The primary visual cortex creates a bottom-up saliency map

Li Zhaoping

University College London, Dept of Psychology, Gower Street, WC1E 6BT, UK

In "Neurobiology of Attention" Eds L. Itti, G. Rees and J.K. Tsotsos, Elsevier, 2005, Chapter 93, page 570-575

Abstract

It has been proposed that the primary visual cortex (V1) creates a saliency map using autonomous intra-cortical mechanisms. This saliency of a visual location describes the location's ability to attract attention without top-down factors. It increases monotonously with the firing rate of the most active V1 cell responding to that location. Given the prevalent feature selectivities of V1 cells (many tuned to more than one feature dimension), no separate feature maps, or any subsequent combinations of them, are needed to create a saliency map. This proposal has been demonstrated in a biologically based V1 model. By relating the saliencies of the visual search targets or object (texture) boundaries to the eases of the visual search or segmentation tasks, the model accounted for behavioral data such as how task difficulties can be influenced by image features and their spatial configurations. This proposal links physiology with psychophysics, thereby making testable predictions some of which are subsequently confirmed experimentally.

Key words: saliency map, theory, V1

1 The bottom up saliency map regardless of visual features signalled by feature selective cells in V1

A saliency map aids the selection of visual inputs for further processing given limited computational resources. To better understand the selection, we separate bottom-up from top-down mechanisms (see [THE FEATUREGATE MODEL OF VISUAL SELECTION] and [GUIDANCE OF VISUAL SEARCH BY PREATTENTIVE INFORMATION] for alternative approaches) 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. The primary visual cortex receives many top-down inputs from higher visual areas. Hence, bottom-up saliency map in V1 is an idealization when the top-down influences are ineffective ([IRRELEVANT SINGLETONS CAPTURE ATTENTION]), such as very shortly after visual presentation ([STIMULUS-DRIVEN GUIDANCE OF VISUAL ATTENTION IN NATURAL SCENES])and without specific top-down knowledge, or when the animal is under aneathesia. Furthermore, the saliency value is regardless of the visual features like color and orientation (Treisman and Gelade 1980) such that, e.g., the saliency of a red dot can be compared with that of a vertical moving bar (see [SALIENCE OF FEATURE CONTRAST]). This property may have led to a common belief, as implicitly or explicitly expressed in previous works (Koch and Ullman 1985, Itti et al 1998) on saliency maps, that saliency must be signalled by cells untuned to features, such as cells in parietal cortex (Gottlieb et al 1998, see [MODELS OF BOTTOM-UP ATTENTION AND SALIENCY]) and that the saliency map must be outside V1 whose cells are feature tuned. However, just like the purchasing power of an UK sterling is regardless of the holder's nationality or gender, the firing rate of V1 cells could be an universal currency for saliency with or without simultaneously decoding the input features from them. Finally, using V1's output for saliency signal (to direct eye movements perhaps by sending outputs to the superior colliculus) does not preclude V1 from sending its outputs to other visual areas and contributing to other visual computations such as object recognition.

The response of a V1 cell to inputs within its classical receptive field (CRF) can be influenced by contextual inputs near but outside the CRF, due to the long but finite range intra-cortical interactions linking nearby cells (Knierim and van Essen 1992, Kapadia et al 1995). Hence, the saliency of a location is determined both by the input strength (or contrast) at that location *and* by its context, as expected (see [SALIENCE OF FEATURE CONTRAST], [SCENE STATISTICS AND SALIENT FEATURES]). Furthermore, any visual location can evoke responses from many V1 cells whose CRFs overlap. For instance, a small vertical red bar may excite cells tuned to vertical orientation, or cells tuned to red but untuned to orientation, or cells whose optimal orientation is 5 degrees from vertical and whose tuning width is 15^o, etc. The proposed saliency of a location is determined by the firing rate of the most responsive cell to it, regardless of the cell's optimal feature value. (This way, no and minimal computation is needed to decide how cells contribute to signaling the saliency of a location). Hence, the saliency of the red vertical bar are likely signalled by a cell tuned to vertical, or a cell tuned to red, or a cell tuned to both, depending on the context, but less likely by a cell tuned to 10 degrees from vertical. Given the population firing rates from all responding cells (regardless of their optimal features) to the whole image, the saliency of a location may be phenomenologically measured by the z score $z \equiv (S - \bar{S})/\sigma$ where S is the (highest) evoked response to that visual location, $\bar{S}$ is the mean

response to the image, and σ is the standard deviation in the population response (Li 2002).

2 Demonstrating the saliency map by a V1 model

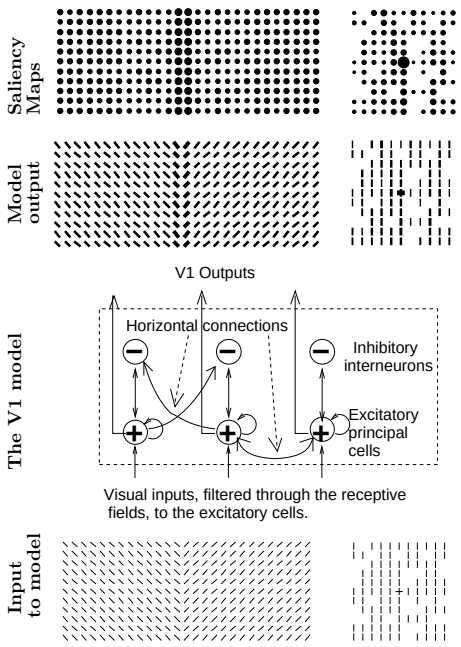

Fig. 1. The V1 model and its function. Our model focuses on the part of V1 responsible for contextual influences: layer 2-3 pyramidal cells, interneurons, and horizontal intracortical connections. Pyramidal cells and interneurons interact with each other locally and reciprocally. A pyramidal cell can 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. Shown are also two input images to the model, and their output response maps. The input strengths are determined by the bar's contrast. Each input bar in each image has the same contrast. A principal (pyramidal) cell can only receive direct visual input from an input bar in its CRF. The output responses depend on both the input contrasts and the contextual stimuli of each bar due to contextual influences. Each input/output image plotted is only a small part of a large extended input/output image. In all figures in this paper, the thicknesses of the bars in each plot are plotted as proportional to their input/output strengthes for visualization. At top are saliency maps where the size of the the circle at each location represents the firing rate of the most active cell responding to that visual location.

A biologically based V1 model (2002) is used to demonstrate and validate the saliency map. The model (Fig. 1) focuses on layer 2-3 of V1 where intra-cortical

connections are prevalent. Each model pyramidal cell receives direct visual inputs within its CRF, mono-synaptic excitation and di-synaptic inhibition from local pyramidal cells tuned to similar orientations, and general orientation unspecific local surround suppression. The model produces the usual contextual influences observed physiologically. In particular, the response of a cell to an optimally oriented bar within its CRF is suppressed if the CRF is surrounded by contextual bars, with the strongest suppression from contextual bars oriented parallel to the central bar within the cell's CRF (termed iso-orientation suppression) and weakest suppression from contextual bars oriented orthogonally to the central bar (Knierim and vanEssen 1992). The cell's response can be enhanced under low input contrast when contextual bars align with the central bar to form a smooth contour — colinear facilitation (Kapadia et al 1995).

When the model is presented with visual stimuli resembling those in visual search and texture segmentation experiments, the strongest responses are located at or near the pop out targets or texture boundaries. In Fig (1) right, the cross pops out among the bars since its horizontal bar, the only one that does not experience any iso-orientation suppression from other (vertical) bars in the image, evokes the highest response in the image. Hence, iso-feature suppression (Li and Li 1994) is the neural basis for the ease of feature search (Treisman and Gelade 1980). Similarly, sufficient orientation contrast at the border between two textures of uniformly oriented bars can pop out because a border bar, having half as many iso-oriented contextual neighbors as those of bars away from the border, evoke relatively higher responses (Fig. (1) left). In Fig. (2A) the target vertical bar does not pop out among crosses since it is suppressed by other vertical bars in the neighboring crosses — hence Treisman's phenomenological rule that a target lacking a feature present in the background does not pop out (Treisman and Gelade 1980). Also, a unique conjunction of two orientations (bars) is difficult to search in a background that includes both orientations, since neither oriented bar escapes from the iso-orientation suppression experienced by all background bars (on average). Fig(2A) and the right image in Fig (1) compose a trivial pair of search asymmetry, when the ease of search changes upon a target-distractor identity swap. Our V1 model also agrees (Li 2002) with human vision in subtler or weaker examples of search asymmetry (used by Treisman and Gormican 1988), such as searching for a circle among elipses and vice versa. These subtle asymmetries provided a severe test to our proposal and model since they can not be explained simply by the simple mechanism of iso-orientation suppression alone. Colinear facilitation and general surround suppression also contribute in these examples. The relative ease to search for a target in a more homogeneous background (with more similar distractors or more regular spatial configurations, Duncan and Hymphreys 1989, Rubinstein and Sagi 1990) is also understood in the model (Fig. (2C,D,E). Iso-feature suppression between the distractors is stronger when the distractors are the same or similar, thus

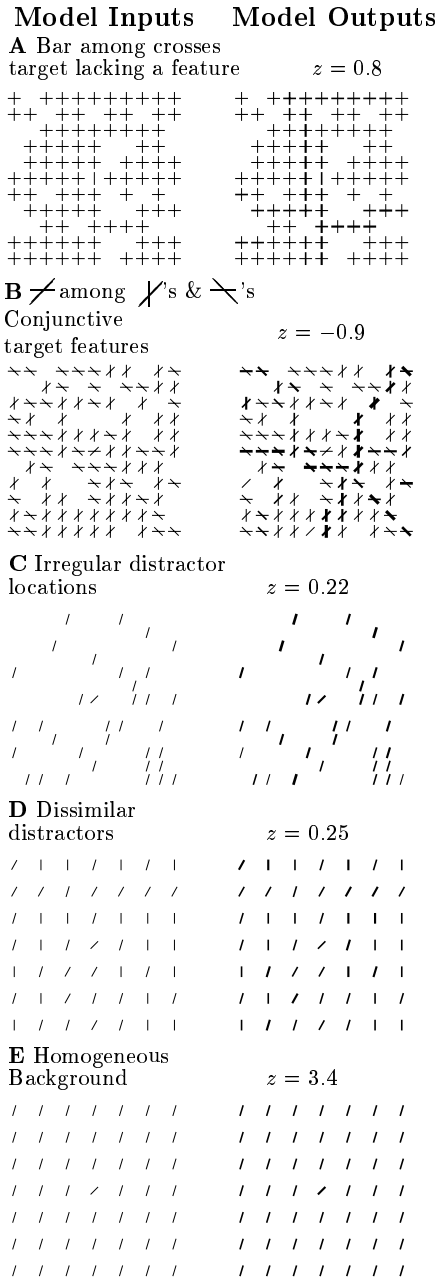

Fig. 2. More examples of model performances. All search targets are at the center of the stimulus patterns. Left column: stimulus inputs; right column: model outputs. The z scores of the target, a measure of the target saliency, are indicated. A and the right example in Fig. 1 compose a trivial example pair of search asymmetry. C, D, and E illustrate how distractor dissimilarity or irregular distractor locations impair the search for a 45^o oriented target bar.

reducing background responses to make the target relatively more conspicu-

ous. Regular spatial locations of the distractors enable each distractor to have the same or similar contextual environment. This makes the background responses homogeneous, or gives small standard deviation σ in the population responses, thus giving high z score $z \propto 1/\sigma$ for the target (Li 2002).

3 Basic features and conjunction searches explained by V1 saliency mechanism

According to our model, a feature dimension is a basic dimension to enable pop out when the following two neural bases are present: (1) V1 cells should be tuned to this feature dimension (e.g., orientation) in order to signal it, (2) the intra-cortical connections should be tuned to that dimension, such that they only link cells tuned to similar optimal feature values (e.g., orientation) to achieve the iso-feature (e.g., iso-orientation) suppression essential for popout. Thus, if conjunctions of features also meet these two criteria, i.e., both the cells and the intracortical connections are tuned to conjunctions of features, pop out of a target defined by a conjunction of features, say a red-vertical target among green vertical and red horizontal distractors, is feasible. Hence, a conjunction of two orientations (or more generally, two sufficiently different features within a single feature dimension) does not pop out since individual V1 cells are not simultaneously tuned to two sufficiently different orientations. Color-orientation conjunction search is typically difficult (Wolfe 1998), since V1 cells are more likely to be either color or orientation tuned and less likely to be conjunctively tuned to both (Livingstone and Hubel 1984, Li 2002). That a conjunction of motion and orientation (or stereo and motion, Wolfe 1998) can pop out is consistent with V1 physiology that cells are conjunctively tuned to motion and orientation, or stereo and motion. Furthermore, these psychophysical observations predict that the intra-cortical connections must be conjunctively tuned to optimal features of the linked cells in both feature dimensions e.g., motion direction and orientation (Li 2002). With input of a unique conjunction target among many distractor conjunctions, the conjunction cell tuned to the target has the highest response by escaping iso-conjunction suppression that are present on cells responding to distractors.

4 Saliency and interactions between feature dimensions

Visual search of a 45^o target bar among 135^o distractor bars is easy (even for infants [VISUAL SEARCH AND POP-OUT IN INFANCY]), so is the segmentation between two textures of uniformly oriented bars at 45^o and 135^o respectively. Snowden (1998) observed that, although these tasks depend only on

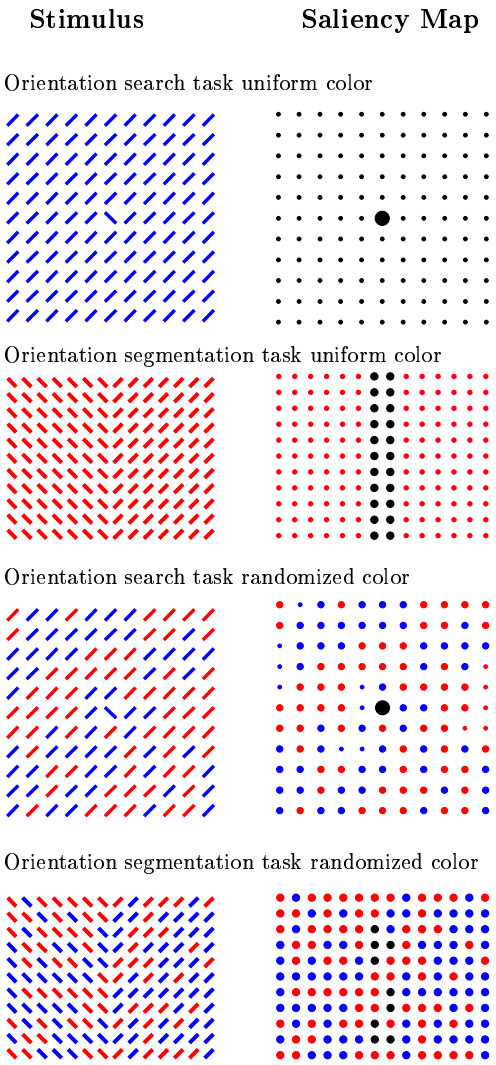

Fig. 3. Color inference in orientation feature based tasks. The two random colors of the stimulus bars are visualized in the grey scale images as black and white bars. Each bar evokes responses from color tuned cells *and* from orientation tuned cells. Randomizing colors of the stimulus bars increases the response levels (and variations) in the color tuned cells, submerging the responses from the orientation tuned cells to the texture border but not that to the search target. Hence, texture segmentation task, but not the search task, is impaired. Note that in uniform color stimuli, the saliency of the target is much higher than that of the texture border against their respective background.

the orientation feature of the bars, the texture segmentation became difficult when each stimulus bar was randomly assigned a color from two choices (say red and green), whereas the orientation search task remained easy under the same color randomization (Fig. 3). (Nothdurft (1997) also observed interference of luminance variations on orientation based texture segmentation.) Such color interference is comprehensible in our framework by noting the changes in saliencies of the target bar or the texture border under the color randomization. Let each colored bar evoke responses from both orientation tuned cells insensitive to color and color tuned cells insensitive to orientation (omitting for simplicity, without changing our conclusion, the minority of cells tuned conjunctively to color and orientation). We consider stimuli such that each colored bar, when presented alone, evoke comparable responses in the corresponding color and orientation tuned cells, and assume that the iso-orientation and iso-color suppressions have comparable strength. With uniform color stimuli, both the color and orientation tuned cells responding to a background bar (i.e., away from the target or texture border) experience iso-feature (iso-color or iso-orientation) suppression and give suppressed responses of similar levels. Meanwhile, the orientation tuned cells responding to the target or the texture border bars are relatively more active since they experience no or weaker iso-orientation suppression, respectively, making the target or the border pop out. Note, however, the orientation tuned cell responding to the target bar, the only one with no contextual neighbors of the same orientation, is much more active than those responding to the texture border bars (Fig. 3). When the bar colors are randomized, the number of iso-color contextual neighbors of any bar is halved on average, making the color tuned cell less suppressed. Furthermore, their iso-color suppression is of similar magnitude as the iso-orientation suppression experienced by the orientation tuned cells responding to the texture border, since each border bar has also half of its contextual neighbors of the same feature (orientation). Thus the response to the border (from the orientation tuned cells) is submerged by the background responses from the color tuned cells, making the border less conspicuous. Meanwhile, in the single target search stimulus, the orientation tuned cell responding to the target is still the most active against the background of the more active color tuned cells, since it is the only excited cell not experiencing any iso-feature suppression (Fig. 3). Hence, pop-out is not impaired. Therefore, the essential reasons for color interference in these orientation feature based tasks are (1) object saliency rather than subject scrutiny plays a larger role in such tasks and (2) saliency is regardless of the feature dimension(s) of cells signalling it — hence the activity of a color tuned cell signalling saliency of one bar is compared with the activity of an orientation tuned cell signaling saliency of another bar to see which bar is more salient. Note that activities of the color tuned cell and the orientation tuned cell responding to the same location (bar) are not summed up linearly or nonlinearly to signal the saliency of this location (bar) (See [SALIENCE OF FEATURE CONTRAST] and Nothdurft 1997 for a different perspective). Otherwise, the border should be more salient than

predicted, since the border highlight from orientation activities superposed on a planar background of color activities, albeit enhanced, leads still to the relative border highlight. From our analysis above, we can predict that drawing color randomly from more color choices should make the tasks even more difficult, by further reducing the iso-color suppresion on the color tuned cells. This prediction is recently confirmed (Zhaoping and Snowden 2003).

5 Discussions

V1 is the largest visual area in the brain. The cells' CRFs are smaller and under weaker attentional influences than those in higher visual areas. Hence the saliency map should be roughly stable and have a satisfactory spatial resolution. Neurons beyond V1 were found to have their activities correlate with saliencies (Gottlieb 1998). It is desirable to answer whether these signals are relayed from lower visual areas or whether there is a hierarchy of saliency maps from different visual areas ([VISUAL SALIENCY AND SPIKE TIMING IN THE VENTRAL VISUAL PATHWAY]). By guiding top-down visual attention, the bottom-up saliency map should have its effects visible in tasks with significant top down influences. Furthermore, a well understood bottom-up saliency map should definitely help to elucidate the mechanisms of attention.

References

[1] Treisman, A. and Gelade, G. (1980) A feature integration theory of attention. *Cogn. Psychol.* 12, 97-136.

[2] Koch, C. and Ullman, S. (1985) Shifts in selective visual attention: towards the underlynig neural circuitry. *Hum. Neurobiol.* 4, 219-227.

[3] 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.

[4] Knierim J.J. and van Essen D. C. (1992) Neuronal responses to static texture patterns in area V1 of the alert macaque monkeys. *J. Neurophysiol.* **67**, 961-980.

[5] Kapadia, M. K., Ito, M. , Gilbert, C. D., and Westheimer G. (1995) Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys. *Neuron.* **15**(4), 843-56.

[6] Li, Z. A saliency map in primary visual cortex *Trends in Cog. Sci.* 2002. 6(1)9-16.

[7] Li, C.Y., and Li, W. (1994) Extensive integration field beyond the classical receptive field of cat's striate cortical neurons — classification and tuning properties. *Vision Res.* **34** (18), 2337-55.

[8] Treisman A. and Gormican S. (1988) Feature analysis in early vision: evidence for search asymmetries. *Psychological Rev.* **95**, 15-48.

[9] Duncan J. Humphreys G. "Visual search and stimulus similarity". *Psychological Review* 96: p1-26, (1989).

[10] Rubenstein B. and Sagi D. "Spatial variability as a limiting factor in texture discrimination tasks: implications for performance asymmetries" *J. Opt. Soc. Am. A* 9: 1632-1643 (1990).

[11] Wolfe J. M. Visual Search, a review, in *Attention* edited by H. Pashler (p 13-74) Psychology Press Ltd. (1998)

[12] Livingstone M. S. and Hubel, D. H. (1984) Anatomy and physiology of a color system in the primate visual cortex. *J. Neurosci.* Vol. 4, No.1. 309-356.

[15] Snowden, R.J. (1998). Texture segregation and visual search: a comparison of the effects of random variations along irrelevant dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 1354-1367. See also Zhaoping L. & Snowden R.J. (2003) A psychophysical test of the saliency map in V1. Vision Science Society annual meeting, Sarasota, Florida, May, 2003.

[15] Nothdurft H-C Different approaches to the coding of visual segmentation. In: Harris L. and Jenkins M. eds. p. 20-43. "Computational and psychophysical mechanisms of visual coding". Cambridge University Press, New York (1997).

[15] Gottlieb, J.P. et al (1998) The representation of visual salience in monkey parietal cortex *Nature* 391, 481-484.