Cerebral Cortex Advance Access published June 19, 2015

OXFORD

Cerebral Cortex, 2015, 1–19

doi: 10.1093/cercor/bhv129
Original Article

ORIGINAL ARTICLE

There's Waldo! A Normalization Model of Visual Search Predicts Single-Trial Human Fixations in an Object Search Task

Thomas Miconi[1,4], Laura Groomes[1] and Gabriel Kreiman[1,2,3]

[1]Children's Hospital, Harvard Medical School, Boston, MA, USA, [2]Center for Brain Science, [3]Swartz Center for Theoretical Neuroscience, Harvard University, Cambridge, MA, USA and [4]The Neurosciences Institute, La Jolla, CA 92037, USA

Address correspondence to Thomas Miconi, The Neurosciences Institute, 800 Silverado Street, La Jolla, CA 92037, USA. Email: miconi@nsi.edu

Abstract

When searching for an object in a scene, how does the brain decide where to look next? Visual search theories suggest the existence of a global "priority map" that integrates bottom-up visual information with top-down, target-specific signals. We propose a mechanistic model of visual search that is consistent with recent neurophysiological evidence, can localize targets in cluttered images, and predicts single-trial behavior in a search task. This model posits that a high-level retinotopic area selective for shape features receives global, target-specific modulation and implements local normalization through divisive inhibition. The normalization step is critical to prevent highly salient bottom-up features from monopolizing attention. The resulting activity pattern constitutes a priority map that tracks the correlation between local input and target features. The maximum of this priority map is selected as the locus of attention. The visual input is then spatially enhanced around the selected location, allowing object-selective visual areas to determine whether the target is present at this location. This model can localize objects both in array images and when objects are pasted in natural scenes. The model can also predict single-trial human fixations, including those in error and target-absent trials, in a search task involving complex objects.

Key words: computational modeling, normalization, object recognition, visual attention, visual search

Introduction

Searching for an object in a crowded scene constitutes a challenging task. Yet, we can detect target objects significantly faster than would be expected by random search, even in a complex scene (Wolfe, Alverez et al. 2011). How does the brain identify the locations that might contain a target object? An influential concept suggests that the brain computes one or more "priority maps," which allocate a certain attentional value to every point in the visual space (Itti and Koch 2000). A large body of evidence shows that this attentional selection involves the frontal eye field (FEF), the lateral intraparietal cortex (LIP), and sub-cortical structures such as the pulvinar and the superior colliculus (Reynolds and Heeger 2009; Noudoost et al. 2010; Bisley 2011). How these areas interact with those involved in shape recognition is poorly understood. To understand the interactions between bottom-up visual inputs and top-down task influences during visual search, we seek a model with 3 main characteristics: (i) computationally implemented so that it can perform search on images, (ii) consistent with state-of-the-art understanding of neural circuit function and cognitive science of visual search, and (iii) capable of capturing human behavioral performance during visual search.

In most models of visual search, the salience of an object is defined by the contrast between the object and its local surround along various features (Koch and Ullman 1985; Tsotsos et al. 1995; Itti and Koch 2000; Rao et al. 2002; Hamker 2005; Navalpakkam and Itti 2005, 2007; Walther and Koch 2007; Wolfe 2007; Chikkerur

et al. 2010). Additionally, search behavior is influenced by the characteristics of the sought target (Williams 1967; Findlay 1997; Eckstein et al. 2000; Rao et al. 2002; Beutter et al. 2003; Bichot et al. 2005; Najemnik and Geisler 2005; Navalpakkam and Itti 2007; Zelinsky 2008; Buschman and Miller 2009; Tavassoli et al. 2009; Carrasco 2011; Kowler 2011; Peelen and Kastner 2011; Tatler et al. 2011).

How do the target characteristics influence visual search? One possibility is that the brain could use the object recognition properties of the ventral pathway, culminating in the highly selective responses of inferotemporal (IT) areas, to compare each portion of space with the target's known appearance. However, the ventral pathway is relatively slow compared with the attentional system (Monosov et al. 2010). Furthermore, the selectivity of IT cells is strongly degraded by clutter in the absence of spatial attention (Desimone 1998; Sheinberg and Logothetis 2001; Zoccolan et al. 2007; Zhang et al. 2011) (but see Li et al. (2009); Agam et al. (2010)), making them a poor candidate to drive visual search in complex environments.

Many existing models of visual search propose that the brain selectively biases the activity of low-level visual feature detectors, increasing the gain of cells selective for the target's features (e.g., when looking for a red object, red-selective cells are up-modulated). As a result, on stimulus onset, regions in which the target's features are present elicit higher responses. This idea extends the theoretical proposal of Wolfe's "guided search" (Wolfe 2007) and forms the basis of many computational models of visual search. These models include Chikkerur et al.'s (2010) elegant derivation of attentional effects by applying Bayesian computations to image and target features (see also Yu and Dayan (2005); Vincent et al. (2009)), Navalpakkam and Itti's (2005, 2007) extension of Itti and Koch's (2000) bottom-up saliency computation by introducing object-specific low-level feature biasing, Lanyon and Denham's (2004) proposal based on modulation and biased competition in extra-striate cortex, and Hamker's (2005) detailed biologically motivated model of how attentional selection emerges from the reentrant interactions among various brain areas. In biologically motivated models, the modulated feature detectors tend to be associated with visual area V4, and the source of feature-specific modulation with the highly selective regions of inferotemporal cortex (IT) or prefrontal cortex (PFC), while the priority map itself is associated with areas controlling visual attention and eye movements (FEF and LIP) (Hamker 2005; Chikkerur et al. 2010).

An attractive feature of this proposal is that a similar phenomenon has indeed been observed in a different type of task, namely feature-based attention (attending to a particular feature or object rather than a location [Treue and Martinez-Trujillo 1999]). In these tasks, lower-level cells are indeed modulated according to how much their preferences match the target, rather than the stimulus in their receptive field (RF), and this observation has been described in a "feature-similarity gain" model (Martinez-Trujillo and Treue 2004). A similar effect has also been reported in human imaging and MEG data (O'Craven et al. 1999; Puri et al. 2009; Baldauf and Desimone 2014).

However, recent experimental evidence raises the possibility of an alternative explanation and flow of information processing during visual search. First, in visual search tasks, modulation specific for target similarity was observed in FEF before V4 (Zhou and Desimone 2011) or IT (Monosov et al. 2010), even though target-independent feature selectivity can be detected in V4 before FEF (Zhou and Desimone 2011). The earlier activation in FEF during search suggests that attention-controlling circuits may "find" visual areas that look like the target (and control the next attentional selection) before low-level detectors and high-level identifiers. Furthermore, some studies show that during visual search the target-selective modulation in V4 is actually spatial, rather than feature based: V4 cells seem to be modulated according to the similarity between the local stimulus and target, rather than based on their feature preference (Zhou and Desimone 2011) (see also [Martinez-Trujillo 2011]).

These observations suggest that similarity between local features and target features may be computed elsewhere, and then redistributed as "feature-guided spotlights" of spatial modulation down to V4: The brain determines which locations of the visual input "look like" the target and then applies a spatial modulation on the visual areas around these locations. This leaves open the question of how feature-guidance occurs, that is, how the priority map that tracks local similarity to target features is computed in the first place.

We propose a model for visual search that is motivated by these recent neurophysiological observations. Briefly, the proposed model suggests that the priority map is computed directly in a retinotopic, feature-selective area within the attentional system, tentatively ascribed to LIP/FEF (see Discussion). We show how feed-forward visual input, interacting with object-specific, top-down modulation and local divisive mutual inhibition (normalization) creates a priority map with higher activation around stimuli that are similar to the target. The maximum of this map is then selected as the next locus of attentional selection. Attentional selection (either covert or overt—see Discussion) spatially enhances a small portion of the visual input around this locus and de-emphasizes the rest. This enhancement allows the ventral visual system to determine whether the target is actually present at that location or not. If the target is not found at this location, the process (attentional selection followed by local recognition in the selected location) is iterated until the target is found or search is aborted. We show that the proposed model can locate target objects in complex images. Furthermore, we compare the model with human behavior on a common non-trivial visual search task involving complex objects. The model predicts human behavior during visual search, not only in overall performance, but also in single trials, including error and target-absent trials.

Materials and Methods

Ethics Statement

All the psychophysics experiments (described later) were conducted under the subjects' consent according to the protocols approved by the Institutional Review Board at Children's Hospital Boston.

Model Sketch

We first provide a high-level intuitive outline of our model (Fig. 1); later sections provide a full description of the implementation details.

We consider the problem of localizing a target object in a cluttered scene (e.g., Fig. 2*A–D*). The model posits that a priority map determines the locus of attentional selection, by conjugating local visual inputs with top-down target information (Fig. 1—"Priority map"). This map is computed by a high-level, retinotopic area, selective for complex shape features. The area in our model that performs the necessary computations has properties inspired by macaque LIP and FEFs; we therefore refer to this area as "LIP/FEF" in our description of the model (Fig. 1). Yet, we emphasize that this is a descriptive naming convention rather than a firm biological equivalence (see Discussion).

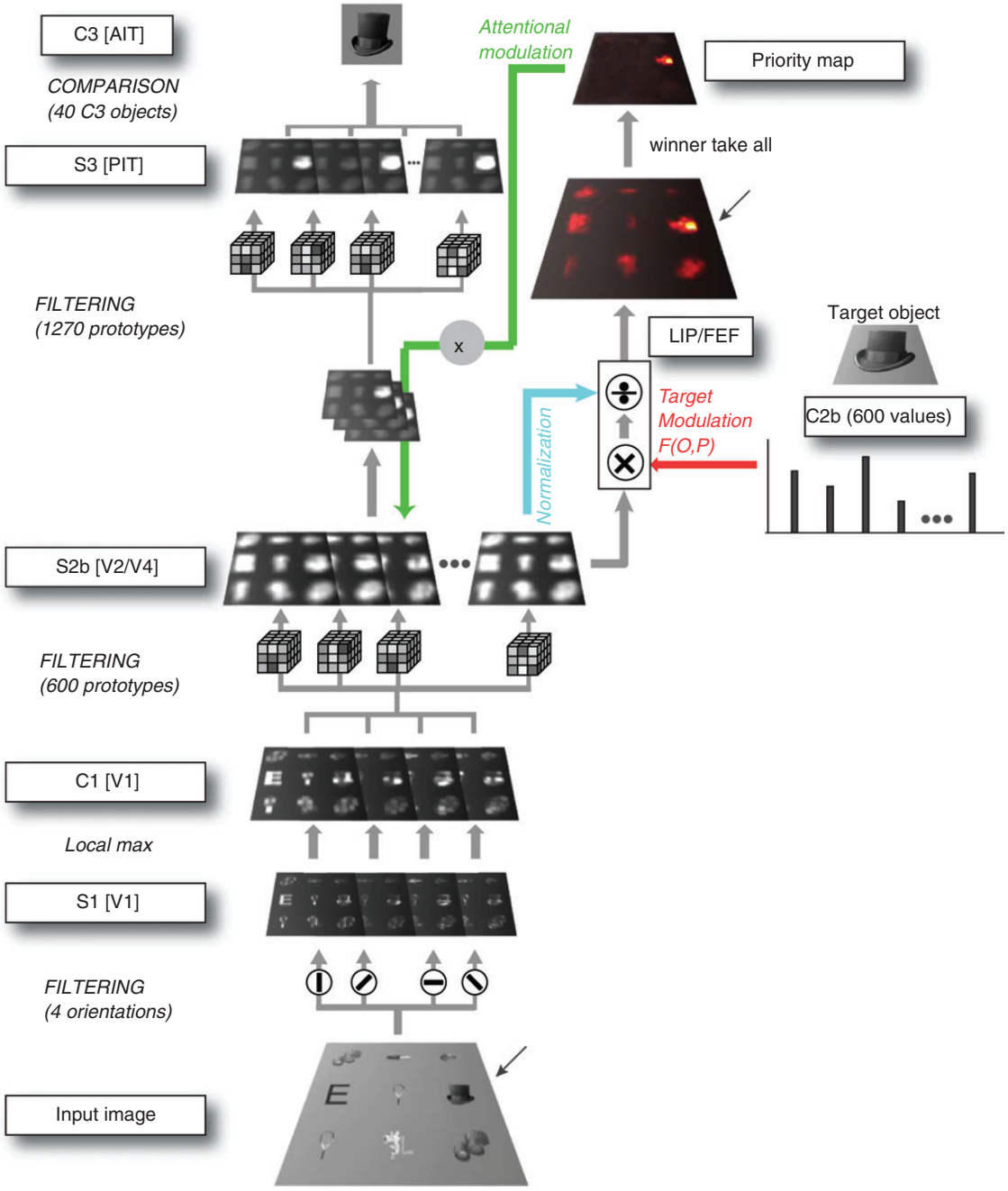

Figure 1. Sketch of the attentional model. When a target object **O** is presented (here a top hat, upper right), the model generates a set of attentional modulation coefficients $F(\mathbf{O}, P)$ (red arrow). During a search task, an image containing a target object among other distractors (bottom) is passed to a cascade of linear filters (Gabor functions [S1] and radial basis functions [S2b]) and non-linear filters (max operation, C1) (Serre et al. 2005; Kouh 2007) (Methods). The output of S2b cells in response to the image is modulated multiplicatively by the attentional signal $F(\mathbf{O}, P)$ and normalized by the total incoming S2b activity at each point (blue arrow). The total resulting activation at every point, summed across all scales and prototypes, constitutes the final priority map $A_O(x, y)$ (top), corresponding to the output of the attentional system for guidance of attentional selection. In this example, as expected, the priority map's maximum value (bright yellow) lies at the location of the target object within the input image. The priority map enhances the selected location at the S2b level (green arrow). The S2b signals are conveyed to a target presence validation step to identify objects at the selected location unhindered by clutter in the scene. In this example, the model identifies the target object at the selected location, successfully completing the search.

The key inputs to this priority map area are (i) bottom-up signals from retinotopic shape-selective cells in earlier visual areas and (ii) top-down signals that modulate the responses according to the identity of the target object. This top-down modulation involves target-specific multiplicative feedback on each cell in the area. This feedback input $F(\mathbf{O}, P)$ (where **O** is the target object and P is the cell's preferred feature) is proportional to how much feature P is present in the target object and is learnt by exposure to images of the object (eq. 4). In addition, the priority map undergoes local normalization through divisive feed-forward inhibition at every point (eq. 4), effectively normalizing the bottom-up inputs. The interaction between feed-forward visual input, top-down object-specific modulation, and local divisive normalization produces a priority map of the visual scene, in which aggregate activity at each location tracks the overall similarity of local visual input with target features.

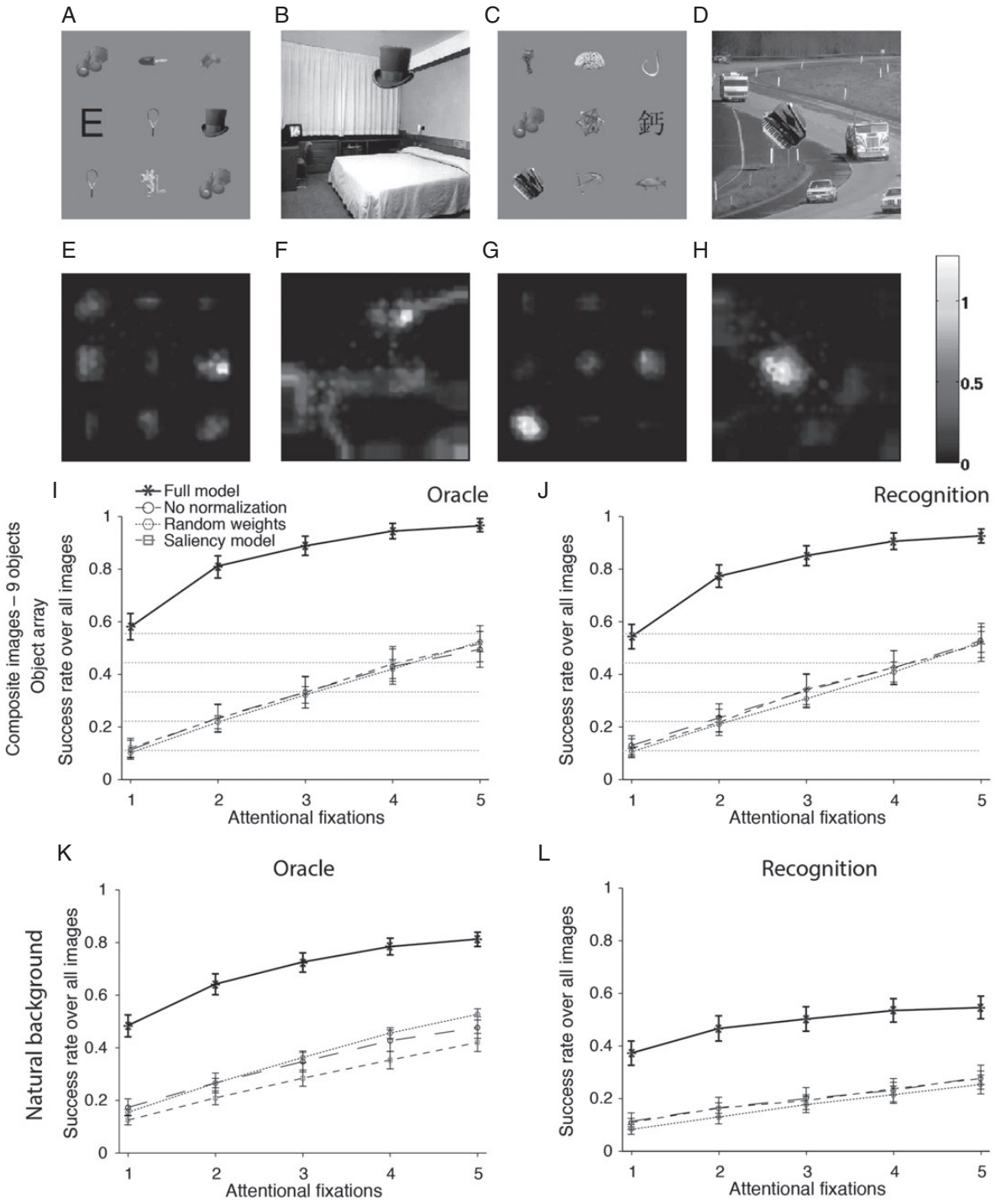

Figure 2. Model performance in object array images and natural-background images. Example object array images (*A*,*C*) and natural-background images (*B*,*D*) used for testing the model. The target object is a top hat in *A*,*B* and an accordion in *C*,*D*. (*E*–*H*) Output of the model for each of the images in *A*–*D*. We show the priority map ($A_O(x, y)$, top layer in Figure 1), after smoothing (see color map on the right, arbitrary units, Methods). (*I*–*K*) Performance of the model (asterisks) in locating 40 target objects in 40 object array images containing 9 objects (*I*,*J*) and 40 natural-background images (*J*,*K*). For each possible number x of fixations ($1 \le x \le 5$)), the y-axis indicates the proportion of all images in which the target was found within the first x fixations. Error bars indicate standard error of the mean across all 40 objects. Dashed line with circles: model performance when the normalization step is omitted (no denominator in eq. 5). Dotted lines with hexagons: model performance when feedback weights are randomly shuffled among objects for each input image. Dashed line with squares: priority maps generated by a purely bottom-up saliency model that has no information about the target object (Walther and Koch 2006). The gray horizontal lines in *I* and *J* indicate increasing multiples of 1/9. *I* and *K* use an "oracle" verification at each fixation point to determine whether the target fixation is correct or not. In contrast, *J* and *L* use a target validation system as illustrated in Figure 1 (Methods).

The attentional focus at any time is defined as the point of maximum local activity in the priority map (eq. 6). This selection is then projected back onto the ventral pathway of the visual system, by "cropping" activity around the current attentional focus. By emphasizing the selected areas, attentional selection allows the ventral pathway to perform fine-grained object

recognition on the selected input (Zhang et al. 2011). The system can then determine whether or not the target object is actually present at the selected location (Fig. 1—"S3/pIT" and "C3/aIT"). This process is iterated over each successive maximum of the priority map, resulting in a series of "fixations." If the recognition pathway model detects that the target is present at the current fixation point, the search concludes successfully. If the target is not found at the current fixation point (either because it is not present at this point, or because the recognition system failed to identify it), the model continues to the next best location in the priority map. Search is arbitrarily stopped at a fixed number of fixations if the target is not found. In those rare cases where the recognition system falsely recognizes the target at a given fixation point even though it is not actually there, the search concludes in failure.

Bottom-Up Architecture

The computational model builds upon the basic bottom-up architecture for visual recognition described in (Serre 2006; Kouh 2007), which is in turn an elaboration of previous feedforward architectures (e.g. Fukushima (1980); Wallis and Rolls (1997); Riesenhuber and Poggio (1999)). This model relies on an _alternation between "simple" cells that compute the match of their inputs with a pre-defined pattern and "complex" cells that return the maximum of their inputs selective for the same pattern but at slightly different positions and scales. Here, we consider 2 layers of simple and complex cells (S1, C1, S2b, C2b, using the same nomenclature as in previous studies) as described later.

We succinctly describe the bottom-up architecture here (for further details, see Serre (2006); Kouh (2007); Serre, Kreiman et al. (2007)). We consider 256 × 256 pixel grayscale images $I(x, y)$ ($1 \le x \le 256$, $1 \le y \le 256$ pixels, $0 \le I(x, y) \le 255$). The model does not include color. The first set of units (S1) convolve the image with Gabor filters at 12 scales S ($S = 1, 2. . . 12$) and 4 orientations θ ($\theta = 45, 90, 135, 180°$). Following Kouh (2007), the activation function for S cells is an L2-normalized inner product between weights and inputs. One difference with previous implementations is that we only use 12 different scales (rather than 16) and do not merge scales at any point: the model is essentially replicated in parallel at all scales. There are also minor differences in the positioning and spacing of cells at successive layers; in particular, the S1 cells do not densely cover the input. These choices result from early experimentation in which these particular arrangements provided a good tradeoff between performance and speed.

Filters at scale S are square matrices of size $D \times D$, with $D = 7 + 2 \times (S-1)$ pixels. S1 cells are evenly spaced every $D/4$ pixels both vertically and horizontally—thus they do not densely cover the image. We enforce complete RFs, which means that a cell's RF cannot overlap the border of its input layer. Because we enforce complete RF, the RF of the top-left-most cell is centered above pixel position $x = D/2$, $y = D/2$. Note that because of difference in RF diameters (and thus in margin and spacing), S1 cell columns of different scales do not generally fall at the same positions. At any given position, there is either no cell at all or a full column of 4 S1 cells (1 per orientation), all of the same scale. This also applies to C1 and S2b cells, replacing orientations with prototypes for S2b cells (see below).

The Gabor filter $G'_{S,\theta}$ of scale S and orientation θ is defined for every row x and column y as $G'_{S,\theta}(x, y) = \exp(-((\hat{x}^2 + \gamma^2\hat{y}^2)/(2\sigma^2)))\cos(2\pi\hat{x}/\lambda)$, (Serre 2006) where $\hat{x} = x\cos\theta + y\sin\theta$, $\hat{y} = -x\sin\theta + y\cos\theta$, $\lambda = 0.8\ \sigma$, $\sigma = 0.0036D^2 + 0.35D + 0.18$, $\gamma = 0.3$. Note that $-D/2 \le x \le D/2$ and $-D/2 \le y \le D/2$. The filter weights are then set to 0 outside of a circle of diameter D: $G'_{S,\theta}(x, y : \sqrt{x^2 + y^2} > D/2) = 0$. Finally, the Gabor filters are normalized to unit norm: $G_{S,\theta}(x, y) = G'_{S,\theta}/\sqrt{\sum_{x,y} G'_{S,\theta}(x, y)^2}$. For a given S1 cell of scale S, orientation θ, centered at position (x_c, y_c), the output is the absolute value of the normalized inner product between the (vectorized) corresponding Gabor filter and the portion of the input image falling within the cell's RF (Kouh 2007):

$$S1_{S,\theta,x_c,y_c} = \frac{\left|\sum_{i,j} G_{S,\theta}(i, j) I(x_c + i, y_c + j)\right|}{\sqrt{\sum_{i,j} I(x_c + i, y_c + j)^2}}. \quad (1)$$

C1 layers take S1 output as their inputs. The output of a C1 cell of scale S and orientation θ is the maximum of S1 cells of identical orientation and scale, within the RF of this C1 cell. At any scale, C1 cells are positioned over "every other" S1 column of the same scale, both vertically and horizontally. Each C1 cell returns the maximum value of all S1 cells of similar scale and orientation within a square of 9 × 9 S1 cells centered at the same position as this C1 cell:

$$C1_{S,\theta}(x_c, y_c) = \mathrm{MAX}_{i,j}(S1_{S,\theta}(x_c + i, y_c + j)), \quad (2)$$

with $-4 \le i \le 4$, $-4 \le j \le 4$. In the previous equation, x_c and y_c refer to position within the S1 layer of scale S, not to image pixel positions.

S2b cells take C1 output as their inputs. The output of an S2b cell depends on the similarity of its inputs with its prototype $P_S(i, j, \theta)$. There are 600 different prototypes, each of which takes the form of a 9 × 9 × 4 matrix as described later (9 × 9 diameter and 4 orientations). The same 600 prototypes are used for all scales. The output of an S2b cell of scale S, prototype P and position x, y is calculated as follows:

$$S2b_{S,P}(x, y) = \frac{\sum_{i,j,\theta} P_S(i, j, \theta) C1_{S,\theta}(x + i, y + j)}{\sqrt{\sum_{i,j,\theta} P_S(i, j, \theta)^2}\sqrt{\sum_{i,j,\theta} C1_{S,\theta}(x + i, y + j)^2 + 0.5}}, \quad (3)$$

with $-4 \le i \le 4$, $-4 \le j \le 4$, and θ ranges over all 4 orientations. Note that the numerator describes a convolution of the entire stack of 4 C1 maps (1 per orientation) with the S2b prototype, whereas the denominator normalizes this output by the norms of the prototype weights and of the inputs. Coordinates x and y refer to positions within the C1 grid of scale S, rather than image pixel positions. Following Serre (2006), each prototype $P_S(i, j, \theta)$ was generated by running the model up to level C1 on a random image, and extracting a patch of size 9 × 9 × 4 (diameter 9, 4 orientations) from the 4 C1 maps (1 per orientation) at a random scale and a random location. Then, 100 randomly selected values from this patch were then kept unmodified, whereas all other values in the patch were set to zero. The resulting patch constituted the actual prototype P. This process was iterated until 600 prototypes were generated. The random images used to set the prototypes were distinct from all the images used in the computational experiments mentioned later (i.e., none of the 40 target objects or 250 natural images used under "Computational experiments" were used to determine $P_S(i, j, \theta)$). Note that the same set of 600 prototypes was used at all scales. The C2b layer returns the global maximum of all S2b cells of any given prototype P, across all positions and scales. Thus, there are 600 C2b cells, 1 for each S2b prototype P. The max operation in C2b as well as that in equation 2 provide tolerance to scale and position changes (Serre, Kreiman et al. 2007).

Attentional Selection

The attentional model considers a situation where we search for object **O** in a complex image *I* that may contain an array of objects or a natural background (e.g., Fig. 2). To search for object **O**, the model considers the bottom-up responses to that object when presented in isolation: C2b(**O**, *P*) ($1 \le P \le 600$) and uses those responses to modulate the bottom-up signals to image *I*. We refer to this modulation as a feedback signal *F*(**O**, *P*), defined by the normalized C2b output for the target object presented in isolation on a blank background:

$$F(\mathbf{O}, P) = \mathrm{C2b}(\mathbf{O}, P)/\overline{\mathrm{C2b}(P)}, \quad (4)$$

where $\overline{\mathrm{C2b}(P)}$ is the average value of the C2b output for prototype *P* over 250 unrelated natural images. The dimension of *F* is given by the number of prototypes (600 in our case). Thus, for each S2b prototype *P*, the value of the feedback modulation *F*(**O**, *P*) when searching for target object **O** is proportional to the maximum response of this prototype to object **O** in isolation, across all positions and scales. *F* is then scaled to the (1, 2) range by subtracting the minimum, dividing by the maximum coefficient, and finally adding 1 to each coefficient. This ensures that the total range of weights is the same for all objects. Note that *F* is not hard-wired; it is task dependent and varies according to the target object.

Equation 4 makes the top-down signals dependent on the ratio between response to target and response to a "mean" stimulus (average response over many unrelated natural images). This is essentially similar to Navalpakkam and Itti's (2007) proposal to base top-down modulation on the signal-to-noise ratio (ratio of activations) of targets versus distractors; the difference is that here "distractors" are unpredictable and approximated by a large set of unrelated natural images.

The value of the feedback signal may be interpreted in a Hebbian framework: *F* represents feedback from "object-selective cells" in a higher area (possibly identified with PFC, see Discussion) that receive inputs from, and send feedback to, all S2B cells. Under Hebbian learning, the connection from any S2B cell to each object-selective cell will tend to be proportional to the activation of this S2B cell when the object is present, and therefore so will the strength of the feedback connection, which determines *F* when the object-specific cell is activated during search. At least in lower visual areas, empirical evidence suggests that feedback connections tend to connect cells with similar feature preferences (Angelucci and Bullier 2003; Shmuel et al. 2005). The learning phase for the *F*(**O**, P) weights is described in Figure 1.

The attentional model combines these target-specific signals with the responses of the bottom-up architecture up to S2b into a so-called LIP map (see Discussion), $\mathrm{LIP}_{S,P,\mathbf{O}}(x, y)$, defined by:

$$\mathrm{LIP}_{S,P,\mathbf{O}}(x, y) = \frac{\mathrm{S2b}_{S,P}(x, y) * F(\mathbf{O}, P)}{\sum_{k=1}^{k=600} \mathrm{S2b}_{S,k}(x, y) + 5}. \quad (5)$$

At every position (*x*, *y*) in the LIP map (which correspond to positions in the S2b map), each LIP cell (of scale S and preferred prototype P) multiplies its S2b input by the attentional coefficient *F* for prototype *P* given the target object **O**. We note again the tentative nature of ascribing this computation to area LIP. The denominator indicates that LIP cells also receive divisive feedforward inhibition, equal to the sum of all incoming S2b inputs at this position. An additive constant in the denominator (corresponding to the "sigma" constant in the canonical normalization equation [Carandini and Heeger 2011]) defines the "strength" of normalization: A large value means that the denominator is dominated by the fixed constant and thus less dependent on local activity, whereas a low value means that the denominator is dominated by the variable, activity-dependent term. We empirically set this parameter to 5 for all simulations. As described later, divisive normalization is crucial to the performance of the model (see Fig. 3*G*–*H* and Discussion).

The final priority map used to determine the location of attentional focus, $A_{\mathbf{O}}(x, y)$, is simply defined as the summed activity of all LIP cells at any given position:

$$A_{\mathbf{O}}(x, y) = \sum_{S,P} \mathrm{LIP}_{S,P,\mathbf{O}}(x, y). \quad (6)$$

At any time, the global maximum of this priority map defines the current fixation/attentional focus. Notice that this map is only defined at discrete positions of the original image—those over which S2b cells are located.

Owing to their discrete support and divisive normalization (which compresses responses at the higher end of the range), the priority maps produced by the system are difficult to interpret visually. For visualization purposes only, these maps are contrast-enhanced by linearly scaling them within the (0.5, 1.5) range, then exponentiating the value of each point 3 times ($x = \exp(\exp(\exp(x)))$); they are then smoothed by filtering the entire map with a Gaussian filter of standard deviation 3 pixels. Importantly, this processing is only used to generate Figures 1 and 2, for the purpose of assisting visualization. All the computations and results in this paper are based on the original, unprocessed $A_{\mathbf{O}}(x, y)$ as defined in equation 6.

Object Search

Searching for a target object **O** in a given image operates by iteratively finding the position of the maximum of the priority map$A_{\mathbf{O}}(x, y)$ defined in equation 6, enhancing the visual input around this location, and validating the presence or absence of the target at the selected location (see Target Presence Validation). If the target object is present within the selected location, the target is deemed to have been found and the search concludes successfully.

If the target is not detected at the selected location (either because it is not present at that point, or because the target validation step failed to identify it), the model sequentially moves to the next strongest location. This is implemented by a strong version of "inhibition-of-return" (IoR, [Klein 2000]) applied to the priority map, decreasing its value around the location of the maximum as described later. The model selects the next maximum, corresponding to a new model fixation. This procedure is iterated until the target is found or the maximum number of fixations (set to 5 unless otherwise noted) has been reached.

The IoR procedure multiplies the current priority map at fixation *f* pointwise by an inverted 2D Gaussian $\mathbf{N}(x_F, y_F, \sigma_{\mathrm{IoR}})$ centered on the position of the current (unsuccessful) fixation (x_F, y_F) and with standard deviation σ_{IoR}:

$$A_{\mathbf{O}}(x, y)[f + 1] = A_{\mathbf{O}}(x, y)[f](1 - k * \mathbf{N}(x_F, y_F, \sigma_{\mathrm{IoR}})). \quad (7)$$

In all simulations, $k = 0.2$ and $\sigma_{\mathrm{IoR}} = 16.667$. We report the proportion of images where the target is found as a function of the

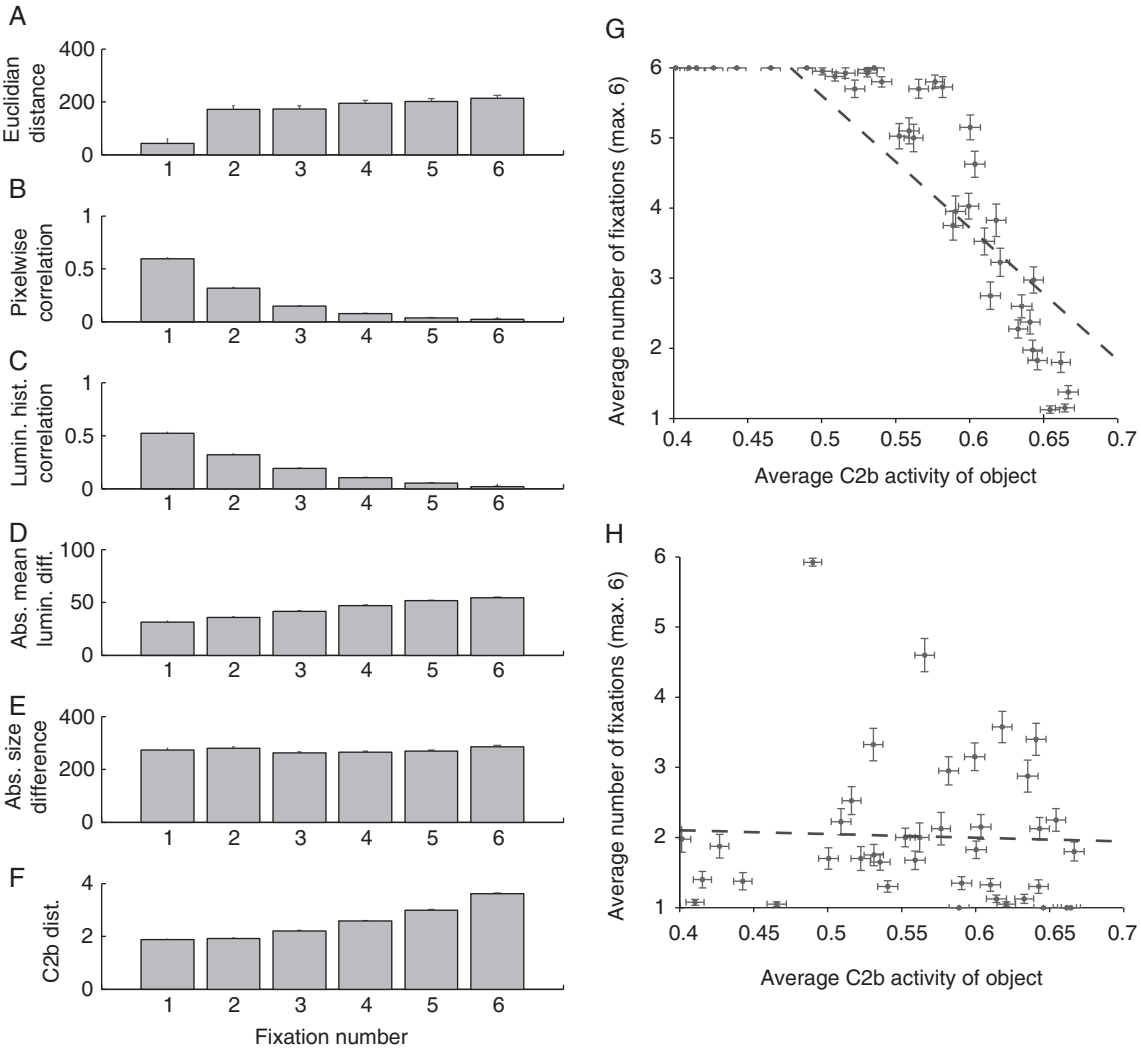

Figure 3. Properties of the model. (*A–F*). Average similarity between target and successively fixated objects, using various measures of similarity or difference between images, excluding fixations to the actual target. The first bar in all graphs indicates the average similarity (or difference) between the target and the first fixated objects across all trials in which the first fixation was erroneous. The similarity or difference measures are as follows: (*A*) pixel-wise Euclidean distance between the object images, (*B*) pixel-wise Pearson correlation between the images, (*C*) correlation between image luminance histograms, (*D*) absolute difference between mean luminance values (excluding background pixels), (*E*) absolute difference in size (i.e., number of non-background pixels within the bounding box), and (*F*) Euclidean distance between C2b vectors in response to the images. A significant correlation between fixation number and similarity or difference exists for all measures, except for size difference (*E*). The barely visible error bars indicate S.E.M. over the number of trials for each particular fixation; because we exclude fixations to the actual target, this number ranges from $n = 1158$ (first column) to $n = 2997$ (last column). These computational data are derived from the same images used for the psychophysics experiment (Fig. 4), using target-present trials only. (*G*,*H*) Effect of normalization on model output. For each one of 40 objects, the *x*-axis indicates the average activity of all C2b cells elicited by the object. The *y*-axis indicates the average number of model fixations necessary to find the object in 40 object array images (if the object is not found after 5 fixations, the number of fixations is set to 6 for this image). Error bars are standard error of the mean over 600 C2b cells (horizontal) or 40 images (vertical). Without normalization (*G*), objects eliciting stronger C2b activity are easier to find, indicating that they attract attention at the detriment of other objects, biasing search (dashed line: $r = -0.82$, $P < 10^{-11}$). With normalization (*H*), this effect disappears ($r = -0.04$, $P = 0.81$).

number of fixations *f* required in the computational experiments in Figures 2–4. In the model, IoR is constant over a trial, a simplification common to other visual search models (Itti and Koch 2001), whereas in reality it has a limited time span (Horowitz and Wolfe 1998). This suggests that IoR only affects the last few selected areas (Klein 2000; Itti and Koch 2001). We only consider a maximum number of successive fixations in our model evaluations (5 in Figs 2 and 6 in Fig. 4C). Furthermore, in the comparisons against human behavior in Figures 5–8, only the first fixation is considered (and therefore IoR does not play a role).

To separately report the efficiency of the attentional system and the target validation step, we consider an alternative version in which an "oracle" determines whether the selected location contains the target object or not. The oracle simply checks whether the attentional maximum falls within the bounding box of the target object or not. The bounding box B_O was defined as the smallest square encompassing all pixels of the object. If $\arg\max[A_O(x, y)] \in B_O$, then the target has been found.

Target Presence Validation

Given a selected region as determined by the maximum of the priority map, the model needs to determine whether it has found the target object or not (Fig. 1). Because our focus lies on the attentional selection system, we chose to use a highly simplified version of the full HMAX system for this purpose.

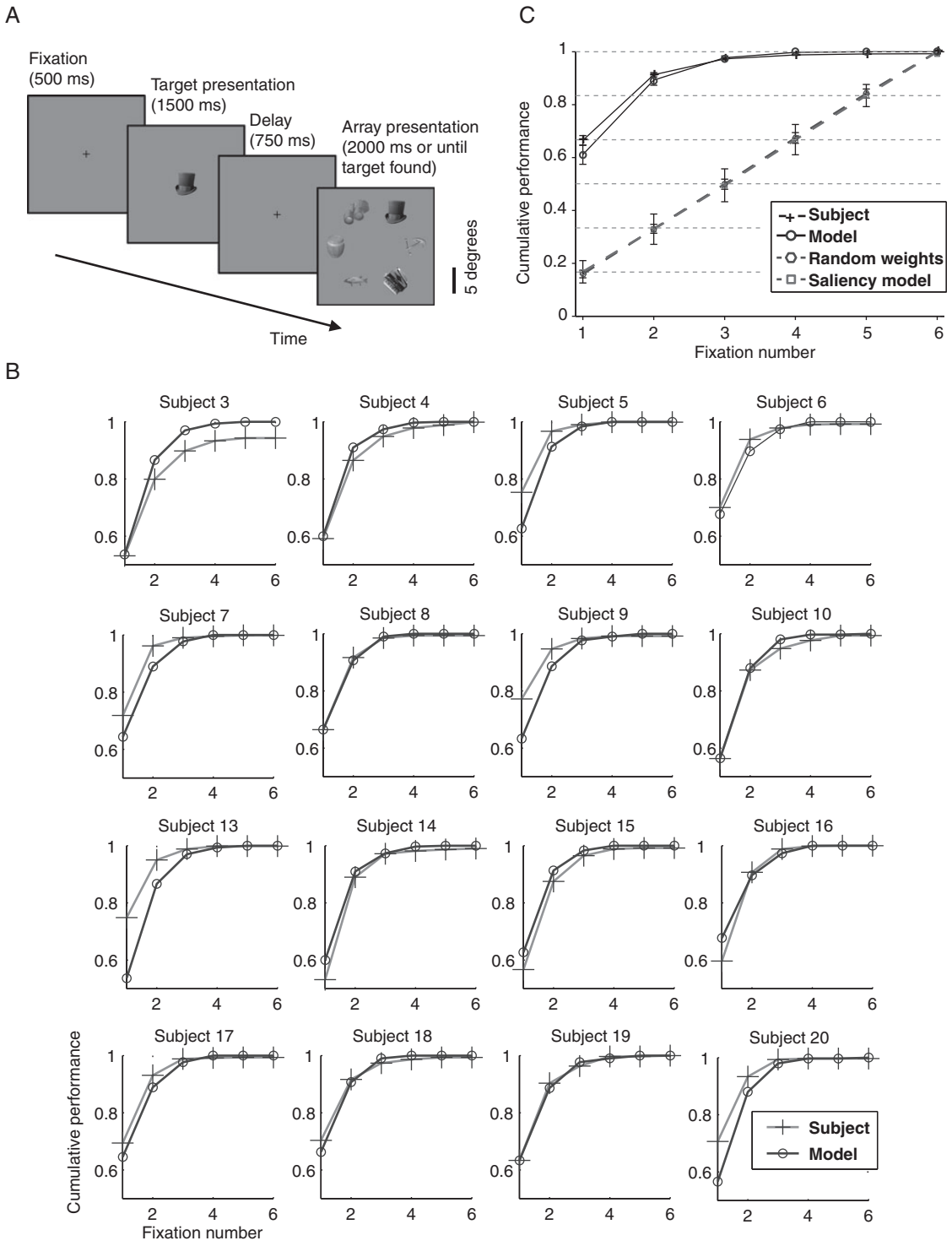

Figure 4. Comparing computer model and humans on a common task. (*A*) Visual search task. After fixation (verified by eye-tracking, Methods), a target object was presented for 1500 ms. The screen was then blanked (except for a central fixation cross) for 750 ms and then a search array consisting of 6 objects was shown. If the subject failed to find the target after 2000 ms, the trial ends and a new trial began. (*B*) Comparison between model performance and individual subjects. Model performance ("o") versus subject performance ("+") on the same stimulus sets (same targets and same array of choice objects, but randomizing object positions within the array) for successive fixations for each individual subject. There are small variations for the model from one plot to another because model performance for each graph is estimated on the same stimulus set shown to the subject, which differs across subjects. (*C*) Average performance for subjects ("+," average of 16 subjects), model ("o"), and control models (random weights model shown with hexagons and saliency model shown with squares) for the task described in (*A*) Only target-present trials averaged across subjects are shown here (see Fig. 5 for target-absent and error trials). Error bars indicate SEM across all 40 target objects. The 2 dashed lines represent the model performance when attentional weights were randomly shuffled across objects for each input image and from a purely bottom-up saliency model that had no information about the target object (Walther and Koch 2006). The horizontal dashed lines represent increasing multiples of 1/6.

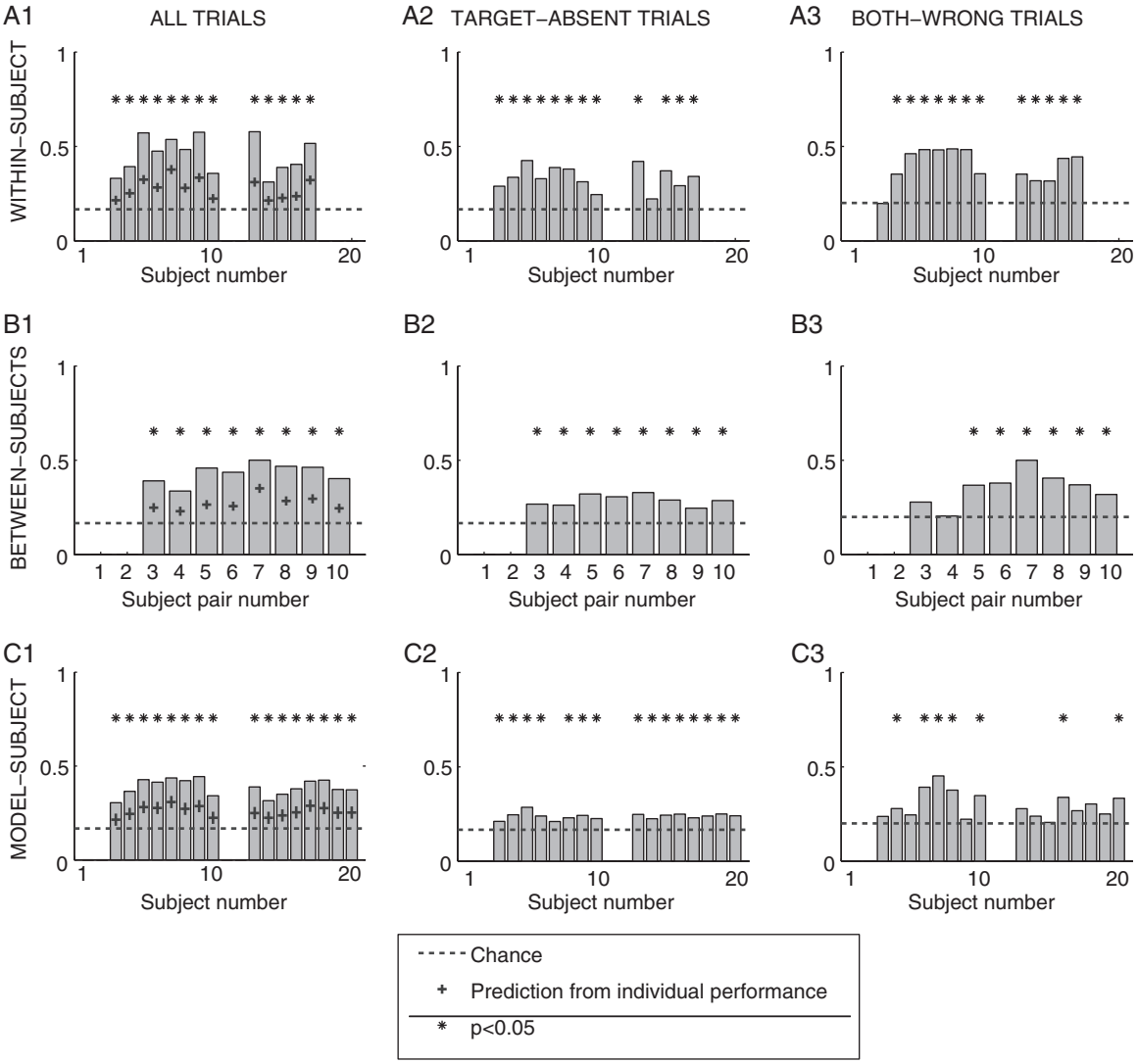

Figure 5. Consistency metrics for individual subjects. We evaluated consistency in "repeated trials" where the same set of stimuli (identical target and same objects in different positions within array) was presented to 2 different subjects or 2 different sessions for the same subject. Within-subject agreement (*A*1–3): proportion of trials in which the subject first fixated the same object in repeated trials (13 subjects). Between-subject agreement (*B*1–3): proportion of repeated trials in which both subjects first fixated the same object (8 subject pairs). Model-subject agreement (*C*1–3): proportion of trials in which both the subject and the model first fixated the same object (16 subjects).Column 1 (*A*1,*B*1,*C*1) includes all trials. Column 2 (*A*2,*B*2,*C*2) includes only target-absent trials. Column 3 (*A*3,*B*3,*C*3) includes only error trials. In all plots, the dotted line indicates the chance level under the assumption of purely random (but non-repeating) fixations (1/6 in Columns 1 and 2 and 1/5 in Column 3). In the "all trials" case (Column 1), we further consider a null model that takes into account the fact that subjects and model were able to locate the target above chance, which affects their level of agreement. If the 2 series being compared (same subject on repeated trials, 2 different subjects, or subject and model) have probability of finding the target at first fixation P_1 and P_2, respectively, then the probability of agreement by chance is $P_1P_2 + (1 - P_1)(1 - P_2)/5$ ("+"). Whenever the degree of consistency was significantly above chance, the comparison was marked with * ($P < 0.05$, binomial cumulative distribution test).

The attentional system enhances processing at the selected location at the S2b stage in the model. The output of the S2b units is then fed to an S3 layer. S3 units emulate the Posterior Inferotemporal cortex (PIT). Each S3 unit receives input from all 600 S2b units sharing a given location and compares the values of its inputs with a stored prototype vector. For simplicity and robustness, this comparison is computed by Spearman rank correlation between the input vector of S2b units and the S3 unit's stored prototype, rather than Euclidean distance (Serre, Wolf et al. 2007) or normalized dot product (Kouh 2007). The value of this correlation constitutes the activation of the S3 unit. For computational efficiency, we only consider the 3 smallest scales of S2b units. Similar to the S2b stage, the entire S3 area consists of the same set of S3 units duplicated over each position in the map; that is, over each S2b column, there is an identical complement of S3 units with the same set of prototypes. These prototypes are also extracted from exposure to isolated pictures of objects, by simply extracting all S2b columns with non-zero activation when presenting a given object and assigning each such column as the prototype of a new S3 cell. This results in a total of 1720 S3 prototypes, representing 43 prototypes per object.

The output of the S3 units is conveyed to a global pooling stage, inspired by the C3 stage in the original HMAX model, and emulating Anterior Inferotemporal Cortex (with high object selectivity and large RFs). There is 1 C3 unit per known object, receiving input from all S3 units associated with that object. C3 units sum the activities of all their S3 inputs, and the object

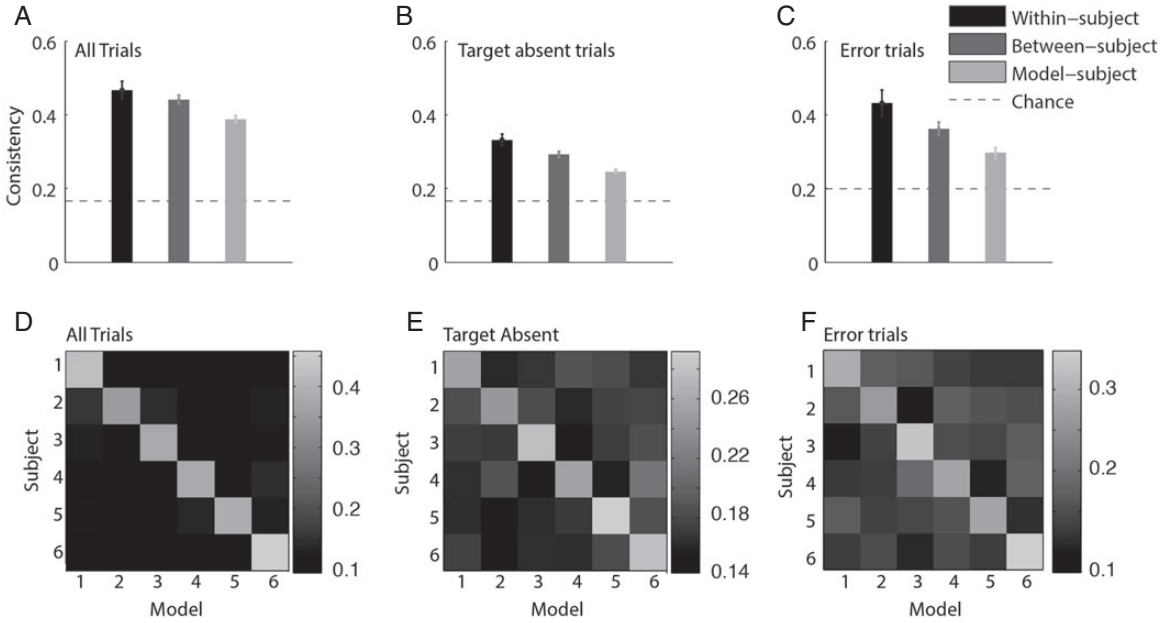

Figure 6. Consistency within subjects, across subjects, and between subjects and model. (*A–C*) Following the format and nomenclature in Figure 5, here we show average consistency values across subjects. Black: within-subject agreement (13 subjects), dark gray: between-subject agreement (8 subject pairs), and light gray: model-subject agreement (16 subjects). Results are shown for all trials (*A*), target-absent trials (*B*), and target-present trials in which both responses were erroneous (error trials, *C*). Error bars indicate SEM across all subjects or pairs. The dashed line indicates chance performance (1/6 in *A*,*B* and 1/5 in *C*). (*D–F*) Subject-model confusion matrix for all trials (*D*), target-absent trials (*E*), and error trials (*D*). The color at row *i* and column *j* shows the conditional probability of the model's response (first saccade) being position *j* when the subject's response (first saccade) was position *i*. These matrices represent the average across all the objects (Methods); individual-specific matrices are shown in Figure 7. The color scales are different in the different panels (there was more consistency and hence a more pronounced diagonal between model and subject in correct target-present trials, which are part of *D* but not *E* or *F*; using the same scale would make it difficult to see the diagonal in *E*,*F*). Diagonal values are significantly higher than non-diagonal values for all 3 matrices ($P < 0.01$, Wilcoxon rank-sum test), reflecting the significant agreement between model and subject first fixations across trials.

associated with the most active C3 unit is the final result of the recognition pathway.

Control Experiments

We performed several controls and comparisons with other models. It is conceivable that in some cases, attentional selection could be purely driven by bottom-up "saliency" effects rather than target-specific top-down attentional modulation implemented via equations 5 and 6. To evaluate this possibility, we compared the performance of the model as described earlier with 2 control conditions. First, we used a modified version of the model, in which attentional modulation used the weights ($F(\mathbf{O}', P)$) of a random object $\mathbf{O}'$ for every input image instead of the specific weights associated with the actual target object **O**. We refer to this control as "Random weights" in Figures 2 and 4. Second, we generated priority maps based on the bottom-up saliency model of Itti and Koch (2000). We used the Saliency Toolbox implemented in Walther and Koch (2006) with default parameters, except for setting "normtype" to "none" (using the default value for normtype results in very sparse saliency maps in which only a few of the objects have a non-zero saliency, leading to worse performance). We refer to this condition as "Saliency model" in Figures 2 and 4. Both control models were applied to the exact same images as the model and following the same procedure outlined under "Object search" above.

In Figure 3, we compared fixated objects and target objects. For this figure, we considered several possible similarity metrics: Euclidean distance (**3A**), pixel-wise correlation (**3B**), correlation between luminance histograms (**3C**), absolute difference between mean luminance values (**3D**), absolute size difference (**3E**), and Euclidean distance between C2b vectors produced by the bottom-up architecture (**3F**).

Psychophysics Experiments

We compared the computer model against human performance in a psychophysics visual search task described in Figure 4 and in the Results. During the psychophysics task, stimuli were presented on a CRT monitor (Sony Trinitron Multiscan G520). We used the Eyelink D1000 system (SR Research, Ontario, Canada) to track eye positions with a temporal resolution of 2 ms and a spatial resolution of ~1° of visual angle. We calibrated the device at the onset of each session by requiring subjects to fixate on visual stimuli located in different parts of the screen. The equipment was re-calibrated as needed during the experiment. A trial did not start if subjects' eyes were not within 1° of the fixation spot for a duration of 500 ms. Failure to detect fixation prompted for eye-tracking recalibration.

Results

We consider the problem of localizing a target object in a cluttered scene (e.g., Fig. 2*A–D*) and propose a computational model constrained by the architecture and neurophysiology of visual cortex (Fig. 1). We start by discussing computational experiments demonstrating the model's ability to localize target objects and subsequently compare the model's performance with human psychophysics measurements.

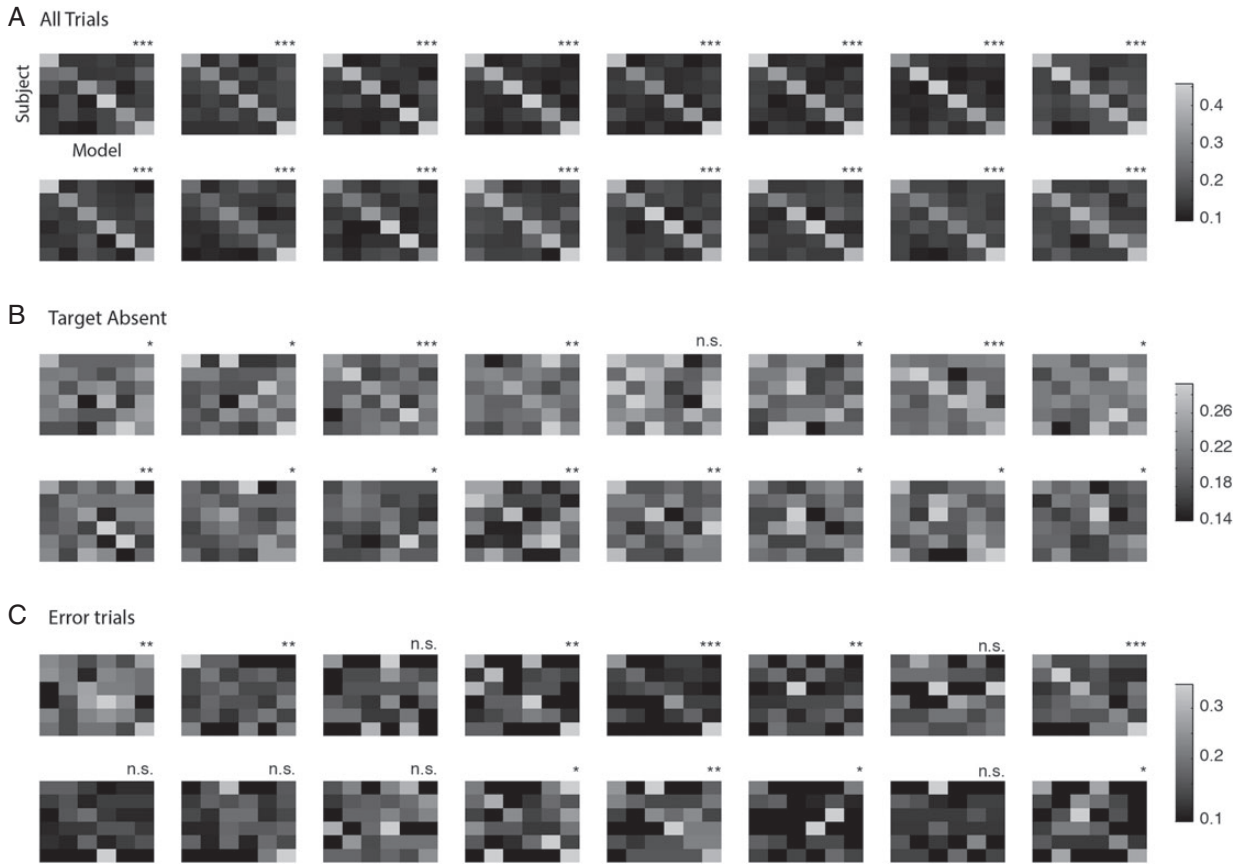

Figure 7. Subject-model comparison for individual subjects. Individual confusion matrices for all 16 subjects, using all trials (*A*), target-absent trials (*B*), or error trials (*C*). The format for each confusion matrix is the same as that in Figure 6*D*–*F*. Matrices with diagonal values significantly higher than non-diagonal values indicate above-chance agreement between model and subject across trials (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$, n.s.: not significant, Wilcoxon rank-sum test).

Can The Model Find an Object in an Array?

Our first objective is to evaluate the performance of the model in localizing target objects in cluttered images. We first considered target objects embedded in multi-object arrays presented on a blank background (e.g., Fig. 2*A*,*C*). We used the same set of 40 different target objects taken from Hung et al. (2005) for all experiments. Object array images consisted of a fixed number of objects ($n = 9$) on a blank background and were generated as follows. Nine objects, comprising the target object plus 8 randomly selected objects different from the target (distractors), were resized so that each would fit within a bounding box of size 43×43 pixels. This size results from seeking to maintain a margin of 20 pixels on all sides around each object, within a 256×256 image. These objects were then regularly placed over a uniform gray background (e.g., Fig. 2*A*,*C*). For each of the 40 objects, we generated 40 images (1600 total images) containing this object as the target, plus 8 other randomly selected objects. The model was trained to learn the appropriate weights for each isolated object, thus generating the modulatory weights $F(\mathbf{O}, P)$ as illustrated for the top hat target in Figure 1.

Examples of the priority map produced by the model for different input images and target objects are shown in Figure 2*E*,*G*. As expected, the priority map depended on the target object. The priority map showed above background activation in multiple locations that contain objects, with enhanced activation in the particular location where the target object was located (Fig. 2*E*,*G*). To assess the performance of the model, we computed the cumulative proportion of successful localizations after 1 to 5 successive attentional "fixations" over all test images. The model found the target object on the first fixation in 54% of the array images (Fig. 2*J*, solid line), whereas randomly selecting 1 of 9 positions would yield a success rate of 11.1%. For 93% of the images, the model found the target within 5 attentional fixations. It is conceivable that in some cases, attentional selection could be driven by bottom-up "saliency" effects rather than target-specific top-down attentional modulation implemented via equations 5 and 6. We performed various control experiments to evaluate this possibility. First, we compared the results with a "randomized" version of the system in which the feedback weights $F(\mathbf{O}, P)$ were taken from a randomly selected object $\mathbf{O}'$ for each image instead of the actual target object $\mathbf{O}$ ("Random weights," dotted line in Fig. 2*J*). The performance of this null model was well below performance of the full model and was essentially similar to what would be expected from randomly fixating successive objects. As an alternative control for purely bottom-up cues, we considered the saliency model of Itti and Koch (2000), as implemented by Walther and Koch (2006), with default parameters (except for setting "normtype" to "none"; using the default value for normtype results in very sparse saliency maps in which only a few of the objects have a non-zero saliency, leading to an even worse performance). This purely bottom-up, task-independent architecture selects parts of an image that attract attention due to local contrast in intensity, orientation, and color. As expected, the performance of this bottom-up model was comparable with that of the random weights model (Fig. 2*J*, dashed line).

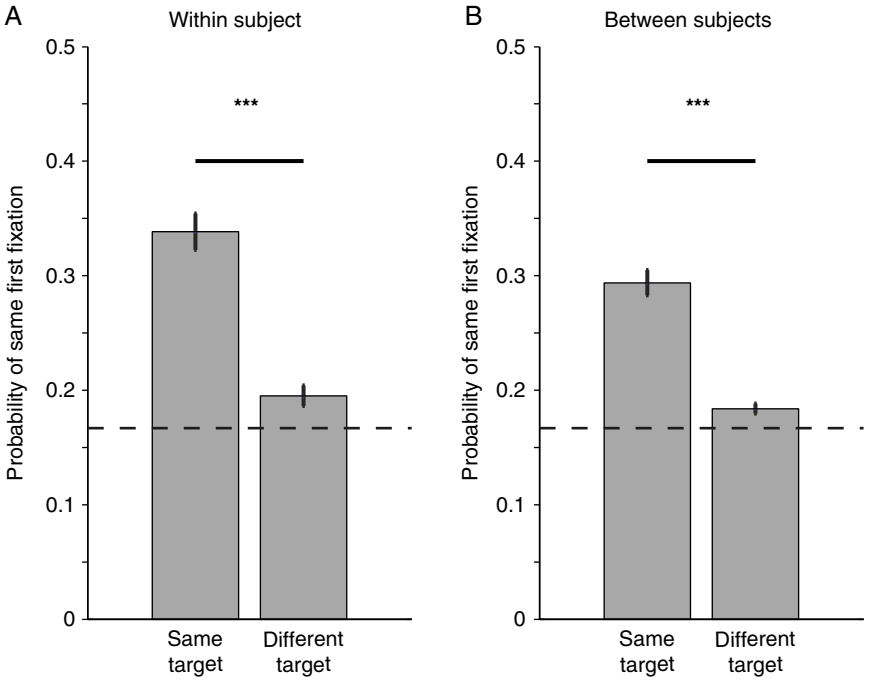

Figure 8. Target identity influences responses in target-absent trials. Average self-consistency (*A*) and between-subject consistency (*B*) in responses to 2 trials where the target was absent, where the same 6-object array was shown (randomized object positions) and the target was either the same (same target) or different (different target). If the first fixation were driven purely by bottom-up signals derived from each object, we would expect similar degrees of consistency in the "same target" versus "different target" conditions. Instead, we observed a significantly higher consistency (both for within-subject comparisons ($P < 10^{-5}$) as well as between-subject comparisons ($P < 10^{-4}$) when the target was the same, suggesting that subjects were using aspects of the target object to dictate their first fixation (Wilcoxon rank-sum tests with 13 and 8 pairs of values, respectively). Error bars indicate S.E.M. across 13 subjects for self-consistency, and 8 pairs of subjects for between-subject consistency. The horizontal dashed line indicates chance levels (1/6). Note that consistency in target-absent trials with different targets at each presentation is slightly, but significantly, above the chance level of 1/6, both within-subject ($P < 0.001$) and between-subjects ($P = 0.03$; signed rank test of median = 1/6 across 13 and 8 values, respectively). This indicates a weak, but significant effect of bottom-up, target-independent features in guiding saccades in these images.

We investigated whether erroneous model fixations were driven by similarity between the fixated object and the target. We plotted the average similarity between successively fixated objects and target, along various measures of similarity, excluding fixations to the actual target (Fig. 3*A*–*F*). Non-target objects attended to during the first fixations were more similar to the target under many similarity measures, including Euclidean distance (**3A**), pixel-wise correlation (**3B**), correlation between luminance histograms (**3C**), absolute difference between mean luminance (**3D**) or size (**3E**) values, and Euclidean distance between C2b vectors produced by the bottom-up architecture (**3F**). We also evaluated whether certain object features correlated with ease of detection independently of the target. Object size (defined as the number of non-background pixels within the object bounding box) significantly correlated with probability of first fixation success ($r = 0.34$, $P = 0.03$). Object contrast (measured as the variance of non-background pixels) did not correlate with first fixation success (object array images: $r = -0.16$, $P > 0.3$).

Can The Model Find an Object in A Natural Scene?

Next, we applied our model to the more challenging task of detecting small target objects embedded within natural scenes (Fig. 2*B*,*D*). Natural images present additional challenges for visual search since the target object and distractors are not segregated from the background. All input images were grayscale squares of size 256 × 256 pixels. A single-target object (1 of the 40 objects above resized to 64 × 64 pixels) was superimposed onto 1 of 250 images of natural scenes from (Serre, Oliva et al. 2007). The insertion position was random (except to ensure that the entire object fell within the natural image). In the real world, there are other constraints on object positions (e.g., there are no accordions suspended in highways as in Fig. 2*D*). While these constraints influence object identification and visual search (Oliva and Torralba 2006), they are not modeled here. We generated 1600 images (40 target objects and 40 natural scenes) where the area of the target object was 1/16 of the area of the whole image (e.g., Fig. 2*B*,*D*). As shown earlier for object arrays, the priority map showed enhanced activation at the target location (Fig. 2*F*,*H*). The model localized the target object on the first fixation in 37% of the images (Fig. 2*L* solid line, 11% for the randomized model and 9% for the saliency model). Performance reached 54% after 5 fixations. Object size correlated with first fixation success ($r = 0.54$, $P < 0.0003$) but object contrast did not ($r = 0.07$, $P > 0.65$). As expected, the model's performance in natural scenes was significantly below performance in object array images (with 9 objects). This in part reflects the additional complexities introduced by natural backgrounds; humans can take advantage of several other cues that are not incorporated in the current model such as contextual and semantic information (Discussion). Despite these limitations, the model performs well above chance under these challenging search conditions.

Dissecting The Model's Performance

To gain further insights into how the model searched for target objects, we considered 2 simplified versions that lacked some of the computational modules. First, the results in Figure 2*J*,*L*

depend on the efficacy of both the attentional selection system *and also* the recognition system. Failure to locate the target may be caused by incorrect fixations, or by incorrect recognition at a given fixation. To dissociate these 2 components, we selectively evaluated the attentional selection system, without interference from the recognition system, by using an "oracle" for verification at every fixation point: The search was deemed successful if the current fixation point was located within the known bounding box of the target object (Methods). The attention + oracle model produced relatively similar results for the object array images: The target was successfully localized in 56% of first fixations (vs. 54% for the attention + verification model) and within 5 fixations for 95% of trials (vs. 93%). In contrast, the attention + oracle model led to a larger improvement on the natural-background images, with success rates of 48% at first fixation (vs. 37% for the attention + verification model) and 81% within 5 fixations (vs. 55%). These results illustrate the challenge that natural backgrounds pose to the recognition system, especially in contrast to isolated objects on a blank background (a known difficulty with computational models of object recognition [Serre, Kreiman et al. 2007]). The results also suggest that the attentional selection model performs relatively well on natural scenes, even though its performance is still noticeably lower than on object array images.

Next, we analyzed the role of the normalization operation introduced in the model (Fig. 1). The normalization operation (divisive feed-forward inhibition by local inputs, eq. 5) played an important role in the model's performance. In the absence of normalization, the system's performance was strongly degraded (Fig. 2, "No normalization"). In the absence of normalization, locations with higher local feature activity tended to dominate the priority map over the task-dependent feedback modulation (Fig. 3*G*). Normalization by local inputs made it easier for object-selective feedback to drive lower-activity locations to prominence, by prioritizing the match between the input signals and top-down modulation over absolute magnitude of local inputs —effectively turning multiplicative modulation into a correlation operation (Fig. 3*H*; see Discussion).

How Fast Can Humans Find An Object in an Array?

Our next objective was to compare the model's output against human behavior on a common visual search task. We designed a psychophysics experiment to (i) evaluate whether human visual search is reproducible, within, and across subjects, under the conditions examined via the computational model, (ii) verify that this reproducible component is actually influenced by target identity and features, rather than being purely driven by bottom-up, target-independent saliency effects or other target-independent biases, and (iii) estimate how much of this reproducible component is captured by the model.

The psychophysics task is illustrated in Figure 4*A*. Subjects were required to maintain fixation within 1° of a central fixation cross during 500 ms to start each trial. After successful fixation, a target object was shown centrally for 1500 ms. Next, another fixation cross was presented for 750 ms. Finally, 6 objects were shown (each subtending ~5° of visual angle), regularly arranged in a circle around the center of the screen (radius ~8°). Objects can be readily recognized at this size and eccentricity (Supplementary Fig. 5*B*). The 6 stimuli were on the screen until the end of the trial. The target was randomly chosen in each trial.

The task was to direct gaze toward the target object "as quickly as possible." Subjects were deemed to have fixated a given object whenever their gaze fell within the bounding box of the object (there was no bounding box on the screen, this was merely to define successful fixations). If the subject's gaze found the target object, the trial ended and the object was surrounded with a white frame for 1000 ms in order to indicate success. If the target was not found within 1800 ms, the trial was aborted and a new trial began. The target was present in 70% of the trials. The 30% of target-absent trials provided us with an additional point of comparison between human and model behavior (see text below and Figs 5–7). Because these trials were randomly interspersed with the target-present trials, subjects could not tell whether the target was present or not without visual search. We used the same set of 40 objects as in the previous computational experiments discussed earlier, resized to 156 × 156 pixels. The same object arrays were used for the psychophysics and computational model in Figures 4–8.

We recruited 16 subjects (10 female, 18 to 35 years old). In each session, the subject was presented with a block of 440 trials (300 target-present, 140 target-absent). Each block of trials was shown to 2 different subjects (but randomizing the temporal order of trials, and the position of objects along the circle in each trial). Using the same target and array of objects allowed us to evaluate the reproducibility of behavioral responses between 2 subjects on single trials—that is, the proportion of trials in which both subject's first fixations were located on the same object (Figs 5–7). At the same time, by randomizing trial order and the positions of objects on the display, we sought to combat the possible influence of various biases that are not directly related to target identity and features, such as persistence effects, or spatial biases toward certain directions (Tatler et al. 2006; Foulsham et al. 2008). In addition, 10 of 16 subjects participated in a second session in which, unbeknownst to them, they were presented with the same block of trials as in their first session (again, randomizing temporal order of trials and object position within each trials). This allowed us to compute subject self-consistency (within-subject agreement).

The single-trial comparisons are based on the first fixation only. This is because after the first fixation, spatial symmetry is broken and various mechanisms (such as different acuity between central and peripheral vision, remapping, etc.) may influence visual search. Because we are mostly interested in the building of the priority map and our model does not capture these additional effects, using only the first fixation allows us to more directly compare human and model performance.

Subjects were able to perform the task well above-chance levels (individual subject data in Fig. 4*B*, average results in Fig. 4*C*, "+" symbols). The task was not trivial as evidenced by the fact that the subjects' first saccade was only 65% correct in target-present trials (randomly selecting an object and other degenerate strategies would yield a performance of 16.6%; perfect detection in the first fixation would yield a performance of 100%). Subjects were essentially at ceiling by the third fixation.

As in other psychophysics tasks, there is a trade-off between speed and accuracy. Given unlimited time, it is possible to make a first fixation to the correct target with almost perfect accuracy (Supplementary Fig. 5*A*). In contrast, the current results pertain to a regime where subjects were urged to make multiple fixations to find the target as soon as possible. The mean latency of the first fixation over all trials was 284 ± 157 ms (mean ± SD, Supplementary Fig. 4*A*). The median latency was 237 ms, reflecting a skewed distribution (Supplementary Fig. 4*B*). Target-absent trials had only slightly longer mean latency at 320 ± 210 ms. This difference was largely caused by a slightly larger "tail" of long latencies, since the early modes of the latency distributions were essentially identical for target-absent ant target-present trials

(Supplementary Fig. 4C). In contrast, among target-present trials, erroneous trials were noticeably faster than correct trials, even in their modes (Supplementary Fig. 4B).

As reported in other studies (Tatler et al. 2006; Foulsham et al. 2008), subjects showed directional biases in their fixation behavior (Supplementary Fig. 3). Our experimental design (symmetrically arranged objects, randomized positions, concentrating on the first fixation only) allowed us to minimize the effects of these biases, as explained earlier.

To compute an upper bound for how well the model could predict subject behavior, we first evaluated the reproducibility of subject responses, both within and across subjects. Subjects showed a high degree of self-consistency, defined as the proportion of repeated trials (same target and distractors) where the subjects first fixated on the same object, both individually (Fig. 5A) and on average (Fig. 6, black bars). In target-present trials, being able to locate the target with probability P above chance suffices to lead to above-chance self-consistency. We evaluated the degree of self-consistency expected purely from the overall performance as $P^2 + (1-P)^2/5$. Subjects showed a stronger degree of self-consistency than predicted from performance in individual target-present trials (Fig. 5A1). Furthermore, subject responses showed significant self-consistency in target-absent trials (Figs 5A2 and 6B), and in trials for which the first fixated object was not the target in both presentations ("error trials," Figs 5A3 and 6C), conclusively showing that consistency was not due to ability to locate the target.

As expected, consistency between subjects was slightly below self-consistency (compare black vs. dark gray bars in Fig. 6; see also Fig. 5A vs. Fig. 5B). Still, between-subject consistency was also well above chance. In other words, different subjects showed consistent first fixation behavior when searching for the same target among the same set of distractor objects.

On target-absent trials, the degree of consistency of first saccades in trials with identical choice objects was significantly higher when the target was the same in both trials compared with when the target was different, both within and between subjects (Fig. 8). This confirmed that the subjects' fixations were guided by target identity even when the target was not present in the image array (as opposed to being driven purely by bottom-up, or other target-independent features of the objects). Furthermore, we evaluated the similarity between objects and targets for the psychophysics experiments using the similarity metrics defined in Figure 3. All similarity metrics (pixel correlation, Euclidean distance, histogram correlation, luminance difference, size difference and C2b correlations) resulted in significant differences between fixated-first and non-fixated-first objects (Supplementary Fig. 2). Consistency in target-absent trials with identical choice objects but different targets at each presentation was slightly, but significantly above the chance level of 1/6 (Fig. 8), both within-subject ($P < 0.001$) and between-subjects ($P = 0.03$; signed rank test across 13 and 8 values, respectively). This indicates a weak, but significant effect of bottom-up or intrinsic, target-independent features in guiding saccades during this task.

Subject fixations were influenced both by target features and target-independent biases. We sought to estimate and compare the relative importance of target-dependent and target-independent factors in determining fixations. There were 2 groups of subjects and each trial was seen by 2 subjects, one from each group, with the same target and object array but, critically, with randomized object positions. We concatenated all objects of all target-absent trials in a single binary vector $\mathbf{x}$, $x(i) = 1$ if object i was fixated first in that trial and zero otherwise (3222 trials × 6 objects = 19 332 entries). We regressed this vector on 2 predictive vectors: (i) a position-based vector which contained the directional bias for each object's position (that is, the proportion of all saccades made toward the position occupied by the object in that trial, Supplementary Fig. 3) and (ii) a feature-based vector that contained a 1 if the object was fixated first by the subject in the other group when seeing the same trial and 0 otherwise (remember that object positions were randomly chosen for each subject; thus, this second vector contains no influence from spatial biases). Both predictor vectors were z-scored. When using the same regression model for prediction, we observed that including spatial biases did not significantly improve prediction of a subject's choice when the other subject's choice is known. We used the output of the regression for each object, B1 × other subject's choice indicator + B2 × spatial position, and then selected the maximum for each trial. The resulting selection showed a consistency of 0.29, which is exactly identical to the inter-subject agreement found in target-absent trials (see below). This exact equality reflects the fact that the output of the regression was entirely determined by the other subject's choice; even though the regression coefficients were close, the spatial bias component's entire range was smaller than the contribution from the other subject's choice, which dominated the predictions when taking the maximum. Arbitrarily increasing the weight for spatial biases (coefficient B2) only decreased consistency. We conclude that, while spatial biases do play an important role in visual search, search behavior in the current task conditions was dominated by the feature-based, target-dependent component of visual search rather than these target-independent spatial biases.

Does The Model Behave Like A Human During Visual Search?

We next directly compared human behavior with model predictions. As expected from Figure 2I, the model was also able to successfully localize the target objects (Fig. 4B,C, circles). The close match between model and subject performance evident in Figure 4B,C is probably dependent on experimental parameters including object eccentricity, similarity and other variables. Nevertheless, it is clear that the task is not trivial, yet both humans and the model can successfully localize the target within a small (and similar) number of fixations.

The overall similarity between psychophysical and model performance (Fig. 4B,C) does not imply direct processing homologies between subjects and model, since both could perform target identification through entirely different mechanisms. To further evaluate the model, we sought to determine whether the model could also predict a more fine-grained aspect of subject behavior, namely single-trial fixations, beyond what could be expected purely from performance in target localization. To this end, we investigated whether the model could predict the subjects' first fixations, including those in target-absent and error trials. We found that the observations about self-consistency and between-subject consistency were also reflected in the agreement between model and subject responses (Fig. 5C and light gray bars in Fig. 6). The model's fixations agreed with the subjects' fixations on a trial-by-trial basis when considering all trials (Fig. 6A), target-absent trials (Fig. 6B), and error trials (Fig. 6C). Because the model uses a normalized multiplication before selecting the maximum (eq. 4–6), increasing the modulation equally across the entire map has minimal effects on the priority map. We confirmed this by running the simulations again, multiplying all top-down modulations by a factor 10 (Supplementary Fig. 1).

As expected, the results in Supplementary Figure 1 are very similar to those in Figure 6. For example, target-absent model-subject agreement was 0.241 ± 0.031 in Supplementary Figure 1, compared with 0.239 ± 0.018 in Figure 6. When comparing the responses of each individual subject with the model, above-chance model-subject consistency was observed in all 16 subjects when considering all trials (Fig. 5C1), in 15/16 subjects for target-absent trials (Fig. 5C2), and 7/16 subjects in error trials (Fig. 5C3). Overall, model-subject agreement was weaker than but qualitatively similar to between-subject agreement.

To further illustrate the agreement between model and subject responses, we plotted confusion matrices for first fixation position, across all images and all target objects (Fig. 6*D*–*F*; see Methods). Each position (row *i*, column *j*) in these 6 × 6 matrices indicates how often the subject's first fixation fell on the *i*-th object in the display, and the model's first fixation was on object *j*, divided by the number of trials in which the subjects first made a saccade toward object *i*. This represents the conditional probability that the model selected object *j*, given that the subject selected object *i*. The diagonal of these matrices measures agreement between subjects and model. The diagonal values in these matrices were significantly higher than the non-diagonal values ($P < 0.001$ for all 3 matrices, Wilcoxon rank-sum test between the diagonal and non-diagonal values), illustrating the single-trial agreement between model and subjects in all trials (Fig. 6*D*), target-absent trials (Fig. 6*E*), and error trials (Fig. 6*F*). Individual confusion matrices for each subject are shown in Figure 7.

Discussion

We have introduced a simple model to explain how visual features guide the deployment of attention during search (Fig. 1). This model proposes that a retinotopic area computes a priority map, through the interaction of bottom-up feature-selective cells with a top-down target-specific modulatory signal and local normalization. An implementation of this model can locate complex target objects embedded in multi-object arrays and even in natural images (Fig. 2). The model's performance also shows a significant degree of concordance with human behavioral performance in a relatively difficult object search task (Fig. 4). The single-trial agreement between model and subjects extends to trials where the target was absent or where both the model and the subjects made a mistake (Figs 5 and 6).

The proposed model integrates visual search with object recognition computations instantiated in the ventral visual stream. There are 2 different but interconnected aspects of the "ventral stream" that are part of the model. First, there are early/mid-level features computed through the S1/S2 layers (Fig. 1). Once the model selects a location, it needs to recognize what is in that location to decide whether to continue searching in Figure 2*J*,*L* (but not in the "Oracle" version presented in Fig. 2*I*,*K*). This recognition component is instituted through the S3/C3 layers depicted in Figure 1.

The proposed model uses a specific type of feature, namely shape selectivity (Serre, Kreiman et al. 2007). Clearly, these are not the only features that guide visual search. Many feature types are known to guide visual search, including color, orientation, and local contrast among others (see Wolfe and Horowitz (2004) for a review). Shape is an important and ubiquitous component of visual search (e.g., consider searching for a face in a crowd), and several studies have shown that shapes can influence visual search at the behavioral and physiological level (Bichot et al. 2005; Zhou and Desimone 2011; Baldauf and Desimone 2014). Additionally, visual areas which feature prominently in the conceptual motivation for the computational steps in the model show selectivity to shape features; these areas include V4 and ITC (Logothetis and Sheinberg 1996; Connor et al. 2007) as well as FEF and LIP (Sereno and Maunsell 1998; Fitzgerald et al. 2011; Lehky and Sereno 2007).

Model performance cannot be explained by an overly easy task, as shown by the fact that human performance was far from perfect under the rapid search conditions studied here (Fig. 4). Also, subjects showed a significant degree of self-consistency and between-subject consistency (Figs 5 and 6) but this consistency was far from 100%, reflecting considerable trial-to-trial and subject-to-subject variability. The agreement between the model and behavioral performance is also significantly above chance but far from 100%. The degree of between-subject consistency bounds how well we can expect the model to perform: It seems unrealistic to expect a model to be a better predictor of human behavior than the performance of other humans themselves. Thus model-subject agreement should be evaluated in comparison with between-subjects agreement, as shown in Figures 5 and 6.

This model is inspired and constrained by current physiological knowledge about the macaque visual system. To search for a given object, we use the pattern of activity at the highest stages of visual processing, represented by the activity of C2b cells (left panel in Fig. 1), which are meant to mimic the output of bottom-up visual information processing along the ventral visual stream (Serre, Kreiman et al. 2007). Target-specific information interacts with bottom-up responses in an attentional area that we tentatively associate with LIP and FEFs as described by equation 5. FEF and LIP have strong reciprocal connections and are involved in controlling both covert and overt visual attention. There is significant evidence implicating LIP in the generation of the priority map (Bisley and Goldberg 2010; Bisley 2011). FEF is known to topographically project to V4 (Barone et al. 2000), and these connections can produce effects similar to those of attention, such as selective gating (Moore and Armstrong 2003). The top-down, target-specific modulatory signal on this area is presumed to originate from object-selective cells in a higher area. Prefrontal cortex (especially ventral lateral PFC) is a strong candidate for the source of the target-specific modulatory signal, since it is known to encode target identity in visual memory experiments (Wilson et al. 1993; Rao et al. 1997). PFC also receives connections from object-selective visual areas such as IT and is also connected to LIP (Medalla and Barbas 2006).

The proposed model focuses on the effect of target shape (the "Waldo-specific" part of "where is Waldo"). Under natural conditions, visual search is influenced by many effects that are not considered in the current model. First, the model does not take into account the loss of acuity and crowding effects in the periphery (e.g., Freeman and Simoncelli (2011)) and the remapping of visual space just before saccades (Bisley and Goldberg 2010). Thus, when comparing the model and humans, we deliberately chose settings that minimized these additional, target-independent effects: We arranged all items on a circle around the fixation point (to ensure maximum symmetry), we only considered the first saccade in Figures 5 and 6 (because after this the symmetry is broken), and we randomized object position between presentations (so as to remove residual "directional" effects, e.g., preference for horizontal directions [Foulsham et al. 2008] or central fixations [Tatler et al. 2006]). Second, the model does not consider color; both visual search and visual recognition can take place with grayscale images (behaviorally and computationally) but color can enhance both tasks. Third, the object array experiment is clearly artificial and does not incorporate the

complexities and cues present in natural scenes. Figure 2 shows that the model can also detect target objects in natural scenes. Yet, the model does not incorporate many cues provided by natural scenes which play an important role in visual search under natural conditions (e.g., Vo and Henderson (2010); Wolfe, Vo et al. (2011)). For example, Vo and colleagues investigated the effects of a short glimpse of a scene on subsequent visual search. Such effects reflect several possible components, including memory and "scene gist" (large-scale properties of visual scenes that influence the direction of search independently of, or in interaction with, target identity [Oliva and Torralba 2006]). Integrating these effects with the target-dependent guidance modeled here will be an important step toward a full understanding of visual search.

The model posits that search terminates after a fixed maximum number of fixations if the target has not been found. Thus, the model leaves open the question of how subjects decide when to give up and abort the search. Explaining and modeling search termination will also be an important component of a complete model of visual search.

Visual search requires computing the match between bottom-up input signals and target modulation. Our experiments suggest the importance of local normalization for successful search, by preventing objects or areas with high bottom-up activity across all features from controlling the priority map (Fig. 3*G*, *H*). Equation 5 resembles a normalized dot product, which measures the alignment between 2 vectors independently of their magnitude. This approximation would be exact (neglecting the spatially constant magnitude of the top-down signal) if the denominator in equation 5 was strictly equal to the Euclidean norm of local activity. The interaction between normalization and modulation thus solves the problem of "pay[ing] attention to stimuli because they are significant, not simply because they are intense" (Abbott 2005), adding a new function to the many known computational properties of normalization in cortex (Carandini and Heeger 2011).

The top-down signals depend on the ratio between response to target and response to a "mean" stimulus (average response over many unrelated natural images, eq. 4). This is essentially similar to Navalpakkam and Itti's (2007) proposal to base top-down modulation on the signal-to-noise ratio (ratio of activations) of targets versus distractors. The main difference is that in the current instantiation, "distractors" are unpredictable and approximated by a large set of unrelated natural images. Thus, the current work shows that Navalpakkam and Itti's proposal works not only with simple features such as orientation and color, but also with more complex shape components of natural objects, and can be implemented using a normalization operation which constitutes a canonical computation across neocortex (Carandini and Heeger 2011). It also generalizes these findings to a situation where the distractors are unpredictable, by using a "mean expected stimulus" as replacement for the unknowable distractors.

The model postulates that the top-down, target-related, feature-based signals (used to compute the priority map) do not target the ventral visual-processing system (used for target presence validation), as suggested by recent neurophysiological evidence (Martinez-Trujillo 2011; Zhou and Desimone 2011). A possible conceptual advantage of such a model, in comparison with models positing direct feature-based modulation of early visual inputs such as V1 or V4 (Lanyon and Denham 2004; Hamker 2005; Chikkerur et al. 2010), is that it allows the ventral visual pathway to operate on a feature-faithful input, without the interference caused by feature biasing. Biasing input cells in a spatially uniform way according to their preference for the target features (e.g., "red" or "vertical") would have the effect of making "everything" look more like the target (e.g., more "red" or more "vertical"). In the proposed model, the higher areas of the ventral pathway receive an undisturbed input, except for the purely spatial, feature-neutral bias introduced by attentional modulation, which actually restores fine object selectivity in higher visual areas in the face of clutter (Zhang et al. 2011). This type of modulation facilitates successful detection in multi-object images and natural-background images.

Attention can be either overt (with eye movements) or covert (without eye movements). Our model seeks to explain the computation of the priority map in "attention-controlling" areas (tentatively linked to LIP/FEF in the scheme in Fig. 1), which are thought to control both covert and overt attention. Although the human experiments involved an eye movement task, the model itself is agnostic about whether attentional selection occurs covertly or overtly: The model seeks to explain how the brain computes where to allocate attention, whether covert or overt.

It is conceivable that, before the first eye movement, observers covertly allocated attention to various items, processing them in turn and selecting one for actual fixation. Even though subjects were asked to respond as fast as possible (reflected in the rapid reaction time, median = 237 ms, first fixation success rate of 66%), the possibility of multiple covert attentional shifts before the first fixation cannot be excluded. The testing protocol, urging subjects to perform rapid behavioral responses, can be expected to reduce (but not necessarily eliminate) the number of covert attentional shifts before the first fixation. The psychophysics results provide a coarse bound on the number of possible covert attentional shifts before 237 ms. In a simpler saccade task, involving only 2 possible alternatives with large stimuli, fixed targets, and well-known stimuli (animals/faces), Kirchner and Thorpe (2006) report a median reaction time of 228 ms, only slightly faster than the median for the first fixation in our study (where there were 6 alternatives, smaller stimuli, different targets in every trial, and different objects). The purely visual recognition and motor aspects of the task require at least 100–150 ms (Thorpe et al. 1996; Kirchner and Thorpe 2006; Liu et al. 2009; Agam et al. 2010). Under situations involving serial visual search as opposed to pop-out, several investigators have shown that reaction times increase with object array set size, with slopes ranging from ~30 to ~90 ms per object (e.g., Treisman and Gelade (1980); Horowitz and Wolfe (1998); Vickery et al. (2005)). The model shows a decrease in performance with increasing the number of objects; conversely, the model requires more fixations to achieve the same performance level when there are more objects in the array (compare Fig. 4C vs. Fig. 2*J*; see also lower performance for objects embedded in natural backgrounds in Fig. 2*L*). Given the visual/motor constraints, the search cost per object and the rapid reaction times, a long sequence of covert attentional selections should produce considerably longer reaction times. We estimate that our reaction times allows for at most 2–3 covert attentional shifts under ultra-rapid conditions (~150 ms vision/motor cost + 2 × 40 ms per object = 230 ms; ~150 ms vision/motor + 3 × 30 ms per object = 230 ms). Additionally, target-absent trials and target-present trials produced similar latency distributions (Supplementary Fig. 4). Under conditions in which subjects are asked to prioritize accuracy and are given ample time to locate the target, target-absent trials produce a rough doubling of the slope of search time vs. number of objects (Wolfe and Horowitz 2004). In sum, while it is conceivable that subjects might covertly shift their attention before their first fixation, the results presented here present an upper bound on the number of such shifts.

In conclusion, our results show that a physiologically motivated model can localize target objects in arrays and natural images. The model not only matches human performance in visual search but also predicts the first fixations in target-absent situations and errors in single trials. In combination with other models purporting to explain the mechanisms of attentional effects in lower visual areas (e.g., Hamker and Zirnsak (2006); Borgers et al. (2008); Womelsdorf et al. (2008); Reynolds and Heeger (2009); Miconi and VanRullen (2011)), this model can provide a component of a global mechanistic understanding of attentional selection in the brain.

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/.

Funding

This work was supported by NIH and NSF.

Notes

Conflict of Interest: None declared.

References

Abbott LF. 2005. Where are the switches on this thing? In: van Hemmen J, Sejnowski T, editors. 23 Problems in Systems Neuroscience. Oxford University Press.

Agam Y, Liu H, Papanastassiou A, Buia C, Golby AJ, Madsen JR, Kreiman G. 2010. Robust selectivity to two-object images in human visual cortex. Curr Biol. 20:872–879.

Angelucci A, Bullier J. 2003. Reaching beyond the classical receptive field of V1 neurons: horizontal or feedback axons? J Physiol Paris. 97:141–154.

Baldauf D, Desimone R. 2014. Neural mechanisms of object-based attention. Science. 344:424–427.

Barone P, Batardiere A, Knoblauch K, Kennedy H. 2000. Laminar distribution of neurons in extrastriate areas projecting to visual areas V1 and V4 correlates with the hierarchical rank and indicates the operation of a distance rule. J Neurosci. 20:3263–3281.

Beutter BR, Eckstein MP, Stone LS. 2003. Saccadic and perceptual performance in visual search tasks. I. Contrast detection and discrimination. J Opt Soc Am A Opt Image Sci Vis. 20:1341–1355.

Bichot NP, Rossi AF, Desimone R. 2005. Parallel and serial neural mechanisms for visual search in macaque area V4. Science. 308:529–534.

Bisley J. 2011. The neural basis of visual attention. J Physiol. 589:49–57.

Bisley J, Goldberg M. 2010. Attention, intention, and priority in the parietal lobe. Annu Rev Neurosci. 33:1–21.

Borgers C, Epstein S, Kopell NJ. 2008. Gamma oscillations mediate stimulus competition and attentional selection in a cortical network model. Proc Natl Acad Sci USA. 105:18023–18028.

Buschman TJ, Miller EK. 2009. Serial, covert shifts of attention during visual search are reflected by the frontal eye fields and correlated with population oscillations. Neuron. 63:386–396.

Carandini M, Heeger DJ. 2011. Normalization as a canonical neural computation. Nat Rev Neurosci. 13:51–62.

Carrasco M. 2011. Visual attention: the past 25 years. Vision Res. 51:1484–1525.

Chikkerur S, Serre T, Tan C, Poggio T. 2010. What and where: A bayesian inference theory of attention. Vision Res. 50:2233–2247.

Connor CE, Brincat SL, Pasupathy A. 2007. Transformation of shape information in the ventral pathway. Curr Opin Neurobiol. 17:140–147.

Desimone R. 1998. Visual attention mediated by biased competition in extrastriate visual cortex. Philos Trans R Soc Lond B Biol Sci. 353:1245–1255.

Eckstein MP, Thomas JP, Palmer J, Shimozaki SS. 2000. A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. Percept Psychophys. 62:425–451.

Findlay JM. 1997. Saccade target selection during visual search. Vision Res. 37:617–631.

Fitzgerald JK, Freedman DJ, Assad JA. 2011. Generalized associative representations in parietal cortex. Nat Neurosci. 14(8):1075–1079.

Foulsham T, Kingstone A, Underwood G. 2008. Turning the world around: patterns in saccade direction vary with picture orientation. Vision Res. 48:1777–1790.

Freeman J, Simoncelli EP. 2011. Metamers of the ventral stream. Nat Neurosci. 14:1195–1201.

Fukushima K. 1980. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern. 36:193–202.

Hamker FH. 2005. The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement. Cereb Cortex. 15:431–447.

Hamker FH, Zirnsak M. 2006. V4 receptive field dynamics as predicted by a systems-level model of visual attention using feedback from the frontal eye field. Neural Netw. 19:1371–1382.

Horowitz TS, Wolfe JM. 1998. Visual search has no memory. Nature. 394:575–577.

Hung C, Kreiman G, Poggio T, DiCarlo J. 2005. Fast read-out of object identity from macaque inferior temporal cortex. Science. 310:863–866.

Itti L, Koch C. 2000. A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res. 40:1489–1506.

Itti L, Koch C. 2001. Computational modelling of visual attention. Nat Rev Neurosci. 2:194–203.

Kirchner H, Thorpe SJ. 2006. Ultra-rapid object detection with saccadic eye movements: visual processing speed revisited. Vision Res. 46:1762–1776.

Klein RM. 2000. Inhibition of return. Trends Cogn Sci. 4:138–147.

Koch C, Ullman S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol. 4:219–227.

Kouh M. 2007. Toward a More Biologically Plausible Model of Object Recognition. MIT. 113 p.

Kowler E. 2011. Eye movements: the past 25 years. Vision Res. 51:1457–1483.

Lanyon LJ, Denham SL. 2004. A model of active visual search with object-based attention guiding scan paths. Neural Netw. 17:873–897.

Lehky SR, Sereno AB. 2007. Comparison of shape encoding in primate dorsal and ventral visual pathways. J Neurophysiol. 97:307–319.

Li N, Cox DD, Zoccolan D, DiCarlo JJ. 2009. What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? J Neurophysiol. 102:360–376.

Liu H, Agam Y, Madsen JR, Kreiman G. 2009. Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex. Neuron. 62:281–290.
Logothetis NK, Sheinberg DL. 1996. Visual object recognition. Annu Rev Neurosci. 19:577–621.
Martinez-Trujillo JC. 2011. Searching for the neural mechanisms of feature-based attention in the primate brain. Neuron. 70:1025–1028.
Martinez-Trujillo JC, Treue S. 2004. Feature-based attention increases the selectivity of population responses in primate visual cortex. Curr Biol. 14:744–751.
Medalla M, Barbas H. 2006. Diversity of laminar connections linking periarcuate and lateral intraparietal areas depends on cortical structure. Eur J Neurosci. 23:161–179.
Miconi T, VanRullen R. 2011. A feedback model of attentional effects in the visual cortex. IEEE Symposium on Computational intelligence for Multimedia, Signal and Vision Processing (CIMSIVP):1–8.
Monosov IE, Sheinberg DL, Thompson KG. 2010. Paired neuron recordings in the prefrontal and inferotemporal cortices reveal that spatial selection precedes object identification during visual search. Proc Natl Acad Sci USA. 107: 13105–13110.
Moore T, Armstrong KM. 2003. Selective gating of visual signals by microstimulation of frontal cortex. Nature. 421:370–373.
Najemnik J, Geisler WS. 2005. Optimal eye movement strategies in visual search. Nature. 434:387–391.
Navalpakkam V, Itti L. 2005. Modeling the influence of task on attention. Vision Res. 45:205–231.
Navalpakkam V, Itti L. 2007. Search goal tunes visual features optimally. Neuron. 53:605–617.
Noudoost B, Chang N, Steinmetz N, Tirin Moore. 2010. Top-down control of visual attention. Curr Opin Neurobiol. 20: 183–190.
O'Craven KM, Downing PE, Kanwisher N. 1999. fMRI evidence for objects as the units of attentional selection. Nature. 401:584–587.
Oliva A, Torralba A. 2006. Building the gist of a scene: the role of global image features in recognition. Prog Brain Res. 155:23–36.
Peelen MV, Kastner S. 2011. A neural basis for real-world visual search in human occipitotemporal cortex. Proc Natl Acad Sci USA. 108:12125–12130.
Puri AM, Wojciulik E, Ranganath C. 2009. Category expectation modulates baseline and stimulus-evoked activity in human inferotemporal cortex. Brain Res. 1301:89–99.
Rao RP, Zelinsky GJ, Hayhoe MM, Ballard DH. 2002. Eye movements in iconic visual search. Vision Res. 42:1447–1463.
Rao S, Rainer G, Miller E. 1997. Integration of what and where in the primate prefrontal cortex. Science. 276:821–824.
Reynolds JH, Heeger DJ. 2009. The normalization model of attention. Neuron. 61:168.
Riesenhuber M, Poggio T. 1999. Hierarchical models of object recognition in cortex. Nat Neurosci. 2:1019–1025.
Sereno AB, Maunsell JH. 1998. Shape selectivity in primate lateral intraparietal cortex. Nature. 395:500–503.
Serre T. 2006. Learning a Dictionary of Shape-Components in Visual Cortex: Comparison with Neurons, Humans and Machines. MIT.
Serre T, Kouh M, Cadieu C, Knoblich U, Kreiman G, Poggio T. 2005. A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. In. Boston: MIT. p CBCL Paper #259/AI Memo #2005–2036.
Serre T, Kreiman G, Kouh M, Cadieu C, Knoblich U, Poggio T. 2007. A quantitative theory of immediate visual recognition. Prog Brain Res. 165C:33–56.
Serre T, Oliva A, Poggio T. 2007. Feedforward theories of visual cortex account for human performance in rapid categorization. PNAS. 104:6424–6429.
Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T. 2007. Robust object recognition with cortex-like mechanisms. IEEE Trans Pattern Anal Mach Intell. 29:411–426.
Sheinberg DL, Logothetis NK. 2001. Noticing familiar objects in real world scenes: the role of temporal cortical neurons in natural vision. J Neurosci. 21:1340–1350.
Shmuel A, Korman M, Sterkin A, Harel M, Ullman S, Malach R, Grinvald A. 2005. Retinotopic axis specificity and selective clustering of feedback projections from V2 to V1 in the owl monkey. J Neurosci. 25:2117–2131.
Tatler BW, Baddeley RJ, Vincent BT. 2006. The long and the short of it: spatial statistics at fixation vary with saccade amplitude and task. Vision Res. 46:1857–1862.
Tatler BW, Hayhoe MM, Land MF, Ballard DH. 2011. Eye guidance in natural vision: reinterpreting salience. J Vis. 11:5.
Tavassoli A, Linde I, Bovik AC, Cormack LK. 2009. Eye movements selective for spatial frequency and orientation during active visual search. Vision Res. 49:173–181.
Thorpe S, Fize D, Marlot C. 1996. Speed of processing in the human visual system. Nature. 381:520–522.
Treisman AM, Gelade G. 1980. A feature-integration theory of attention. Cognit Psychol. 12:97–136.
Treue S, Martinez-Trujillo JC. 1999. Feature-based attention influences motion processing gain in macaque visual cortex. Nature. 399:575–579.
Tsotsos J, Culhane S, Wai W, Lai Y, Davis N, Nuflo F. 1995. Modeling visual attention via selective tuning. Artif Intell. 78: 507–545.
Vickery TJ, King LW, Jiang Y. 2005. Setting up the target template in visual search. J Vis. 5:81–92.
Vincent BT, Baddeley RJ, Troscianko T, Gilchrist ID. 2009. Optimal feature integration in visual search. J Vis. 9:15 11–11.
Vo ML, Henderson JM. 2010. The time course of initial scene processing for eye movement guidance in natural scene search. J Vis. 10:14 11–13.
Wallis G, Rolls ET. 1997. Invariant face and object recognition in the visual system. Prog Neurobiol. 51:167–194.
Walther D, Koch C. 2006. Modeling attention to salient proto-objects. Neural Netw. 19:1395–1407.
Walther DB, Koch C. 2007. Attention in hierarchical models of object recognition. Prog Brain Res. 165:57–78.
Williams LG. 1967. The effects of target specification on objects fixated during visual search. Acta Psychol (Amst). 27:355–360.
Wilson F, O'Scalaidhe S, Goldman-Rakic P. 1993. Dissociation of object and spatial processing domains in primate prefrontal cortex. Science. 260:1955–1958.
Wolfe J, Alverez GA, Rosenholtz R, Kuzmova YI, Sherman AM. 2011. Visual search for arbitrary objects in real scenes. Atten Percept Psychophys. 73:1650–1671.
Wolfe JM. 2007. Guided search 4.0: Current progress with a model of visual search. In: Gray W, editor. Integrated models of cognitive systems. NewYork: Oxford, pp. 99–119.
Wolfe JM, Horowitz TS. 2004. What attributes guide the deployment of visual attention and how do they do it? Nat Rev Neurosci. 5:495–501.
Wolfe JM, Vo ML, Evans KK, Greene MR. 2011. Visual search in scenes involves selective and nonselective pathways. Trends Cogn Sci. 15:77–84.

Womelsdorf T, Anton-Erxleben K, Treue S. 2008. Receptive field shift and shrinkage in macaque middle temporal area through attentional gain modulation. J Neurosci. 28: 8934–8944.

Yu AJ, Dayan P. 2005. Uncertainty, neuromodulation, and attention. Neuron. 46:681–692.

Zelinsky GJ. 2008. A theory of eye movements during target acquisition. Psychol Rev. 115:787–835.

Zhang Y, Meyers EM, Bichot NP, Serre T, Poggio TA, Desimone R. 2011. Object decoding with attention in inferior temporal cortex. Proc Natl Acad Sci USA. 108:8850–8855.

Zhou H, Desimone R. 2011. Feature-based attention in the frontal eye field and area V4 during visual search. Neuron. 70:23.

Zoccolan D, Kouh M, Poggio T, DiCarlo JJ. 2007. Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. J Neurosci. 27:12292–12307.