A saliency map in primary visual cortex

Zhaoping Li

I propose that pre-attentive computational mechanisms in primary visual cortex create a saliency map. This map awards higher responses to more salient image locations; these responses are those of conventional V1 cells tuned to input features, such as orientation and color. Hence no separate feature maps, or any subsequent combination of them, is needed to create a saliency map. I use a model to show that this saliency map accounts for the way that the relative difficulties of visual search tasks depend on the features and spatial configurations of targets and distractors. This proposal links psychophysical behavior to V1 physiology and anatomy, and thereby makes testable predictions.

Given that visual input contains large amounts of data and that the visual apparatus has limited computational resources, it is necessary to limit detailed processing to selected aspects of the input. It is computationally efficient for much of this selection to be carried out by bottom-up mechanisms. To understand the selection process better, we separate bottom-up from top-down mechanisms [1] and consider a saliency map of the visual field constructed by bottom-up mechanisms only, such that a location with a higher scalar value in this map is more likely to attract attention and be further processed. Given its function, the degree of salience at a visual location should be irrespective of the actual feature (e.g. color, depth or orientation) at that location (see Box 1). Hence, the salience of a red spot at one location could be compared with, say, that of a black vertical bar at another location [2]. The map should also depend essentially on the organization of the visual scene. Such a saliency map would have obvious significance for visual functions.

'...firing rates of V1's output neurons increase monotonically with the salience values of the visual input...'

Various groups have hypothesized that a saliency map can be built by combining information from a collection of separate feature maps, representing single visual features, such as color or orientation, across the input [3–6]. The feature maps are combined into a single, master map to represent salience irrespective of the actual features. For instance, Koch and Ullman [5], and later Itti *et al.* [6], suggested that each individual feature map represents the spatial contrast in its associated feature in the input (via a competitive process), and that the outputs from these feature maps are summed to give a master saliency map. In these theories, neither the neural mechanisms nor the exact underlying cortical area responsible for the saliency map have been clearly specified.

Building on previous work [1,3–7], I have recently proposed that primary visual cortex (V1) provides a saliency map such that, for a given visual scene, firing rates of V1's output neurons increase monotonically with the salience values of the visual input in the corresponding classical receptive fields (CRFs) [8–10]. The pyramidal cells in layers 2 and 3, which provide V1 outputs, receive visual inputs (via feedforward pathways) within their relatively small CRFs. Horizontal intracortical interactions make each

Box 1. Signaling saliency regardless of features

Contrary to how it might sound, signaling regardless of features does not mean that the cells reporting salience must be *untuned* to specific features. In other words, here 'regardless of' means that in this saliency map, the meaning of firing rates for saliency is universal, and, given an input scene, the same firing rate from two V1 (output) neurons selective to different features means the same salience value of the two corresponding inputs. Thus, one of the cells might be color selective, responding to a static red bar, and the other cell tuned to motion, responding to a moving black dot, but their salience value is the same.

Usually, an image item, say, a short red bar, evokes responses from many cells with different optimal features and overlapping tuning curves or CRFs. The actual input features have to be decoded in a complex and feature-specific manner from the population responses [a]. However, locating the most responsive cell to a scene by definition locates the most salient item whether or not features can be decoded beforehand or simultaneously from the same cell population. It is economical *not* to use subsequent cell layers or visual areas (whether the cells are feature tuned or not) for a saliency map; the small CRFs in V1 layers 2 and 3 also mean that this saliency map can have a higher resolution.

Ultimately, our proposal needs to be validated by examining whether the responses of feature-selective cells in V1 *do* indeed signal saliences (even though input features could be decoded from responses of the very same cell population). For simplicity (and without loss of generality), we assume that only a single V1 cell responds to inputs within its CRF, unless otherwise stated.

Reference

a Dayan, P. and Abbott, L. (2001) *Theoretical Neuroscience*, MIT Press

Zhaoping Li
Dept of Psychology, University College London, Gower Street, London, UK WC1E 6BT.
e-mail: z.li@ucl.ac.uk

 1364-6613/02/$ – see front matter PII: S1364-6613(00)01817-9

Box 2. A biologically based V1 model to simulate the saliency map

Our model focuses on the part of V1 responsible for contextual influences: layer 2–3 pyramidal cells, interneurons, and horizontal intracortical connections [a–d]. Pyramidal cells and interneurons interact with each other locally and reciprocally. A pyramidal cell can excite other pyramidal cells monosynaptically, or inhibit them disynaptically, by projecting to the relevant inhibitory interneurons. General and local normalization of activities are also included in the model [e].

V1 transforms input to output such that the activity of each cell depends on both its direct input (taken to be the stimuli within its classical receptive field, CRF), and the contextual stimuli outside the CRF. The centers of the CRFs are uniformly distributed in space. The preferred orientations of the cells at a given location span 180°. Images are filtered by edge- or bar-like local CRFs. The results of this processing form the direct inputs to the model excitatory pyramidal cells. The graded responses of the pyramidal cells are initially determined by the direct visual inputs within their CRFs, and are then quickly modulated by contextual influences coming from intracortical interactions (Fig. I). The temporal averages of the responses of the pyramidal cells are the outputs of the model, and report the results of V1 processing. The horizontal connections are designed:

(1) to be consistent with the V1 anatomy, linking cells that prefer similar orientations [a,b] and projecting along the axes corresponding to the preferred orientations of the pre-synaptic cells [c];

(2) such that the resulting model consistently reproduces the usual phenomena of contextual influence that are observed physiologically: iso-orientation and general surround suppression, and contour enhancement [h,i]. (Refs [f,g,j,k] contain more details of the model, including all the model parameters necessary to reproduce our results and discussion of the computational role played by V1 in pre-attentive visual tasks, such as contour enhancement and texture segmentation.)

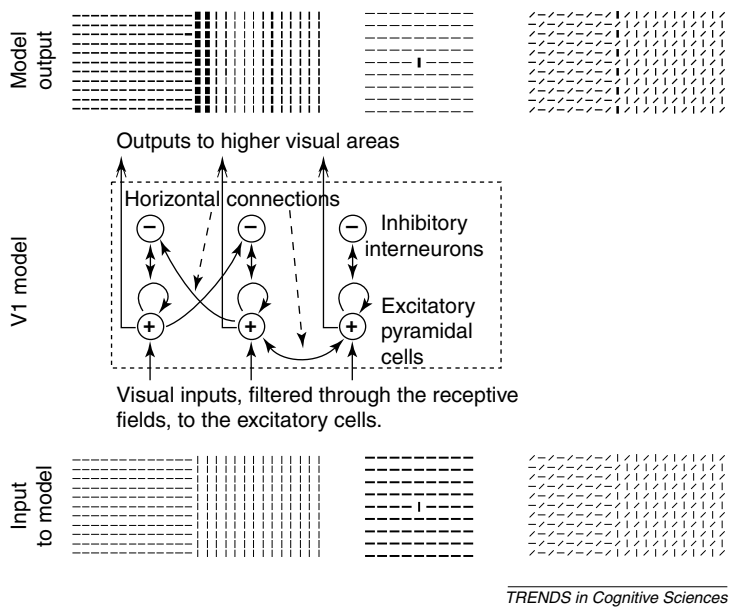

Fig. I. The V1 model and its function. Shown are three input images (bottom) to the model, and their output response maps (top). The thicknesses of the bars in each plot are made to be proportional to their input/output strengths, for the purposes of visualization. The input strength of each bar is determined by its contrast. Note that every input bar in each of these three images has the same contrast. A principal (pyramidal) cell can only receive direct visual input from an input bar in its receptive field. The output responses depend on the input contrasts and on the contextual stimuli of each bar.

References

a Rockland, K.S. and Lund J.S. (1983) Intrinsic laminar lattice connections in primate visual cortex. *J. Comp. Neurol.* 216, 303–318

b Gilbert, C.D. and Wiesel, T.N. (1983) Clustered intrinsic connections in cat visual cortex. *J. Neurosci.* 3, 1116–1133

c Fitzpatrick, D. (1996) The functional organization of local circuits in visual cortex: insights from the study of tree shrew striate cortex. *Cereb. Cortex* 6, 329–341

d Douglas, R.J. and Martin, K.A. (1990) Neocortex. In *Synaptic Organization of the Brain* (3rd edn), (Shepherd, G.M., ed.), pp. 389–438, Oxford University Press

e Heeger, D.J. (1992) Normalization of cell responses in cat striate cortex. *Visual Neurosci.* 9, 181–197

f Li, Z. (1999) Visual segmentation by contextual influences via intracortical interactions in primary visual cortex. In *Netw. Comput. Neural Syst.* 10, 187–212

g Li, Z. (1998) Primary cortical dynamics for visual grouping. In *Theoretical Aspects of Neural Computation* (Wong, K.Y.M. *et al.*, eds), pp. 155–164, Springer-Verlag

h Knierim, J.J. and van Essen, D.C. (1992) Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophysiol.* 67, 961–980

i Kapadia, M.K. *et al.* (1995) Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys. *Neuron* 15, 843–856

j Li, Z. (1998) A neural model of contour integration in the primary visual cortex. *Neural Comput.* 10, 903–940

k Li, Z. (2000) Pre-attentive segmentation in the primary visual cortex. *Spat. Vis.* 13, 25–50

pyramidal cell's response dependent on both the input strength within its CRF and the contextual stimuli, thus mediating the computation of salience. For instance, bars having the same input contrast in an image can evoke different V1 responses depending on their relative positions and orientations (see Box 2), and thus can have different saliences. It is well known that each V1 cell can be tuned to one or more feature dimensions, such as orientation, scale, color, motion and depth. Whereas the cells' 'identities' (the labeled lines to higher visual centers) code the features and locations of the underlying stimuli, according to our proposal, the cells' firing rates report the stimuli's saliences regardless of the actual features represented by the cells. Hence, according to our proposal, no separate feature maps, or indeed any

Box 3. Brief overview of visual search

Visual search is the task of finding a target item among distractors in a visual input [a–f] (Fig. I). Reaction time to find a target typically increases with the number of distractors. The rate of increase (Fig. Ie) characterizes the ease or efficiency of the search. When the target has a feature that is absent in the distractors, the search can be very efficient when the feature is in a basic feature dimension such as color, orientation, depth or motion direction. Such a search is termed a feature search [a] (e.g. Fig. Ia).

When the target is only distinguishable by a particular conjunction of features (e.g. green and vertical in Fig. Ib), each of which is present in the distractors, the search is termed a conjunction search. Some conjunction searches are very difficult, like the example shown here, and others can be easy [g,h]. When the target is characterized by lacking a feature that is present in the distractors (e.g. Fig. Ic), the search is more difficult than a feature search. Dissimilarity between distractors and similarity between the distractors and the target make searches more difficult [d] (e.g. Fig. Id).

Visual search asymmetry occurs when a single target item A among a background of distractor items B is more difficult to find than a target item B among distractor items A [f] (see Figs 2 and 4 in main article for examples). Earlier work [a,b] categorized searches into efficient or parallel searches and inefficient or serial searches, and were presumed to be carried out by pre-attentive and attentive mechanisms respectively. Later work found a continuum of search difficulties, depending on the inputs [c–e].

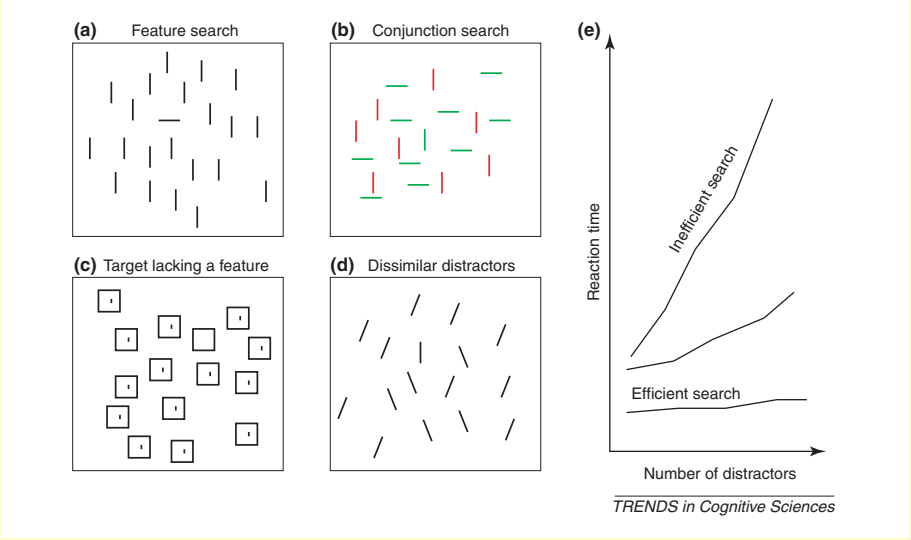

Fig. I. Examples and schematic illustrations of visual search tasks. (a) The target horizontal bar differs from the distractor bars in orientation. (b) The vertical green bar (target) is distinguished from distractor (red vertical or green horizontal) bars by the conjunction of green and vertical features. (c) The target square is without an inner dot. (d) The target is the vertical bar, harder to find because the distractors have dissimilar orientations. (e) The ease or efficiency of the search tasks are often characterized by the slope of the plot of reaction time versus the number of distractors. A steeper slope is taken to indicate a more difficult (inefficient) search.

References

a Treisman, A. and Gelade, G. (1980) A feature integration theory of attention. *Cogn. Psychol.* 12, 97–136
b Julesz, B. (1981) Textons, the elements of texture perception, and their interactions. *Nature* 290, 91–97
c Wolfe, J.M. *et al.* (1989) Guided search: an alternative to the feature integration model for visual search. *J. Exp. Psychol.* 15, 419–433
d Duncan, J. and Humphreys, G. (1989) Visual search and stimulus similarity. *Psychol. Rev.* 96, 1–26
e Wolfe, J.M. (1998) Visual search. In *Attention* (Pashler, H., ed.), pp, 13–74, Psychology Press
f Treisman, A. and Gormican, S. (1988) Feature analysis in early vision: evidence for search asymmetries. *Psychol. Rev.* 95, 15–48
g Nakayama, K. and Silverman, G.H. (1986) Serial and parallel processing of visual feature conjunctions. *Nature* 320, 264–265
h Mcleod, P. *et al.* (1988) Visual search for a conjunction of movement and form is parallel. *Nature* 332, 154–155

subsequent combination of them, is needed to produce a saliency map. In this article, we show that being specific about the underlying neural mechanisms allows us to gain substantial insight into experimental data by linking psychophysical results with V1 physiology and anatomy.

Simulating the saliency map using a V1 model

To test our proposal, we simulate our saliency map using the only existing biologically based V1 model (Box 2) of contextual influences that performs all three tasks of texture segmentation [11], contour enhancement [12] and pop out. We apply the map to visual search tasks (briefly overviewed in Box 3), assuming that the ease of performing each task is determined by the salience of the search target in the task. The model is constructed using known anatomical and physiological facts about the neural elements and horizontal intra-cortical interactions in V1, for instance, that the horizontal connections tend to link V1 cells tuned to similar orientations [13–16]. The same horizontal connection strengths (and all other model parameters) are used for all examples in this article; no special tailoring has been applied.

The model reproduces the known excitatory and inhibitory contextual influences observed in V1 physiology associated with contour integration and texture segmentation [17,18]. In particular, a model cell's response to a high contrast bar within its CRF is suppressed when the bar is surrounded by other bars of the same orientation (iso-orientation suppression). This contextual suppression is reduced when the contextual bars are oriented randomly, and is reduced further when the contextual bars are oriented orthogonally to the bar within the CRF [17]. However, as is also physiologically observed [18], the model cell's response is enhanced when the bar within the CRF and the contextual bars link to form a smooth

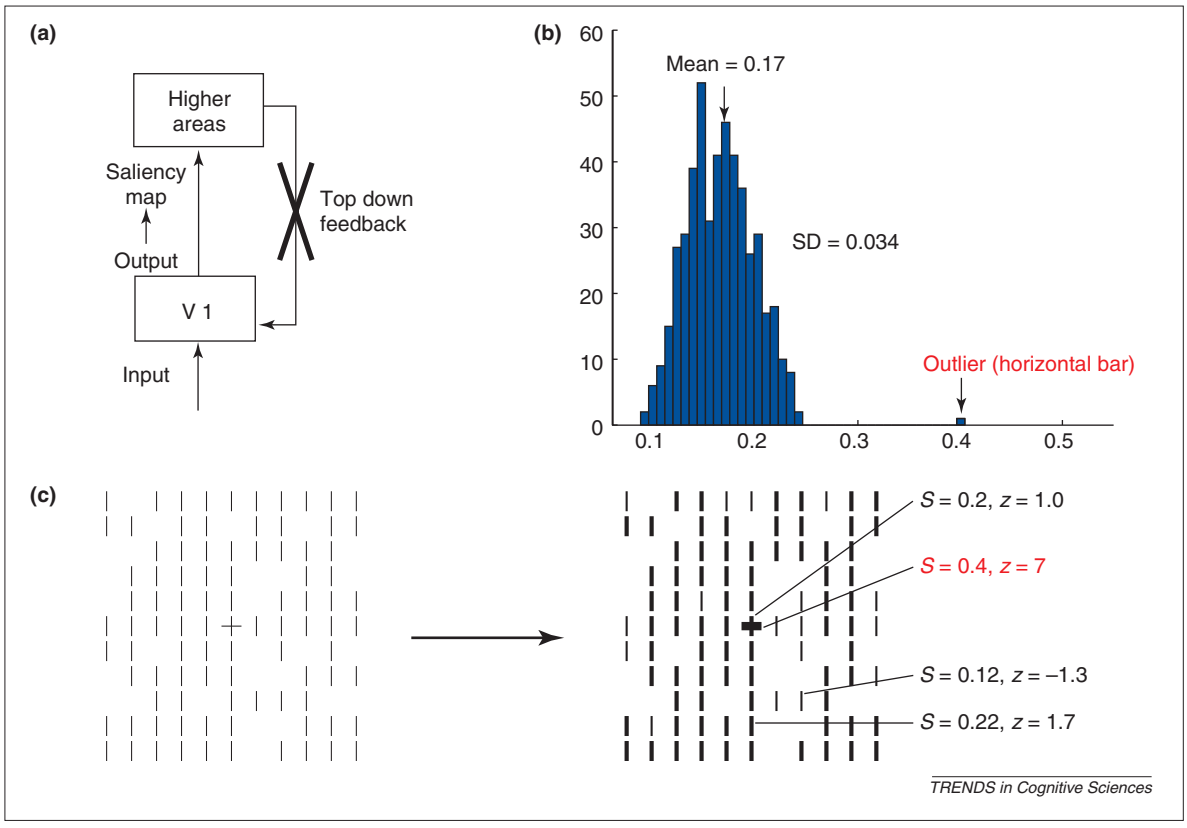

Fig. 1. (a) V1's output as a saliency map is viewed under the idealization of the top-down feedback to V1 being disabled. (b) Histogram of responses S_i to all bars, i, in an input image, a small part of which is plotted in (c). Given a cell response, S_i, the saliency order of the inducing stimulus is assessed by relating to the mean $\bar{S}$ and standard deviation σ_s of the population of responses, $z_i = (S_i - \bar{S})/\sigma_s$. (c) An example of the output (right) from the V1 model, given an input stimulus (left) of horizontal and vertical bars of equal contrast. The input simulates the search task of finding a target cross among distractor vertical bars. The thicknesses of the bars in each plot (as in all the figures in this article) are made to be proportional to their input/output strengths, for the purposes of visualization. The model output values, S_i, and saliency measures, z_i, are given for four bars. The most salient horizontal bar in the cross has $z = 7$, because its evoked response is significantly higher than the population average. (Values of $z \geq 3$ are likely to be for the most salient item in an image.) One of the distractor vertical bars has $z = -1.3$ because its evoked response is below the background average. Another distractor vertical bar has $z = 1.7$; its evoked response is above the background mean but is not an outlier with respect to the population response.

(isolated) contour (i.e. contour enhancement). A vertical bar among horizontal bars (a pop-out target) or the bars at a texture boundary induce higher responses because they have fewer iso-orientation neighbors than other input bars (see Box 2, Fig. I). Furthermore, because of contour enhancement, the vertical bars at the vertical texture border induce higher responses than the horizontal bars at the same texture border. As it is consistent with the known V1 anatomy and physiology, the model is an attractive candidate to explain the psychophysics of salience.

Salience of an image item, i, is viewed in relation to a large or whole input image. It can thus be assessed by comparing the evoked response, S_i, to the population of responses to all visible input items (see Fig. 1b). As salience merely serves to order inputs for further processing, only the order of the saliences is relevant and it is computationally unnecessary to have an absolute saliency measure. Hence, the most active cell points to the most salient item, the second most active cell to the next most salient item, and so on. To relate this to empirical data, a measure of the saliency order, or simply salience, for item i can be assessed by the z-score,

$$z_i \equiv (S_i - \bar{S}) \,/\, \sigma_s \qquad [1]$$

where $\bar{S}$ and σ_s are respectively the mean and standard deviation of the response S_i over all input item. As argued above, the brain need not calculate these z_i scores for sequential attention deployment. z-scores of $z \gg 1$, $z \sim 1$, and $z < 0$ indicate, respectively, input items that are very salient, typical but more salient than average, or less salient than average (Fig. 1c). z-scores for various texture boundaries or smooth contours can also be calculated [19]. The ease of a visual search task (see Box 3) should, excluding top-down factors, such as terminating the search because of special knowledge, increase with the target's z-score [9].

Testing the saliency map on visual search tasks

The highly salient horizontal bar in the target cross in Fig. 1c makes the whole cross conspicuous. We henceforth define the salience of a composite image item, such as a cross, as the salience of its most salient component. When a target is defined by a unique feature, such as the horizontal bar in Fig. 2a and c, that is absent in the distractors, it pops out because this unique feature suffers less iso-orientation (or iso-feature) suppression than other image features [9].

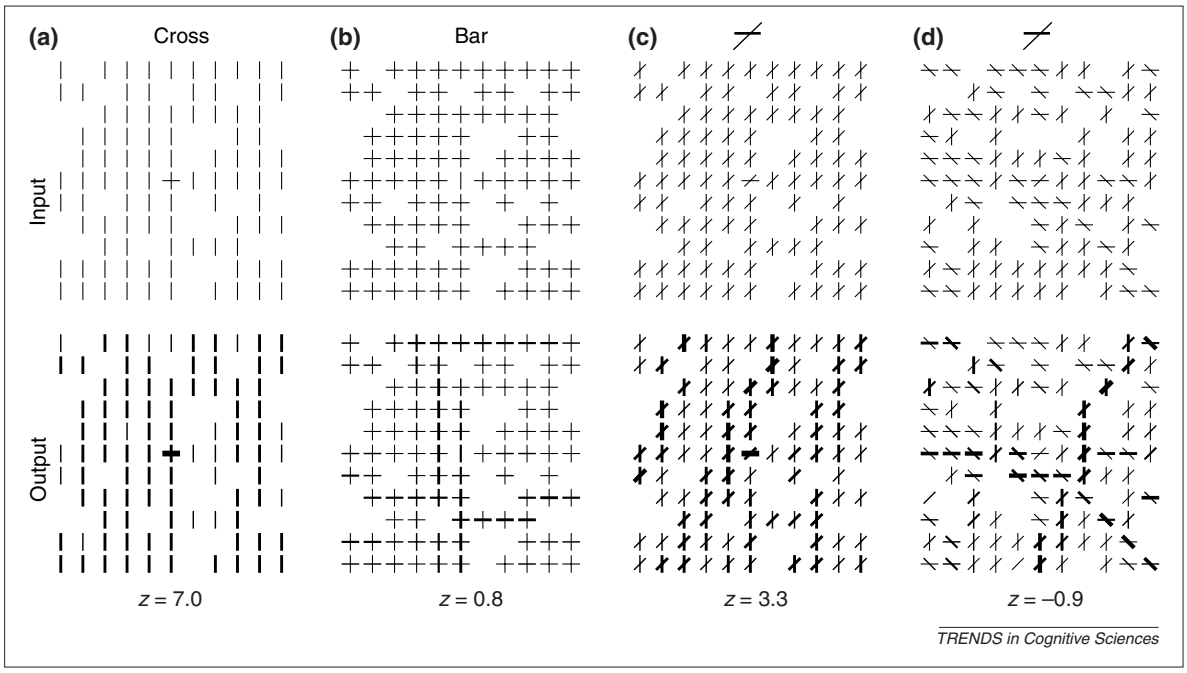

Fig. 2. Four examples of a target at the center of an image full of distractors. The target is shown above each input. The *z*-scores for the targets are shown below each output plot. A simple example of visual search asymmetry is provided by (a) and (b): a cross among bars is easier to find than a single bar among crosses.

When the target is distinguished by a lack of a feature, as in Fig. 2b, or a conjunction of features present in the background, as in Fig. 2d, the target does not pop out, because the target features experience similar iso-orientation (or iso-feature) suppression to the background features [9]. Hence, the orientation- or feature-specific contextual influences in V1 provide a plausible neural basis underlying feature-integration theory [4], or the related texton theory [20].

Whereas a target with $z \geq 3$ pops out and another with $z < 0$ requires serial search, targets with *z*-scores between, and in particular, in the middle of, the two extremes, will need searches that are neither purely parallel nor definitely serial. As previously noted, the separation between parallel and serial search in earlier work is probably an idealization of the actual visual processes [1,4,20–22].

A homogeneous background makes the population of responses *S* less variable than otherwise, leading to a smaller σ_s. This should make a moderately salient target, with $0 < z < 2$, more salient with a boosted *z*-score. Background irregularity can result from random positioning of distractors and/or dissimilarities between distractors, making contextual influences noisy. Fig. 3a–c, presents an example of this in the model where either of the two irregularities reduces the target's *z*-score from $z = 3.4$ to $z = 0.22, 0.25$. In Fig. 3d and e the target vertical bar has a negative *z* in both the homogeneous and inhomogeneous backgrounds. Interestingly, the target is easier to spot in Fig. 3e than in Fig. 3d, even though its *z*-score in Fig. 3e is slightly lower. This is because a homogeneous background increases the *z*-scores of the nearest neighbors of the target. These neighbors are on average more salient than other distractors because the target, by lacking a horizontal bar, exerts weaker general and iso-orientation suppression on them. The highest *z*-score among the nearest neighbors of the target increases from $z = 0.68$ in Fig. 3d to $z = 3.7$ in Fig. 3e. This attracts visual attention to locations near the target and makes the target easier to spot.

Subtle examples of asymmetries in visual search

Figure 2a and b show a simple example of search asymmetry: the cross is easier to spot among vertical bars than is a vertical bar among crosses. Other examples of asymmetry can be much subtler, when neither target nor distractors has a basic feature (e.g. a particular orientation) that is absent in the other. In these cases, the phenomena are psychophysically quite weak, and can often no longer be understood in the model simply by iso-orientation (or iso-feature) suppression. Local co-linear excitation and general (orientation non-specific) surround suppression also play roles.

Such search examples, with their small changes in search difficulties with input stimuli, can thus provide a more severe test of our saliency map proposal. Figure 4 shows that the model can account for the signs of the typical examples of such asymmetry, using stimuli modeled after those of Treisman and Gormican [23]. The responses to different items differ only by small fractions (that is, $S_i/\bar{S} \sim 1$), and are hard to visualize in a figure. However these fractions are significant for the more salient targets when the background saliences (responses *S*) are sufficiently homogeneous (i.e. σ_s is sufficiently small) to make the *z*-score large.

Predicting V1 physiology/anatomy from psychophysics

The idea that V1 produces a saliency map suggests a mechanistic definition of a 'basic visual feature'. This has been defined psychophysically in terms of 'pop out':

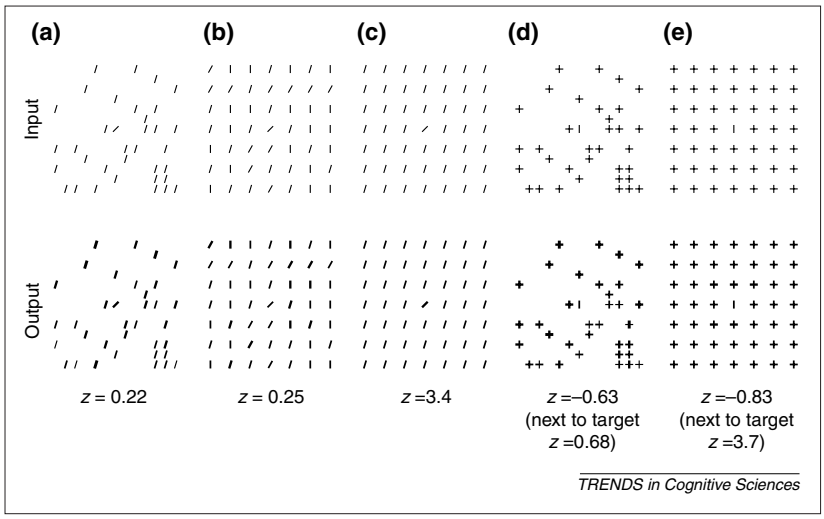

Fig. 3. The model's account of the effect of background homogeneity on search difficulty, which was observed by Duncan and Humphreys [21]. Rubenstein and Sagi have suggested a related idea that random background textural variability acts as noise and limits search performance [41]. (a–c) Searches for a target bar tilted 45° clockwise from the vertical, among distractors which are: (a) irregularly placed identical bars tilted 15° clockwise from vertical; (b) regularly placed but dissimilar bars randomly tilted 0°, 15°, or 30° clockwise from vertical; or (c) regularly placed identical bars tilted 15° clockwise from vertical. The *z*-scores for the targets are listed immediately below each example. (d,e) Search for a vertical target bar among crosses. The *z*-scores for a distractor next to the target are shown below those for the target. Although a homogeneous background decreases the *z*-score of the target in (e), the target is easy to spot simply because its nearest neighbor has an increased salience to attract attention.

a target having a 'basic feature' will pop out of a background of distractors that lack this feature [4]. Our model suggests that the following two neural substrates are necessary for such a basic feature: (1) a population of V1 cells selective or tuned to various values along this feature dimension to sample the features; (2) selectivity or tuning of the horizontal intra-cortical connections to the optimal feature values of both the pre-synaptic and post-synaptic cells in this feature dimension, such that iso-feature suppression, or the lack of it, can be manifested in response levels.

For instance, because the horizontal connections mediating suppression predominantly link cells preferring the same or similar orientations, iso-orientation suppression within an iso-orientation background will be strong, but iso-orientation suppression on a target bar whose orientation is sufficiently different will be weak. Whereas tuning of the CRFs is a familiar concept, the tuning of the intracortical connections is relatively new, and we advocated it in our framework. This connection tuning should be the main cause for the 'just noticeable difference' [26] and 'orientation categories' [27] in pre-attentive vision.

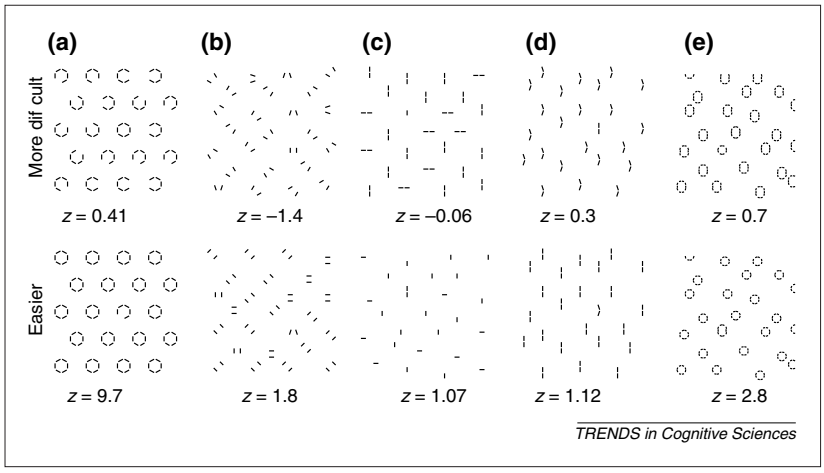

Fig. 4. The model behavior on five examples of asymmetries in visual search, with the target *z*-scores from the model under each search example. The model agrees with human visual behavior on the signs of these asymmetries, that is, on which search of each pair is relatively easier. (a) Closed vs open circles. The gap in a circle reduces co-linear facilitation as well as reducing the general and iso-orientation suppression between the circle segments. Apparently the decrease in suppression outweighs the decrease in facilitation, thus making the gapped circle more salient. The *z*-score is further boosted by the artificially regular background. (b) Parallel vs convergent pairs. A pair of parallel bars is less salient because stronger suppression occurs between the two (iso-oriented) bars. (c) Short vs long lines. Co-linear excitation makes longer lines more salient than shorter ones. (d) Straight vs curved lines. Co-linear excitation within and between image items is not so sensitive to a slight change in item curvature, but iso-orientation suppression is stronger in a background of straight (than curved) lines. Thus, the curved target is more salient. (e) Circle vs ellipse. Whereas interaction between circles (via the circle segments) depends only on the circle–circle distance, interaction between ellipses depends additionally on another random factor, the orientation of the ellipse–ellipse displacement. Hence, noisier cortical responses (larger σ_s) are evoked from a background of ellipses than from circles, submerging responses (i.e. reducing *z*) to a target circle.

'A particular strength of [the] model is that it links V1 physiology and anatomy with psychophysics.'

More importantly, the insight of the two types of tuning helps to predict when some conjunctions of two features (e.g. orientation and motion) will enable pop-out or become 'basic' [9,28,29]. The requirements are: (a) if V1 cells are simultaneously (or conjunctively) tuned to feature values of both feature dimensions, such as orientation and motion; (b) if the horizontal connections are simultaneously (or conjunctively) tuned to optimal feature values in both feature dimensions (e.g. vertical orientation *and* rightward motion direction) of the pre- and post-synaptic cells. Requirement (a) is certainly not possible, for instance, for a conjunction of two orientations (as in Fig. 2d), as few V1 cells respond to combinations of two sufficiently different orientations. Indeed orientation conjunctions produce among the most difficult search tasks [22].

Conjunction of color and orientation is also more difficult to search than a basic feature search [30] because most V1 cells selective to color are not orientation selective [31,32]. Our original V1 model was augmented to include color-selective cells untuned to orientation, as well as cells with a broader tuning to both color and orientation – the conjunction cells. The intracortical connections preferentially link cells tuned to similar feature values in color and/or orientation. This is a natural extension, backed by physiological data [32], of the connection structure in orientation dimension.

Figure 5 demonstrates that it is the cells tuned to color/orientation conjunctions, as well as the corresponding intracortical connections, that make a search for such a conjunction easier. However, cells simultaneously tuned to both orientation and motion

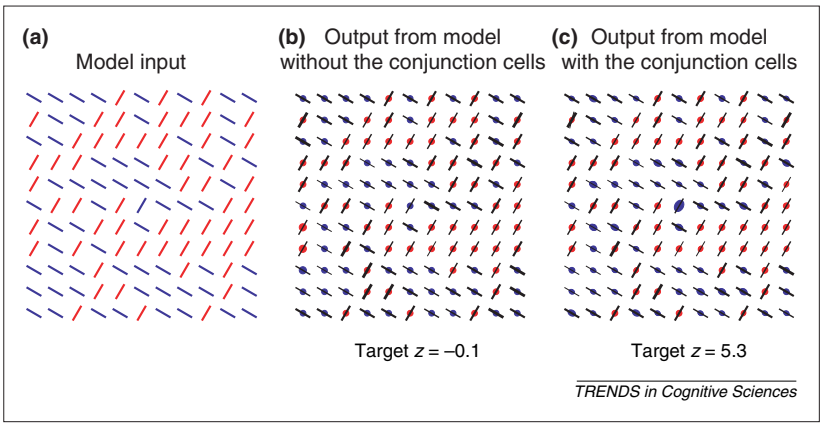

Fig. 5. Demonstrating the importance of conjunction cells for conjunction searches. (a) The target is the blue, right-tilted, bar (in the middle of the input image) among red, right-tilted, and blue, left-tilted, distractor bars. A colored and oriented bar gives input to cells whose optimal orientation and/or optimal color are close to the bar's feature(s). The model responses are visualized by the thickness of the black bars for non-color selective cells, the size of the colored circles for non-orientation selective cells, or the size of the colored and oriented ellipses for the broadly tuned conjunction cells. The most responsive cell to each bar, whether or not it is color and/or orientation tuned, reports the saliency signal and results in a *z*-score. The *z*-scores for the target when the color/orientation conjunction cells are silenced or removed from the model (b), or are available (c) are indicated below each output. Decreasing the input sensitivity of the conjunction cells, and/or increasing the feature tuning widths of the corresponding intracortical connections, decrease the target saliency. The example illustrated in (c) used parameters that give very efficient searches [30] to illustrate an extreme example.

direction, or orientation and depth, are abundant in V1 [33,34]. As psychophysical data suggest that searches are easy for targets defined by a conjunction of motion and form (orientation) [29], or by a conjunction of depth and either motion or color [28], we predict that there should be (1) horizontal connections that link pre- and post-synaptic cells preferring similar motion directions *and* similar orientations, and (2) connections linking cells preferring similar disparities *and* motion directions (and/or color). Most recent physiological evidence is consistent with prediction (1), showing that contextual suppression from an iso-orientation surround is reduced in many V1 cells when the surround moves in the opposite direction from the center stimulus (H.E. Jones *et al.*, pers. commun.).

Understanding psychophysics from V1 physiology/anatomy

In contrast to conjunction searches are the double-feature searches, for which the target differs from distractors in more than one feature dimensions. Using the example of a red vertical target bar among green horizontal bars, the target evokes responses in three cell types: (1) non-orientation selective cells tuned to red, (2) non-color-selective cells tuned to vertical, and (3) cells tuned to both red color and vertical orientation. Types (1) and (2) are single-feature tuned cells and type (3) is double-feature tuned. The most responsive of them should determine the salience of the target.

We assume that the cells tuned to single features determine the ease of the single-feature searches (e.g. for a red target among green distractors or a vertical bar among horizontal ones). We thus predict that double-feature search should be somewhat easier than each of the two corresponding single-feature searches. Furthermore, a lack of sufficient double-feature-tuned cells should diminish the advantage of the double-feature search over the easier one of the two single-feature searches. Because V1 has fewer cells doubly tuned to color and orientation than to motion and orientation, the double-feature advantage should be stronger for motion-and-orientation than color-and-orientation, as recently observed psychophysically [2].

Questions for future research

- Reaction times in search tasks depend both on the bottom-up saliency map and on top-down attentional effects, as well as on the specific search algorithms used. For instance, it will make a difference whether the subject knows what the target is, whether the search depends stochastically or deterministically on the saliences of image items, and whether there is infinite or limited short-term memory for the image locations that have already been visited [38]. These questions are being studied intensively by the vision community, but until we have detailed answers, our saliency map can provide only relative measures of task difficulties given the same top-down and algorithmic factors.
- If V1 provides a bottom-up saliency map in its initial responses to the visual input, what is the functional role of its later responses, which are affected by top-down feedback?
- Some examples of 'pop-out' and search asymmetries arise from seemingly higher-order features such as 3-D shape or character familiarity [22,39]. They have led to proposals that attribute 'pop-out' to higher cortical areas (S. Hochstein and M. Ahissar, pers. commun.). Is it possible that these higher-order features are built into the intracortical connections in V1? Can 'higher-order' saliency maps co-exist with the one in V1, and if so, which saliency map dominates?
- Physiological experiments have found neurons in cortical areas beyond V1 whose activities correlate with object salience [6,40]. Are these 'saliency' signals generated within the cortical area where they are measured or relayed from lower cortical areas?
- What are the roles of higher visual areas that directly receive the V1 outputs? How should the higher-visual areas extract the feature values of the visual inputs in addition to, and without being corrupted by, the salience information from V1?

Physiological data [31,32], together with computational considerations [35], suggest that the color/orientation conjunction (double-feature) cells in V1 are tuned to only a limited range of spatial scales. We thus predict that manipulating the scale of visual stimuli (at a given visual eccentricity) could make the conjunction search more or less difficult, by decreasing or increasing the direct activation to the conjunction cells.

A particular strength of our model is that it links V1 physiology and anatomy with psychophysics. In comparison, the popular two-stage psychophysical models, which perform operations such as local surround inhibition on the non-linearly transformed and spatially filtered input, cannot make specific enough physiological predictions because they use abstract contextual interactions [36,37]. We also predicted, and subsequently tested and confirmed (A. Popple and Z. Li, pers. commun.), a perceptual bias in the location of the border between two iso-orientation textures, owing to the asymmetric distribution of the strongest outputs with respect to the border [19] (see Box 2).

Conclusion

By applying our V1 model to visual search tasks, we have demonstrated that the primary visual cortex could provide a saliency map. This map is signaled in the responses of feature-selective cells, and no separate feature maps or any subsequent combination of them is necessary. By relating V1 responses to saliences, and saliences to the ease of visual search, our framework can account, at least qualitatively, for how and why the search difficulty depends on visual features and their conjunctions or combinations, and on the spatial arrangements of the image items. Not only can our V1 model explain large differences in search difficulties, such as between the feature and conjunctive search tasks, but it also accounts for the more subtle differences in examples of search asymmetry such as those of Treisman and Gormican [23] (although there is much to be improved quantitatively from our current model). Our computational framework provides a unique link between psychophysics and V1 physiology/anatomy, and makes testable predictions.

Acknowledgements
I thank Peter Dayan for many helpful discussions and conversations, Todd Horowitz, Jeremy Wolfe, and the anonymous reviewers for very useful comments.

References

1 Wolfe, J.M. *et al.* (1989) Guided search: an alternative to the feature integration model for visual search. *J. Exp. Psychol.* 15, 419–433
2 Nothdurft, H.C. (2000) Salience from feature contrast: additivity across dimensions. *Vis. Res.* 40, 1183–1202
3 Itti, L. and Koch, C. (2001) Computational modelling of visual attention. *Nat. Rev. Neurosci.* 2, 194–203
4 Treisman, A. and Gelade, G. (1980) A feature integration theory of attention. *Cogn. Psychol.* 12, 97–136
5 Koch, C. and Ullman, S. (1985) Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* 4, 219–227
6 Itti, L. *et al.* (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Patt. Anal. Mach. Intell.* 20, 1254–1259
7 Olshausen, B.A. *et al.* (1993) A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* 13, 4700–4719I
8 Li, Z. (1999) Visual segmentation by contextual influences via intracortical interactions in primary visual cortex. In *Netw. Comput. Neural Syst.* 10, 187–212
9 Li, Z. (1999) Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proc. Natl. Acad. Sci. U. S. A.* 96, 10530–10535
10 Li, Z. (2001) Saliency and figure-ground effects. In *Visual Attentional Mechanisms* (Cantoni, V. and Petrosino, A. eds), Plenum Press
11 Malik, J. and Perona, P. (1990) Preattentive texture discrimination with early vision mechanisms. *J. Opt. Soc. Am.* 7, 923–932
12 Yen, S-C. and Finkel, L.H. (1998) Extraction of perceptually salient contours by striate cortical networks. *Vis. Res.* 38, 719–741
13 Rockland, K.S. and Lund, J.S. (1983) Intrinsic laminar lattice connections in primate visual cortex. *J. Comp. Neurol.* 216, 303–318
14 Gilbert, C.D. and Wiesel, T.N. (1983) Clustered intrinsic connections in cat visual cortex. *J. Neurosci.* 3, 1116–1133
15 Fitzpatrick, D. (1996) The functional organization of local circuits in visual cortex: insights from the study of tree shrew striate cortex. *Cereb. Cortex* 6, 329–341
16 Douglas, R.J. and Martin, K.A. (1990) Neocortex. In *Synaptic Organization of the Brain* (3rd edn), (Shepherd, G.M., ed.), pp. 389–438, Oxford University Press
17 Knierim, J.J. and van Essen, D.C. (1992) Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophysiol.* 67, 961–980
18 Kapadia, M.K. *et al.* (1995) Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys. *Neuron* 15, 843–856
19 Li, Z. (2000) Pre-attentive segmentation in the primary visual cortex. *Spat. Vis.* 13, 25–50
20 Julesz, B. (1981) Textons, the elements of texture perception, and their interactions. *Nature* 290, 91–97
21 Duncan, J. and Humphreys, G. (1989) Visual search and stimulus similarity. *Psychol. Rev.* 96, 1–26
22 Wolfe, J.M. (1998) Visual search. In *Attention* (Pashler, H., ed.), pp, 13–74, Psychology Press
23 Treisman, A. and Gormican, S. (1988) Feature analysis in early vision: evidence for search asymmetries. *Psychol. Rev.* 95, 15–48
24 Li, Z. (1998) Primary cortical dynamics for visual grouping. In *Theoretical Aspects of Neural Computation* (Wong, K.Y.M. *et al.*, eds), pp. 155–164, Springer-Verlag
25 Li, Z. (1998) A neural model of contour integration in the primary visual cortex. *Neural Comput.* 10, 903–940
26 Foster, D.H. and Ward, P.A. (1991) Horizontal–vertical filters in early vision predict anomalous line-orientation identification frequencies. *Proc. R. Soc. London Ser. B Biol. Sci.* 243, 83–86
27 Wolfe, J.M. *et al.* (1992) The role of categorization in visual search for orientation. *J. Exp. Psychol. Hum. Percept. Perform.* 18, 34–49
28 Nakayama, K. and Silverman, G.H. (1986) Serial and parallel processing of visual feature conjunctions. *Nature* 320, 264–265
29 Mcleod, P. *et al.* (1988) Visual search for a conjunction of movement and form is parallel. *Nature* 332, 154–155
30 Wolfe, J.M. (1992) 'Effortless' texture segmentation and 'parallel' visual search are not the same thing. *Vis. Res.* 32, 757–763
31 Livingstone, M.S. and Hubel, D.H. (1984) Anatomy and physiology of a color system in the primate visual cortex. *J. Neurosci.* 4, 309–356
32 Ts'o, D.Y. and Gilbert, C.D. (1988) The organization of chromatic and spatial interactions in the primate striate cortex. *J. Neurosci.* 8, 1712–1727
33 Hubel, D.H. and Wiesel, T.N. (1959) Receptive fields of single neurons in the cat's visual cortex. *J. Physiol.* 148, 574–591
34 Barlow, H.B. *et al.* (1967) The neural mechanism of binocular depth discrimination. *J. Physiol.* 193, 327–342
35 Li, Z. and Atick, J.J. (1994) Toward a theory of the striate cortex. *Neural Comput.* 6, 127–146
36 Graham, N. (1994) Non-linearities in texture segmentation. In *Higher-Order Processing in the Visual System* (Ciba Foundation Symposium 184), pp. 309–329, John Wiley & Sons
37 Sagi, D. (1995) The psychophysics of texture segmentation. In *Early Vision and Beyond* (Papathomas, T. *et al.*, eds), pp. 69–78, MIT Press
38 Horowitz, T.S. and Wolfe, J.M. (1998) Visual search has no memory. *Nature*, 394, 575–577
39 Enns, J. and Rensink, R. (1990) Influence of scene-based properties on visual search. *Science* 247, 721–723
40 Gottlieb, J.P. *et al.* (1998) The representation of visual salience in monkey parietal cortex. *Nature* 391, 481–484
41 Rubenstein, B. and Sagi, D. (1990) Spatial variability as a limiting factor in texture discrimination tasks: implications for performance asymmetries. *J. Opt. Soc. Am.* 9, 1632–1643