Some fingerprints of V1 mechanisms in the bottom up saliency for visual selection

Li Zhaoping, Keith A. May, Ansgar Koene

Department of Computer Science, University College London, UK

Published as Chapter 7 (page 137-164) in *Computational modelling in behavioural neuroscience: closing the gap between neurophysiology and behaviour* Editted by Dietmar Heinke and Eirini Mavritsaki, Psychology Press, 2009. ISBN 978-1-84169-738-3

Abstract:

A unique vertical bar among horizontal bars is salient and pops out perceptually regardless of the observer's goals. Physiological data have suggested that mechanisms in the primary visual cortex (V1) contribute to the high saliency of such a unique basic feature, but fail to indicate whether V1 plays an essential or peripheral role in input-driven or bottom up saliency. Meanwhile a biologically based V1 model has suggested that V1 mechanisms can also explain bottom up saliencies beyond the pop out of basic features (Li 1999a, 2002). For instance the low saliency of a unique conjunction feature like a red-vertical bar among red-horizontal and green-vertical bars is explained, under the hypothesis that the bottom up saliency at any location is signalled by the activity of the most active cell responding to it regardless of the cell's preferred features such as color and orientation. While some recent experimental data have provided support for this V1 saliency hypothesis, higher visual areas such as V2 and V4 also contain neurons tuned to similar basic features that can pop out in the bottom up manner. Furthermore, previous saliency models can capture much of the visual selection behavior using generic rather than V1 specific neural mechanisms. It is therefore important to ascertain V1's role in saliency by identifying visual selection behavior that show specific identifying characteristics, i.e., fingerprints, of V1 or other cortical areas. In this paper, we present our recent findings on bottom-up saliency based behavior of visual search and segmentation that directly implicate V1 mechanisms. The three specific fingerprints are: (1) ocular singleton captures attention despite being elusive to awareness, (2) V1's collinear facilitation manifested in texture segmentation, and (3) a match between the redundancy gains in double feature singleton search and V1's conjunctive cells.

Abbreviated Title: Fingerprints of V1 in bottom up saliency.

1 Introduction

Limitations in cognitive resources force us to select only a fraction of the visual input for detailed attentive processing. Naturally, we are more aware of intentional selections such as directing our gaze to text while reading or being attracted to red colors when looking for a red cup. Indeed, many models of visual attentional mechanisms, such as the stimulus similarity framework (Duncan and Humphreys 1989), the selective tuning theory of attention (Tsotsos 1990), and biased competition model (Desimone and Duncan 1995), focus mainly on goal-directed or top-down attention, and treat selection based on bottom up saliency as something given without detailed exploration of its mechanisms. Nevertheless, much of the visual selection is carried out in a bottom up manner, which can be dominant in selections very soon after visual stimulus onset (Jonides 1981, Nakayama and Mackeben 1989, Yantis 1998). For instance, a vertical bar among horizontal ones or a red dot among green ones automatically pops out to attract perceptual attention (Treisman and Gelade 1980), typically regardless of the task demands. Such pop-out stimuli are said to be highly salient pre-attentively. Indeed, goal-directed or top-down attention has to work with or even against the bottom up selection (Zhaoping and Dayan 2006, Zhaoping and Guyader 2007). In this paper, we focus on understanding the mechanisms underlying bottom up saliency that automatically guides visual selection.

Physiologically, a neuron in the primary visual cortex (V1) gives a higher response to its preferred feature, e.g., a specific orientation, color, or motion direction, within its receptive field (RF) when this feature is unique within the display, rather than when it is surrounded by neighbors identical to itself (Allman, Miezin, & McGuinness 1985, Knierim and van Essen 1992, Li and Li 1994, DeAngelis, Freeman, and Ohzawa 1994, Sillito, Grieve, Jones, Cudeiro, & Davis 1995, Nothdurft, Gallant, & Van Essen 1999, 2000, Jones, Grieve, Wang, & Sillito 2001, Wachtler, Sejnowski, & Albright 2003, Webb, Dhruv, Solomon, Tailby, & Lennie 2005). This is the case even when the animal is under anesthesia (Nothdurft et al 1999), suggesting bottom up mechanisms. The responsible mechanism is iso-feature suppression, in particular iso-orientation or iso-color suppression, so that nearby neurons tuned to the same feature suppress each other's activities via intra-cortical connections between nearby V1 neurons (Gilbert and Wiesel 1983, Rockland and Lund 1983, Hirsch and Gilbert 1991). The same mechanisms also make V1 cells respond more vigorously to an oriented bar when it is at the border, rather than the middle, of a homogeneous orientation texture, as physiologically observed (Nothdurft et al 2000), since the bar has fewer iso-orientation neighbors at the border. These observations have prompted suggestions that V1 mechanisms contribute to bottom up saliency for pop out features like the unique orientation singleton or the bars at an orientation texture border (e.g., Knierim and van Essen 1992, Sillito et al 1995, Nothdurft et al 1999, 2000). This is consistent with observations that highly salient inputs can bias responses in extrastriate areas receiving inputs from V1 (Reynolds and Desimone 2003, Beck and Kastner 2005).

Behavioral studies have extensively examined bottom up saliencies in visual search and segmentation tasks (Treisman and Gelade 1980, Wolfe, Cave, & Franzel 1989, Duncan and Humphreys 1989), showing more complex, subtle, and general situations beyond basic feature pop outs. For instance, a unique feature conjunction, e.g., a red-vertical bar as a color-orientation conjunction among red-horizontal and green-vertical bars, is typically less salient; ease of searches can change with target-distractor swaps; and target salience decreases with background irregularities. However, few physiological recordings in V1 have used stimuli of comparable complexity, leaving it open as to how generally V1 mechanisms contribute to bottom up saliency.

Recently, a model of contextual influences in V1 (Li 1999ab, 2000, 2002), including physiologically observed iso-feature suppression and collinear facilitation (Kapadia, Ito, Gilbert, & Westheimer 1995), has demonstrated that V1 mechanisms can feasibly explain the complex behaviors mentioned above, assuming that the highest response among V1 cells to a target, relative to all other responses to the scene, determines its salience and thus the ease of a task. Accordingly, V1 has been proposed to create a bottom up saliency map, such that the RF location of the most active V1 cell is most likely selected for further detailed processing (Li 1999a, 2002). We call this proposal the V1 saliency hypothesis. This hypothesis is consistent with the observation that microstimulation of a V1 cell can drive saccades, via superior colliculus, to the corresponding RF location (Tehovnik, Slocum, & Schiller 2003), and that higher V1 responses are associated with quicker saccades to the corresponding receptive fields (Super, Spekreijse, & Lamme 2003). This can be clearly expressed algebraically. Let $(O_1, O_2, ..., O_M)$ denote outputs or responses from V1 output cells indexed by $i = 1, 2, ...M$, and let each cell cover receptive field location $(x_1, x_2, ..., x_M)$ respectively. Then, the highest response among all cells is $\hat{O} \equiv \max_i O_i$. Let this response $\hat{O}$ be from a cell indexed by $\hat{i}$, mathematically $\hat{i} \equiv \text{argmax}_i O_i$, this cell's receptive field is then at $\hat{x} \equiv x_{\hat{i}}$ and is the most salient or most likely selected by the bottom-up visual selection. The receptive field location of the second most responsive cell is the second most salient or second most likely selected by the bottom-up mechanism, and so on. Note that the interpretation of $x_i = x$ is that the receptive field of cell i covers location x and is centered near x. Defining $\hat{O}(x) \equiv \max_{x_i=x} O_i$ as the highest response among neurons whose receptive field covers location x, then $\hat{O} = \max_x \hat{O}(x)$, i.e., the highest response among all cells is the maximum of $\hat{O}(x)$ among all x. Now define SMAP(x) as the saliency of a visual location x, such that the value of SMAP(x) increases with the likelihood of location x to be selected by bottom-up

mechanisms. From the definitions above, it is then clear that, given an input scene,

(1) SMAP(x) increases with the maximum response $\hat{O}(x) = \max_{x_i=x} O_i$ to x, and regardless of their feature preferences, the less activated cells responding to x do not contribute (1)

(2) $\hat{O}(x)$ is compared with $\hat{O}(x')$ at all x' to determine SMAP(x), since $\hat{O} = \max_x \hat{O}(x)$, (2)

(3) the most likely selected location is $\hat{x} = \text{argmax}_x \text{SMAP}(x)$, where SMAP($x$) is maximum (3)

As salience merely serves to order the priority of inputs to be selected for further processing, only the order of the salience is relevant. However, for convenience we could write equation (1) as SMAP$(x) = \hat{O}(x)/\hat{O}$, with the denominator $\hat{O}$ as the normalization.

Meanwhile, some experimental observations have raised doubts regarding V1's role in determining bottom up saliency. For instance, Hegde and Felleman (2003) found that, from V1 cells tuned to both orientation and color to some degree, the responses to a uniquely colored or uniquely oriented target bar among a background of homogeneously colored or oriented bars are not necessarily higher than the responses to a target bar defined by a unique conjunction of color and orientation. According to the V1 hypothesis, the saliency of a target is determined by its evoked response relative to that evoked by the background. Hence, the response to the most salient item in one scene is not necessarily higher than the response to the intermediately salient item in a second scene, especially when the second scene, with less homogeneity in the background, evokes higher responses in a population. Hence, Hegde and Felleman's finding does not disprove the V1 hypothesis, although it does not add any confidence to it. Furthermore, neurons in the extrastriate cortical areas, such as V2 and V4, are also tuned to many of the basic features that pop out pre-attentively when uniquely present in a scene, so it is conceivable that much of the bottom up selection may be performed by these visual areas higher than V1. Previous frameworks on pre-attentive visual selection (Treisman and Gelade 1980, Wolfe et al 1989, Duncan and Humphreys 1989, Koch and Ullman 1985, Itti and Koch 2000, Itti and Koch 2001) have assumed separate feature maps which process individual features separately. These feature maps are considered to be more associated with the extra-striate areas, some of which seem to be more specialized for some features than others, compared with V1. Since the different basic features in these models, such as orientation and size, act as apparently separable features in visual search, and early (striate and extra-striate) visual areas have cells tuned to conjunctions of features (such as orientation and spatial frequency), the previous frameworks suggest a relatively late locus for the feature maps. Additionally, previous models for bottom up saliency (e.g., Koch and Ullman 1985, Wolfe et al 1989, Itti and Koch 2000) assume that the activities in the feature maps are summed into a master saliency map that guides attentional selection, implying that bottom up saliency map should be in a higher cortical area such as the parietal cortex (Gottlieb, Kusunoki, & Goldberg 1998).

1.1 Conceptual characteristic of the V1 saliency hypothesis

Based on the previous experimental observations and the previous models, it is not clear whether or not V1 contributes only marginally to visual saliency, so that the computations of saliency are significantly altered in subsequent brain areas after V1. If this were the case, the behavior of bottom up selection would be devoid of characteristics of V1 other than some generic mechanisms of iso-feature surround suppression, which, necessary for basic feature pop out, could perhaps be implemented just as well in extra-striate or higher cortical areas (Allman et al 1985). To address this question, we identify the characteristics of V1 and the V1 saliency hypothesis which are different from other visual areas or from other saliency models. First, let us consider the conceptual characteristics. The V1 hypothesis has a specific selection mechanism to select the most salient location from the V1 responses, namely that: the RF location of the most activated cell is the most salient location regardless of the preferred feature of this cell. This means, the activities of the neurons are like universal currency bidding for selection, regardless of the neuron's preferred features (Zhaoping 2006, Zhaoping & Snowden 2006). As argued above, this character led to the saliency value SMAP$(x) \propto \max_{x_i=x} O_i$ at location x, and we will call this the MAX rule to calculate saliency.

It therefore contrasts with many previous bottom up saliency models (Koch and Ullman 1985, Wolfe et al 1989, Itti and Koch 2000) which sum activities from different feature maps, e.g., those tuned to color, orientation, motion directions etc., to determine saliency at each location, and we will refer to this as the SUM rule, SMAP$(x) \propto \sum_{x_i=x} O_i$. Recently, Zhaoping and May (2007) showed that the MAX rule predicts specific interference by task irrelevant inputs on visual segmentation and search tasks, and that such predictions are confirmed psychophysically, see Figure (1). Note that the MAX rule acts on the responses from V1 rather than imposing mechanisms within V1 for creating the responses to select from. It arises from the unique assumption in the V1 saliency hypothesis that no separate feature maps, nor any combination of them, are needed for bottom up saliency (Li 2002). In other words, the MAX rule would not preclude a saliency map in, say, V2, as long as no separate feature maps or any summation of them are employed to create this saliency map. Hence, while the MAX rule supports the V1 hypothesis, this rule by itself cannot be a fingerprint of V1.

1.2 Neural characteristics that can serve as fingerprints of V1

We identify three neural characteristics of V1. First is the abundance of monocular cells. These cells carry the eye of origin information. Most V1 neurons are monocular (Hubel and Wiesel 1968), whereas any higher visual area has only few monocular cells (Burkhalter and van Essen 1986). Second is collinear facilitation, i.e., a neuron's response to an optimally oriented bar within its RF is enhanced when a neighboring bar outside the RF is aligned with the bar within the RF such that they could be seen as segments of a smooth contour. It has been observed in V1 since the 1980s (Nelson and Frost 1985, Kapadia et al 1995), and is inherited by V2 (von der Heydt, Peterhans, & Baumgartner 1984, Bakin et al 2000), but could not exist in visual stages before V1 without any cells tuned to orientation. Collinear facilitation is observed psychophysically only when target and flankers are presented to the same eye, suggesting that the phenomenon depends on links between monocular cells (Huang, Hess, & Dakin 2006), and thus a V1 origin. Third is the feature-specific conjunctive cells in V1. V1 has cells tuned conjuntively to a specific orientation and motion direction, or conjunctively to specific orientation and color (Hubel & Wiesel, 1959; Livingstone & Hubel, 1984; Ts'o & Gilbert, 1988), but has almost no cells tuned conjunctively to specific color and motion direction (Horwitz & Albright, 2005). This is not the case in V2 or V3, where there are cells tuned conjunctively to all of the three pairwise possible conjunctions of feature dimensions (Tamura, Sato, Katsuyama, Hata, & Tsumoto 1996, Gegenfurtner, Kiper, & Fenstemaker 1996, Gegenfurtner, Kiper, Levitt 1997, Shipp 2007 Private communication), and the higher cortical neurons are expectedly selective to more complex input features than V1.

If V1's responses indeed dictate saliency by evoking responses (e.g., via superior colliculus to drive eye movement and selection) before the involvement of the subsequent visual areas, these V1 specific characteristics should be reflected in the corresponding selection behavior. These fingerprints, corresponding to the three neural characterisics of V1, are specifically as follows. Firstly, given V1's monocular cells, and its mechanism of iso-ocular suppression (DeAngelis et al 1994, Webb et al 2005) as an instantiation of iso-feature suppression responsible for a feature singleton to pop out, V1 saliency hypothesis predicts that an ocular singleton should capture attention automatically. It is known that eye of origin information is typically elusive to visual awareness (Wolfe and Franzel 1988, Kolb and Braun 1995, Morgan, Mason, & Solomon 1997). This is consistent with the idea that, unlike higher cortical areas, information available in V1 is usually at most weakly associated with awareness (see reviews by Crick & Koch 1995 and Tong 2003). Hence, attention capture by an ocular singleton even without awareness would be a hallmark of V1, and perhaps the ultimate bottom-up or exogenous visual selection. Secondly, collinear facilitation suggests that, between oriented bars, contextual influences that determine saliency are not isotropic. Consequently, the selection behavior in stimuli consisting of orientation textures should depend on the spatial configuration in these stimuli in specific non-isotropic ways that are signatures of the collinear facilitation mechanism in V1. Thirdly, consider the saliency of a red vertical bar among green vertical bars, and of a red vertical bar among red horizontal bars, and of a red-vertical bar among green-horizontal bars. We will refer to the first two as single-feature (color or orientation)

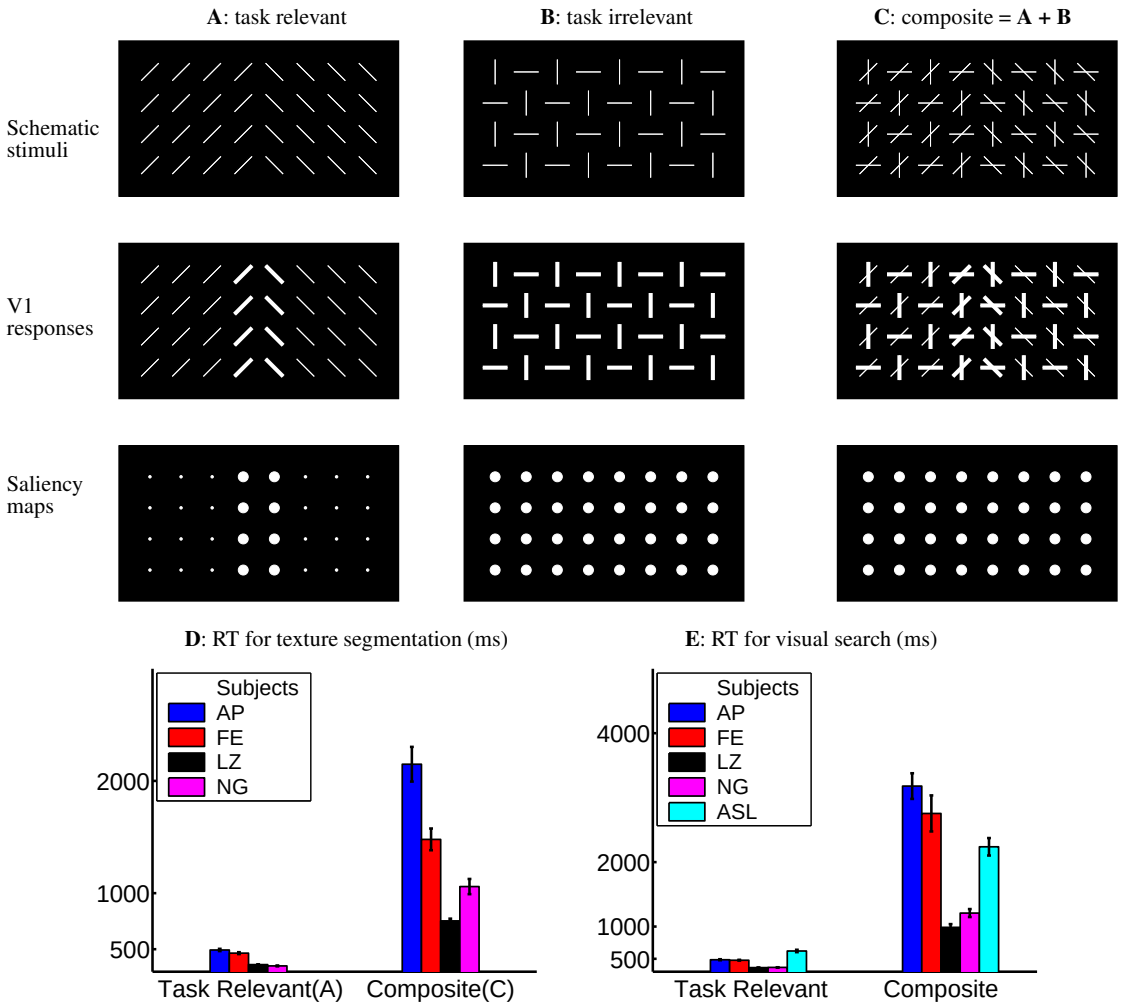


Figure 1: Prediction of the MAX rule by the V1 saliency hypothesis — interference by task irrelevant features, and its psychophysical test (adapted from Zhaoping and May 2007). **A, B, C** are schematics of texture stimuli (extending continuously in all directions beyond the portions shown), each followed by schematic illustrations of its V1 responses, in which the orientation and thickness of a bar denote the preferred orientation and response level, respectively, of the most activated neuron by an input bar. Below each V1 response pattern is a saliency map, in which the size of a disk corresponds to the response of the most activated neuron at the texture element location. The orientation contrasts at the texture border in **A** and everywhere in **B** lead to less suppressed responses to the stimulus bars since these bars have fewer iso-orientation neighbours to evoke iso-orientation suppression. The composite stimulus **C**, made by superposing **A** and **B**, is predicted to be difficult to segment, since the task irrelevant features from **B** interfere with the task relevant features from **A**, giving no saliency highlights to the texture border. **D**: reaction times for texture segmentation testing the prediction (differently colored bars denote different subjects). **E:** like **D**, but for a task to search for an orientation singleton. The stimuli were made from those in the segmentation task by shrinking one of the two texture regions into a single texture element. RT for the composite condition is significantly higher ($p < 0.001$). Stimuli for experiments in Fig. 1,3,4, and 5 consist of 22 rows $\times$ 30 columns of items (of single or double bars) on a regular grid with unit distance 1.6^o of visual angle.

saliency and the last as the (color-orientation) double-feature saliency, and expect that the double-feature singleton should be somewhat more salient than the single-feature ones. The magnitude of this double-feature advantage, or the feature redundancy gain, should depend on whether the conjunctive cells for the two features concerned exist. Hence, the existence of V1's conjuntive cells in some combinations of feature dimensions and not others should create a corresponding, feature dimension specific, pattern of double-feature advantages.

In the next section, we review the behavoral fingerprints of V1 mechanisms in detail, and illustrate how they arise as predictions of the V1 hypothesis. The experimental data confirming these predictions are then shown. All the details of the experiments have been published (Zhaoping and May 2007, Koene and Zhaoping 2007, Zhaoping 2008). The presentation in this paper not only reviews the published results for the purpose of summarizing and highlighting the fingerprints of V1 in saliency, but also presents some different perspectives and analysis of the published results. We summarize with discussions in Section 3.

2 Predicted fingerprints and their experimental tests

To predict behavioral consequences of V1's responses which are used for bottom-up saliency, we need to know the most relevant V1 characteristics, which are summarized as follows: (1) neural tuning to basic features within its receptive fields (RF), such that, e.g., a neuron tuned to color responds more to preferred than to non-preferred colors; (2) iso-feature suppression that suppresses a neuron's response to a preferred feature within its RF when there are inputs of the same feature outside and yet near its RF; (3) general surround suppression, i.e., a neuron's response is suppressed by activities in all nearby neurons regardless of their feature preferences (this suppression is weaker than the iso-feature suppression but introduces interactions between neurons tuned to different features); (4) collinear facilitation — enhancement of a neuron's response to an optimally oriented bar within its RF when a contextual bar outside its RF is aligned with the bar within; (5) neural tuning to conjunctions of orientation and motion direction (OM), or color and orientation (CO), but not to color and motion direction (CM); (6) some V1 neurons are monocular and thus are tuned to eye of origin. Mechanisms (1) and (2) are essential for unique feature pop-out, e.g., a singleton red pops out of many green items since a red-tuned cell responding to the singleton does not suffer from the iso-color suppression imposed on the green-tuned neurons responding to the background items (Li 1999ab). Mechanism (3) will modify the contextual influences to modulate but typically not dictate the saliency outcome, as will be discussed later. One may argue that mechanisms (1-3) (except for the neural tuning to eye of origin and iso-ocular suppression, as specific examples of (1) and (2)) are generic and also present in higher visual areas (Allman et al 1985). V1's fingerprints on saliency behavior will have to arise from mechanisms (4), (5), and (6), which we will show to manifest in the saliency outcome in a predictable way. (Even though V2 also manifests mechanism (4), we consider mechanism (4) as special for V1 given psychophysical evidence (Huang et al 2006) for its V1 origin).

2.1 The fingerprint of V1's monocular cells

V1 is the only cortical area that has a substantial number of cells tuned to ocular origin, i.e., being differentially sensitive to inputs from the different eyes or receiving inputs dominantly from one eye only. Since a V1 neuron's response is suppressed more by contextual inputs presented to the same rather than a different eye (DeAngelis et al 1994, Webb et al 2005), a V1 neuron responding to an ocular singleton, i.e., an input item with a different eye of origin from all other input items, is expected to give a higher response than the V1 neurons responding to any of many identical input items seen through the other eye. In other words, an ocular singleton should be salient to capture attention.

Since inputs that differ only in their eyes of origin typically appear identical to human subjects, it is difficult to directly probe whether an ocular singleton pops out by asking subjects to search for it (Wolfe and Franzel 1988, Kolb and Braun 1995). This fingerprint can however be tested by making the ocular singleton task irrelevant and

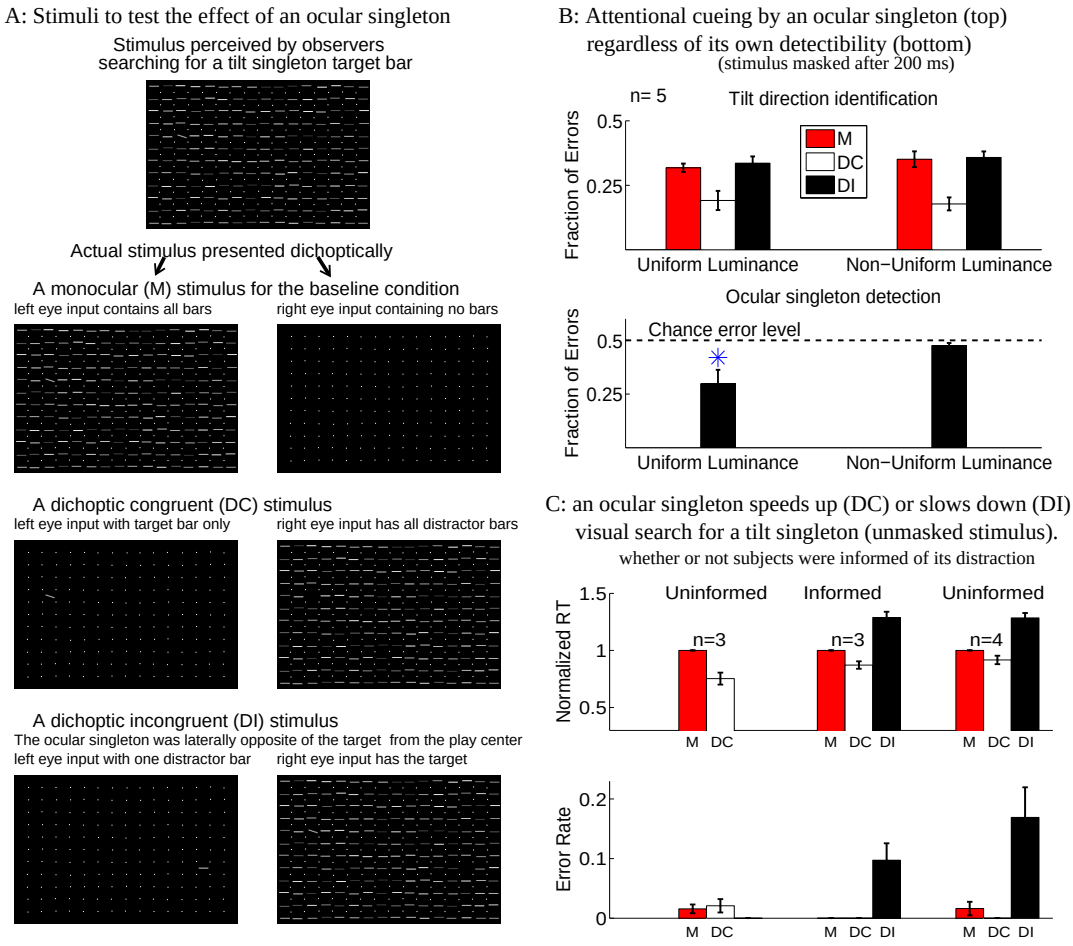


Figure 2: An ocular singleton captures attention automatically. A: Illustrations of the stimulus for visual search for an orientation singleton, and the various dichoptic presentation conditions: dichoptic congruent (DC), dichoptic incongruent (DI), and monocular (M). The actual stimuli used had 22 rows $\times$ 30 columns of bars, spanning $34^o \times 46^o$ in visual angle. From the display center, the search target had an eccentricity of $\sim 15^o$, and at least 12^o horizontal eccentricity. The eye of origin of the task critical bar(s) was random. B: fractions of error trials for reporting the tilt direction of the tilt singleton (top) in a brief (200 ms) display (like in A, except that in half of the trials, all bars had the same (uniform) luminance), and for reporting whether the ocular singleton was present in the same stimuli without the tilt singleton (bottom, '*' denotes significant difference from the chance error level). The left and right halves of the (top and bottom) plots are for when the stimulus bars had uniform or non-uniform (as in A) lumiance values respectively. Tilt identification was best in the DC condition (top), independent of the ability (depending on whether the bars had uniform luminance values) to detect the ocular singleton beyond the chance level (bottom). Data are averages from $n = 5$ subjects, who, before the ocular singleton detection task, were acquainted with the ocular singleton in an example stimulus displayed for as long as necessary. C: Reaction times (top, RT_M, RT_{DC}, and RT_{DI} in M, DC, and DI conditions respectively) and fractions of error trials (bottom) for reporting whether the tilt singleton was in the left or right half of the display which stayed unmasked before subjects' reports. Each subject's RT was normalized by his/her RT_M ($\sim$ 700 ms). All data are averages among subjects. Stimuli were as in A except that all bars had the same (uniform) luminance, and the target and non-target bars were tilted 25^o from horizontal in opposite directions. The left, middle, and right parts of the (top and bottom) plots are results from three different experiments respectively, employing $n = 3$, 3, and 4 subjects respectively. The first experiment (left) included M and DC conditions, both the 2nd (middle) and 3rd (right) experiments included M, DC, and DI conditions. In the 1st and 3rd experiments, subjects were uninformed (and unaware, except for one subject in 3rd experiment who became aware of an attention capturing distractor) of the existence of the various task irrelevant dichoptic conditions. In the 2nd experiment, subjects were informed (before data taking) of a possible attention capturing distractor in some trials. Note that $RT_{DC} < RT_M < RT_{DI}$. The DI condition, when included, caused the most errors.

observe its effect on the performance a task that requires attention to a task relevant location (Zhaoping 2007, 2008). In one experiment, observers searched for an orientation singleton among background horizontal bars. The search display was binocularly masked after only 200 milliseconds, and the subjects were asked to report at their leisure whether the tilt singleton was tilted clockwise or anticlockwise from horizontal. The display was too brief for subjects to saccade about the display looking for the target, which was only tilted 20^o from the background bars. Hence, this task was difficult unless subjects' attention was somehow covertly guided to the target. Unaware to the subjects (except one, the first author), some trials were dichoptic congruent (DC), when the target was also an ocular singleton, some were dichoptic incongruent (DI), when a distractor on the opposite lateral side of the target from the display center was an ocular singleton, and the other trials were monocular (M) when all bars were seen by the same single eye (see Fig. 2A). If the ocular singleton can exogenously cue attention to itself, subjects' task performance should be better in the DC condition. This was indeed observed (Fig. 2B). A control experiment was subsequently carried out to probe whether the same observers could detect the attention capturing ocular singleton if they were informed of its existence. It had the same stimuli except that all bars were horizontal. In randomly half of the trials an ocular singleton was at one of the same locations as before, and the observers were asked to report whether an ocular singleton existed by forced choice. Their performance was better than the chance level only when all bars had the same (uniform) luminance value (Fig. 2B, bottom). Meanwhile, the same ocular singleton, whether it was detectable by forced choice or not, had demonstrated the same deree of cueing effect (Fig. 2B, top) . This suggests that the ocular singleton cued attention completely exogenously to its location, facilitating the identification of the tilt singleton in the DC condition. The M and DI conditions can be seen as the uncued and invalidly cued conditions respectively. Note that the tilt singleton in the M condition, with a 20^o orientation contrast from the background bars, should pop out in a typical visual search task when the search stimulus stays on unmasked, at least when the bars had uniform luminance. Our data suggest that the ocular singleton was more salient than the orientation singleton.

In three additional experiments, the search display stayed unmasked until the subjects responded, and the orientation contrast between the target and distractors was 50^o. Observers were asked to report as soon as possible whether the tilt singleton was in the left or right half of the display. Their reaction times (RTs) in reporting were shorter in the DC, and longer in the DI, than the M condition, regardless of whether the observers were aware or informed of the existence of the different task irrelevant dichoptic conditions (Fig. 2C). Note that RT_{DI}, the RT in the DI condition, was about 200 ms longer than RT_M, the RT in the M condition. This 200 ms difference is about an average fixation duration in typical visual search tasks (Hooge & Erkelens 1998). Hence, our findings suggest that, in typical trials, attention was more quickly attracted to the target in the DC condition, and initially distracted from the target in the DI condition. In particular, our data suggest that, in a typical DI trial, subjects saccaded to the ocular singleton distractor first, realizing that it was not the target, before shifting attention to the target. Hence, an ocular singleton, though elusive to awareness, can effectively compete for attention with an orientation singleton of even 50^o contrast from the background bars. This is consistent with the finding that subjects also made more errors in the DI condition: presumably, in a hurry to respond, they easily mistook the ocular singleton distractor as the target. Furthermore, the high error rates persisted even when the subjects were informed that an attention capturing distractor could appear in some trials (Fig. 2C, the middle group of the data). This suggests that it is not easy to suppress the saliency by an ocular singleton by top-down control.

If the eye of origin feature was as visible to visual awareness as some of the basic features such as color and orientation, it would be considered a basic feature, defined as one when a singleton in the feature dimension has a negligible set size effect in visual search (i.e., when RT does not depend on the number of background items). Its elusiveness to awareness means that subjects cannot do a visual search for an ocular singleton effectively, as shown by Wolfe and Franzel (1988). However, an ocular singleton should make a difficult search easier by eliminating the set size effect. Indeed, we found that the set size effect in searching for a T among L can be eliminated when the target was also an ocular singleton (Zhaoping 2008).

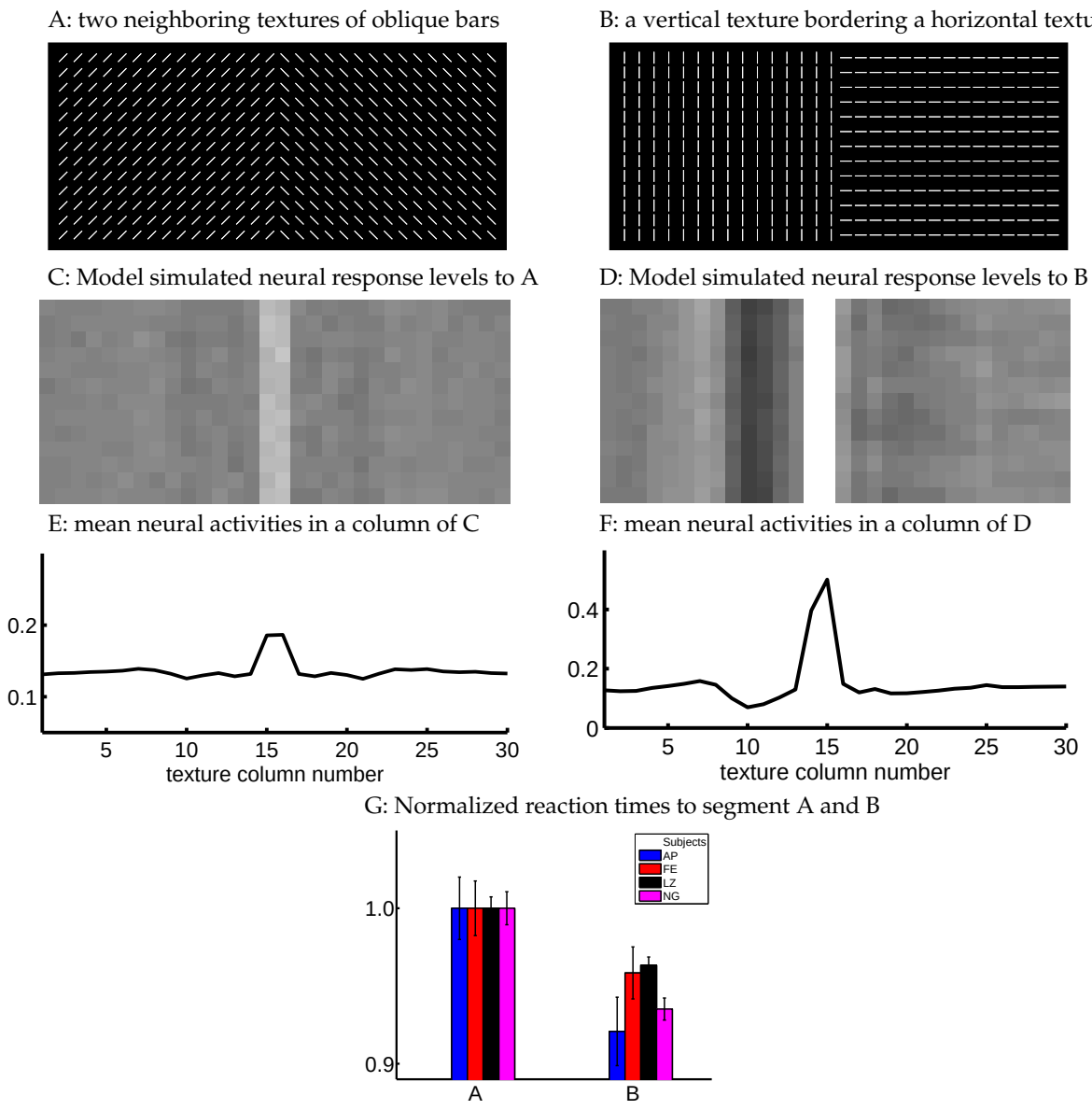


Figure 3: Fingerprint of the collinear facilitation in V1: a texture border with texture bars parallel to the border is more salient. A and B: stimulus patterns for texture segmentation; each contains two neighboring orientation textures with a 90^o orientation contrast at the texture border. The texture border in B appears more salient. C and D: simulation results from a V1 model (Li 1999b, 2000, used in all model simulations in this paper) on the neural activity levels in space for stimulus patterns A and B respectively. Higher activities are visualized by a lighter luminance at the corresponding image location. E and F: neural activities in C and D respectively averaged in each texture column. G: Normalized reaction times of human observers to locate the texture border, in the units of the reaction time RT_A of the subject for stimulus A. The RT_A for various subjects are respectively, 493, 465, 363, 351 milliseconds. For each subject (same as in Fig. 1D), it is statistically significant that $\text{RT}_A > \text{RT}_B$ ($p < 0.05$).

2.2 Fingerprints of V1's collinear facilitation in texture segmentation behavior

The experiments showing the fingerprints in this subsection of the paper are part of a previous study (Zhaoping and May 2007). Here we illustrate the fingerprints with a simulation of V1's behavior using a previously developed

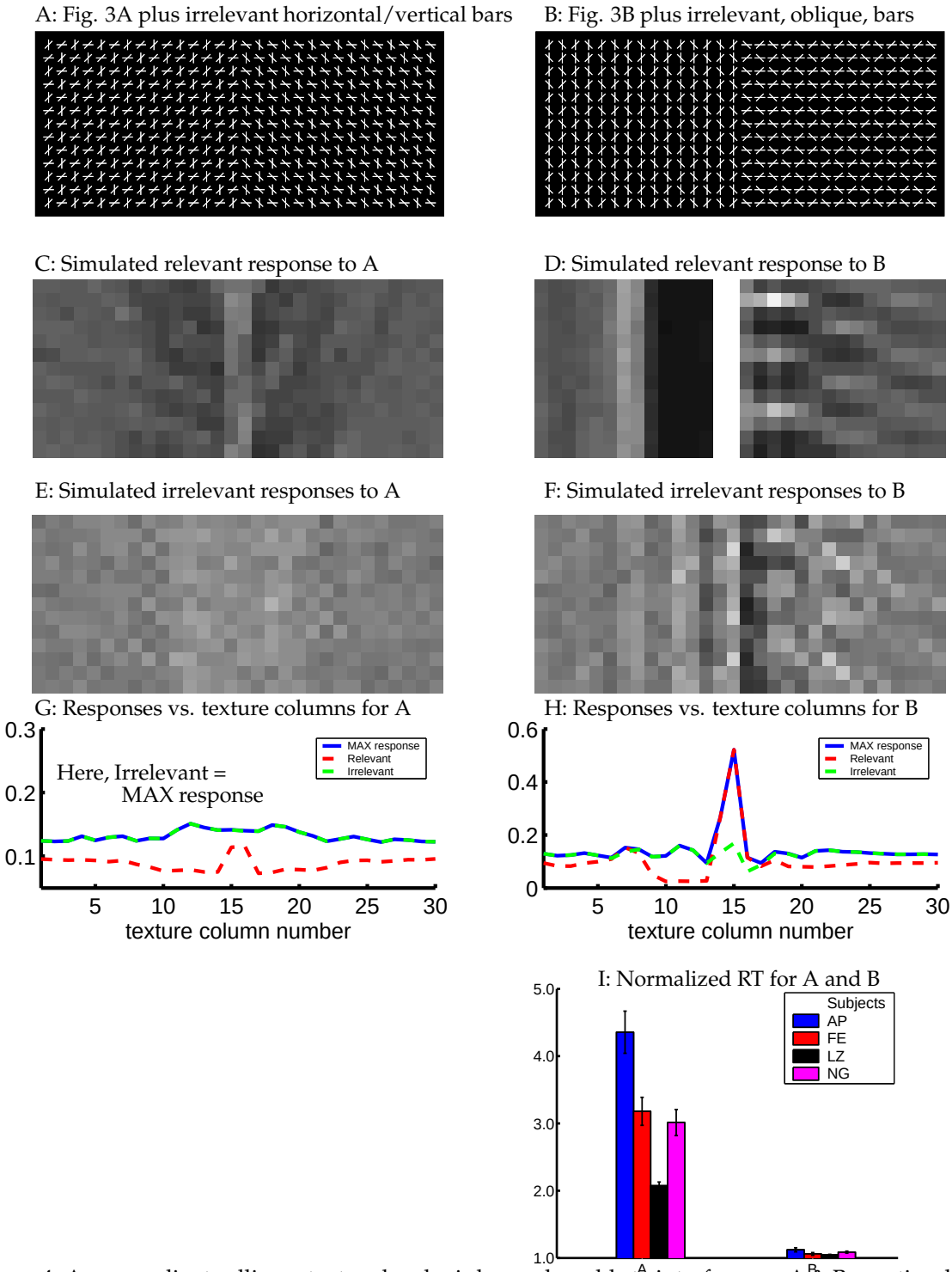


Figure 4: A more salient collinear texture border is less vulnerable to interference. A & B are stimuli in Fig. 3AB after superposing task irrelevant bars which form a checkerboard pattern. The simulated relevant responses respectively are in C & D, and the irrelevant responses in E & F, using the same format as Fig 3CD. G & H plot the responses vs. texture columns, for relevant, irrelevant, and the maximum of them, i.e., saliency. I: Normalized reaction times to A and B. The subjects are the same as in Fig. 1D, Normalized RT for each subject is obtained by dividing the RT for A and B, respectively, by the RT of the subject for the corresponding stimulus without irrelevant bars (i.e., Fig. 3A and Fig. 3B, respectively). The interference in B, even though significant (i.e., the normalized RT is significantly larger than 1 for each subject with $p < 0.02$), is much less than in A.

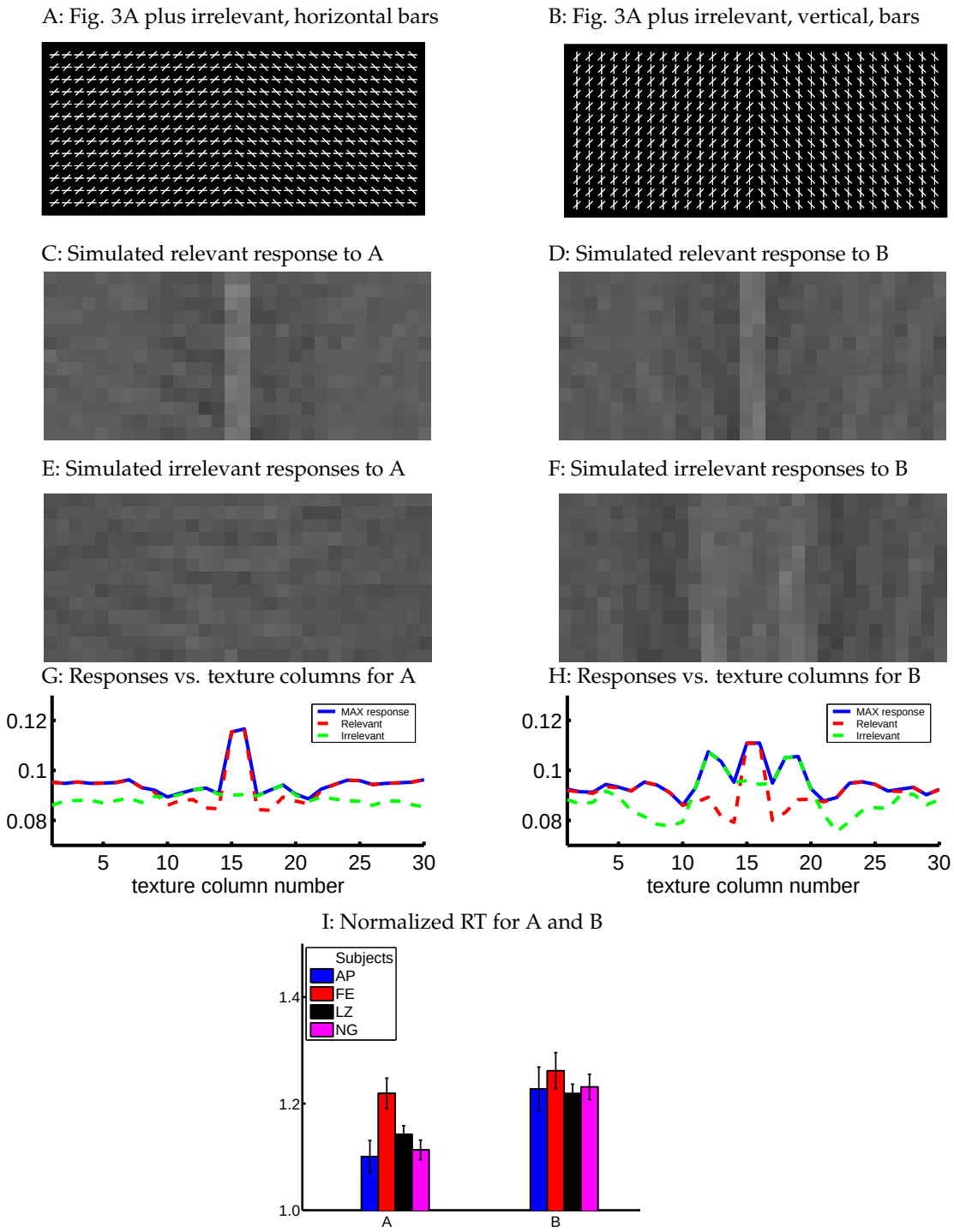


Figure 5: Differential interference by irrelevant bars due to collinear facilitation. A-H have the same format as Fig. 4A-H. Their contents differ due to the change of stimuli in A & B, which have Fig. 3A as the relevant stimulus and uniformly horizontal (A) or vertical bars (B) as irrelevant stimuli. I: Normalized reaction times to A and B: each is the RT divided by the RT of the same subject for Fig. 3A (the stimulus without irrelevant bars). Subjects are the same as in Fig. 1D. In three out of four subjects, RT_B for B is significantly longer than that RT_A for A ($p < 0.01$). By matched sample t-test across subjects, the $RT_B > RT_A$ significantly ($p < 0.01$). For each subject, RTs for both A and B are significantly longer ($p < 0.0005$) than that for Fig. 3A (the stimulus without irrelevant bars).

V1 model (Li 1999b, 2000), and compare the model's behavior with human data. Figure (3) shows this fingerprint. Fig. (3A) and (3B) both have two orientation textures with a 90^o contrast between them. The texture borders pop out automatically, as the saliency of such texture borders increases with the orientation contrast at the border (Nothdurft 1992). A texture border bar is salient since it has fewer iso-orientation neighbors than the texture bars away from the border, and hence the neuron responding to it experiences weaker iso-orientation suppression. However, in Fig. (3B), the vertical texture border bars in addition enjoy full collinear facilitation, since each has more collinear neighbors than other texture border bars in either Fig. (3A) or Fig. (3B). The vertical texture border bars are thus more salient than other border bars. In general, given an orientation contrast at a texture border, the border bars parallel to the texture border are predicted to be more salient than other border bars (Li 1999b, 2000), and we call these border bars collectively as a collinear border.

We hence predict that the border in Fig. (3A) takes longer to locate than the border in Fig. (3B). This is tested in an experiment with such stimuli in which the texture border is sufficiently far from the display center to bring performance away from ceiling. We asked human subjects to press a left or right button as soon as possible after stimulus onset to indicate whether the border is in the left or right half of the display. Our prediction is indeed confirmed (Fig. (3G)). Higher saliency of a collinear border is likely the reason why Wolfson and Landy (1995) observed that it is easier to discriminate the curvature of a texture border when it is collinear than otherwise.

Note that, since both texture borders in Fig. (3A) and Fig. (3B) are salient enough to require only short RTs, and since RTs can not be shorter than a certain minimum for each subject, a large difference in the degrees of border highlights in our two stimuli can only give a small difference in their required RTs. We can unveil this predicted large difference in V1 responses by *interference*, explained in Fig. (4), thereby demonstrating another manifestation of the fingerprint in figure (3). Fig. (4A) is made by superposing onto Fig. (3A) a checkerboard pattern of horizontal and vertical bars, just like in Fig. (1), and analogously, Fig. (4B) by superposing onto Fig. (3B) left-oblique and right-oblique bars. The superposed checkerboard patterns are irrelevant to the task of segmenting the textures. We refer to the responses to the task relevant and irrelevant stimuli as "relevant" and "irrelevant" responses, respectively; similarly, the neuron populations tuned to the relevant and irrelevant orientations are referred to as "relevant" and "irrelevant" neuron populations, respectively. By the MAX rule in the V1 saliecy hypothesis, the irrelevant responses compete with the relevant ones to dictate saliency at each location. If they win the competition at some locations, they can interfere with segmentation by misleading visual attention and thus prolong the RT. As illustrated in Fig. (1), the irrelevant response level to any texture element location is comparable to that of the relevant response to the border, since an irrelevant bar has as few iso-orientation neighbors as a relevant texture border bar. Consequently, the maximum neural response at each texture element location is roughly the same across space, and the texture border highlight is now reduced or diminished. Indeed, RTs (Fig. (4)I) for the same texture segmentation task are much longer for stimuli Fig. (4)A and (4)B than those for stimuli without irrelevant bars (Fig. (3)). Meanwhile, it is clear that the RT for Fig. (4)B is much shorter than the RT for Fig. (4)A, as the interference is much weaker in Fig. (4)B. The extra salient, collinear, vertical border bars evoke responses that are much higher than the irrelevant responses, and are thus less vulnerable to being submerged by the higher background saliency levels, even though the relative border salience is somewhat reduced due to the raised background salience levels.

The arguments above are qualitative since we included only iso-orientation suppression and collinear facilitation in our argument, and have omitted for simplicity the effect of general surround suppression which, although weaker than the iso-orientation suppression, causes nearby neurons responding to different orientations to suppress each other and thus modulate the overall spatial patterns of the responses. To verify our qualitative arguments, we simulated the V1 responses using our previously developed V1 model (see Li 1998, 1999b for details sufficient for the reproduction of the model behavior), which includes all three forms of the contextual influences: iso-orientation suppression, collinear facilitation, and general suppression. The model behavior, shown in Fig. (3C-F) and Fig. (4C-F), confirmed our qualitative analysis. In viewing the model responses, note that the highest possible responses from the model neurons (at saturation) are set to 1, and that the model includes some levels of noise simulating in-

put or intrinsic noise in the system. Also note that, without knowledge of quantitative details of the V1 mechanisms, the quantitative details of our model should be seen only as an approximation of the reality to supplement our qualitative predictions. Nevertheless, as the model parameters were previously developed, fixed, and published, our predictions and simulation results were produced without model parameter tuning[1]

Additional qualitative details, although not affecting our conclusions here, are also visible in the model behavior. For example, a local suppression of relevant responses near the texture border is due to the stronger iso-orientation suppression from the more salient (relevant) border bars. This local suppression is particularly strong next to the most salient vertical border bars (in Fig. (3)D and (3)F). We call this local suppression region next to the border the *border suppression region* (Zhaoping 2003).

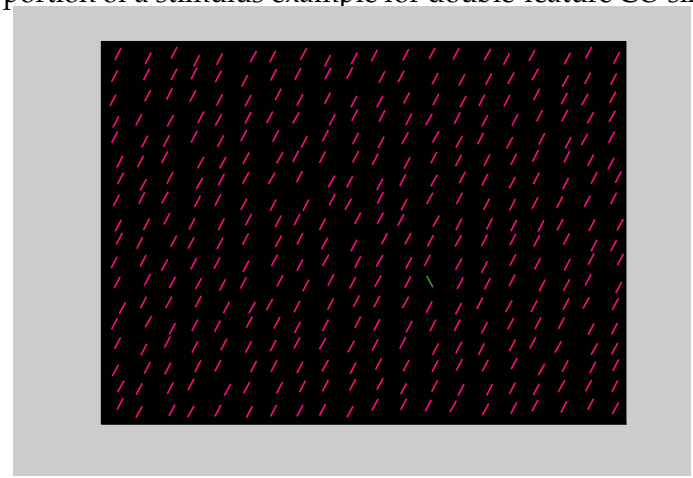
Figure (5) demonstrates another fingerprint of the collinear facilitation. Fig. (5)A-H are analogous to Fig. (4)A-H. The task relevant stimulus component is that of Fig. (3)A, while the task irrelevant stimulus components are the horizontal bars in Fig. (5A) and vertical bars in Fig. (5B). Without orientation contrast among the task irrelevant bars, the irrelevant responses have a similar level to relevant responses in the background, since the level of iso-orientation suppression is about the same among the irrelevant bars as that among the relevant bars in the background. Based on the MAX rule, if there were no general surround suppression enabling interaction between differently oriented bars, there would be no interference to segmentation based on the relevant bars, which evoke a response highlight at the texture border. However, general surround suppression induces interactions between local relevant and irrelevant neurons. Thus spatially inhomogeneous relevant responses induce inhomogeneity in the irrelevant responses, despite the spatial homogeneity of the irrelevant stimulus. In particular, because the relevant responses in the border suppression region generate weaker general suppression, the local irrelevant responses are slightly higher (or less suppressed). Hence, the irrelevant response as a function of the texture column number exhibits local peaks next to the texture border, as apparent in Fig. (5GH) (and Fig. (4GH)). These irrelevant response peaks not only dictate the local saliencies, but also reduce the relative saliency of the texture border, thereby inducing interference. Fig. (5A) and Fig. (5B) differ in the direction of the collinear facilitation among the irrelevant bars: it is in the direction across the border in Fig. (5A) and along the border in Fig. (5B). Mutual facilitation between neurons tends to equalize their response levels, i.e., smooth away the response peaks or variations in the direction along the collinear facilitation. Consequently, the irrelevant response peaks near the border are much weaker for Fig. (5A) (see Fig. (5EG)) than for Fig. (5B) (see Fig. (5FH)), predicting a stronger interference in Fig. (5B) than in Fig. (5A). This is indeed confirmed in our data for the same segmentation task (Fig. (5I)).

2.3 Fingerprints of V1's conjunctive cells in bottom up saliency

In figure (6), among a background of purple-right-tilted bars, a unique green-left-tilted bar is salient due to its unique color *and* its unique orientation. We call such a singleton a double-feature singleton, and a singleton unique in only one feature is called a single-feature singleton. By measuring the reaction times to search for the singletons, one can measure the amount of the double-feature advantage, i.e., how much more salient the double-feature singleton is compared to the corresponding single-feature singletons. We will explain below that the double-feature advantage depends in specific ways on the existence of conjunctive cells or neurons tuned conjunctively to features in both of the relevant feature dimensions, e.g., color and orientation. Since V1 has neurons tuned conjunctively to color (C) *and* orientation (O), or to orientation *and* motion direction (M), but none conjunctively to the color *and* motion direction, the V1 saliency hypothesis predicts specific double-feature advantages among various feature dimensions.

[1]Methods for all the model simulations for this paper are as follows. Each displayed model response area of $30\times$ 13 texture grid locations is in fact only a central small portion of a sufficiently large area of textures without the wrap around or periodic boundary condition, in order to avoid the artifacts of the boundary conditions. The model inputs to each visual texture bar was set at a level $\hat{I} = 1.9$ (in notations used in Li 1999b), corresponding to intermediate contrast level condition. For each input image, the model simulates the neural responses for a duration of at least 12 time constants. A model neuron's output was temporally averaged to get the actual outputs displayed in the figures.

A: a portion of a stimulus example for double-feature CO singleton search



B: Normalized RT for double-feature singletons

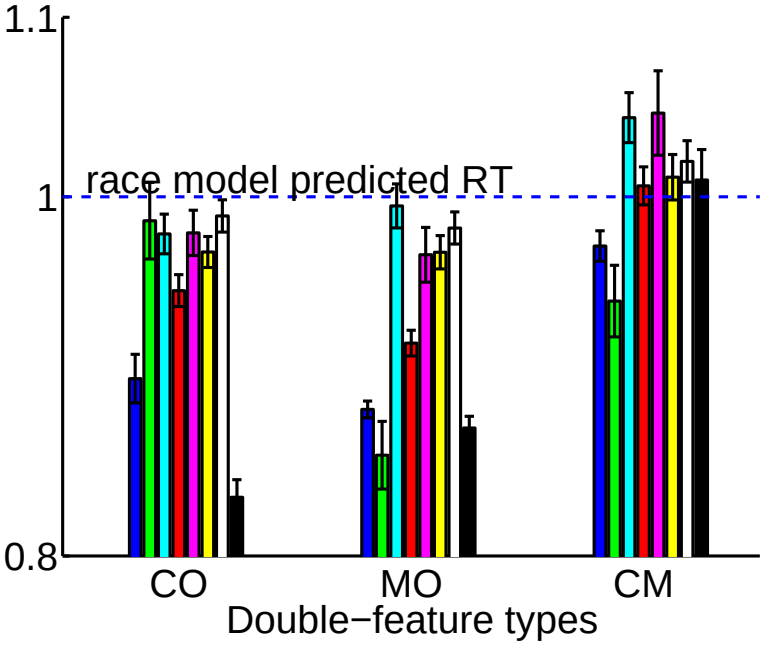


Figure 6: Fingerprint of the types of the conjunctive cells in V1. A: a portion of an example stimulus to search for a CO singleton. B: the normalized RTs (by the race model predicted RTs, which are of order 500 ms) for the double-feature singletons for seven subjects. Different subjects are denoted by the differently colored bars, only two subjects, denoted by blue and green colors (the first two subjects from the left in each double feature group), are non-naive. Error bars denote standard errors of the mean. By matched sample 2-tailed t-tests, the observed RT^{CO} and RT^{MO} for the double-feature singletons CO and MO are signficantly ($p = 0.03$ and 0.009 respectively) shorter than predicted by the race model, whereas the observed RT^{CM} for the double feature singleton CM is not significantly ($p = 0.62$) different from the race model prediction. More details are available in Koene and Zhaoping 2007.

Take the example of a color and orientation double-feature, denoted as CO, and the corresponding single features as C and O respectively. To each colored bar, let the neurons respond with outputs O_C, O_O, and O_{CO} respectively, from neurons (or neural populations) tuned only to C, only to O, or conjunctively to CO. We use superscript to denote the nature of the bar, so (O^C_C, O^C_O, O^C_{CO}) is the triplet of responses to a color singleton, (O^O_C, O^O_O, O^O_{CO}), to an orientation singleton, $(O^{CO}_C, O^{CO}_O, O^{CO}_{CO})$ to a double-feature singleton, and (O^B_C, O^B_O, O^B_{CO}) to one of the many bars in the background.

For a neuron tuned only to color or orientation, its response should be independent of feature contrast in other

feature dimensions. Hence

$$O_C^{CO} \approx O_C^C, \quad O_O^{CO} \approx O_O^O, \quad O_C^O \approx O_C^B, \quad O_O^C \approx O_O^B. \tag{4}$$

Furthermore, iso-color and iso-orientation suppression implies

$$O_C^C > O_C^B \quad \text{and} \quad O_O^O > O_O^B. \tag{5}$$

And generalizing iso-feature suppression to the conjunctive cells, we expect

$$O_{CO}^{CO} > O_{CO}^O, \quad O_{CO}^{CO} > O_{CO}^C, \quad O_{CO}^O > O_{CO}^B, \quad O_{CO}^C > O_{CO}^B \tag{6}$$

The MAX rule states that the maximum response $O_{\text{max}}^{\alpha} \equiv \max(O_C^{\alpha}, O_O^{\alpha}, O_{CO}^{\alpha})$ determines the saliency of the bar for $\alpha = C, O, CO$, or B. With and without the conjunctive cells, we denote O_{max} by O_{max} (conj) and O_{max}(base) respectively, hence

$$O_{\text{max}}^{\alpha}(\text{base}) = \max[O_C^{\alpha}, O_O^{\alpha}] \quad \text{and} \quad O_{\text{max}}^{\alpha}(\text{conj}) = \max[O_C^{\alpha}, O_O^{\alpha}, O_{CO}^{\alpha}] \geq O_{\text{max}}^{\alpha}(\text{base}) \tag{7}$$

Since the singletons pop out, we have, with or without the conjunctive cells,

$$O_{\text{max}}^C, O_{\text{max}}^O, O_{\text{max}}^{CO} \gg O_{\text{max}}^B. \tag{8}$$

Without conjunctive cells, we note with equation (4) that

$$O_{\text{max}}^B(\text{base}) = \max(O_C^B, O_O^B) \approx \max(O_C^O, O_O^C) \tag{9}$$

Then, combining equalities and inequalities (4), (5), (7), (8), and (9) gives

$$O_{\text{max}}^C(\text{base}) = O_C^C, \qquad O_{\text{max}}^O(\text{base}) = O_O^O \tag{10}$$

$$O_{\text{max}}^{CO}(\text{base}) = \max[O_C^C, O_O^O] = \max[O_{\text{max}}^C(\text{base}), O_{\text{max}}^O(\text{base})] \tag{11}$$

So the double-feature singleton is no less salient than either single-feature singleton. With conjunctive cells, combining the equalities and inequalities (4 - 8)

$$\begin{aligned} O_{\text{max}}^{CO}(\text{conj}) &= \max[O_C^{CO}, O_O^{CO}, O_{CO}^{CO}] \\ &= \max[O_C^C, O_O^O, O_{CO}^{CO}] \\ &= \max[\max(O_C^C, O_C^O), \max(O_O^O, O_O^C), \max(O_{CO}^C, O_{CO}^O, O_{CO}^{CO})] \end{aligned}$$

The last equality arises from noting $O_C^C > O_C^O$, $O_O^O > O_O^C$, and $O_{CO}^{CO} > O_{CO}^C, O_{CO}^O$. Now re-arranging the variables in the various max(...) gives

$$\begin{aligned} O_{\text{max}}^{CO}(\text{conj}) &= \max[\max(O_C^C, O_O^C, O_{CO}^C), \max(O_C^O, O_O^O, O_{CO}^O), O_{CO}^{CO}] \\ &= \max[O_{\text{max}}^C(\text{conj}), O_{\text{max}}^O(\text{conj}), O_{CO}^{CO}] \geq \max[O_{\text{max}}^C(\text{conj}), O_{\text{max}}^O(\text{conj})] \end{aligned} \tag{12}$$

The double-feature singleton can be more salient than both the single-feature singletons if there are conjunctive cells whose response O_{CO}^{CO} has a non-zero chance of being the dictating response.

Due to the variabilities in the neural responses, the actual neural output in a single trial may be seen as drawn randomly from probability distributions (pdfs). So O_{max}^C, O_{max}^O, and O_{CO}^{CO} are all random variables from their respective pdfs, making O_{max}^{CO} (which is the maximum of these three random variables) another random variable. As O_{max}^{α} determines RT by some monotonically decreasing function RT(O_{max}^{α}) to detect the corresponding input item α, variabilities in neural responses give variabilities in RT^C, RT^O, or RT^{CO} to detect, respectively, the singleton unique in color, in orientation, or in both features. Hence, equations (11) and (12) lead to

$$RT^{CO}(\text{base}) = \min(RT^C, RT^O) \tag{13}$$

$$RT^{CO}(\text{conj}) = \min[RT^C, RT^O, RT(O_{CO}^{CO})] \leq \min(RT^C, RT^O) = RT^{CO}(\text{base}) \tag{14}$$

Hence, without conjunctive cells, RT^{CO} to detect a double-feature singleton can be predicted by a race model between two racers $O^C_{\max}$ and $O^O_{\max}$, with their respective racing times, RT^C and RT^O, as the RTs to detect the corresponding single-feature singletons. With conjunctive cells, RT^{CO} can be shorter than predicted by this race model. Averaged over trials, as long as the additional racer O^{CO}_{CO} has a non-zero chance of winning the race, the mean RT^{CO} should be shorter than predicted by the race model based only on the RTs for detecting the two single-feature singletons.

Hence, the fingerprints of V1's conjunctive cells are predicted as follows: compared to the RT predicted by the race model from the RTs for the corresponding single-feature singletons, RTs for the double-feature singleton should be shorter if the singleton is CO or OM, but should be the same as predicted if the singleton is CM.

We tested for these fingerprints in a visual search task for a singleton bar among 659 background bars (Koene and Zhaoping 2007). Any bar, singleton or not, is about 1×0.2^o in visual angle, takes one of the two possible iso-luminant colors (green and purple), tilted from vertical to either left or right by a constant amount, and moves left or right by a constant speed. All the background bars are identical to each other by color, tilt, and motion direction, and the singleton pops out by unique color, tilt, or motion direction, or any combination of them. The singleton had a 10^o eccentricity from the display center. The subjects had to press a button as soon as possible to indicate whether the singleton was in the left or right half of the display regardless of the singleton conditions which were randomly interleaved and unpredictable by the subjects. To test the predictions, we compare the RTs, e.g., RT^{CO}, for the double-feature singletons with the predictions from the race model, e.g., RT^{CO}(base). The RTs predicted by the race model were calculated from the RTs for the single-feature singletons using Monte Carlo simulation methods by equation (13) as follows. For instance, with features C, O, and CO, we randomly obtain one sample each from the collected data of RT^C and RT^O respectively, and equation (13) is then used to obtain a simulated sample of RT^{CO}(base). Sufficient number of samples can be generated by these Monte Carlo methods to obtain a histogram distribution of RT^{CO}(base) to compare with the human data RT^{CO} to test whether RT^{CO} $<$ RT^{CO}(base).

Figure (6) plots the observed RTs normalized by the race model predicted RTs for the double-feature singletons. The results confirm the predicted fingerprint. By matched sample 2-tailed t-tests, the observed RT^{CO} and RT^{OM} for the double-feature singletons CO and OM are signficantly shorter than predicted by the race model, whereas the observed RT^{CM} for the double feature singleton CM is not significantly different from the race model prediction. The normalized RT^{CO} and RT^{OM} are not significantly different from each other, but are significantly shorter than the normalized RT^{CM}. Double-feature advantage for the CO singleton has also been observed previously (Krummenacher, Muller, & Heller 2001). Nothdurft (2000) used a discrimination task, without requiring subjects to respond as soon as possible, and found no qualitative dependence of the double-feature advantages on the feature dimensions. We believe that reaction time tasks like ours are better suited for probing bottom up selection which by nature acts quickly and transiently (Jonides 1981, Nakayama and Mackeben 1989, van Zoest and Donk 2004).

3 Summary and Discussion

We modelled and derived the predictions of visual saliency behavior from neural properties known to be specific to V1, namely: (1) the existence of cells tuned to eye-of-origin of the inputs, (2) collinear facilitation between neurons, and (3) the existence of only certain types of conjunctively tuned neurons. Our predictions are consequences of combining (1) these V1 neural properties and the iso-feature suppression via intra-cortical interactions in V1, (2) the hypothesis that the receptive field location of the most responsive V1 neuron is the most likely to be selected in the bottom up manner, and (3) the assumed shorter RTs for higher saliencies of the target locations in visual segmentation and search tasks. We presented experimental data confirming these predictions, thereby lending support to the V1 saliency hypothesis.

Previous frameworks for visual saliency and selection (Treisman and Gelade 1980, Koch and Ullman 1985,

Duncan and Humphreys 1989, Wolfe et al 1989, Itti and Koch 2000) have relied on the asumption that each feature dimension is independently processed in the early vision. While it has been known that neural coding in early visual cortices are not independent, it has been assumed or hoped that at a functional level the feature independence would be achieved, or that neural properties specific to V1 would not be manifested so precisely in the visual selection behavior, such as the reaction time based segmentation and search tasks. For example, some of these works (Koch and Ullman 1985, Wolfe et al 1989, Itti and Koch 2000) have assumed seperate feature maps to process visual inputs of various features, and that the activations from different feature maps are then summed, by the SUM rule, into a master saliency map to guide selection. Such a framework implies that the visual saliency map should be in higher cortical areas such as lateral intraparietal cortex (LIP) (Gottlieb et al 1998).

In a previous work by two of us (Zhaoping and May 2007), behavioral data on the interference by irrelevant features, like the ones in Fig. 1 and 3, were presented to confirm the MAX rule which, as shown in Introduction, arises directly from the V1 saliency hypothesis implying that no separate feature maps, nor any combination of them are needed for bottom up selection. However, the MAX rule in itself does not preclude another cortical area from being the locus for the visual saliency map. In fact it does not even preclude the separate processings of different features, as long as the selection is done by attending to the receptive field location of the most activated feature unit regardless of its feature map origin. One could even, in principle, modify the previous saliency framework to suit the MAX rule without V1 specific neural properties, simply by replacing the SUM rule of the previous models by the MAX rule when combining activations of the separate feature maps to create a master saliency map.

It is therefore important, for our purpose, to affirm or refute the V1 saliency hypothesis by identifying the predicted fingerprints of V1 in bottom up saliency behavior. In this paper, the behavioral fingerprints specifically identify V1 since they rely on neural mechanisms, the existence of monocular cells, collinear facilitation, particular types of conjunctive cells, specific to V1 only. In particular, automatic attraction to attention by ocular discontinuity exclude V2 and higher visual areas since V1 is the only visual cortical area with a substantial number of monocular cells, and our finding of zero double-feature advantage (over the race model prediction) of the color and motion double-feature singleton cannot be explained by V2 mechanisms, since V2 and V3 contains neurons tuned conjunctively to color and motion direction (Tamura et al 1996, Gegenfurtner, Kiper, & Fenstemaker 1996, Gegenfurtner, Kiper, Levitt 1997, Shipp private communication 2007) and would create the double-feature advantage by our arguments in section 2.3.

It is likely that V1's saliency map is read by the superior colliculus which receives input from V1 and directs gaze and thus attention. Indeed, microstimulation of the V1 cells can make monkey saccade to the receptive field location of the stimulated cell, presumably via V1's drive to Superior collicus (Tehovnik et al 2003). The selection of the receptive field of the most active V1 neuron could be done in principle in a single step by the Superior colliculus. In practice, it is likely that this selection is partially or at a local level carried out by the deeper layers of V1 which receive inputs from layer 2-3 V1 neurons (where intra-cortical interactions for contextual influences are implemented) and send outputs to the Superior Colliculus. Specifically, it is possible that some V1 neurons in layer 5-6 carry out a local MAX rule to relay the local maximum responses (of the layer 2-3 cells) to the superior colliculus which carries out a global MAX rule to identify the selected location — this is an empirical question to be answered experimentally.

The V1 saliency hypothesis, however, does not preclude V1 from contributing to other functional goals such as object recognition and learning. Nor does it preclude higher cortical areas, such as V2, from contributing additionally to bottom up saliency. Indeed, the Superior colliculus receives inputs from many higher cortical areas (Shipp 2004). It is likely that V1's contribution to bottom up saliency is mainly dominant for the time duration immediately after exposure to visual inputs. Even though V2 and higher cortical areas should have the neural signals and information that would provide the double feature advantage for the color-motion singleton, our finding of a lack of this advantage implies that the Superior Colliculus or some other brain area made the decision for attention shift without waiting for such information to arrive in our task and stimulus arrangement. This is not surprising since

being fast is presumably one of the priorities of bottom-up attentional shifts — as long as there is sufficient neural signal or information to arrive at a clear decision for a winner for attention, it is not imperitive to ponder or dawdle for a refined decision. With a longer latency, especially for inputs when V1 signals alone are too equivocal to select the salient winner within that time duration, it is likely that the contribution from higher visual areas will increase relatively. These contributions from higher visual areas to bottom up saliency are in addition to the top-down selection mechanisms that further involve mostly higher visual areas (Tsotsos 1990, Desimone and Duncan 1995, Yantis and Serences 2003). Meanwhile, the bottom-up saliency signals observed in higher level visual areas, such as LIP (Gottlieb et al 1998) and frontal eye field (FEF) (Schall and Thompson 1999), are likely relayed from lower visual areas, particularly V1, rather than computed or created within these higher areas.

The feature-blind nature of the bottom up V1 selection also does not prevent top-down selection and attentional processing from being feature selective (Wolfe et al 1989, Treue and Martinez-Trujillo 1999, Chelazzi, Miller, Duncan, & Desimone 1993), so that, for example, the texture border in Fig. 3A could be located through feature scrutiny or recognition rather than saliency. By exploring the potentials and limitions of the V1 mechanisms for bottom up selection, it could position us better to understand the roles of the higher visual areas and top-down attention. After all, what V1 could not do must be carried out by higher visual areas, and the top-down attentional selection must work with or against the bottom up selectional mechanisms in V1 (Zhaoping and May 2007, Zhaoping and Dayan 2006, Zhaoping and Guyader 2007).

Acknowledgement Work supported in part by the Gatsby Charitable Foundation and by the UK Research Council.

References

[1] Allman J, Miezin F, McGuinness E. (1985) Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu Rev. Neurosci.* 8:407-30

[2] Bakin JS, Nakayama K, Gilbert CD. (2000) Visual responses in monkey areas V1 and V2 to three-dimensional surface configurations. *J Neurosci.* 20(21):8188-98.

[3] Beck DM, Kastner S. (2005) Stimulus context modulates competition in human extra-striate cortex. *Nature Neuroscience* 8(8):1110-6.

[4] Burkhalter A, van Essen DC (1986) Processing of color, form, and disparity information in visual areas VP and V2 of ventral extrastriate cortex in the macaque monkey. *J. Neurosci.* 6(8):2327-51.

[5] Chelazzi L, Miller EK, Duncan J, Desimone R. (1993) A neural basis for visual search in inferior temporal cortex. *Nature* 363(6427):345-7.

[6] Crick F, Koch C. (1995) Are we aware of neural activity in primary visual cortex? *Nature* 375(6527):121-3.

[7] DeAngelis GC, Freeman RD, Ohzawa I. (1994) Length and width tuning of neurons in the cat's primary visual cortex. *J Neurophysiol.* 71(1):347-74.

[8] Desimone R., Duncan J. (1995) Neural mechanisms of selective visual attention. *Ann. Rev. Neuroscience.* 18:193-222.

[9] Duncan J., Humphreys G.W. (1989) Visual search and stimulus similarities. *Psychological Rev.* 96, 433-58.

[10] Gegenfurtner KR, Kiper DC, Fenstemaker SB. (1996) Processing of color, form, and motion in macaque area V2. *Vis Neurosci.* 13(1):161-72.

[11] Gegenfurtner KR, Kiper DC, Levitt JB. (1997) Functional properties of neurons in macaque area V3. *J. Neurophysiol.* 77(4):1906-23.

[12] Gilbert C.D., Wiesel T.N. (1983) Clustered intrinsic connections in cat visual cortex. *J. Neurosci.* 3(5):1116-33.

[13] Gottlieb JP, Kusunoki M, Goldberg ME. (1998) The representation of visual salience in monkey parietal cortex. *Nature* 391(6666):481-4.

[14] Hegde J. Felleman DJ (2003) How selective are V1 cells for pop-out stimuli? *J. Neurosci.* 23(31):9968-80.

[15] Hirsch JA, Gilbert CD. (1991) Synaptic physiology of horizontal connections in the cat's visual cortex. *J. Neurosci.* 11(6):1800-9.

[16] Hooge IT, Erkelens CJ. (1998) Adjustment of fixation duration in visual search. *Vision Res.* 38(9):1295-302.

[17] Horwitz GD, Albright TD. (2005) Paucity of chromatic linear motion detectors in macaque V1. *Journal of Vision 5(6):525-33*.

[18] Huang PC, Hess RF, Dakin SC. (2006) Flank facilitation and contour integration: different sites. *Vision Res.* 46(21):3699-706.

[19] Hubel DH, Wiesel TN. (1959) Receptive fields of single neurones in the cat's striate cortex. *J. Physiology,* 148:574-91.

[20] Hubel DH Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. *J Physiol.* 195(1):215-43.

[21] Itti L, Koch C. (2001) Computational modelling of visual attention *Nature Rev. Neurosci.* 2(3):194-203.

[22] Itti L., Koch C. (2000) A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* 40(10-12):1489-506.

[23] Jones HE, Grieve KL, Wang W, Sillito AM. (2001). Surround suppression in primate V1. *J. Neurophysiol.* 86(4):2011-28.

[24] Jonides J. (1981) Voluntary versus automatic control over the mind's eye's movement In J. B. Long & A. D. Baddeley (Eds.) *Attention and Performance IX* (pp. 187-203). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

[25] Kapadia MK, Ito M, Gilbert CD, 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.

[26] Knierim JJ., Van Essen DC (1992) Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophysiol.* 67(4): 961-80.

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

[28] Koene AR and Zhaoping L. (2007) Feature-specific interactions in salience from combined feature contrasts: Evidence for a bottom-up saliency map in V1. , Journal of Vision, 7(7):6, 1-14, http://journalofvision.org/7/7/6/, doi:10.1167/7.7.6

[29] Kolb FC, Braun J. (1995) Blindsight in normal observers. *Nature* 377(6547):336-8

[30] Krummenacher J., Muller H.J., Heller D. (2001) Visual search for dimensionally redundant pop-out targets: evidence for parallel-coactive processing of dimensions. *Percept Psychophys.* 63(5):901-17, (2001).

[31] Li Z. (1998) A neural model of contour integration in the primary visual cortex *Neural Computation* 10. 903-940,

[32] Li Z. (1999a) Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proc. Natl Acad. Sci USA*, 96(18):10530-5.

Li Z. (1999b) Visual segmentation by contextual influences via intracortical interactions in primary visual cortex. *Network: Computation in Neural Systems* 10(2):187-212

[33] Li Z (2000) Pre-attentive segmentation in the primary visual cortex. *Spatial Vision,* 13(1) 25-50.

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

[35] Li CY 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.

[36] Livingstone MS, Hubel DH (1984) Anatomy and physiology of a color system in the primate visual cortex. *J. Neurosci.* 4(1):309-56.

[37] Morgan MJ, Mason AJ, Solomon JA. (1997) Blindsight in normal subjects? *Nature* 385(6615):401-2.

[38] Nakayama, K. & Mackeben M. (1989) Sustained and transient components of focal visual attention. *Visual Research* 29: 1631-1647.

[39] Nelson JI, and Frost BJ. (1985) Intracortical facilitation among co-oriented, co-axially aligned simple cells in cat striate cortex. *Exp Brain Res.* 61(1):54-61.

[40] Nothdurft HC (1992) Feature analysis and the role of similarity in preattentive vision. *Percept Psychophys.* 52(4):355-75.

[41] Nothdurft H.C. (2000) Salience from feature contrast: additivity across dimensions. *Vision Research* 40:1183-1201.

[42] Nothdurft HC, Gallant JL, Van Essen DC. (1999) Response modulation by texture surround in primate area V1: correlates of "popout" under anesthesia. *Vis. Neurosci.* 16, 15-34.

[43] Nothdurft HC, Gallant JL, Van Essen DC. (2000) Response profiles to texture border patterns in area V1. *Vis. Neurosci.* 17(3):421-36.

[44] Reynolds JH, Desimone R. (2003) Interacting roles of attention and visual salience in V4. *Neuron* 37(5):853-63.

[45] Rockland KS., Lund JS. (1983) Intrinsic laminar lattice connections in primate visual cortex. *J. Comp. Neurol.* 216(3):303-18.

[46] Schall JD and Thompson KG (1999) Neural selection and control of visually guided eye movemnts. *Annual Review Nuroscience* 22:241-259.

[47] Shipp S. (2004) The brain circuitry of attention. *Trends Cogn. Sci.* 8(5):223-30.

[48] Sillito AM, Grieve KL, Jones HE, Cudeiro J, Davis J. (1995) Visual cortical mechanisms detecting focal orientation discontinuities. *Nature* 378, 492-496 (1995).

[49] Super H, Spekreijse H, Lamme VA. (2003) Figure-ground activity in primary visual cortex (V1) of the monkey matches the speed of behavioral response. *Neurosci Lett.* 344(2):75-8

[50] Tong F. (2003) Primary visual cortex and visual awareness. *Nat Rev Neurosci.* 4(3):219-29.

[51] Tamura H, Sato H, Katsuyama N, Hata Y, Tsumoto T. (1996) Less segregated processing of visual information in V2 than in V1 of the monkey visual cortex. *Eur. J. Neurosci.* 8(2):300-9.

[52] Tehovnik EJ, Slocum WM, Schiller PH. (2003) Saccadic eye movements evoked by microstimulation of striate cortex. *Eur J. Neurosci.* 17(4):870-8.

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

[54] Treue S. and Martinez-Trujillo JC (1999) Feature-based attention influences motion processing gain in macaque visual cortex *Nature* 399: 575-79 (1999).

[55] Ts'o DY, Gilbert CD. (1988) The organization of chromatic and spatial interactions in the primate striate cortex. *J Neurosci.* 8(5):1712-27.

[56] Tsotsos, J.K. (1990) Analyzing Vision at the Complexity Level. *Behavioral and Brain Sciences* 13-3:423 - 445.

[57] van Zoest W, Donk M. (2004) Bottom-up and top-down control in visual search *Perception* 33(8):927-37

[58] von der Heydt R, Peterhans E, Baumgartner G. (1984) Illusory contours and cortical neuron responses. *Science* 224(4654):1260-2.

[59] Wachtler T., Sejnowski TJ., Albright TD. (2003) Representation of color stimuli in awake macaque primary visual cortex. *Neuron*, 37(4):681-91.

[60] Webb BS, Dhruv NT, Solomon SG, Tailby C, Lennie P.(2005) Early and late mechanisms of surround suppression in striate cortex of macaque. *J. Neurosci.* 25(50):11666-75.

[61] Wolfe JM, Franzel SL. (1988) Binocularity and visual search. *Percept Psychophys.* 44(1):81-93

[62] Wolfe J.M., Cave K.R., Franzel S. L. (1989) Guided search: an alternative to the feature integration model for visual search. *J. Experimental Psychol.* 15, 419-433.

[63] Wolfson SS, Landy MS. (1995) Discrimination of orientation-defined texture edges. *Vision Res.* 35(20):2863-77.

[64] Yantis S. (1998) Control of visual attention. in *Attention*, p. 223-256. Ed. H. Pashler, Psychology Press.

[65] Yantis S, Serences JT. (2003) Cortical mechanisms of space-based and object-based attentional control. *Curr Opin Neurobiol.* 13(2):187-93.

[66] Zhaoping L. (2003) V1 mechanisms and some figure-ground and border effects. *Journal of Physiology, Paris* 97:503-515.

[67] Zhaoping L. and Guyader N. (2007) Interference with bottom-up feature detection by higher-level object recognition in *Current Biology*, 17:26-31.

[68] Zhaoping L. May K. (2007) Psychophysical tests of the hypothesis of a bottom up saliency map in primary visual cortex. *PLoS Computational Biology*. 3(4):e62. Epub 2007 Feb 20.

[69] Zhaoping L. and Snowden RJ (2006) A theory of a salency map in primary visual cortex (V1) tested by psychophysics of color-orientation interference in texture segmentation. *Visual Cognition* 14(4/5/6/7/8):911-933.

[70] Zhaoping L. (2006) Theoretical Understanding of the early visual processes by data compression and data selection *Network: Computation in neural systems* 17(4):301-334.

[71] Zhaoping L, Dayan P. (2006) Pre-attentive visual selection. *Neural Network* 19(9):1437-9

[72] Zhaoping L. (2007) Popout by unique eye of origin: A fingerprint of the role of primary visual cortex in bottom-up saliency. Presented at Annual meeting of Society for Neuroscience, 2007, Nov. 3-7, 2007, San Diego, USA. Program No. 717.8.

[73] Zhaoping L. (2008) Attention capture by eye of origin singletons even without awareness — a hallmark of a bottom-up saliency map in the primary visual cortex. *Journal of Vision* 8(5):1, 1-18, http://journalofvision.org/8/5/1/, doi:10.1167/8.5.1.