in Fort Hunter Liggett, California, and were provided to us, along with all human data, by the TNO Human Factors Research In-

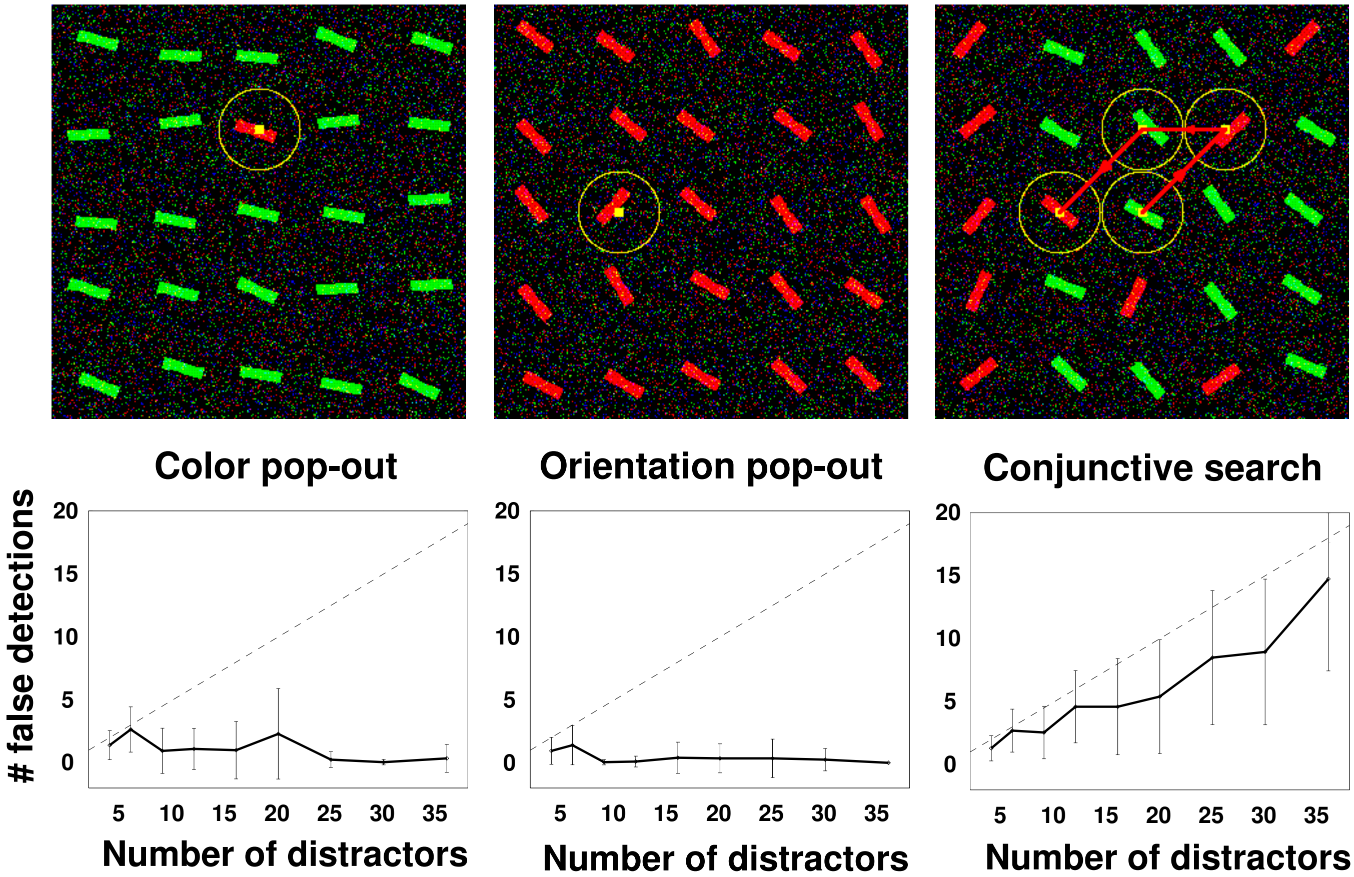


Figure 5: Model performance on noisy versions of pop-out and conjunctive tasks of the type pioneered by Treisman and Gelade [9]. Dark and light bars in the figure represent isoluminant red and green colored bars in the original displays with strong speckle noise added. Dashed lines: chance value, based on the size of the simulated visual field and the size of the candidate recognition area (corresponds to the performance of an ideal observer who scans, on average, half of the distractors prior to target detection). Solid lines: performance of the model. Error bars: one standard deviation. The typical search slopes of human observers in feature search and conjunction search, respectively, are successfully reproduced by the model. Each stimulus was drawn inside a 64×64 pixels box, and the radius of the focus of attention was fixed to 32 pixels. For a fixed number of stimuli, we tested twenty randomly generated images in each task; the saliency map and winner-take-all were initialized to zero (corresponding to a uniformly black visual input) prior to each trial.

stitute in the Netherlands [7]. The field of view for each image is $6.9 \times 4.6°$. Each scene contained one of nine possible military vehicles, at a distance ranging from 860 to 5822 meters from the observer. Each slide was digitized at 6144×4096 pixels resolution. Sixty two human observers aged between 18 and 45 years and with visual acuity better than 1.25 arcmin^{-1} participated to the experiment (about half were women and half men). Subjects were first presented with 3 close-up views of each of the 9 possible target vehicles, followed by a test run of 10 trials. A Latin square design [64] was then used for the randomized presentation of the images. The slides were projected such that they subtended $65 \times 46°$ visual angle to the observers (corresponding to a linear magnification by about a factor 10 compared to the original scenery). During each trial, observers pressed a button as soon as they had detected the target, and subsequently indicated at which location on a 10×10 projected grid they had found the target. Further details on these experiments can be found in [7, 65].

The model was presented with each image at full resolution. Contrary to the human experiment, no close-ups or test trials were presented to the model. The most generic form of the model described above was used, without any specific parameter adjustment for this experiment. Simulations for up to 10,000 ms of simulated time (about 200-400 attentional shifts) were done on a Digital Equipment Alpha 500 workstation. With these high-resolution images, the model comprised about 300 million simulated neurons. Each image was processed in about 15 minutes with a peak memory usage of 484 megabytes (for comparison, a 640×480 scene was typically processed in 10 seconds, and processing time approximately scaled linearly with the number of pixels). The focus of attention (FOA) was represented by a disk of radius 340 pixels **(Figs. 6 and 7)**. Full coverage of the image by the FOA would hence require 123 shifts (with overlap); a random search would thus be expected to find the target after 61.5 shifts on average. The target was considered detected when the focus of attention intersected a binary mask representing the outline of the target, which was provided with the images. Two examples of scenes and model trajectories are presented in **Figs. 6 and 7**; In the first image, the target was immediately found by the model, while, in the

Figure 6: Example of image from the database of 44 scenes depicting a military vehicle in a rural background. The algorithm operated on 24-bit color versions of these 6144×4096 pixel images (here only reproduced in grey-scale) and took on the order of 15 min real time on Dec Alpha workstation to carry out the saliency computation. **(a)** Original image; humans found the location of the vehicle in 2.6 sec. on average. **(b)** The vehicle was determined to be the most salient object in the image, and was attended first by the model. Such a result indicates strong performance of the algorithm in terms of artificial vision using complex natural color scenes. After scaling of the model's simulated time such that it scans two to four locations per second on average, and adding an 1.5 sec. period to account for the human's latency in motor response, the model found the target in 2.2 sec.

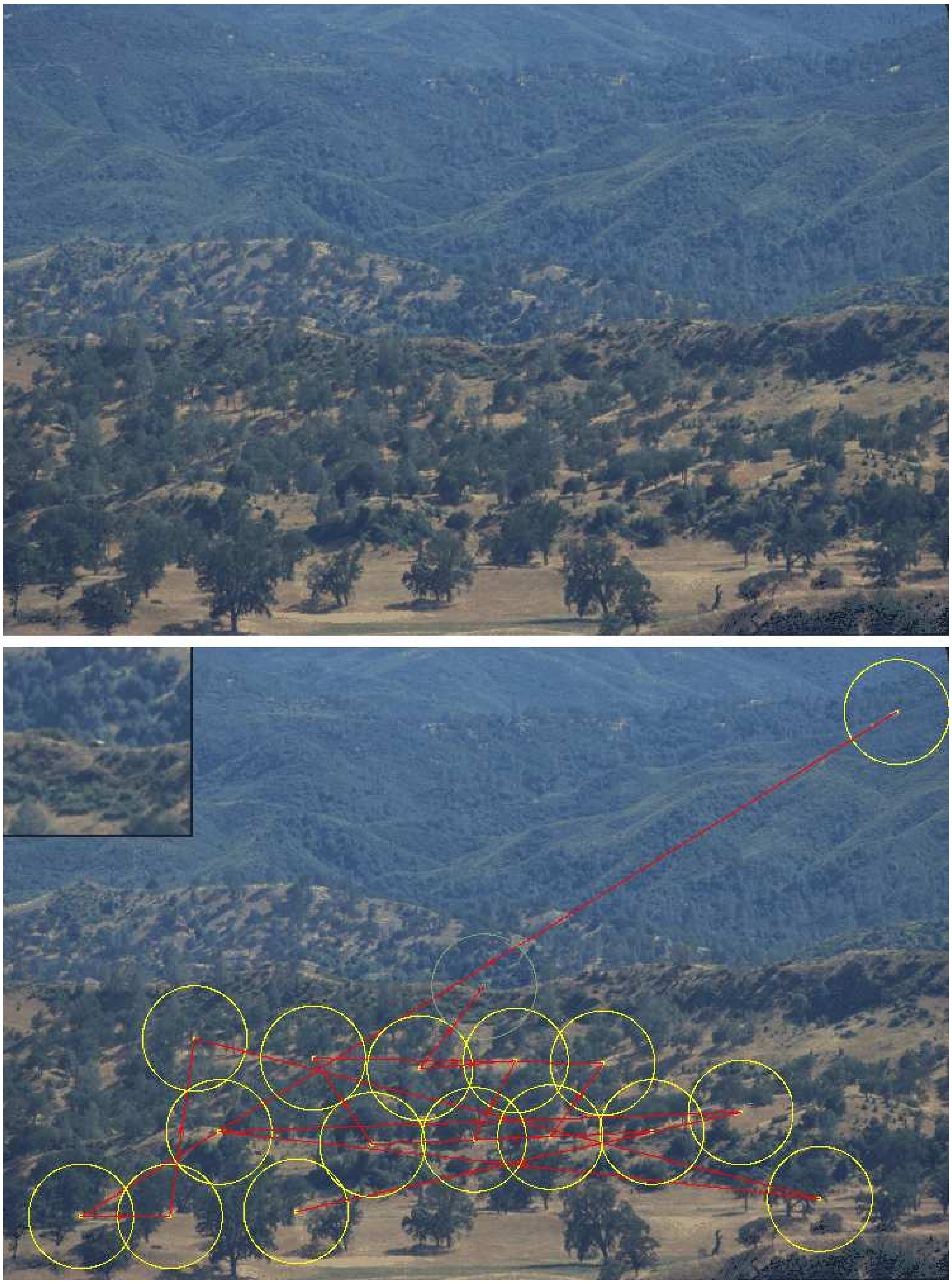


Figure 7: A more difficult example from the image database studied. **(a)** A grey-scale rendition of the color image. Humans found the location of the vehicle in 7.5 sec. on average. **(b)** The target is not the most salient object, and the model searches the scene in order of decreasing saliency. The algorithm came to rest on the location of the target on the 17^{th} shift, after 6.1 sec (using same time scaling as in the previous figure).

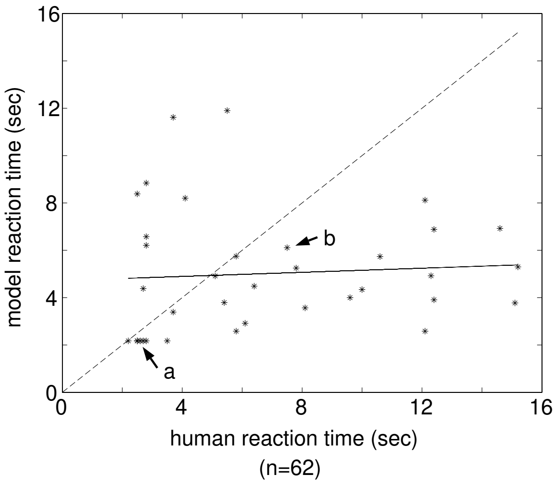


Figure 8: Mean reaction time to detect the target for 62 human observers and for our deterministic algorithm. Eight of the 44 original images are not included, in which either the model or the humans failed to reliably find the target. For the 36 images studied, and using the same scaling of model time as in the previous two figures, the model was faster than humans in 75% of the images. In order to bring this performance down to 50% (equal performance for humans and model), one would have to assume that no more than two locations can be visited by the algorithm each second. Arrow **(a)** indicates the "pop-out" example of **Fig. 6**, and arrow **(b)** the more difficult example presented in **Fig. 7**.

second, a serial search was necessary before the target could be found.

The model immediately found the target (first attended location) in seven of the 44 images. It quickly found the target (fewer than 20 shifts) in another 23 images. It found the target after more than 20 shifts in 11 images, and failed to find the target in 3 images. Overall, the model consequently performed surprisingly well, with a number of attentional shifts far below the expected 61.5 shifts of a random search in all but 6 images. In these 6 images, the target was extremely small (and hence not conspicuous at all), and the model cycled through a number of more salient locations.

The following analysis was performed to generate the plot presented in **Fig. 8**: First, a few outlier images were discarded, when either the model did not find the target within 2000 ms of simulated time (about 40-80 shifts; 6 images), or when half or more of the humans failed to find the target (3 images), for a total of 8 discarded images. An average of 3 overt shifts per second was assumed for the model, hence allowing us to scale the model's simulated time to real time. An additional 1.5 second was then added to the model time to account for human motor response time. With such calibration, the fastest reaction times for both model and humans were approximately 2 seconds, and the slowest approximately 15 seconds, for the 36 images analyzed.

The results plotted in **Fig. 8** overall show a poor correlation between human and model search times. Surprisingly however, the model appeared to find the target faster than humans in 3/4 of the images (points below the diagonal), despite the rather conservative scaling factors used to compare model to human time. In order to make the model's performance equal (on average) to that of humans, one would have to assume that humans shifted their gaze not faster than twice per second, which seems unrealistically slow under the circumstances of a speeded search task on a stationary, non-masked scene. Even if eye movements were that slow, most probably would humans still shift covert attention at a much faster rate between two overt fixations.

4 Discussion

We have demonstrated that a relatively simple processing scheme, based on some of the key organizational principles of pre-attentive early visual cortical architectures (center-surround receptive fields, non-classical within-feature inhibition, multiple maps) in conjunction with a single saliency map performs remarkably well at detecting salient targets in cluttered natural and artificial scenes.

Key properties of our model, in particular its usage of inhibition-of-return and the explicit coding of saliency independent of feature dimensions, as well as its behavior on some classical search tasks, are in good qualitative agreement with the human psychophysical literature.

It can be argued, based on the tentative scaling between simulated model time and human time described above (disregarding the fact that our computer implementation required on the order of 15 minutes to converge for the 6144×4096 pixel images versus search times on the order of a 2-20 seconds for human observers, and disregarding the fact that our algorithm did not deal with the problem of detecting the target in the focus of attention), that the bottom-up saliency-based algorithm outperforms humans in a demanding but realistic target detection task involving camouflaged military vehicles.

One paradoxical explanation for this superior performance might be that top-down influences play a significant role in the deployment of attention in natural scenes. Top-down cues in humans might indeed bias the attentional shifts, according to the progressively constructed mental representation of the entire scene, in inappropriate ways. Our model lacks any high-level knowledge of the world and operates in a purely bottom-up manner.

This does suggest that for certain (possibly limited) scenarios, such high-level knowledge might interfere with optimal performance. For instance, human observers are frequently tempted to follow roads or other structures, or may "consciously" decide to thoroughly examine the surroundings of salient buildings that have popped-out, while the vehicle might be in the middle of a field or in a forest.

4.1 Computational Implications

The main difficulty we encountered was that of combining information from numerous feature maps into a unique scalar saliency map. Most of the results described above do not hold for intuitively simple feature combination schemes, such as straight summation. In particular, straight summation fails to reliably detect pop-outs in search arrays such as those shown in **Fig. 5**. The reason for this failure is that almost all feature maps contain numerous strong responses (e.g., the intensity maps show strong activity at all target and distractor elements, because of their high contrast with the black background); the target consequently has a very low signal-to-noise ratio when all maps are simply summed. Here, we proposed a novel solution, which finds direct support in the human and animal studies of non-classical receptive-field interactions.

The first computational implication of our model is that a simple, purely bottom-up mechanism performs surprisingly well on real data in the absence of task-dependent feedback. This is in direct contrast to some of the previous models of visual search, in which top-down bias was almost entirely responsible for the relative weighting between the feature types used [22].

Further, although we have implemented the early feature extraction mechanisms in a comparatively crude manner (e.g., by approximating center-surround receptive fields by simple pixel differences between a coarse and a fine scale versions of the image), the model demonstrates a surprising level of robustness, which allows it to perform in a realistic manner on many complex natural images. We have previously studied the robustness of a pop-out signal in the presence of various amounts of added speckle noise (using a far less elaborate and biologically implausible approximation of our non-classical interactions), and have found that the model is almost entirely insensitive to noise as long as such noise is not directly masking the main feature of the target in spatial frequency or chromatic frequency space [23]. We believe that such robustness is another consequence of the within-feature iterative scheme which we use to allow for the fusion of information from several dissimilar sources.

That our model yields robust performance on natural scenes is not too surprising when considering the evidence from a number of state-of-the-art object recognition algorithms [66, 67, 68, 69]. Many of these demonstrate superior performance when compared to classical image processing schemes, although these new algorithms are based on very simple feature detection filters, similar to the ones found in biological systems.

4.2 Neurobiological Implications

While our model reproduces certain aspects of human search performance in a qualitative fashion, a more quantitative comparison is premature for several reasons.

Firstly, we have yet to incorporate a number of known features. For instance, we did not include any measure of saliency based on temporal stimulus onset or disappearance, or on motion [70]. We also have not yet integrated any retinal non-uniform sampling of the input images, although this is likely to strongly alter the saliency of peripherally-viewed targets. Nor have we addressed the well-known asymmetries in search tasks [71]. When targets and non-targets in a visual search tasks are exchanged, visual search performance often changes too (e.g., it is easier to search for a curved line among straight distractors than for a straight line among curved distractors). Spatial "grouping" acting among stimuli is also known to dramatically affect search time performance [72] and has not been dealt with here. In principle, this can be addressed by incorporating excitatory, cooperative center-surround interactions among neurons both within and across feature maps. And, as discussed above, our model is completely oblivious to any high-level features in natural scenes, including social cues.

More importantly, a number of electrophysiological findings muddy the simple architecture our model operates under **(Fig. 1b)**. Single-unit recordings in the visual system of the macaque indicate the existence of a number of distinct maps of the visual environment that appear to encode the saliency and/or the behavioral significance of targets. These include neurons in the superior colliculus, the inferior and lateral subdivisions of the pulvinar, the frontal-eye fields and areas within the intraparietal sulcus [73, 26, 74, 28, 29]. What remains unclear is whether these different maps emphasize saliency for different behaviors or for different visuo-motor response patterns (for instance, for attentional shifts, eye or hand movements). If saliency is indeed encoded across multiple maps, this raises the question of how competition can act across these maps to ensure that only a single location is chosen as the next target of an attentional or eye shift.

Following Koch and Ullman's [21] original proposal that visual search is guided by the output of a selection mechanism operating on a saliency map, it now seems plausible that such a process does characterize processing in the entire visual system. Inhibition-of-return (IOR) is a critical component of such search strategy, which essentially acts as memory. If its duration is reduced, the algorithm fails to find less salient objects because it endlessly cycles through the same number of more salient objects. For instance, if the time scale of IOR was reduced from 900 ms to 50 ms, the model would detect the most salient object, inhibit its location, then shift to the second most salient location, but it would subsequently come back to the most salient object, whose inhibition would have ceased during the attentional shift from first to second object. Under such conditions, the algorithm would never focus on anything else than the two most salient locations in the image. Our finding that IOR plays a critical role in purely bottom-up search may not necessarily disagree with recently suggested evidence that humans appear to use little or no memory during search [75]; while these authors do not refute the existence of IOR, a precise understanding of how bottom-up and top-down aspects of attention interact in human visual search remains to be elucidated.

Whether or not this implies that saliency is expressed explicitly in one or more visual field maps remains an open question. If saliency is encoded (relatively) independently of stimulus dimensions, we might be able to achieve a dissociation between stimulus attributes and stimulus saliency. For instance, appropriate visual masks might prevent the attributes of a visual stimulus to be read out without affecting its saliency. Or we might be able to directly influence such maps, for instance using reversible pharmacological techniques in animals or transcranial magnetic stimulations in human volunteers (TMS)?

Alternatively, it is possible that stimulus saliency is not expressed independently of feature dimensions but is encoded implicitly within each specific feature map as proposed by Desimone and Duncan [30]. This raises the question of how interactions among all of these maps gives rise to the observed behavior of the system for natural scenes. Such an alternative has not yet been analyzed in depth by computational work (see, however, [31]).

Mounting psychophysical, electrophysiological, clinical and functional imaging evidence [76, 77, 78, 74, 79, 29] strongly implies that the neuronal structures underlying the selection and the expression of shifts in spatial attention and occulomotor processing are tightly linked. These areas include the deeper parts of the superior colliculus; parts of the pulvinar; the frontal eye fields in the macaque and its homologue in humans, the precentral gyrus; and areas in the intraparietal sulcus in the macaque and around the intraparietal and postcentral sulci and adjacent gyri in humans.

The close relationship between areas active during covert and during overt shifts of attention raises the issue of how information in these maps is integrated across saccades, in particular given the usage of both retinal and occulo-motor coordinate systems in the different neuronal maps (see, for instance, [80]). This is an obvious question that will be explored by us in future computational work.

Finally, we can now wonder about the relationship between the saliency mechanism, the top-down volitional attentional selection process, and awareness. We have recently proposed a quantitative account of the action of spatial attention on various psychophysical thresholds for pattern discrimination, in terms of a strengthening of cooperative and competitive interactions among early visual filters [81]. How can such a scheme be combined with the current selection process based on purely bottom-up sensory data? Several possibilities come to mind. First, both processes might operate independently and both mediate access to visual awareness. Computationally, this can be implemented in a straightforward manner. Second however, top-down attention might also directly interact with the single saliency map, for instance by influencing its constitutive elements via appropriate synaptic input. If the inhibition-of-return could be selectively inactivated at locations selected under volitional control, for example by shunting [82], then the winner-take-all and the attentional focus would remain at that location, ignoring for a while surrounding salient objects. Although such feedback to the saliency map seems plausible and is functionally useful, it certainly does not constitute all of the top-down attentional modulation of spatial vision [81]. Finally, independent saliency maps could operate for the different feature maps and both saliency and volitional forms of attention could access them independently. Current experimental evidence does not allow us to unambiguously choose among these possibilities.

Acknowledgements: We thank Dr. Toet from the TNO Human Factors Research Institute, The Netherlands, for providing us with the database of military images and human search times on these images. This research was supported by NSF-ERC, NIMH and ONR.

References

[1] Heisenberg, M. & Wolf, R. *Studies of Brain Function, Vol. 12: Vision in Drosophila* (Springer-Verlag, Berlin, Germany, 1984).

[2] Rensink, J. H., Vanderven, J., Vanpamelen, G., Fedder, F. & Majoor, E. The modified renphosystem - a high biological nutrient removal system. *Water Science and Technology* **35**, 137–46 (1997).

[3] O'Regan, J. K., Rensink, R. A. & Clark, J. J. Change-blindness as a result of 'mudsplashes'. *Nature* **398**, 34 (1999).

[4] Simons, D. J. & Levin, D. T. Failure to detect changes to attended objects. *Investigative Ophthalmology & Visual Science* **38**, 3273 (1997).

[5] Crick, F. & Koch, C. Constraints on cortical and thalamic projections: the no-strong-loops hypothesis. *Nature* **391**, 245–50 (1998).

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

[7] Toet, A., Bijl, P., Kooi, F. L. & Valeton, J. M. *A high-resolution image dataset for testing search and detection models (TNO-TM-98-A020)* (TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1998).

[8] James, W. *The Principles of Psychology* (Harvard UP, Cambridge, MA, 1890/1981).

[9] Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cognit Psychol* **12**, 97–136 (1980).

[10] Bergen, J. R. & Julesz, B. Parallel versus serial processing in rapid pattern discrimination. *Nature* **303**, 696–8 (1983).

[11] Treisman, A. Features and objects: the fourteenth bartlett memorial lecture. *Q J Exp Psychol [A]* **40**, 201–37 (1988).

[12] Nakayama, K. & Mackeben, M. Sustained and transient components of focal visual attention. *Vision Res* **29**, 1631–47 (1989).

[13] Braun, J. & Sagi, D. Vision outside the focus of attention. *Percept Psychophys* **48**, 45–58 (1990).

[14] Hikosaka, O., Miyauchi, S. & Shimojo, S. Orienting a spatial attention–its reflexive, compensatory, and voluntary mechanisms. *Brain Res Cogn Brain Res* **5**, 1–9 (1996).

[15] Braun, J. & Julesz, B. Withdrawing attention at little or no cost: detection and discrimination tasks. *Percept Psychophys* **60**, 1–23 (1998).

[16] Braun, J. Vision and attention: the role of training [letter; comment]. *Nature* **393**, 424–5 (1998). Comment on: Nature 1997 Jun 19;387(6635):805-7.

[17] Gallant, J. L., Connor, C. E. & Essen, D. C. V. Neural activity in areas v1, v2 and v4 during free viewing of natural scenes compared to controlled viewing. *Neuroreport* **9**, 85–90 (1998).

[18] Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J. & Davis, J. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature* **378**, 492–6 (1995).

[19] Sillito, A. M. & Jones, H. E. Context-dependent interactions and visual processing in v1. *J Physiol Paris* **90**, 205–9 (1996).

[20] Levitt, J. B. & Lund, J. S. Contrast dependence of contextual effects in primate visual cortex. *Nature* **387**, 73–6 (1997).

[21] Koch, C. & Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol* **4**, 219–27 (1985).

[22] Wolfe, J. M. Visual search in continuous, naturalistic stimuli. *Vision Res* **34**, 1187–95 (1994).

[23] Itti, L., Koch, C. & Niebur, E. A model of saliency-based visual-attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 1254–9 (1998).

[24] Niebur, E. & Koch, C. Control of selective visual attention: Modeling the 'where' pathway. In *Neural Information Processing Systems (NIPS*8)* (eds. Touretzky, D., Mozer, M. & Hasselmo, M.), 802–808 (MIT Press, Cambridge, MA, 1996).

[25] Olshausen, B. A., Anderson, C. H. & Van Essen, D. C. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J Neurosci* **13**, 4700–19 (1993).

[26] Robinson, D. L. & Petersen, S. E. The pulvinar and visual salience. *Trends Neurosci* **15**, 127–32 (1992).

[27] Rockland, K. S., Andresen, J., Cowie, R. J. & Robinson, D. L. Single axon analysis of pulvinocortical connections to several visual areas in the macaque. *J Comp Neurol* **406**, 221–50 (1999).

[28] Gottlieb, J. P., Kusunoki, M. & Goldberg, M. E. The representation of visual salience in monkey parietal cortex. *Nature* **391**, 481–4 (1998).

[29] Colby, C. L. & Goldberg, M. E. Space and attention in parietal cortex. *Annu Rev Neurosci* **22**, 319–49 (1999).

[30] Desimone, R. & Duncan, J. Neural mechanisms of selective visual attention. *Annu Rev Neurosci* **18**, 193–222 (1995).

[31] Hamker, F. H. The role of feedback connections in task-driven visual search. In *Connectionist Models in Cognitive Neuroscience, Proc. of the 5th Neural Computation and Psychology Workshop (NCPW'98)* (eds. von Heinke, D., Humphreys, G. W. & Olson, A.) (Springer Verlag, London, 1999).

[32] Beymer, D. & Poggio, T. Image representations for visual learning. *Science* **272**, 1905–9 (1996).

[33] Burt, P. & Adelson, E. The laplacian pyramid as a compact image code. *IEEE Trans. on Communications* **31**, 532–540 (1983).

[34] Leventhal, A. *The Neural Basis of Visual Function (Vision and Visual Dysfunction Vol. 4)* (CRC Press, Boca Raton, FL, 1991).

[35] Luschow, A. & Nothdurft, H. C. Pop-out of orientation but no pop-out of motion at isoluminance. *Vision Research* **33**, 91–104 (1993).

[36] Engel, S., Zhang, X. & Wandell, B. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature* **388**, 68–71 (1997).

[37] DeValois, R. L., Albrecht, D. G. & Thorell, L. G. Spatial-frequency selectivity of cells in macaque visual cortex. *Vision Research* **22**, 545–559 (1982).

[38] Tootell, R. B., Hamilton, S. L., Silverman, M. S. & Switkes, E. Functional anatomy of macaque striate cortex. i. ocular dominance, binocular interactions, and baseline conditions. *J Neurosci* **8**, 1500–30 (1988).

[39] Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S. & Anderson, C. H. Overcomplete steerable pyramid filters and rotation invariance. In *Proc. IEEE Computer Vision and Pattern Recognition (CVPR), Seattle, WA*, 222–228 (Jun 1994).

[40] Itti, L. & Koch, C. A comparison of feature combination strategies for saliency-based visual attention systems. In *SPIE Human Vision and Electronic Imaging IV (HVEI'99), San Jose, CA*, (in press) (1999).

[41] Rockland, K. S. & Lund, J. S. Intrinsic laminar lattice connections in primate visual cortex. *J Comp Neurol* **216**, 303–18 (1983).

[42] Gilbert, C. D., Das, A., Ito, M., Kapadia, M. & Westheimer, G. Spatial integration and cortical dynamics. *Proc Natl Acad Sci U S A* **93**, 615–22 (1996).

[43] Gilbert, C. D. & Wiesel, T. N. Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *J Neurosci* **9**, 2432–42 (1989).

[44] Gilbert, C. D. & Wiesel, T. N. Clustered intrinsic connections in cat visual cortex. *J Neurosci* **3**, 1116–33 (1983).

[45] Weliky, M., Kandler, K., Fitzpatrick, D. & Katz, L. C. Patterns of excitation and inhibition evoked by horizontal connections in visual cortex share a common relationship to orientation columns. *Neuron* **15**, 541–52 (1995).

[46] Polat, U. & Sagi, D. The architecture of perceptual spatial interactions. *Vision Res* **34**, 73–8 (1994).

[47] Polat, U. & Sagi, D. Spatial interactions in human vision: from near to far via experience-dependent cascades of connections. *Proc Natl Acad Sci U S A* **91**, 1206–9 (1994).

[48] Zenger, B. & Sagi, D. Isolating excitatory and inhibitory nonlinear spatial interactions involved in contrast detection. *Vision Res* **36**, 2497–513 (1996).

[49] Cannon, M. W. & Fullenkamp, S. C. Spatial interactions in apparent contrast: inhibitory effects among grating patterns of different spatial frequencies, spatial positions and orientations. *Vision Res* **31**, 1985–98 (1991).

[50] Knierim, J. J. & van Essen, D. C. Neuronal responses to static texture patterns in area v1 of the alert macaque monkey. *J Neurophysiol* **67**, 961–80 (1992).

[51] Ts'o, D. Y., Gilbert, C. D. & Wiesel, T. N. Relationships between horizontal interactions and functional architecture in cat striate cortex as revealed by cross-correlation analysis. *J Neurosci* **6**, 1160–70 (1986).

[52] Malach, R., Amir, Y., Harel, M. & Grinvald, A. Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex. *Proc Natl Acad Sci U S A* **90**, 10469–73 (1993).

[53] Malach, R. Cortical columns as devices for maximizing neuronal diversity. *Trends Neurosci* **17**, 101–4 (1994).

[54] Horiuchi, T., Morris, T., Koch, C. & DeWeerth, S. Analog vlsi circuits for attention-based, visual tracking. In *Neural Information Processing Systems (NIPS*9)* (eds. Mozer, M., Jordan, M. & Petsche, T.), 706–712 (MIT Press, Cambridge, MA, 1997).

[55] Yuille, A. L. & Grzywacz, N. M. A mathematical-analysis of the motion coherence theory. *International Journal of Computer Vision* **3**, 155–75 (1989).

[56] Tsotsos, J. K., Culhane, S. M., Wai, W. Y. K., Lai, Y. H., Davis, N. & Nuflo, F. Modeling visual-attention via selective tuning. *Artificial Intelligence* **78**, 507–45 (1995).

[57] Posner, M. I., Cohen, Y. & Rafal, R. D. Neural systems control of spatial orienting. *Philos Trans R Soc Lond B Biol Sci* **298**, 187–98 (1982).

[58] Kwak, H. W. & Egeth, H. Consequences of allocating attention to locations and to other attributes. *Percept Psychophys* **51**, 455–64 (1992).

[59] Motter, B. C. & Belky, E. J. The guidance of eye movements during active visual search. *Vision Res* **38**, 1805–15 (1998).

[60] Saarinen, J. & Julesz, B. The speed of attentional shifts in the visual field. *Proc Natl Acad Sci U S A* **88**, 1812–4 (1991).

[61] Yarbus, A. L. *Eye Movements and Vision* (Plenum Press, New York, 1967).

[62] Noton, D. & Stark, L. Scanpaths in eye movements during pattern perception. *Science* **171**, 308–11 (1971).

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

[64] Wagenaar, W. A. Note on the construction of digram-balanced latin squares. *Psychol Bull* **72**, 384–86 (1969).

[65] Bijl, P., Kooi, F. K. & van Dorresteijn, M. *Visual search performance for realistic target imagery from the DISSTAF field trials* (TNO Human Factors Research Institute, Soesterberg, The Netherlands, 1997).

[66] Poggio, T. Image representations for visual learning. *Lecture Notes in Computer Science* **1206**, 143 (1997).

[67] Niyogi, P., Girosi, F. & Poggio, T. Incorporating prior information in machine learning by creating virtual examples. *Proceedings of the IEEE* **86**, 2196–209 (1998).

[68] Malik, J. & Perona, P. Preattentive texture discrimination with early vision mechanisms. *J Opt Soc Am [A]* **7**, 923–32 (1990).

[69] Simoncelli, E. P., Freeman, W. T., Adelson, E. H. & Heeger, D. J. Shiftable multiscale transforms. *Ieee Transactions on Information Theory* **38**, 587–607 (1992).

[70] Hillstrom, A. P. & Yantis, S. Visual-motion and attentional capture. *Perception & Psychophysics* **55**, 399–411 (1994).

[71] Treisman, A. & Gormican, S. Feature analysis in early vision: evidence from search asymmetries. *Psychol Rev* **95**, 15–48 (1988).

[72] Driver, J., Mcleod, P. & Dienes, Z. Motion coherence and conjunction search - implications for guided search theory. *Perception & Psychophysics* **51**, 79–85 (1992).

[73] Laberge, D. & Buchsbaum, M. S. Positron emission tomographic measurements of pulvinar activity during an attention task. *Journal of Neuroscience* **10**, 613–9 (1990).

[74] Kustov, A. A. & Robinson, D. L. Shared neural control of attentional shifts and eye movements. *Nature* **384**, 74–7 (1996).

[75] Horowitz, T. S. & Wolfe, J. M. Visual search has no memory. *Nature* **394**, 575–7 (1998).

[76] Shepherd, M., Findlay, J. M. & Hockey, R. J. The relationship between eye movements and spatial attention. *Q J Exp Psychol* **38**, 475–491 (1986).

[77] Andersen, R. A., Bracewell, R. M., Barash, S., Gnadt, J. W. & Fogassi, L. Eye position effects on visual, memory, and saccade-related activity in areas lip and 7a of macaque. *J Neurosci* **10**, 1176–96 (1990).

[78] Sheliga, B. M., Riggio, L. & Rizzolatti, G. Orienting of attention and eye movements. *Exp Brain Res* **98**, 507–22 (1994).

[79] Corbetta, M. Frontoparietal cortical networks for directing attention and the eye to visual locations: identical, independent, or overlapping neural systems? *Proc Natl Acad Sci U S A* **95**, 831–8 (1998).

[80] Andersen, R. A. Multimodal integration for the representation of space in the posterior parietal cortex. *Philos Trans R Soc Lond B Biol Sci* **352**, 1421–8 (1997).

[81] Lee, D. K., Itti, L., Koch, C. & Braun, J. Attention activates winner-take-all competition among visual filters. *Nat Neurosci* **2**, 375–81 (1999).

[82] Koch, C. *Biophysics of Computation: Information Processing in Single Neurons* (Oxford University Press, Oxford, England, 1998).