Relative contributions of 2D and 3D cues in a texture segmentation task, implications for the roles of striate and extrastriate cortex in attentional selection

Li Zhaoping

Department of Computer Science,
University College London, UK

Nathalie Guyader

Department Image and Signal,
Grenoble Image Parole Signal Automatique Lab
(GIPSA-Lab, CNRS UMR 5216), France

Alex Lewis

Goldman Sachs, London, UK

Experimental evidence has given strong support to the theory that the primary visual cortex (V1) realizes a bottom–up saliency map (A. R. Koene & L. Zhaoping, 2007; Z. Li, 2002; L. Zhaoping, 2008a; L. Zhaoping & K. A. May, 2007). Unlike the conventional models of texture segmentation, this theory predicted that segmenting two textures in an image I_{rel} comprising obliquely oriented bars would become much more difficult when a task-irrelevant texture I_{ir} of spatially alternating horizontal and vertical bars is superposed on the original texture I_{rel}. The irrelevant texture I_{ir} interferes with I_{rel}'s ability to direct attention. This predicted interference was confirmed (L. Zhaoping & K. A. May, 2007) in the form of a prolonged task reaction time (RT). In this study, we investigate whether and how 3D depth perception, believed to be processed mostly beyond V1 and starting in V2 (J. S. Bakin, K. Nakayama, & C. D. Gilbert, 2000; B. G. Cumming & A. J. Parker, 2000; F. T. Qiu & R. von der Heydt, 2005; R. von der Heydt, H. Zhou, & H. S. Friedman, 2000), contribute additionally to direct attention. We measured the reduction of the interference or the RT when the position of the texture grid for I_{ir} was offset horizontally from that for I_{rel}, forming an offset, 2D, stimulus. This reduction was compared with that when this positional offset was only present in the input image to one eye, or when it was in the opposite directions in the images for the two eyes, creating a 3D stimulus with a depth separation between I_{ir} and I_{rel}. The contribution by 3D processes to attentional guidance would be manifested by any extra RT reduction associated with the 3D stimulus over the offset 2D stimulus. This 3D contribution was not present unless the task was so difficult that RT (by button press) based on 2D cues alone was longer than about 1 second. Our findings suggest that, without other top–down factors, V1 plays a dominant role in attentional guidance during an initial window of processing, while cortical areas beyond V1 play an increasing role in later processing. Subject-dependent variations in the manifestations of the 3D effects also suggest that this later, 3D, contribution to attentional guidance can be easily influenced by top–down control.

Keywords: saliency map, V1 saliency hypothesis, attention, depth perception, psychophysics, bottom–up attention, top–down

Citation: Zhaoping, L., Guyader, N., & Lewis, A. (2009). Relative contributions of 2D and 3D cues in a texture segmentation task, implications for the roles of striate and extrastriate cortex in attentional selection. *Journal of Vision, 9*(11):20, 1–22, http://journalofvision.org/9/11/20/, doi:10.1167/9.11.20.

Introduction

Background

Since our cognitive resources are limited, it is necessary to constrain visual processing to only a fraction of visual inputs. The process of selecting this fraction is often called attentional guidance to certain input locations or characteristics (or features). This selection can be by top–down, or task-driven, factors, as when one directs gaze to a book while reading, and by bottom–up, or stimulus-driven, factors, as when one is distracted from the book by a sudden appearance of a predator. There have been many psychological (Duncan & Humphreys, 1989; Julesz, 1981; Treisman & Gelade, 1980; Wolfe, Cave, & Franzel, 1989), physiological (Chelazzi, Miller, Duncan, & Desimone, 1993; Corbetta & Shulman, 2002; Desimone & Duncan, 1995; Moore & Armstrong, 2003; Motter, 1993; Reynolds & Desimone, 2003; Treue & Martinez-Trujillo, 1999), and computational (Itti & Koch, 2000; Koch & Ullman, 1985; Li, 1999a, 2002) studies of visual attentional guidance and their consequences or effects on behavior and neural responses. This paper concerns the neural substrates of the visual selection of spatial locations, and not the mechanisms and effects of information processing as a consequence of the selection.

Most previous studies into the neural bases of visual selection focus on the top–down, task-dependent factors and have implicated a network of cortical areas including

doi: 10.1167/9.11.20 Received June 1, 2009; published October 21, 2009 ISSN 1534-7362

dorsal posterior parietal and frontal cortex. Parts of the network respond to cues specifying the tasks even before the appearance of the task-relevant stimuli, and lesions in these areas cause deficits in the appropriate direction of attentions. Meanwhile a right lateralized network including temporoparietal junction and the ventral frontal cortex has been proposed to detect, or orient attention to, sensory stimuli relevant to the task or contingent on task demands (Corbetta & Shulman, 2002; Desimone & Duncan, 1995). There are also proposals for the subcortical thalamus (pulvinar) and superior colliculus to gate the entry of sensory inputs into cortical areas (Crick, 1984; Olshausen, Anderson, & Van Essen, 1993) because of the widespread cortical connections of these subcortical areas.

However, bottom–up selection is often faster and more potent than top–down selection (Jonides, 1981; Nakayama & MacKeben, 1989; van Zoest & Donk, 2006). Therefore, understanding bottom–up selection regardless of top–down task demands is essential to understanding selection as a whole. It has thus been proposed (Li, 1999a, 2002) that the primary visual cortex (V1) itself computes a bottom–up saliency map. According to this proposal, the most likely location to attract bottom–up attention is at the receptive field of the V1 cell that is most activated by the scene. We call this proposal the V1 saliency hypothesis. The superior colliculus, which receives inputs from V1 and gives motor commands for gaze shifts (Tehovnik, Slocum, & Schiller, 2003), could read out this saliency map rather than computing it. It could do so simply by finding the highest response among the V1 neurons and transforming this to a gaze shift to the corresponding visual location. In natural viewing, when gaze is not constrained, this gaze shift achieves selection, since directing attention to a spatial location is mandatorily associated with gaze shifts to this location (Hoffman, 1998). To transform local contrasts in visual input to saliencies which depend on global context (e.g., a red item is salient in a background of green but not red items), the intra-cortical interactions in V1 make the response of a V1 cell dependent on the input context outside the corresponding receptive field (e.g., Allman, Miezin, & McGuinness, 1985; Knierim & van Essen, 1992). Various lines of supporting evidence (Jingling & Zhaoping, 2008; Koene & Zhaoping, 2007; Zhaoping, 2008a; Zhaoping & May, 2007; Zhaoping & Snowden, 2006), detailed in the Discussion section, have emerged for this V1 saliency proposal. For example, a task-irrelevant eye of origin singleton apparently identical to background items can capture attention and shift gaze away from task-relevant locations (Zhaoping, 2008a, 2008b). To visual awareness, this singleton does not appear distinctive from background items since its distinctiveness, i.e., its unique eye of origin, is barely represented in any visual cortical areas except V1 (the only cortical area with a substantial fraction of neurons tuned to eye of origin; Burkhalter & Van Essen, 1986; Hubel & Wiesel, 1968). The strong attraction to attention by this singleton, despite the blindness of extrastriate areas to its distinctive feature, is thus a strong evidence for the V1 saliency hypothesis.

Motivation and design

The V1 saliency hypothesis, however, does not preclude additional contributions from other cortical areas to selection at a bottom–up or at a higher perceptual or cognitive level. The superior colliculus, which directs gaze, also receives inputs from cortical areas beyond V1, including cortical areas such as V2 on the sensory pathway from V1 to frontal regions. Hence, we ask whether the extrastriate cortex, defined as the higher order or associative visual cortical areas (including V2, V3, V4, MT/V5, and IT) beyond and receiving inputs from the primary visual cortex (Orban, 2008), contribute additionally to visual selection. (A small fraction of retinal cells also project to the superior colliculus. However, retinal control of saccades is limited since the neurons associated with saccades in superior colliculus can not be driven by visual inputs after V1 lesion (Schiller, 1998)). Let us imagine a scenario in which the decision on where to direct attention is decided by a commander, such as superior colliculus or some other attention directing center, and the various brain areas sending their contributing signals to this commander. Then the contribution by each brain area to the decision will be determined by various factors including the strength, task relevance, and timeliness of its input to the commander. Hence, some decisions can be easily dominated by the top down contributions while others by bottom–up ones. Since extrastriate areas receive much of the bottom–up inputs from V1, their outputs based on bottom–up inputs will generally lag behind that of V1 (Bullier & Nowak, 1995; Schmolesky et al., 1998). Hence, their contribution to the decision could be too slow to have an impact when V1's contribution is sufficiently strong and fast. Their contribution could also be ignored if it is relatively too weak or redundant with contribution already available from V1. Conversely, extrastriate contribution could be substantial when V1's contribution is too weak to reach a quick decision. The current study aims to address this question of how and when extrastriate cortex contributes to selection. By "contribution" we mean the contribution contingent on the outcomes of processing the bottom–up sensory inputs. Hence, this contribution could include top–down task-driven factors which nevertheless can only exert their effects after the bottom–up inputs have been processed to become useful for the task, as will be made clearer later in the paper. In the rest of this Introduction section, we introduce texture segmentation as a task to probe the answers to our question and motivate our experiments using depth perception for this purpose.

Consider the task of finding the location of the border between two neighboring textures comprising uniformly oriented bars. The speed of doing so depends on the

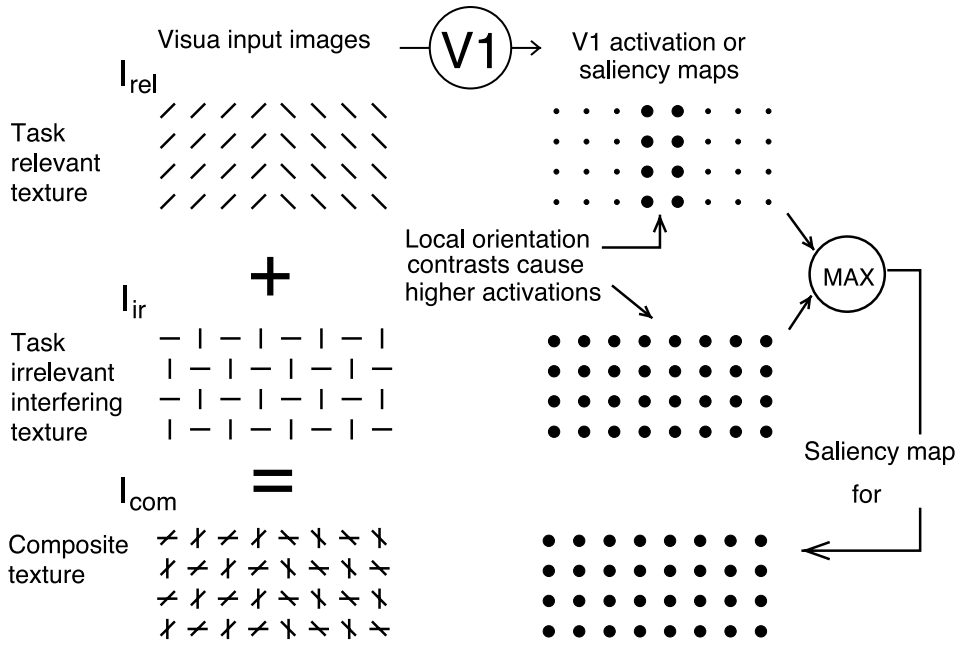

Figure 1. Texture segmentation, interference, and their interpretation under the V1 saliency hypothesis. Left: three textures: task-relevant I_{rel}, task-irrelevant I_{ir}, and the composite texture $I_{com} = I_{rel} + I_{ir}$. Each texture region plotted is only a small part of a much bigger texture. Right: corresponding saliency maps, in which saliency values are visualized by the sizes of the disks. In the maps for I_{rel} and I_{ir}, the saliency value for each location matches the response from the most vigorously responding V1 cell to the texture bar. This response is relatively higher, or less suppressed by iso-orientation suppression, at locations of high orientation contrast. Segmenting I_{rel} is easy since the texture border has a higher saliency. In the map for I_{com}, the saliency value is the maximum V1 response to each texture location, regardless of which of the two bars at this location evokes this response. A lack of a saliency highlight at the texture border makes segmenting I_{com} difficult.

saliency of the texture border bars to attract attention, when the location of this border is unknown before the stimulus onset. Here saliency is defined as the ability to attract attention, in the bottom–up or stimulus-driven manner. If the orientation contrast is large at the border, see the texture I_{rel} in Figure 1, the border pops out perceptually, attracting attention automatically, resulting in short reaction times (RTs) to report its location. Hence, texture segmentation tasks have been used extensively to study bottom–up selection (Julesz, 1981; Nothdurft, 1991). According to the V1 saliency hypothesis, the saliency of the texture border can be understood as follows. Intra-cortical interactions in V1 lead to suppression between nearby V1 cells tuned to similar orientations, giving iso-orientation suppression as observed physiologically (Allman et al., 1985; Knierim & Van Essen, 1992; Li & Li, 1994; Nothdurft, Gallant, & van Essen, 1999; Sillito, Grieve, Jones, Cudeiro, & Davis, 1995). A neuron responding to a bar at the texture border is less suppressed, since the border bar has fewer iso-orientation neighbors than a texture bar far away from the border and surrounded by iso-orientation neighbors. Looking at the spatial map of neural responses from all V1 neurons, regardless of their feature preferences, we should see a spatial map like shown in Figure 1, with the response highlight at the texture border. Selection of the texture border can then be achieved by locating the most responsive cell to the scene, i.e., locating the highest value in the response map as the saliency map. The segmentation of I_{rel} can also be explained by the standard "back-pocket" model, also called filter–rectify–filter model, of texture segmentation (Bergen & Landy, 1991; Landy & Graham, 2004; Malik & Perona, 1990). This standard model filters the image with filters of different spatial orientations and frequencies, takes the energies (or rectifications) of the filter responses, and the texture border is where the energy of any particular filter changes with spatial location at a coarse scale. Accordingly, the texture border in I_{rel} is where the energy outputs of the left or right tilted filters change. In this sense, the V1 saliency hypothesis may be seen as identifying the intra-cortical interactions in V1 as the neural implementation of the phenomenological non-linear operation (rectify-filter) in the filter–rectify–filter model.

Zhaoping and May (2007) found that segmenting the simple texture I_{rel}, i.e., to locate the texture border, can be made very difficult if a task-irrelevant texture I_{ir} comprising horizontal and vertical bars is superimposed on it, giving the composite texture I_{com} shown in Figure 1. In other words, the task-irrelevant texture interferes dramatically with the task. This strong interference cannot be explained by the standard "back-pocket" model, since the energy outputs of the left and right tilted filters still change at the texture border. More specifically, the orientation features with strong stimulus energies are

horizontal, vertical, and right tilted in the texture left of the border, while they are horizontal, vertical, and left tilted in the texture right of the border. Any texture segmentation mechanism or algorithm based on calculating feature values of a (local) image area would easily identify the texture border as where the feature values (left or right tilted orientation) change at a coarse scale. Even if, due to the presence of I_{ir}, the filter output value change at the border in I_{com} may be quantitatively weaker than that in I_{rel}, there is no obvious reason within these models why this quantitative weakening should make the segmentation so dramatically more difficult—after all segmentation is only slightly or quantitatively more difficult if a similar weakening is produced instead by reducing the 90° orientation contrast in I_{rel} to 20° (see Figure 10G; Zhaoping & May, 2007). The interference by I_{ir} is also not accountable as simply a result of masking by I_{ir}, since a I_{ir} texture made of uniformly oriented vertical bars will barely mask the salient border (Zhaoping & May, 2007; see Figure 10E). Nor can it be explained by the non-uniformity of orientations within a texture, since replacing each composite texture element in I_{com} by a single bar oriented at the average orientation of the two intersecting bars will practically eliminate the interference (in terms of RTs) while retaining the non-uniformity of orientations within a texture (Zhaoping & May, 2007; see Figure 10F). Nevertheless, this interference can be predicted by the V1 saliency hypothesis as follows. In the task-irrelevant texture I_{ir}, each bar has half of its contextual neighbors sharing the same orientation as itself, just as for a texture border bar in the task-relevant texture I_{rel}. Hence, a task-irrelevant bar in I_{ir} and a texture border bar in I_{rel} experience comparable levels of iso-orientation suppression and thus evoke comparable levels of V1 responses. Let us say, for convenience, this response level is 10 in the most vigorously responding cells to the respective bars, while each task-relevant texture bar in I_{rel} away from the border evokes a response level 5. Then, in the composite texture I_{com}, the highest response is 10 and is present at every texture element location. By the V1 saliency hypothesis, selection looking for the highest V1 response will find it at all locations. Hence, all locations are equally salient, and so the texture border will not pop out. This makes it difficult for attention to find the texture border, prolonging the reaction time to locate the border from around half a second to typically more than one second for untrained subjects. This interference is strong and persists even when observers know ahead of time that the interfering texture I_{ir} is task irrelevant and that they should pay attention only to the oblique bars (in I_{rel}) in the composite texture I_{com}. This demonstrates that the strength of bottom–up selection can be too strong to be eliminated by top–down control. Note from above that, according to the V1 saliency hypothesis, the saliency at each texture element location is determined by the maximum response to that location, regardless of which of the two bars at that location evokes this response. We call this way to compute saliency by the maximum of the responses as the MAX rule.

To probe extrastriate contributions to attentional selection, we take advantage of some physiological observations indicating that cortical areas after V1, and starting at V2, are responsible for 3D and surface perception. Even though V1 cells are tuned to binocular disparities between two matching features to the two eyes, stereo-matching processes for 3D perception by false match suppression and visual grouping process for surface perception occur more outside V1, notably in V2 (Bakin, Nakayama, & Gilbert, 2000; Cumming & Parker, 2000; Qiu & von der Heydt, 2005; von der Heydt, Zhou, & Friedman, 2000). Hence, attentional guidance by depth or three-dimensional (3D) input cues should reflect the contribution by saliency computations that occur beyond V1. It has been shown (He & Nakayama, 1995; Nakayama & Silverman, 1986) that searching for a target defined by an unique conjunction of depth and another feature is much easier than typical conjunction searches without depth feature. This suggests that 3D cues can help to direct attention to task-relevant locations. However, to isolate the extrastriate contribution to attentional guidance, we need to separate, in the input stimuli, the 2D cues, which can be processed by V1, from the 3D cues. For this purpose, we modify the 2D image I_{com} in Figure 1 to create two new types of stimuli which have the same 2D cues but only one of them has the 3D cue, as illustrated in Figure 2. The first type is made from I_{com} by offsetting the position of the texture grid for I_{ir} horizontally from that for I_{rel}, forming an offset, 2D, stimulus, see the middle column of Figure 2. This positional offset reduces the interference by I_{ir}, and thus the RT for our texture segmentation task, as confirmed by data (shown later) and by a simulation of the V1 saliency map in a V1 model (Li, 1999b, 2000; data not shown). The second type of the new stimuli has this positional offset present only in the input image to one eye (while the image I_{com} to the other eye has no such offset) or have this offset in the opposite directions in the images for the two eyes, creating a 3D stimulus with a depth separation between I_{ir} and I_{rel}. Hence, the 3D stimulus and the corresponding 2D offset stimulus have the same 2D cues, in particular the 2D positional offset between the task-relevant and -irrelevant textures. However, the 3D stimulus has an additional 3D cue, the depth separation between the two textures, which is extracted when the 3D processes in the brain stereo-match the two monocular images. If 3D processes do not contribute to attentional guidance, the RTs for the 2D offset stimulus and the 3D stimulus would be no different. Otherwise, the contribution by 3D processes would be manifested by any extra RT reduction for our task associated with the 3D stimulus over the offset 2D stimulus.

Figure 3 summarizes the stimulus types and the desired quantities to be measured in our experiments. There are eight stimulus types. Four of them are bioptic or 2D stimuli, in which two eyes receive identical inputs,

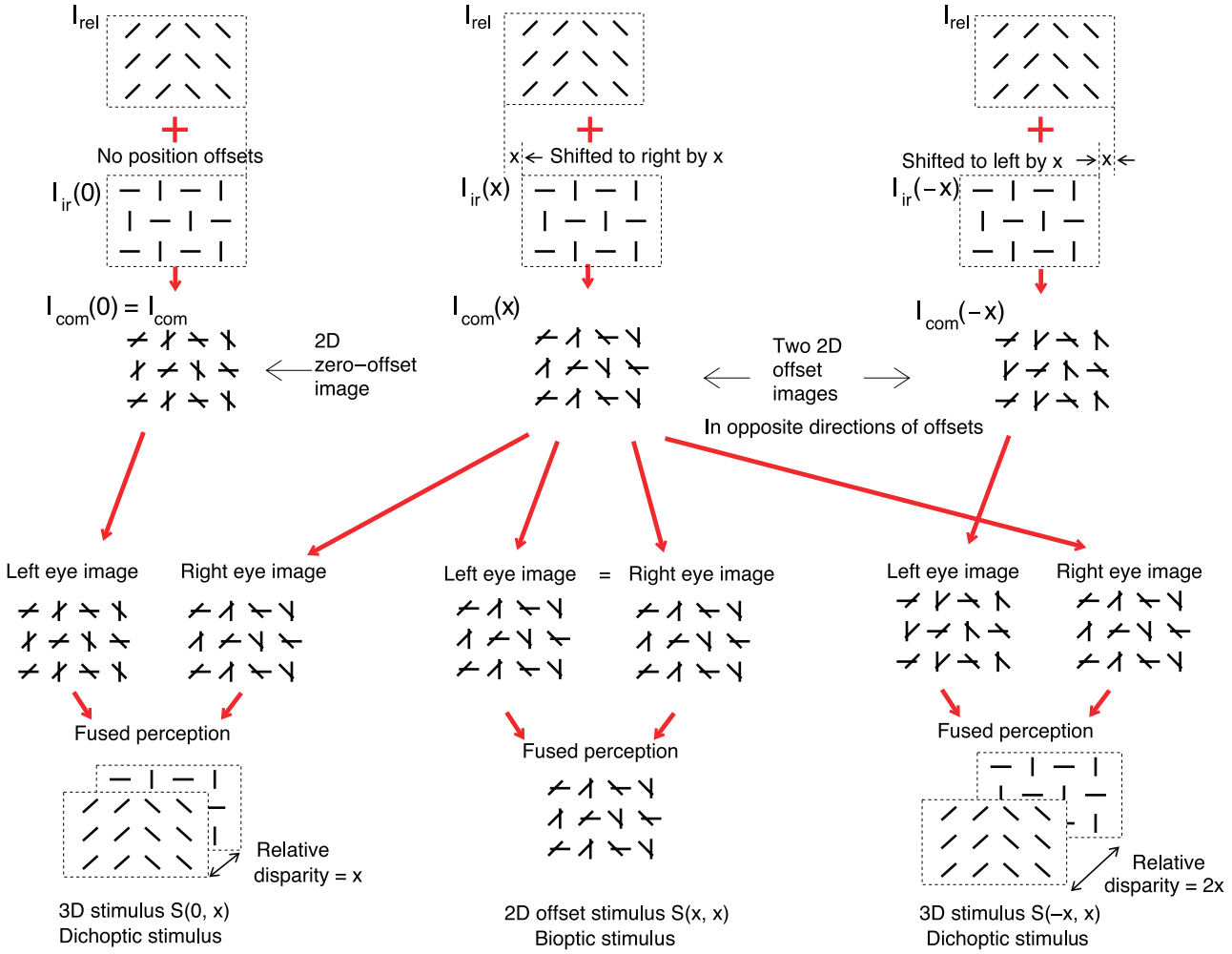

Figure 2. Schematic illustrations of the constructions of some 3D and 2D stimuli used in the study. Top shows the original composite image I_{com} and the 2D offset images $I_{com}(x)$ and $I_{com}(-x)$ modified from I_{com}, by a horizontal position offset of I_{ir} to the right or left by x to get $I_{ir}(x)$ and $I_{ir}(-x)$, respectively. Hence, $I_{com}(\pm x) = I_{rel} + I_{ir}(\pm x)$. Bottom shows the 2D offset stimulus $S(x, x)$ when the 2D offset image $I_{com}(x)$ is presented to both eyes, the 3D stimulus $S(0, x)$, when the positional offset is present in only the right eye image, and 3D stimulus $S(-x, x)$ when the positional offset is in the opposite directions in the two eyes. The relative disparity between I_{rel} and I_{ir} in the 3D stimuli is x for $S(0, x)$ and $2x$ for $S(-x, x)$.

although in fact these are stimuli with textures at the depth of the display screen (but without a depth difference between the relevant and irrelevant texture). They are called, respectively, baseline, $2D_0$, $2D_a$, and $2D_{2a}$. In the baseline stimulus, the input image contains I_{rel} only. In the other 2D stimuli $2D_x$, for $x = 0$, a, and $2a$, the input image is a 2D offset image, with an absolute offset 0 (i.e., no offset as in the original composite image I_{com}), a, and $2a$, respectively (for a particular a value given in the Methods section). The four other stimulus types, Figure_a, Ground_a, Figure_{2a}, and Ground_{2a}, are 3D or dichoptic stimuli, in which two eyes receive different inputs. In these terms, Figure or Ground denotes whether the task-relevant texture I_{rel} is the figure (foreground) or the ground (background) surface, and the subscript a or $2a$ denotes whether the absolute positional offset is a or $2a$, respectively, in the 2D offset image to one of the eyes. Note that a Figure_x stimulus becomes a Ground_x stimulus when the images to the two eyes are swapped. The relative disparity between the two textures is always a constant value $2a$ in a 3D stimulus. It is created either by an absolute offset a in opposite directions in the two eyes, as in Figure_a and Ground_a, or by an absolute offset $2a$ in only one eye (and no offset in the other eye), as in Figure_{2a} and Ground_{2a}. The 3D contribution would be manifested in the following three RT differences

$$\Delta(x) \equiv \text{RT}(2\text{D}_x) - \text{RT}(\text{Figure}_x), \quad (1)$$

$$\delta_1(x) \equiv \text{RT}(2\text{D}_x) - \text{RT}(\text{Ground}_x), \quad (2)$$

$$\delta_2(x) \equiv \text{RT}(\text{Ground}_x) - \text{RT}(\text{Figure}_x), \quad (3)$$

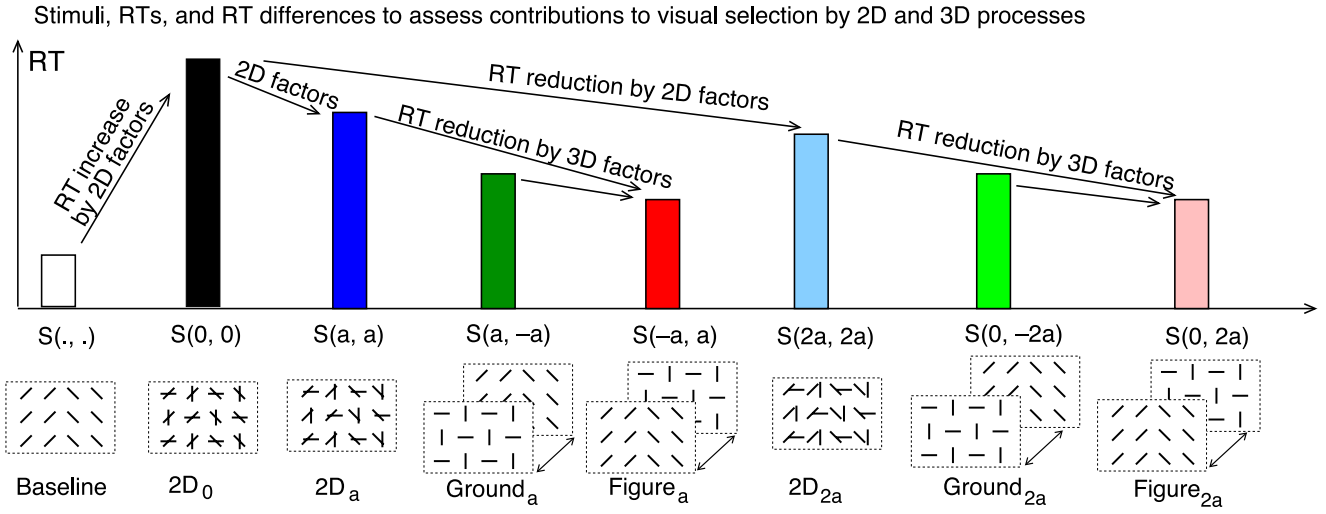

Figure 3. Schematic illustrations of the types of stimuli, as named in the bottom, used in the experiments, and how the RTs to locate the texture border in the relevant I_{rel} can be used to assess the contributions from 2D and 3D factors in visual selection.

where $x = a$ or $2a$, and each RT(stimulus type) denotes the RT averaged over trials belonging to that stimulus type (see Methods). If depth separation between I_{rel} and I_{ir} makes it easier to allocate attention to I_{rel} than otherwise, $\Delta(x)$ and $\delta_1(x)$ should be positive. Meanwhile, since it is easier to direct attention to the foreground than background surface (Mazza, Turatto, & Umilta, 2005), RT(Ground$_x$) is typically (or by default) longer than RT(Figure$_x$) when depth perception is playing a role in the task. This figure–ground factor should make $\delta_2(x)$ = RT(Ground$_x$) − RT(Figure$_x$) positive and make $\delta_1(x)$ smaller than $\Delta(x)$. However, this default figure–ground factor can be reduced by top–down attentional control as follows. In our stimuli, I_{rel} is always on the depth plane of the display screen, whether it is in front of or behind the irrelevant texture I_{ir}. Hence, subjects could force their attention to the display screen so that their attention is more likely to be on the task-relevant I_{rel} when the depth perception emerges from viewing the stimulus, even when I_{rel} is the background surface. This top–down attentional control should thus reduce RT(Ground$_x$), consequently making $\delta_1(x)$ larger and $\delta_2(x)$ smaller at the same time. The opposite effects of the top–down control on $\delta_1(x)$ and $\delta_2(x)$ should make their sum $\delta_1(x) + \delta_2(x)$ less sensitive to the degree of this top–down control, and in fact $\delta_1(x) + \delta_2(x) = \Delta(x)$ does not contain RT(Ground$_x$). Since different observers have different reaction speeds, or RT(2D$_x$), characteristic of themselves, a $\Delta(x)$ = 100 milliseconds (ms) for RT(2D$_x$) = 700 ms is quite likely to indicate a more substantial 3D effect than the same $\Delta(x)$ for RT(2D$_x$) = 2000 ms. Hence, we use the index

$$E(x) \equiv \frac{\Delta(x)}{\mathrm{RT}(2\mathrm{D}_x)} = \frac{\mathrm{RT}(2\mathrm{D}_x) - \mathrm{RT}(\mathrm{Figure}_x)}{\mathrm{RT}(2\mathrm{D}_x)}, \quad (4)$$

to assess the overall contribution of the 3D cue to attentional guidance. Later on, we will use $\delta_1(x)/\Delta(x)$ to assess the top–down dependent factors and use

$$\frac{2\delta_2(x)}{\mathrm{RT}(\mathrm{Ground}_x) + \mathrm{RT}(\mathrm{Figure}_x)}, \quad (5)$$

as another assessment of the 3D contribution.

Since the 3D contribution index

$$E(x) = 1 - \frac{\mathrm{RT}(\mathrm{Figure}_x)}{\mathrm{RT}(2\mathrm{D}_x)}, \quad (6)$$

we expect that the 3D contribution is likely to increase as RT(2D$_x$) becomes longer. This can be understood as follows. If the task is sufficiently easy with 2D cues only, RT(2D$_x$) will be short. Then the 3D contribution is more likely redundant or unnecessary in the face of sufficiently clear 2D cues, such that RT(Figure$_x$) will not be any shorter than RT(2D$_x$). By contrast, the 3D cue is more likely to be useful for the task when the task is sufficiently difficult without it, i.e., when RT(2D$_x$) is long. Also, the processes beyond V1 should have a longer latency than those within V1. Thus, contributions from higher visual areas should take longer to manifest. If RT(2D$_x$) is shorter than the time needed to manifest the 3D contribution behaviorally, we will see a near zero $E(x)$. We will investigate such temporal dependence of $E(x)$ by making the task more or less difficult to lengthen or shorten reaction times RT(2D$_x$). This is done by having two different—large and small—orientation contrasts between the neighboring task-relevant textures in I_{rel}.

The large contrast is 90° as in Figures 1–3, investigated by Experiment 1, and the smaller contrast is 14° as in Figure 6A, investigated by Experiment 2.

Methods

Experiments 1 and 2 differ only in the orientations of the texture bars in the task-relevant pattern I_{rel}. In Experiment 1, these bars were oriented 45° and −45° respectively from vertical, as in Figure 1. In Experiment 2, they were α + 7° from vertical or α − 7° from vertical, with α = 45° or −45° with equal chances, as in Figure 6A. Hence, the orientation contrast was 90° and very salient in Experiment 1, but in Experiment 2 it was only 14° and near the just-noticeable difference for the border to pop out (without the I_{ir}). Experiment 2 should produce much longer RTs than Experiment 1, allowing us to probe the temporal characteristics of the contributions to attentional guidance of lower and higher visual areas.

Stimuli were displayed on a Clinton Monoray Monitor (20-in. flat profile DP104 fast-phosphor CRT, with CIE chromaticity coordinates (0.43, 0.54), with phosphor decaying time of 200 μs for more than 90% of energy for light outputs). A pair of FE-1 (Ferro-electric) stereo goggles was used to view the display dichoptically. The input images to the two eyes were displayed in temporally alternating video frames on the monitor, with the video frames synchronized with the opening and closing of the shutters to the respective goggles (using ViSaGe video card from the Cambridge Research System www.crsltd.com). With a 100-μs shutter switching times, 3-log unit contrast ratios between the open-shut states, and 25% open shutter transmission, there was very little cross talk between stimuli intended for the two eyes.

The texture elements in both the task-relevant and interfering patterns were placed on a regular grid of 30 columns and 22 rows spanning 46 × 34 degrees in visual angle. The screen refresh rate was 150 frames per second. The texture bars, when displayed constantly (in every frame) on screen and viewed without the stereo goggle, were yellow, 0.12 × 1.1 degree rectangles, and of luminance 48 candela/m^2 on a black background. The texture border in the task-relevant pattern I_{rel} was located randomly at 7, 9, or 11 columns left or right of the center. In each trial, the orientation of the texture bars in the left texture of the task-relevant I_{rel} was randomly chosen as tilted left or right from vertical, and for Experiment 2, it was further randomly chosen as tilted left or right from α = ±45°. In the interfering texture I_{ir}, the top-left bar was randomly chosen as horizontal or vertical. Accordingly, the subjects were not able to guess beyond chance which stimulus would appear. Each trial started with a central fixation dot of radius 0.3 degrees at zero disparity, defined as the depth of the display screen. The relative disparity between the task-relevant I_{rel} and irrelevant I_{ir} pattern was $2a$ = 0.46 degrees when they were separated in depth. The task-relevant I_{rel} was always at the zero disparity on the display screen. There were four anchoring dots of radius 0.6 degrees at the four corners of the screen just outside the texture patterns. They were at zero disparity and were always present throughout the experimental session in order to anchor the vergence eye positions. Pilot experiments showed that the 3D effect increased with the relative disparity between the two surfaces and then saturated at a sufficiently large relative disparity, which was used for our experiments.

To describe our stimuli in detail, we introduce some notations, see some examples of them in Figures 2 and 3. Parametrically, let $I_{ir}(x)$ denote I_{ir} position shifted x horizontally, $I_{com}(x) = I_{rel} + I_{ir}(x)$ the resulting composite, 2D offset, image, and $S(x, y)$ the dichoptic stimulus with $I_{com}(x)$ in the left and $I_{com}(y)$ in the right eye. Accordingly, 2D zero-offset stimulus $2D_0$ is $S(0, 0)$, the 2D offset stimulus $2D_a$ can be either $S(a, a)$ or $S(-a, -a)$, and $2D_{2a}$ stimulus can be either $S(2a, 2a)$ or $S(-2a, -2a)$. The baseline stimulus without the I_{ir} component is denoted as $S(., .)$. Meanwhile, Figure$_a$ stimulus is $S(-a, a)$, Ground$_a$ stimulus is $S(a, -a)$, Figure$_{2a}$ stimulus can be either $S(0, 2a)$ or $S(-2a, 0)$, and Ground$_{2a}$ stimulus can be either $S(0, -2a)$ or $S(2a, 0)$. We note that $I_{com}(a)$ and $I_{com}(-a)$ are related by symmetry such as mirror reflection and should be equally easy to segment. Hence, the text in this paper often does not explicitly distinguish between these symmetrically related stimuli, such as between $S(a, a)$ and $S(-a, -a)$ of $2D_a$ or between $S(0, 2a)$ and $S(-2a, 0)$ of Figure$_{2a}$, even though all the symmetrically related stimuli were used with equal chance for their corresponding stimulus type in our experiments.

Experiments were conducted in a dark room. All subjects were adults and younger than 40 years old, had normal or corrected-to-normal vision, and participated in four sessions in each experiment, with 200 trials per session. Trials of different stimulus types, baseline, $2D_0$, $2D_a$, etc., were randomly interleaved within each session, and different stimulus variations within each stimulus type (e.g., $S(a, a)$ and $S(-a, -a)$ for stimulus type $2D_a$) were randomly chosen for each trial of this type. The stimulus types included the ones shown in Figure 3 and an additional type, called "mismatch" type, when I_{rel} and I_{ir} were presented to different eyes. We omit the presentation of data on the "mismatch" type since they are less relevant to our scope in this paper, although interleaving the "mismatch" trials during a session may affect performance in other trials, particularly since the "mismatch" trials can be quite difficult if binocular rivalry onsets before the button response. On average, each stimulus type had 25 trials in each session, except the stimulus types Figure$_a$ and Ground$_a$ each of which had on average only half as many trials due to a programming error (which was not critical

for the purpose of our study). After a subject pressed a button, each trial started with the central fixation display for 1.2 second before the test stimulus was displayed. The subjects were instructed to fixate centrally before the test stimulus onset, freely move their eyes to search for the texture border once the test stimulus appeared, and press the left or right button as quickly as possible to indicate whether the texture border in I_{rel} was in the left or right half of the display while maintaining performance accuracy. The test stimulus stayed on the screen till after the button press. Before the first experimental session for each subject, two examples of each stimulus type were shown while the task was explained to the subject. We made sure that each subject was able to see depth in the 3D stimuli in these examples and thereby excluded one subject. Only one example of each stimulus type was shown before each of the subsequent sessions to refresh the subject's memory about the task. Subjects may find some stimulus types (such as $2D_0$ and "mismatch") quite difficult, especially in Experiment 2. Hence, we told the subjects that if they could not locate the texture border in some trials after sufficient effort they could simply respond by guessing. The mean RT of each subject for a particular stimulus type was calculated using trials (from all four sessions) which were not only correctly performed but also had RTs within three standard deviations from the average RT over all trials by this subject for this stimulus type. The mean RT across subjects for a particular stimulus type is the average of the mean RTs of individual subjects. The error bars displayed are the standard errors of the means. Two-tailed *t*-tests were used for assessing statistical significance of any quantity when needed. In addition to the RT(stimulus type) for each stimulus type, we also obtain the error rate *e*(stimulus type), the fraction of error trials, to find out if our conclusions may be compromised by a speed-accurate trade-off. Across subjects, experiments, and the eight stimulus types in Figure 3, the mean and maximum error rates are 0.0575 and 0.267, respectively, and the mean and maximum fraction of trials excluded in the RT calculation (because the trial was an error trial or had an outlier RT) are 0.0723 and 0.283, respectively. Inter-subject differences will be addressed.

Results

Experiment 1

Figure 4 shows mean RTs in Experiment 1, when I_{rel} had a high 90° orientation contrast, for the various stimulus types listed in Figure 3. RTs vary substantially between subjects, e.g., the mean RT($2D_a$) of subjects varied between 460 and 1768 milliseconds (ms). Hence, for better visualization, the RTs in Figure 4 are normalized by the mean RT($2D_a$) of the corresponding subject. It is clear that RT is overwhelmingly determined by 2D stimulus variations. Superposing the task-irrelevant texture I_{ir} to I_{rel} at exactly the same position and depth causes RT to roughly double, whereas small shifts of I_{ir} relative to I_{rel} within the depth plane roughly halved the RT increase due to the interference by I_{ir}. In comparison, 3D cues by different images in the two eyes cause only small additional changes in the normalized RTs. Averaged across subjects, most of these small changes are not statistically significant.

However, a closer look, see Figure 5, reveals large inter-subject differences. The 3D effect tends to be bigger for subjects with longer RTs. For each of the six subjects LJ, LZ, NG, PA, EL, and AL whose RT($2D_x$) for $x > 0$ are shorter than about 1 second, the RTs for $2D_x$, Figure_x, and Ground_x are almost the same. In particular, none of them has RT($2D_x$) > RT(Figure_x) significantly for $x = a$ or $2a$. Their average 3D indices, $E(a) = -0.036 \pm 0.013$ and $E(2a) = -0.048 \pm 0.026$, are not significantly different from zero ($p > 0.05$). Meanwhile the average 3D indices $E(a) = 0.22 \pm 0.05$ and $E(2a) = 0.13 \pm 0.03$ of the four slower subjects are significantly different from zero ($p < 0.041$) and from those of the faster subjects ($p < 0.01$).

Note that for the slower subjects, their non-zero 3D effects, i.e., the RT differences RT($2D_x$) − RT(Figure_x), are not the results of a speed-accuracy trade-off (defined here as when a positive/negative difference in RT is caused by a negative/positive difference in the corresponding error rates). To show this, we examine the error rates $e(2D_x)$, $e(\text{Figure}_x)$, and $e(\text{Ground}_x)$ for the stimulus types $2D_x$, Figure_x, and Ground_x respectively, and define, analogous to the 3D index $E(x)$, the error index $E_{error}(x) \equiv e(2D_x) - e(\text{Figure}_x)$ (note that the $E_{error}(x)$ is not normalized by $e(2D_x)$ since each error rate is already normalized by the total number of trials). For each of the four slower subjects, $E_{error}(x)$ is either zero or has the same sign as $E(x)$, i.e., $E_{error}(x)$ is consistent with $E(x)$. Hence, the positive $E(x)$ cannot arise from a speed-accuracy trade-off. Similarly, for the faster group of subjects, the lack of a positive 3D effect on average is also not due to a speed accuracy trade-off. For shift $x = 2a$, $E_{error}(x)$ is consistent with $E(x)$ for each faster subject; for shift $x = a$, the largest inconsistency between $E_{error}(x)$ and $E(x)$ is for subject EL with $E_{error}(x) \sim 0.05$, which is not a huge error rate difference for typical reaction time tasks, and in any case the average $E_{error}(a) = 0.009 \pm 0.15$ for the faster group is not significantly different from zero. The qualitative findings of Experiment 1 still hold when we exclude data from the two subjects (PA and EL) whose overall error rates (averaged over all the eight stimulus conditions) are larger than 9% (while other subjects have an overall error rates no more than 7%).

There are two possible reasons for the difference between the faster and slower subjects. The first reason may simply be that the difference is caused by the difference in the RTs. Even though both the 2D and 3D cues could guide attention

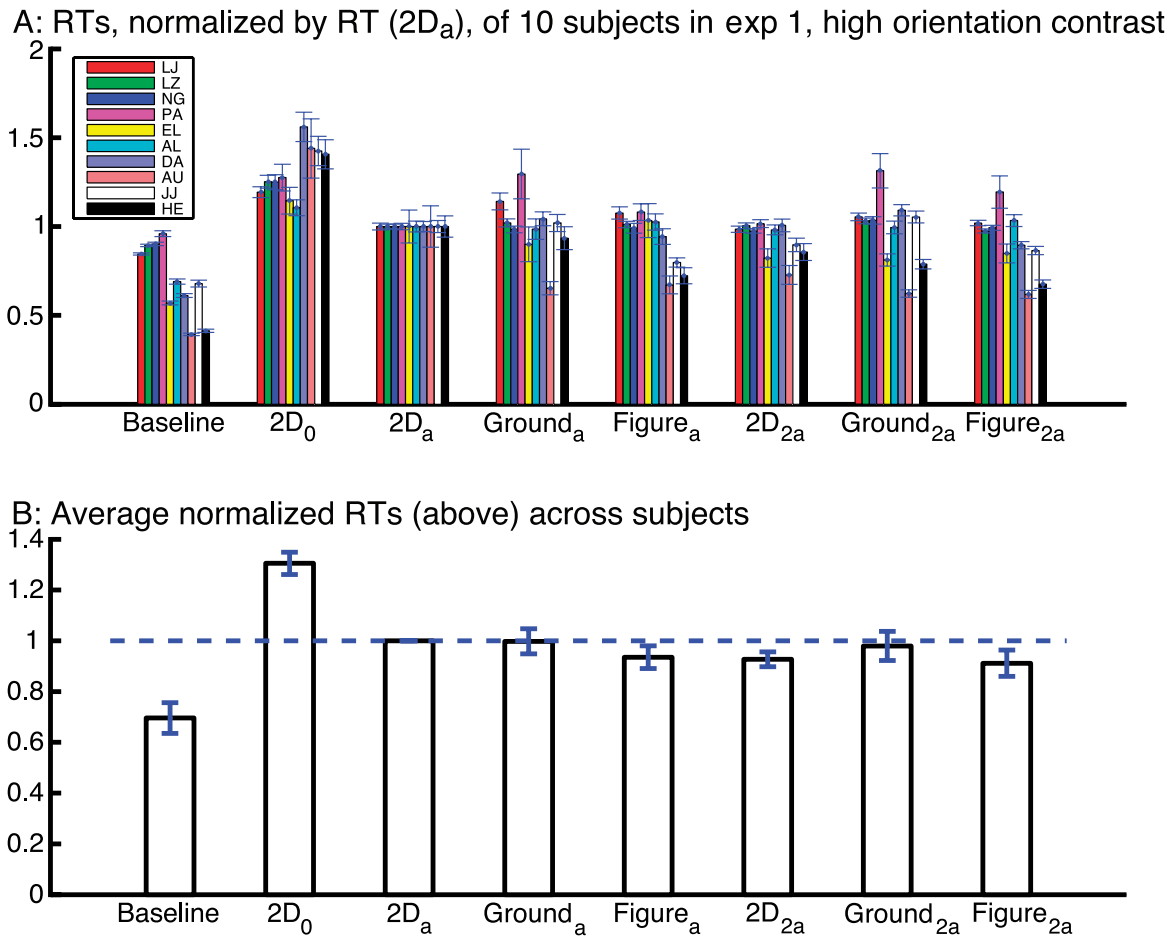

Figure 4. (A) RTs in Experiment 1, of each stimulus type in Figure 3 for each of the 10 subjects (each denoted by a different color). Each RT is normalized by the mean RT(2D$_a$) of the subject. (B) The averages across subjects of the results in panel A. Matched sample *t*-tests indicate that the average (across subjects) normalized RT(2D_0) and RT(Baseline) are significantly higher and lower ($p < 0.005$), respectively, than those of every other stimulus types. Meanwhile, the average normalized RT(2D$_x$), RT(Figure$_x$), and RT(Ground$_x$) for x = a or 2a are not significantly different from each other ($p > 0.05$) except between the average normalized RT(Figure$_{2a}$) and RT(Ground$_{2a}$) ($p = 0.031$). All error bars in all figures indicate the standard errors of the mean.

to locate the texture border quickly, the decision process may not wait for the slower contribution from the 3D cue. In such a case, a shorter RT should be the outcome. The 3D contribution could be ignored if the 2D cue is sufficient for the task before the 3D contribution is available, or if the 3D contribution when available is more or less redundant with the 2D contribution. The second reason could be that the faster group of subjects has no 3D effects regardless of their RTs. Four of the faster subjects (LJ, NG, LZ, and AL) had previously gained extensive experience in the baseline and 2D$_0$ stimulus types due to their participation in the study by Zhaoping and May (2007), three of these four subjects (NG, LZ, and AL) were not naive to the purpose of the experiment, while LJ, one of these four subjects, was not aware of the purpose of the difference between Experiment 1 and Experiment 2. To distinguish between these two reasons, we ask whether the faster subjects continue to demonstrate a lack of contribution from 3D cues, i.e., a zero index E, even when the task is made more difficult so that longer RTs are required.

Experiment 2

Therefore, in Experiment 2, we made a smaller orientation contrast (14°) at the texture border in I_{rel}, rendering the task much more difficult, see Figure 6A. Consequently, the RTs were indeed prolonged, such that RT(2D$_x$) is at least 1200 ms. As seen in Figure 6B, for both I_{ir} shift of $x = a$ and $x = 2a$, the average normalized RT(2D$_x$) is significantly different (and higher) than those of RT(Ground$_x$) and RT (Figure$_x$). Figure 7 shows that the average 3D index $E(x)$ across eight subjects is significantly different (and positive) from zero. In particular, each of the three fastest subjects, LJ, LZ, and NG (among them LJ was not aware of the purpose of the difference between Experiment 1 and Experiment 2), in Experiment 1 demonstrated a significant

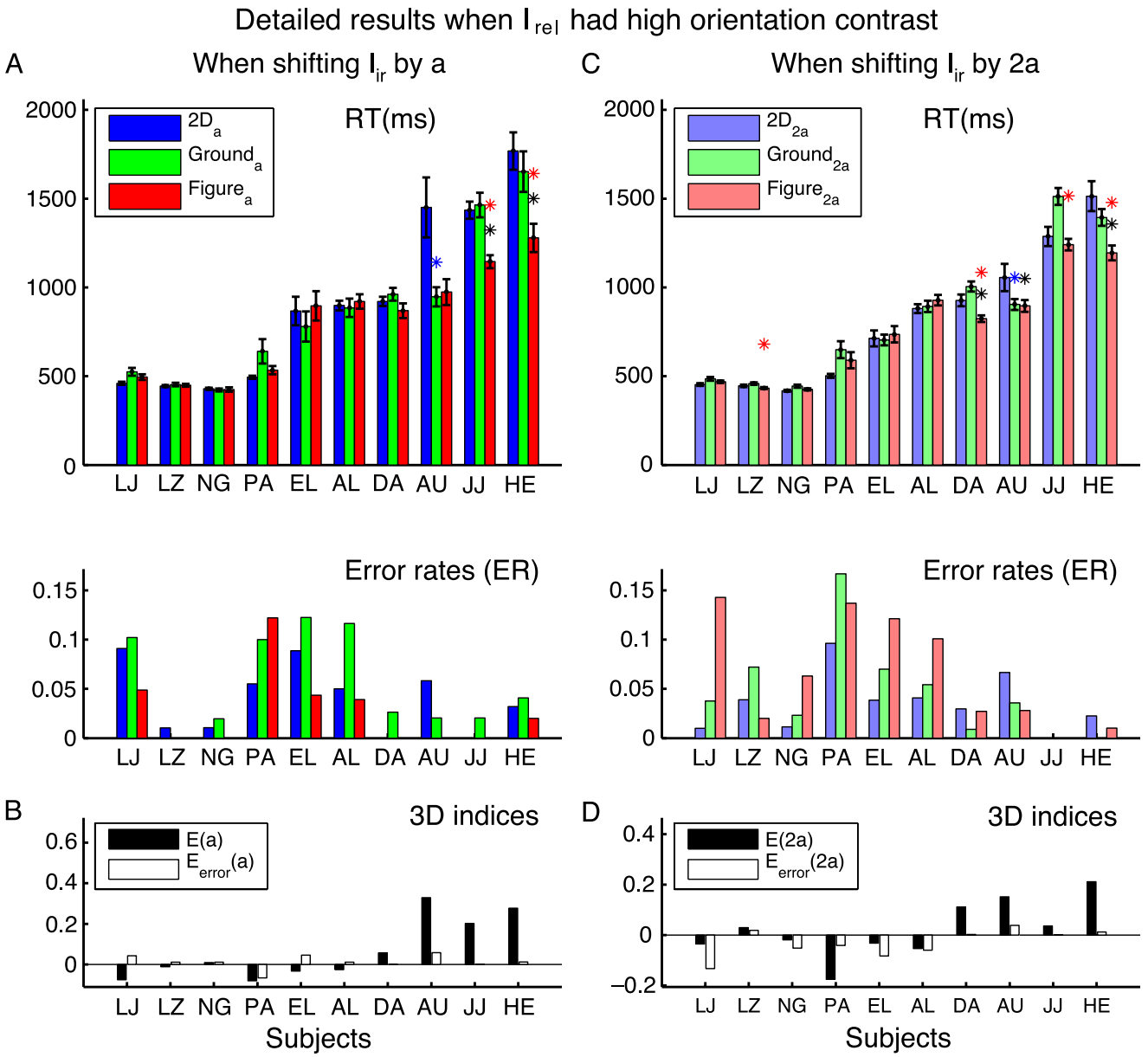

Figure 5. Detailed results from Experiment 1 when I_{rel} contained high orientation contrast, with subjects presented in a rough order of their RTs. Data for absolute offsets of I_{ir} by $x = a$ and $x = 2a$ are plotted in the left and right columns, respectively. The top and middle rows plot, respectively, RTs and error rates for stimulus types $2D_x$, $Figure_x$ and $Ground_x$; the bottom row plots the 3D indices $E(x)$ and the Error indices $E_{error}(x) \equiv e(2D_x) - e(Figure_x)$. For each subject, the black, blue, or red asterisks (*), respectively, indicate that $\Delta(x) = RT(2D_x) - RT(Figure_x)$, $\delta_1(x) = RT(2D_x) - RT(Ground_x)$, or $\delta_2(x) = RT(Ground_x) - RT(Figure_x)$ is significantly larger than zero (two tailed t-test). While the 3D index $E(x)$ tends to be larger when subjects have longer RTs, their means $E(a) = 0.07 \pm 0.05$ and $E(2a) = 0.02 \pm 0.03$ are not significantly different from zero ($p > 0.2$).

3D effect, i.e., their $RT(Figure_x)$ was significantly smaller than their $RT(2D_x)$ for both $x = a$ and $x = 2a$. These findings based on RTs are again not the results of a speed-accuracy trade-off (see Figure 7) since, in individual subjects, the analogous index $E_{error}(x)$ from the error rates is either consistent, or negligible in comparison, with the 3D index $E(x)$ from the RTs. Furthermore, the qualitative findings of Experiment 2 hold when the subjects (AL, AU, and SI) with overall error rates larger than 9% are excluded.

One may be cautious about possible reasons, other than a zero 3D contribution, that could make $RT(2D_x)$ approximately the same as $RT(Figure_x)$ and $RT(Ground_x)$ when RT is short. For instance, it might be possible that a positive 3D contribution is canceled by a cost possibly arising because the two different monocular images in a

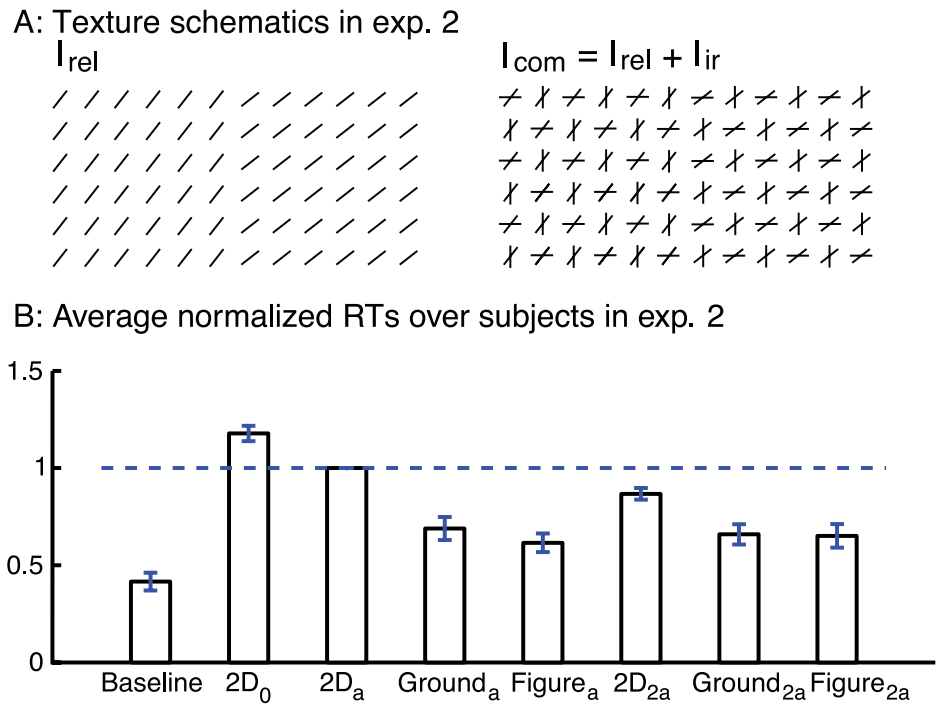

Figure 6. Stimulus characteristics and average RTs in Experiment 2. (A) The schematics of the textures I_{rel} and $I_{com}(0)$, with the vertical texture border in the middle. (B) The RTs, each normalized by RT($2D_a$) of the corresponding subject, averaged over eight subjects, in the same format as that of Figure 4B. By matched sample *t*-test, the average normalized RT($2D_x$) is significantly different ($p < 0.003$) from those of RT($Ground_x$) and RT($Figure_x$), for both $x = a$ and $2a$.

dichoptic display cannot be simply summed up to increase the signal-to-noise ratio of the input strength in V1 (compared to the case of the 2D, bioptic, stimuli when two eyes receive identical inputs). (It is interesting to note that, by the MAX rule of saliency computation depicted in Figure 1, a V1 saliency benefit due to an input change from I_{com} to $I_{com}(x)$ in one eye should be the same as the saliency benefit by this change in both eyes, if any effects by the association and interactions between the inputs to the two eyes are not considered in a very coarse approximation.) For this reason, we compare RTs in the stimulus types $Figure_a$ and $Ground_a$. Both stimulus types have $I_{com}(a)$ in one eye and $I_{com}(-a)$ in the other eye, and they differ only in which eye has $I_{com}(a)$. Hence, without any 3D processing, these two stimulus types are related by a reflection symmetry and should have the same RT. This is so even considering the possible effects of eye dominance, since $I_{com}(a)/I_{com}(-a)$ in the dominant/non-dominant eye should be symmetric with $I_{com}(-a)/I_{com}(a)$ in the dominant/non-dominant eye, and that $I_{com}(a)$ and $I_{com}(-a)$ should be equally easy or equally difficult to segment. This is not so when comparing the stimulus types $Figure_{2a}$ and $Ground_{2a}$, since $I_{com}(2a)/I_{com}(0)$ in the dominant/non-dominant eye is not symmetric with $I_{com}(0)/I_{com}(2a)$ in the dominant/non-dominant eye. Therefore, any RT difference $\delta_2(a) = RT(Ground_a) - RT(Figure_a)$ should arise purely from the 3D processes to separate the two textures into foreground and background surfaces. Denoting the average of $RT(Ground_a)$ and $RT(Figure_a)$ as $RT(3D_a) \equiv [RT(Ground_a) + RT(Figure_a)]\ /\ 2$, we define a figure–ground index

$$\text{Figure}-\text{ground index}$$

$$G(a) \equiv 2\frac{RT(Ground_a) - RT(Figure_a)}{RT(Ground_a) + RT(Figure_a)} = \frac{\delta_2(a)}{RT(3D_a)}, \quad (7)$$

as an additional assessment of any 3D effects. Figure 8 shows that this index value tends to be larger when $RT(3D_a)$ is larger. From both experiments, we obtain 18 figure–ground index values, 10 from Experiment 1 and 8 from Experiment 2. We group them into two groups, the faster and the slower groups, which include subject/experiment with $RT(3D_a)$ shorter and longer than 1000 ms, respectively. The average $G(a) = 0.01 \pm 0.03$ and $G(a) = 0.152 \pm 0.046$ of the faster and slower groups are insignificantly ($p = 0.72$) and significantly ($p = 0.01$) different from zero, respectively, with the average $G(a)$ of the faster group significantly ($p = 0.02$) different from that of the slower group. The weak or negative $G(a)$ values in the two slowest subject/experiments (FA/2 and SI/2) are likely caused by better top–down strategies, enabled by longer RTs, to overcome the attentional distraction in the stimulus $Ground_a$, see Figure 9 and the accompanying analysis. The $G(a)$ difference between the faster and slower groups is not caused by a speed-accuracy trade-off, since

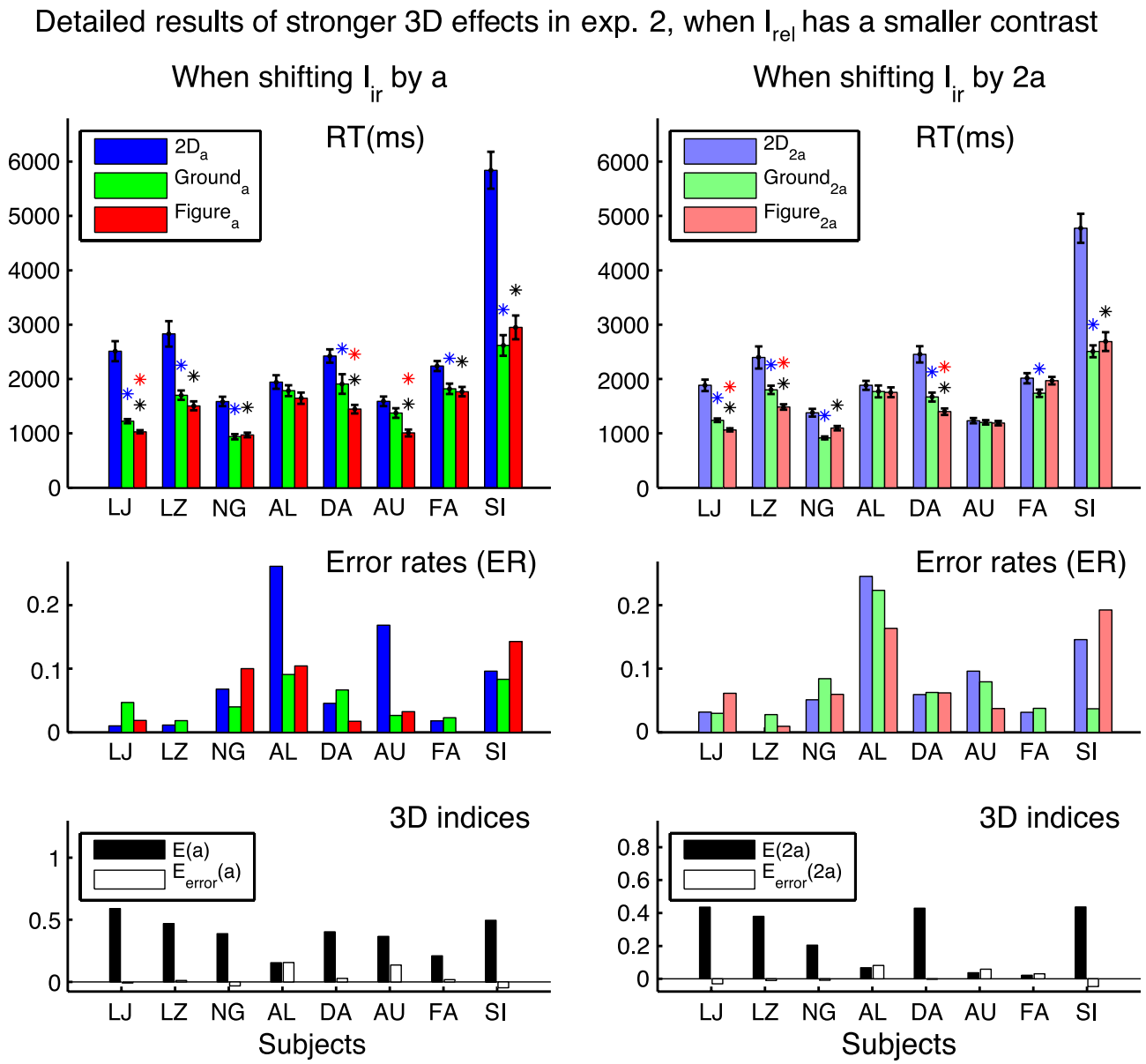

Figure 7. Stronger 3D contributions in Experiment 2, when I_{rel} has an 14° orientation contrast. Shown are RTs, error rates, and 3D indices $E(x)$ and the corresponding error index $E_{error}(x)$ for I_{ir} shifts $x = a$ and $x = 2a$, respectively. The plots are in the same format as Figure 5. Averaged over eight subjects, the 3D indices $E(a) = 0.39 \pm 0.05$ and $E(2a) = 0.25 \pm 0.06$ are both significantly different from zero ($p < 0.01$). The error index $E_{error}(x)$ of each subject is either consistent with the corresponding 3D index $E(x)$ in direction or is negligible in comparison. The average $E_{error}(x)$ over subjects is positive but not significantly different from zero ($p > 0.25$) for both $x = a$ and $2a$.

the analogous figure–ground index $G_{error}(a) \equiv e(\text{Ground}_a) - e(\text{Figure}_a)$ from the corresponding performance error rates is not significantly different from zero for either group. These observations are consistent with the idea that 3D contribution to visual attentional guidance appears only after sufficient visual processing time.

From Equations 1–4 and the accompanying arguments, we see that the relative proportion of $\delta_1(x) = \text{RT}(\text{2D}_x) - \text{RT}(\text{Ground}_x)$ within the RT difference $\Delta(x) = \delta_1(x) + \delta_2(x) = \text{RT}(\text{2D}_x) - \text{RT}(\text{Figure}_x)$ can depend on the top–down control of attention. In particular, when subjects focus their attention on the depth plane of the display screen which contains the task-relevant I_{rel} in all 3D stimuli, they can be less distracted by the task-irrelevant I_{ir} even when it is in front of I_{rel}. In such a case, $\text{RT}(\text{Ground}_x)$ can be made smaller and $\delta_1(x) = \text{RT}(\text{2D}_x) - \text{RT}(\text{Ground}_x)$ can be increased. Hence, the ratio

$$F(x) \equiv \delta_1(x)/\Delta(x), \tag{8}$$

can be a top–down strategy index. An $F(x) \approx 1$ suggests a better strategy. A small or even negative $F(x)$ suggests that the subject is distracted by the default deployment of attention to surface I_{ir} in the front. If subjects used an

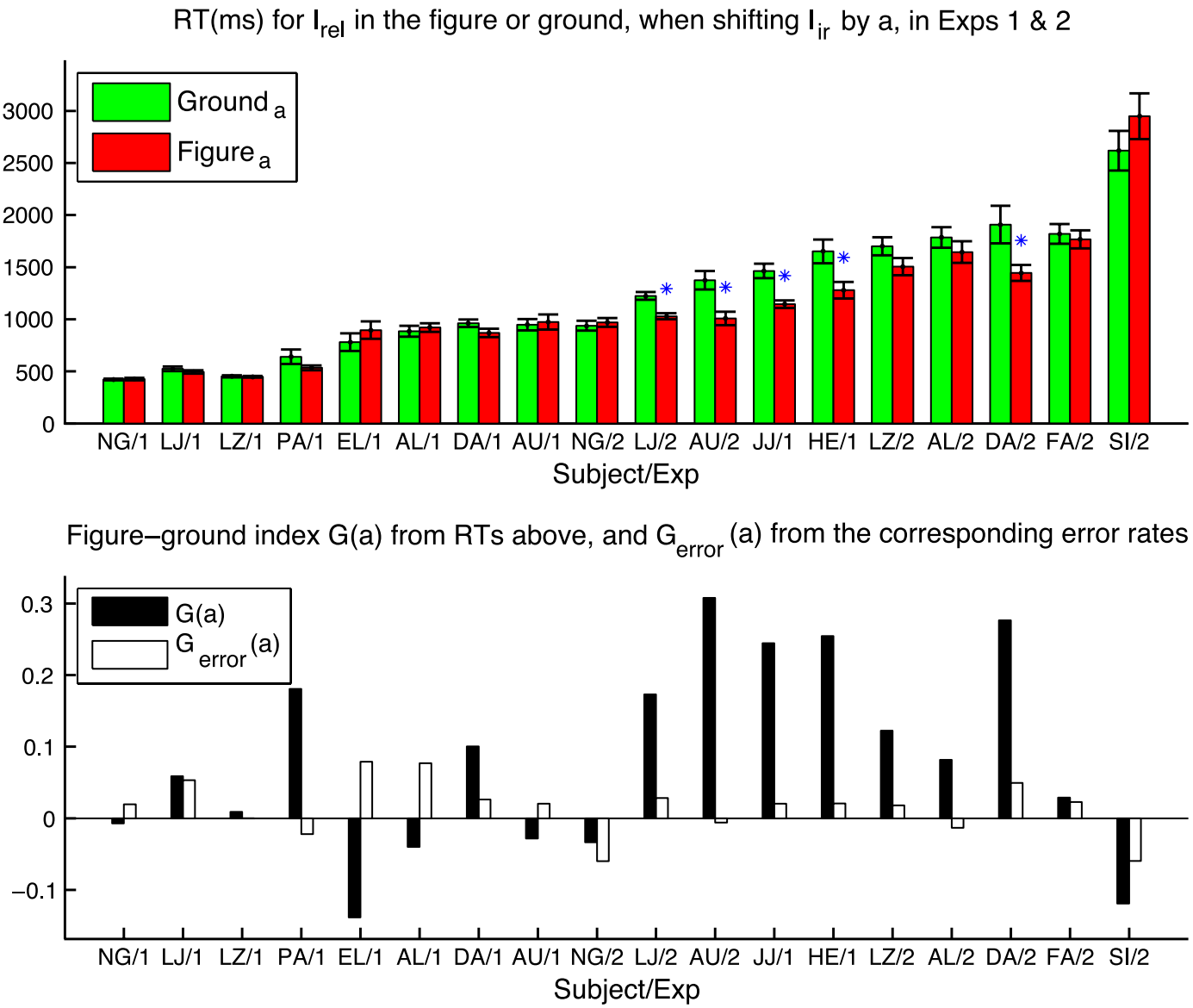

Figure 8. Figure–ground effects tend to increase with RT. Top: RT(Ground$_a$) and RT(Figure$_a$) for each case (subject/experiment), listed in the rough order of the average RT(3D$_a$) ≡ [RT(Ground$_a$) + RT(Figure$_a$)] / 2. The RT difference $\delta_2(a)$ = RT(Ground$_a$) − RT(Figure$_a$) tends to increase with RT(3D$_a$) and is marked by an asterisk (*) when it is significantly different from zero ($p < 0.05$). Bottom: figure–ground indices $G(a) \equiv \delta_2(a)$/RT(3D$_a$) of the corresponding cases and the analogous indices $G_{error}(a) \equiv e$(Ground$_a$) − e(Figure$_a$). For the faster and slower groups of subject/experiments, respectively (with the faster group containing nine subject/experiments, NG/1, LJ/1, LZ/1, PA/1, EL/1, AL/1, DA/1, AU/1, and NG/2, whose RT(3D$_a$) < 1000 ms), the average $G(a) = 0.01 \pm 0.03$ and $G(a) = 0.152 \pm 0.046$ are insignificantly ($p = 0.72$) and significantly ($p = 0.01$) different from zero. Both groups have an average index $G_{error}(a)$ insignificantly different from zero ($p > 0.19$).

over-compensating strategy by preparing to ignore the foreground surface even before a trial started, it might incur an RT cost in RT(Figure$_x$) since subjects did not know ahead of each trial whether I_{rel} would be in the foreground or background. This strategy could even lead to RT(Ground$_x$) < RT(Figure$_x$) so that $F(x) > 1$. Figure 9A plots $F(x)$ in the order of ascending RT(2D$_x$) (plotted in Figure 9B) for all cases (any subject, either experiment, and for either shift $x = a$ or $2a$) when there was a significant 3D contribution index $E(x)$, i.e., when a subject's RT(2D$_x$) is significantly different from, and larger than, RT(Figure$_x$). It is apparent that when RT(2D$_x$) is small, $F(x)$ varies substantially between subjects and cases, ranging from $F(x) < -0.5$ to $F(x) > 1.5$. By contrast, when RT(2D$_x$) is large, $F(x)$ varies much less between cases and clusters around $F(x) \sim 0.84$. The mean $F(x)$ is 0.47 ± 0.25 for the eight cases with shorter RT(2D$_x$) < 2000 ms but is 0.84 ± 0.06 for the nine cases with longer RT(2D$_x$). These observations suggest the following. First, guidance of attention by 3D processes carries some degree of top–down control (He & Nakayama, 1995), which by nature is expected to vary between subjects and easily affected by

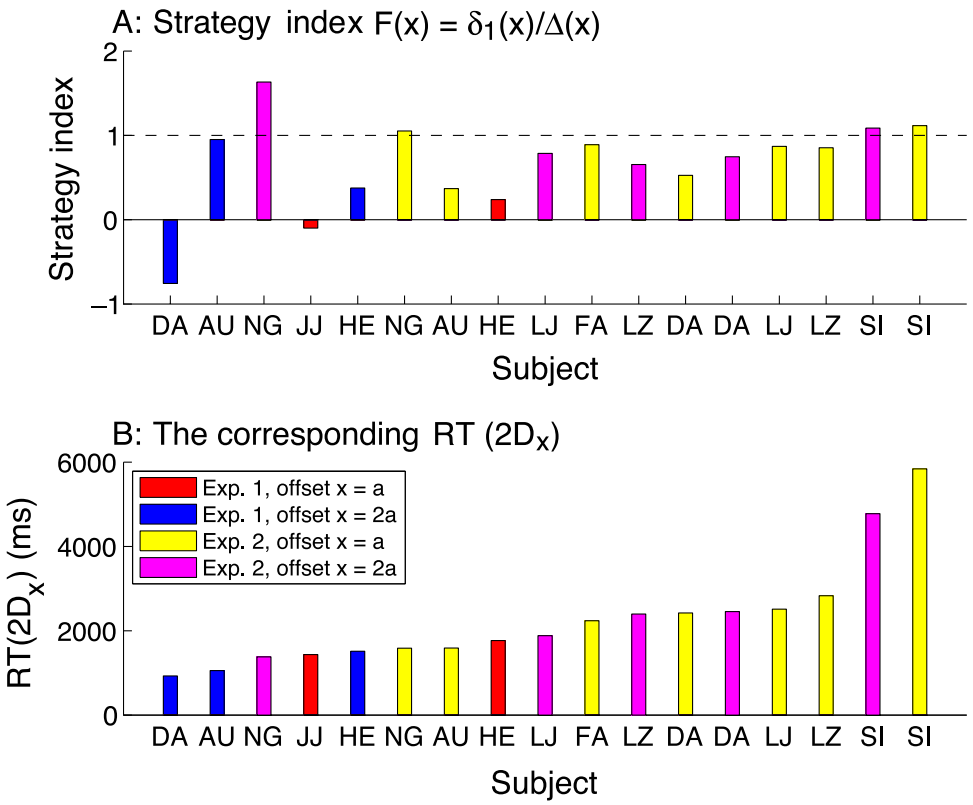

Figure 9. (A) Strategy index $F(x) = \delta_1(x)/\Delta(x) = [RT(2D_x) - RT(Ground_x)] / [RT(2D_x) - RT(Figure_x)]$, plotted in the ascending order of the corresponding $RT(2D_x)$ plotted in panel B. The plotted data come from all cases when $RT(2D_x) - RT(Figure_x)$ is significantly different and larger than zero for any subject in either experiment for either offset $x = a$ or $x = 2a$. Note that $F(x)$ is less variable and on average larger when $RT(2D_x)$ is long enough.

learning. Second, a longer processing time associated with a longer RT makes it easier to deploy top–down control effectively.

Discussion

The current findings in relation to hierarchical visual functions

Attention can be attracted to a spatial position, such as that of a red item among green items. It can also be guided by 3D surface structures in the scene, as demonstrated by the ease of visual search within one of several parallel surfaces in depth (He & Nakayama, 1995). Our task of locating a texture border in a textured surface, in the presence of another, irrelevant, texture surface, can be executed with or without separating the two texture surfaces in depth. Therefore, attention can be guided to the relevant location on the task-relevant texture surface with or without 3D depth processing. This task thus provides us with a means of probing the relative contributions to attentional guidance by 2D and 3D visual processes, by measuring the reduction in the task RT when the 3D cues are available from that when the 3D cues are unavailable. This is measured in our Experiment 1 and Experiment 2, which are the same except that the task in Experiment 1 is easier and requires shorter RTs due to a larger orientation contrast in the task-relevant texture surface. We find that the RT reduction by depth cues in Experiment 1 is present only for subjects whose RTs without the depth cues are longer than about 1 second. In contrast, in Experiment 2, in which RTs for all subjects are longer than 1 second without depth cues, the RT reduction by depth cues is evident for all subjects, including those having no RT reductions (due to their shorter RTs) in Experiment 1. These findings suggest that 3D depth processes do not contribute to attentional guidance until a substantial time after visual input onset. Our findings are consistent with the past observation that the RT is typically longer than 1 second to find a visual search target when the depth information is essential rather than dispensable—in a conjunction search in which depth was one of the defining features of the target (McCarley & He, 2001).

3D and object-based attention

Previous work (He & Nakayama, 1995; Nakayama & Silverman, 1986) found that 3D surface depth information can facilitate attentional guidance to a depth and within a depth plane, such as in conjunctive searches involving depth. Since 3D depth information is associated with visual surface or object attributes, object surfaces should also facilitate attentional guidance. Indeed, the focus of attention spreads within an object surface much more readily than across object boundaries (Duncan, 1984), and this is known as the object-based attention. By comparison, our study shows that the contribution from 3D depth processes to attentional guidance only appears after an initial temporal window since visual input. If the outcome of visual processes (such as the bottom–up saliency mechanism from 2D processes) within this initial window is sufficient for the task, the task decision could be made without a contribution from 3D or object processes. It was recently shown that the object-based attentional effect is absent if the visual inputs are viewed too briefly (below a second; Chen & Cave, 2008). This implies that object is not formed when processing time is too short, and that attention can be allocated before the stimulus has been processed into separate objects. The latter is in line with our observations suggesting that attentional guidance for a task can be completed before image features are attributed to separate depths. That formation of visual objects from image features takes a long time is also implied by the observation that the task-irrelevant feature of a prime object does not affect the RT of a task unless there is enough processing time (Chen, 2005), presumably to bind the irrelevant feature with the relevant feature to form a single prime object.

"Where" and "what" vision in low and mid-level visual processes

It is quite common to rely only on a subset of dissociable visual processes to carry out various visual tasks. For instance, some visual detection tasks can be executed relying mainly on the visual processes that extract "where" but not "what" information in visual inputs (Sagi & Julesz, 1985). Often human observers can quickly locate a salient location, such as that of a red item among green ones, or that of a texture border with a high orientation contrast (and without any interfering textures), even though without a conscious or attentive effort they often cannot identify the visual features (e.g., color or orientation) of the salient item immediately afterwards if the visual input is quickly terminated (Koene & Zhaoping, 2007; Zhaoping & May, 2007). Human attention can even locate, or be attracted to, an eye of origin singleton, which is apparently identical in appearance to the background items, without the observer being aware of any visual distinction between the singleton and background items (Zhaoping, 2008a). In these cases, the bottom–up saliency computed by V1 carries much of the "where" information. This is also consistent with the finding that V2 lesions in monkeys can disrupt region content discrimination but not region border detection (Merigan, Nealey, & Maunsell, 1993).

In these "where" tasks above, the visual features concerned (color, orientation, eye of origin) are all low level features processed by V1. Information about these features is used to locate or detect "where" an input location is, even though it may not be used by the "what" processes to enter awareness. In comparison, the visual depth feature investigated by the current study lies at a somewhat higher perceptual level and is extracted beyond V1 starting as early as V2 (see the next paragraph for more detail). Our subjects were aware of the surface depths when they used them to guide their task performance. He and Nakayama (1995) referred to this 3D surface information as coming from an intermediate level of perceptual organization.

Neural substrates

Extrastriate cortex for depth and surface processes

Converging physiological evidence suggests that the neural substrates for 3D and surface/object processes are most likely beyond V1 and are in the extrastriate cortex starting from V2 in the visual pathway. Depth perception by stereo cues requires making true matches between image features in the two eyes and eliminating false matches. A true match is between corresponding visual features in the two eyes that arise from the same object location in the perceived 3D scene. A false match arises from coincidental image correspondence between the input features to the two eyes that do not arise from a single object location in the perceived 3D scene. Cumming and Parker (2000) showed that V1 neurons respond to false matches as strongly as they do to true matches. Binocular rivalry, an outcome of (an unsuccessful) stereo-matching for 3D perception, can also be used as an indicator of 3D processes. In this regard, responses in V1 are much less correlated with the fluctuating perceptions (the outcomes of stereo matching) during rivalry than those in V4, and responses from almost all monocular cells in V1 are unaffected by perceptual suppression of the inputs to the preferred eye (Leopold & Logothetis, 1996), suggesting again that V1 is less involved in stereo-matching processes. By contrast, V2 and V4 neurons show stronger responses to true matches (Parker, 2007; Tanabe, Umeda, & Fujita, 2004), responses of MT neurons are linked with the depth perception in behavior (DeAngelis, Cumming, & Newsome, 1998), and by the level of IT (inferior temporal cortex, the far extrastriate cortex) it is suggested that the stereo correspondence problem is solved since IT neurons do not respond to binocularly anti-correlated stimuli which do not correspond to real world objects

(Janssen, Vogels, Liu, & Orban, 2003). V2 neurons in monkeys also use contextual depth information to integrate occluded contours, signal the presence of object boundaries, and segment surfaces, such that V2 responses manifest amodal completion, illusory contour completion, and disparity capture associated with surface boundaries and depth orders (Bakin et al., 2000). This means that the responses from V2 neurons reflect depth perception, which arises from true but not false matches, rather than merely replicating direct stimulus inputs, which include both false and true matches. V2 neurons in monkeys respond to the boundaries of surfaces formed by random-dot stereograms such that these surfaces are not visible in 2D monocular images alone (von der Heydt et al., 2000). As stereo process is closely associated with surface and object processes, it is not surprising that there have been many observations implicating V2's involvement in surface and object processes without necessarily involving stereo or dichoptic inputs. These observations are as follows: V2 but not V1 neurons can respond to illusory contours as boundaries for surfaces (von der Heydt, Peterhans, & Baumgartner, 1984); V2 and V4 cells are more likely than V1 cells to carry information about border ownership, which is the information regarding the depth orders of the surfaces (Zhou, Friedman, & von der Heydt, 2000); and V2 neurons are involved in inferring 3D figure–ground structures from 2D or dichoptic displays (Qiu & von der Heydt, 2005, 2007). Consistent with our findings implicating top–down control in attentional guidance by 3D cues, it has been shown that the processes of figure–ground separation and attentional control combine to impact on the neural activities in V2 neurons (Qiu, Sugihara, & von der Heydt, 2007). Computational models of figure–ground processing in V2 involving stereo (Zhaoping, 2002) or non-stereo inputs (Craft, Schutze, Niebur, & von der Heydt, 2007; Zhaoping, 2005) have also been constructed to illustrate the algorithmic feasibilities of the presumed V2 and extrastriate neural circuits for depth and surface processes.

V1 for image feature-based attentional selection

Meanwhile, there has also been converging evidence from physiological, psychophysical, and computational studies that V1 creates a bottom–up saliency map to guide attention in 2D stimuli. Physiologically, V1 responses exhibit contextual influences, such as the iso-orientation suppression (e.g., Allman et al., 1985; Knierim & van Essen, 1992; Li & Li, 1994) that enables the texture border pop out in Figure 1 and colinear facilitations (Kapadia, Ito, Gilbert, & Westheimer, 1995) to make a smooth contour (see Figure 10B) more salient, mediated by intra-cortical interactions. Psychophysical studies (Jingling & Zhaoping, 2008; Koene & Zhaoping, 2007; Zhaoping, 2008a; Zhaoping & May, 2007; Zhaoping & Snowden, 2006) using various 2D texture stimuli have tested and confirmed various behavioral predictions (on visual search and segmentation) from the theoretical hypothesis (Li, 1999a, 2002) that V1 creates a saliency map, such that bottom–up attention is most likely attracted to the receptive field location of the most active V1 neuron. Of particular relevance to our current study, the V1 saliency hypothesis can explain the dramatically different RTs in segmenting I_{rel} and I_{com}, as illustrated in Figure 1. Variations in the eases of texture segmentation and visual search due to variations in stimulus characteristics in many other textures have also been shown by computational modeling (Li, 1999a, 1999b, 2000) or psychophysical tests to follow the predictions of the V1 saliency hypothesis. For instance, Koene and Zhaoping (2007) showed that finding a bar unique in color, orientation, moving direction, or a combination of them, in a background texture of uniformly oriented, colored, and moving bars can be accounted for by the V1 hypothesis using a property of V1 but not extrastriate cortex—that there are cortical neurons jointly tuned to orientation and color, or jointly to orientation and motion direction, but not jointly to color and motion direction (Gegenfurtner, Kiper, & Fenstemaker, 1996; Horwitz & Albright, 2005; Livingstone & Hubel, 1984). Zhaoping (2008a) showed that texture segmentation or singleton detection can also be caused by a spatial contrast in the eye of origin feature which is non-distinctive to visual awareness and is coded mainly by V1 rather than higher cortical neurons (Burkhalter & Van Essen, 1986; Hubel & Wiesel, 1968). Zhaoping and May (2007) and Zhaoping and Snowden (2006) showed how texture segmentation by orientation features can be interfered by irrelevant color features, but less so in reverse, according to the V1 hypothesis and V1 mechanisms.

Combining (1) the above evidence for V1's role in guiding attention in 2D stimuli, (2) the physiological evidence outlined above on the divide between V1 and extrastriate cortex in the 3D processes, and (3) our current finding of a delayed manifestation of 3D contribution to attentional guidance, we can arrive at the following implication: contributions by the extrastriate cortex to attentional guidance (from the outcomes of processing the sensory inputs) are sufficiently weak and delayed relative to the contributions by V1 such that they will not manifest in the subject's report until about 1000 ms after visual input onsets.

From neural response latencies to behavioral reaction times

It is known that latencies of V1 responses are around 60 ms and those of V2 responses are only about 15 ms longer (Bullier & Nowak, 1995; Schmolesky et al., 1998). It may then seem puzzling that a small latency difference between V1 and V2 or extrastriate responses could lead to a much longer RT difference, given that the RT to report a

V1 feature singleton target (e.g., an unique vertical bar among horizontal background bars) can often be shorter than 400 ms. This can be understood by noting that the latency of a sensory neural response and that of a decision making (for the consequent motor response) based on the sensory response are two different matters. The 400-ms latency to report a V1 feature singleton is itself much longer than the 60-ms latency of V1 responses, not only because of a motor latency to execute a motor command, but also because of the time needed to make a decision for the motor command. A more salient singleton target requires a shorter reporting latency. In comparison, the difference between the latencies of V1 responses to targets of different saliencies is negligible. Using a standard model of decision making (Smith & Ratcliff, 2004), one may view the latency for decision making as the time needed to integrate the saliency signal so that the integrated signal reaches a sufficient amount called a decision threshold. Accordingly, longer decision latencies are needed for less salient targets. Therefore, the long latency of the behavioral manifestation of the proposed extrastriate contribution to attentional guidance may be attributed mainly to a weak saliency signal generated by the extrastriate processes, and negligibly to a slightly longer neural response latency. Note that V1 processes can also generate weak saliency signals in response to non-salient inputs like our stimulus type $2D_0$ and should consequently lead to longer behavioral RTs which may be longer than some RTs attributed to the extrastriate processes (for example, in our data when $RT(2D_x) > RT(Figure_x)$). However, while much shorter behavioral RTs can be attributed to V1 processes, such as when in responses to our salient baseline stimulus, our findings in this paper suggest that, in response to or contingent on external input stimulus (rather than top–down factors), extrastriate processes are not as capable as V1 processes to generate short behavioral RTs.

A network of brain processes through time

In typical visual behavior, attention is guided by both top–down and bottom–up factors, and a network of many cortical areas will be involved for attentional control. Many previous studies focus on controls by top–down factors, involving cortical areas, such as the parietal and frontal cortex, much further downstream in the visual pathway from V1 (Corbetta & Shulman, 2002; Desimone & Duncan, 1995). By contrast, our work started from the bottom–up guidance of attention by V1 and used time and levels of difficulties of processing to probe the role of the next cortical areas downstream in the visual pathway for attentional control. It is natural to expect that more top–down components of attentional control will manifest themselves as one moves further downstream towards cortical areas known to be involved in top–down control. In our RT data demonstrating the contributions of 3D processes, there are some patterns, as shown in Figure 9, already suggestive of the top–down component. It is difficult to quantify the respective contributions from the top–down and bottom–up factors in the RT difference $\Delta(x) = RT(2D_x) - RT(Figure_x)$, $\delta_1(x) = RT(2D_x) - RT(Ground_x)$, and $\delta_2(x) = RT(Ground_x) - RT(Figure_x)$ caused by the 3D processes, although one can perhaps qualitatively state that $\Delta(x)$ is the least influenced by top–down control among the three. More extensive investigations, beyond the scope of this study, will be necessary to answer this question fully. Our approach, using successive temporal windows of processing or behavior to investigate the involvements of brain areas in guiding attention, can be extended further along the visual pathway. According to this approach, V1 appears to play a dominant role in bottom–up attentional guidance for behavior within about 1 second after the onset or a sudden unpredictable change of visual inputs. Meanwhile, attentional guidance by object processes or contributions from the extrastriate cortex could still be substantial or even dominant after this initial temporal window. This is consistent with the observations that while human saccades on static photographs are better predicted by visual objects than by saliency (Einhäuser, Spain, & Perona, 2008), the first few saccades are very similar to those made by observers with visual object agnosia (Mannan, Kennard, & Husain, 2009), suggesting that the early saccades are dominantly controlled by bottom–up saliency rather than object processes.

Theories and models

As pointed out in Introduction, the traditional "back-pocket" model (Bergen & Landy, 1991; Landy & Graham, 2004; Malik & Perona, 1990) can also explain many texture segmentation phenomena in 2D textures, such as the texture I_{rel}. These models, also called filter–rectify–filter models, are characterized by three phenomenologically serial processing stages: (1) a linear filtering of the input image by spatial arrays of filters, each array contains filters tuned to a particular feature (or feature combination) such as a particular orientation and/or spatial frequency; (2) a non-linear operation, such as squaring, of the filter outputs to obtain a spatial map of response levels, such as energy, from each array of filters, e.g., a uniform texture of 45° oriented bars should generate uniform energy levels in the map for the 45° oriented filters; and (3) a linear coarse scale spatial filtering (by a spatial differentiating filter) to find locations of spatial contrast in each map of responses from Stage 2. Stage 3 is treated by the "back-pocket" model as phenomenological, without any particular neural justifications. In this sense, the V1 saliency hypothesis—with V1's intra-cortical mechanisms to highlight the salient locations (which are often the borders of texture regions)—could be seen as proposing the neural substrates for the filter–rectify–filter

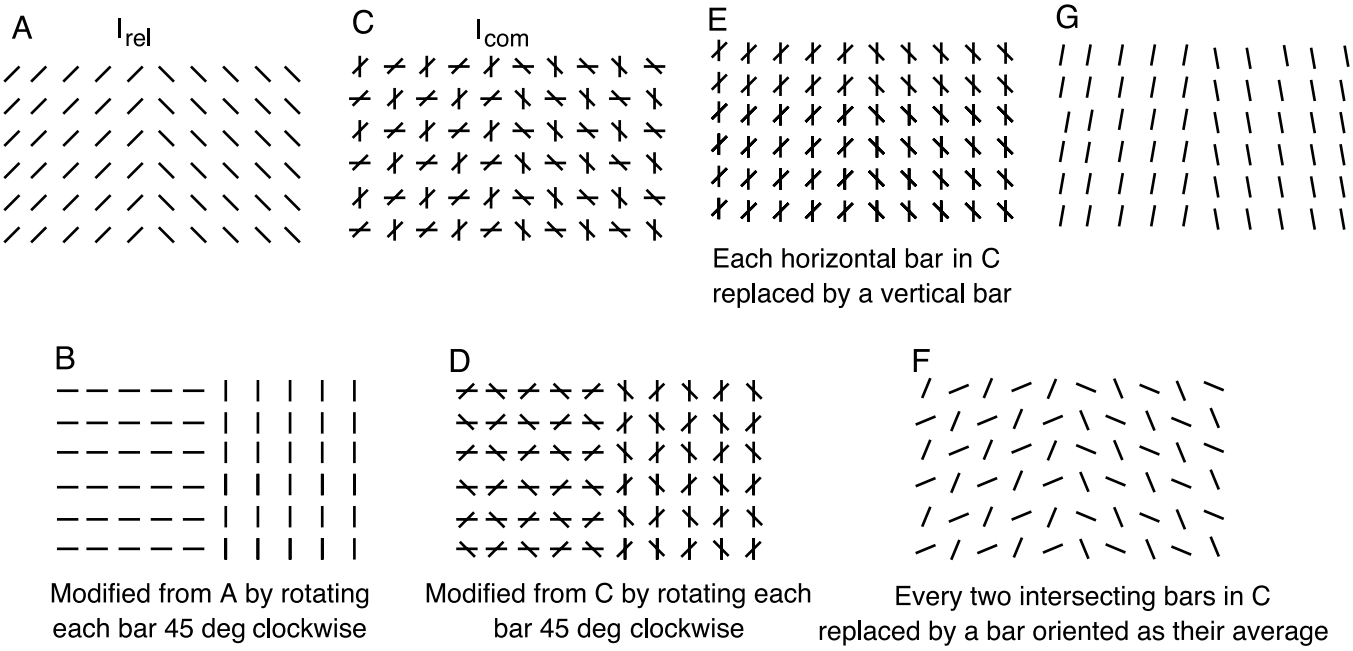

Figure 10. Illustration that V1 saliency hypothesis explains the relative eases in finding the vertical texture border in the middle of each texture A–G. Textures A and C are I_{rel} and I_{com} in Figure 1 used in our study; B and D are modified from A and C, respectively, by rotating each bar 45° clockwise; replacing each horizontal bar in C by a vertical bar makes E; and replacing each pair of intersecting bars in C by a single bar oriented at the average orientation of the two bars makes F. Reducing the orientation contrast in A to 20° makes G. The V1 hypothesis predicts that C is the most difficult and definitely more difficult than D, E, F, and G, since it creates no saliency highlights at the border as explained in Figure 1; E is easy because the V1 responses to the irrelevant vertical bars are as iso-orientation suppressed as those to the background (non-border) relevant bars and thus producing little interference; F is easy because all bars are now relevant and the border bars should evoke higher, or less suppressed, responses than non-border bars by having fewer iso-orientation neighbors; B is easier than A since V1's colinear facilitation (not included in most region feature-based models) makes a (vertical) texture border made of (vertical) bars parallel to the border more salient; D is easier than C since the relevant border bars evoke higher responses in D than in C due to the colinear facilitation, so these bars in D are less susceptible to interference by the irrelevant bars (since the responses to the relevant bars are no longer completely buried among the responses to the irrelevant bars); G is easier than C since the reduced iso-orientation at the border leads to a saliency highlight. These V1 predictions have been confirmed (Zhaoping & May, 2007). That B is easier than A to segment had been observed by Wolfson and Landy (1995).

model. In particular, the V1's neural receptive fields implement Stage 1 and V1's intra-cortical interactions effectively realize Stage 3. The filter–rectify–filter model is just one example of the general class of (image) region feature-based algorithms or models, which segment textures by identifying the differences between image features in different image regions. Traditional psychological models of bottom–up visual attention (e.g., Itti & Koch, 2000; Koch & Ullman, 1985; Treisman & Gelade, 1980; Wolfe et al., 1989) are also examples of region feature-based models, and they have also been applied to texture segmentation. These models assume that visual inputs are analyzed simultaneously by various feature maps, each devoted to a particular feature value such as red color, green color, left tilted orientation, and upward movements. They also assume that locations of spatial contrasts in these feature maps give higher activations, and that these activations from various feature maps are summed into a single master map to direct bottom–up attention to locations of higher total activations. These models are also not so concerned with the underlying neural substrates. They are similar to the filter–rectify–filter model phenomenologically since the feature map activations are like the outcomes of Stage 2 in the filter–rectify–filter model, while the highlights in the master map correspond to the outcomes in Stage 3.

Connecting with the discussions above on the "where" and "what" processes in vision, we note that the region feature-based models are essentially doing texture segmentation by processing the "what" information, since they work by finding a difference between the "what" feature values in different image regions. Indeed, given enough time, subjects in our experiments can eventually locate the non-salient texture border by scrutinizing the feature values and their differences. Hence, the region feature-based models may be more applicable to the slower visual processes associated with top–down attention that identifies visual features. In the context of our study, these models may be more applicable to the segmentation behavior in Experiment 2 than Experiment 1. In comparison, the V1

saliency hypothesis operates on the faster "where" processes (Sagi & Julesz, 1985), aimed to find "where" to direct attention to without explicitly decoding the "what" information from the visual inputs, even though the image feature values are implicitly used by the V1 mechanisms (through the feature specific intra-cortical interactions) to compute the "where" information. This process of segmentation without the explicit "what" processing has been referred to as *segmentation without classification* (Li, 1999b). Not so involved in depth processing, the V1 mechanisms should be more applicable for 2D texture segmentation on a faster time scale.

Aimed at investigating differential functional roles of V1 and extrastriate cortex in attentional guidance, this paper is not focused on contrasting the V1 saliency hypothesis from the region feature-based models. Interested readers can find in a previous work (Zhaoping & May, 2007) texture segmentation and visual search behavioral data arguing that the V1 hypothesis provides a more parsimonious account of how eases in segmentation and search tasks vary with 2D visual input stimuli. Here we briefly outline the findings by Zhaoping and May (2007) using the seven indicative texture examples A–G in Figure 10. A texture segmentation model should explain why it is much easier to locate the vertical texture border in some of these seven textures than others. Note that there is a feature difference, in terms of the difference in the orientations of the texture bars, on a coarse spatial scale across the middle vertical texture border in each of the seven examples. It is thus a particular challenge to explain why texture G is easier to segment than texture C, even though a 90° orientation contrast occurs across the border in C but only a 20° contrast in G, or to explain why E is easier than C even though both contain task-irrelevant bars and have the same orientation contrast across the border, or to explain why F is easier than C even though both have irrelevant orientation contrasts within the left and right half textures. Without hypothesizing any additional mechanisms beyond those known to exist in V1, the V1 saliency hypothesis predicts the following relative ease of segmenting textures A–G: texture C is definitely more difficult to segment than D, E, F, and G, while textures A and B are the easiest with texture B easier than A, as explained in the caption of Figure 10. These predictions have been subsequently confirmed, in particular, the RTs for textures D–G are typically no more than 50% higher than the RT for A regardless of whether the subjects are trained, while the RT for texture C more than double the RT for A for untrained subjects (Zhaoping & May, 2007).

One may modify the region feature-based models to construct a new psychological or phenomenological model which behaves according to the V1 saliency hypothesis applied on the known V1 mechanisms. This new phenomenological model would be able to explain our data on the 2D texture segmentation. To explain our data on the 3D contribution to the texture segmentation, this phenomenological model could be augmented by a subsequent stage for processing 3D information for the goal of the task. Building such a phenomenological model is however not within the scope of this study.

Conclusion

We measured the speed of segmenting two textures in a surface in the presence of a superposing irrelevant texture surface. Given the same 2D visual cues, a depth separation between the relevant and irrelevant texture surfaces does not make the segmentation faster, unless the task is so difficult that the segmentation without the depth separation takes more than about 1 second for subjects to report the task outcome. Provided that the speed of the task performance is mainly determined by the speed of attentional guidance to the task-relevant visual inputs and that V1 and extrastriate cortex are mainly responsible for 2D and depth processes, respectively, and considering a temporal latency from the completion of effective attentional guidance to subjects' task report, our findings suggest that V1 dominates the input-driven attentional guidance within perhaps the initial several hundreds milliseconds after visual input onset, and extrastriate cortex starts to contribute afterwards.

Acknowledgments

This work is supported in part by the Gatsby Charitable Foundation and a Cognitive Science Foresight grant BBSRC #GR/E002536/01. I wish to thank the two anonymous reviewers, Zhe Chen, Mike Landy, and Ning Qian for very helpful comments, and Zhou Pei-Yuan Center for Applied Math and Institute for Advanced Study for hosting my sabbatical visit to Tsinghua University in 2008–2009.

Commercial relationships: none.
Corresponding author: Li Zhaoping.
Email: z.li@ucl.ac.uk.
Address: Department of Computer Science, University College London, London WC1E 6BT, UK.

References

Allman, J., Miezin, F., & McGuinness, E. L. (1985). Stimulus specific responses from beyond the classical receptive field: Neurophysiological mechanisms for local-global comparison in visual neurons. *Annual Review of Neuroscience, 8,* 407–430. [PubMed]

Bakin, J. S., Nakayama, K., & Gilbert, C. D. (2000). Visual responses in monkey areas V1 and V2 to

three-dimensional surface configurations. *Journal of Neuroscience, 20,* 8188–8198. [PubMed]

Bergen, J. R., & Landy, M. S. (1991). Computational modeling of visual texture segregation. In M. S. Landy & J. A. Movshon (Eds.), *Computational models of visual processing* (pp. 253–271). Cambridge, MA: MIT Press.

Bullier, J., & Nowak, L. G. (1995). Parallel versus serial processing: New vistas on the distributed organization of the visual system. *Current Opinion in Neurobiology, 5,* 497–503. [PubMed]

Burkhalter, A., & Van Essen, D. C. (1986). Processing of color, form and disparity information in visual areas VP and V2 of ventral exrastriate cortex in the macaque monkey. *Journal of Neuroscience, 6,* 2327–2351. [PubMed]

Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature, 363,* 345–347. [PubMed]

Chen, Z. (2005). Selective attention and the perception of an attended non-target object. *Journal of Experimental Psychology: Human Perception and Performance, 31,* 1495–1509. [PubMed]

Chen, Z., & Cave, K. R. (2008). Object-based attention with endogenous cuing and positional certainty. *Perception & Psychophysics, 70,* 1435–1443. [PubMed]

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Review: Neuroscience, 3,* 201–215. [PubMed]

Craft, E., Schutze, H., Niebur, E., & von der Heydt, R. (2007). A neural model of figure–ground organization. *Journal of Neurophysiology, 97,* 4310–4326. [PubMed] [Article]

Crick, F. (1984). The function of the thalamic reticular complex: The searchlight hypothesis. *Proceedings of the National Academy of Sciences of the United States of America, 81,* 4586–4590. [PubMed] [Article]

Cumming, B. G., & Parker, A. J. (2000). Local disparity not perceived depth is signaled by binocular neurons in cortical area V1 of the macaque. *Journal of Neuroscience, 20,* 4758–4767. [PubMed]

DeAngelis, G. C., Cumming, B. G., & Newsome, W. T. (1998). Cortical area MIT and the perception of stereoscopic depth. *Nature, 394,* 677–680. [PubMed]

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective attention. *Annual Review of Neuroscience, 18,* 193–222. [PubMed]

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 113,* 501–517. [PubMed]

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. *Psychological Review, 96,* 433–458. [PubMed]

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14):18, 1–26, http://journalofvision.org/8/14/18/, doi:10.1167/8.14.18. [PubMed] [Article]

Gegenfurtner, K. R., Kiper, D. C., & Fenstemaker, S. B. (1996). Processing of color, form, and motion in macaque area V2. *Visual Neuroscience, 13,* 161–172. [PubMed]

He, Z. J., & Nakayama, K. (1995). Visual attention to surfaces in three-dimensional space. *Proceedings of the National Academy of Sciences of the United States of America, 92,* 11155–11159. [PubMed] [Article]

Hoffman, J. E. (1998). Visual attention and eye movements. In H. Pashler (Ed.), *Attention* (pp. 119–154). Hove, UK: Psychology Press.

Horwitz, G. D., & Albright, T. D. (2005). Paucity of chromatic linear motion detectors in macaque V1. *Journal of Vision, 5*(6):4, 525–533, http://journalofvision.org/5/6/4/, doi:10.1167/5.6.4. [PubMed] [Article]

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195,* 215–243. [PubMed]

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40,* 1489–1506. [PubMed]

Janssen, P., Vogels, R., Liu, Y., & Orban, G. A. (2003). At least at the level of inferior temporal cortex, the stereo correspondence problem is solved. *Neuron, 37,* 693–701. [PubMed]

Jingling, L., & Zhaoping, L. (2008). Change detection is easier at texture border bars when they are parallel to the border: Evidence for V1 mechanisms of bottom–up salience. *Perception, 37,* 197–206. [PubMed]

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.

Julesz, B. (1981). Textons, the elements of texture perception, and their interactions. *Nature, 290,* 91–97. [PubMed]

Kapadia, M. K., Ito, M., Gilbert, C. D., & 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,* 843–856. [PubMed]

Knierim, J. J., & van Essen, D. C. (1992). Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *Journal of Neurophysiology, 67,* 961–980. [PubMed]

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology, 4,* 219–227. [PubMed]

Koene, A. R., & 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. [PubMed] [Article]

Landy, M. S., & Graham, N. (2004). Visual perception of texture. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 1106–1118). Cambridge, MA: MIT Press.

Leopold, D. A., & Logothetis, N. K. (1996). Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature, 379,* 549–553. [PubMed]

Li, C. Y., & Li, W. (1994). Extensive integration field beyond the classical receptive field of cat's striate cortical neurons—Classification and tuning properties. *Vision Research, 34,* 2337–2355. [PubMed]

Li, Z. (1999a). Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proceedings of the National Academy of Sciences of the United States of America, 96,* 10530–10535. [PubMed]

Li, Z. (1999b). Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex. *Network, 10,* 187–212. [PubMed]

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

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6,* 9–16. [PubMed]

Livingstone, M. S., & Hubel, D. H. (1984). Anatomy and physiology of a color system in the primate visual cortex. *Journal of Neuroscience, 4,* 309–356. [PubMed]

Malik, J., & Perona, P. (1990). Preattentive texture discrimination with early vision mechanisms. *Journal of the Optical Society of America A, Optics and Image Science, 7,* 923–932. [PubMed]

Mannan, S. K., Kennard, C., & Husain, M. (2009). The role of visual salience in directing eye movements in visual object agnosia. *Current Biology, 19,* R247–R248. [PubMed]

Mazza, V., Turatto, M., & Umilta, C. (2005). Foreground-background segmentation and attention: A change blindness study. *Psychological Research, 69,* 201–210. [PubMed]

McCarley, J. S., & He, Z. J. (2001). Sequential priming of 3-D perceptual organization. *Perception & Psychophysics, 63,* 195–208. [PubMed]

Merigan, W. H., Nealey, T. A., & Maunsell, J. H. (1993). Visual effects of lesions of cortical area V2 in macaques. *Journal of Neuroscience, 13,* 3180–3191. [PubMed]

Moore, T., & Armstrong, K. M. (2003). Selective gating of visual signals by microstimulation of frontal cortex. *Nature, 421,* 370–373. [PubMed]

Motter, B. C. (1993). Focal attention produces spatially selective processing in visual cortical areas V1, V2, and V4 in the presence of competing stimuli. *Journal of Neurophysiology, 70,* 909–919. [PubMed]

Nakayama, K., & Mackeben, M. (1989). Sustained and transient components of focal visual attention. *Vision Research, 29,* 1631–1647. [PubMed]

Nakayama, K., & Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions. *Nature, 320,* 264–265. [PubMed]

Nothdurft, H. C. (1991). Texture segmentation and pop-out from orientation contrast. *Vision Research, 31,* 1073–1078. [PubMed]

Nothdurft, H. C., Gallant, J. L., & van Essen, D. C. (1999). Response modulation by texture surround in primate area V1: Correlates of 'popout' under anesthesia. *Visual Neuroscience, 16,* 15–34. [PubMed]

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience, 13,* 4700–4719. [PubMed]

Orban, G. A. (2008). Higher order visual processing in macaque extrastriate cortex. *Physiological Reviews, 88,* 59–89. [PubMed]

Parker, A. J. (2007). Binocular depth perception and the cerebral cortex. *Nature Review: Neuroscience, 8,* 379–391. [PubMed]

Qiu, F. T., Sugihara, T., & von der Heydt, R. (2007). Figure–ground mechanisms provide structure for selective attention. *Nature Neuroscience, 10,* 1492–1499. [PubMed]

Qiu, F. T., & von der Heydt, R. (2005). Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. *Neuron, 47,* 155–166. [PubMed]

Qiu, F. T., & von der Heydt, R. (2007). Neural representation of transparent overlay. *Nature Neuroscience, 10,* 283–284. [PubMed]

Reynolds, J. H., & Desimone, R. (2003). Interacting roles of attention and visual salience in V4. *Neuron, 37,* 853–863. [PubMed]

Sagi, D., & Julesz, B. (1985). "Where" and "what" in vision. *Science, 228,* 1217–1219. [PubMed]

Schiller, P. H. (1998). The neural control of visually guided eye movements. In J. E. Richards (Ed.), *Cognitive neuroscience of attention, a developmental perspective* (pp. 3–50). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Schmolesky, M. T., Wang, Y. C., Hanes, D. P., Thompson, K. G., Leutgeb, S., Schall, J. D., et al. (1998). Signal timing across the macaque visual system. *Journal of Neurophysiology, 79,* 3272–3278. [PubMed]

Sillito, A. M., Grieve, K. L., Jones, H. E., Cudeiro, J., & Davis, J. (1995). Visual cortical mechanisms detecting focal orientation discontinuities. *Nature, 378,* 492–496. [PubMed]

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in Neuroscience, 27,* 161–168. [PubMed]

Tanabe, S., Umeda, K., & Fujita, I. (2004). Rejection of false matches for binocular correspondence in macaque visual cortical area V4. *Journal of Neuroscience, 24,* 8170–8180. [PubMed]

Tehovnik, E. J., Slocum, W. M., & Schiller, P. H. (2003). Saccadic eye movements evoked by microstimulation of striate cortex. *European Journal of Neuroscience, 17,* 870–878. [PubMed]

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12,* 97–136. [PubMed]

Treue, S., & Martinez-Trujillo, J. C. (1999). Feature-based attention influences motion processing gain in macaque visual cortex. *Nature, 399,* 575–579. [PubMed]

van Zoest, W., & Donk, M. (2006). Saccadic target selection as a function of time. *Spatial Vision, 19,* 61–76. [PubMed]

von der Heydt, R., Peterhans, E., & Baumgartner, G. (1984). Illusory contours and cortical neuron responses. *Science, 224,* 1260–1262. [PubMed]

von der Heydt, R., Zhou, H., & Friedman, H. S. (2000). Representation of stereoscopic edges in monkey visual cortex. *Vision Research, 40,* 1955–1967. [PubMed]

Wolfe, J. M., Cave, K. R., & Franzel, S. L. (1989). Guided search: An alternative to the feature integration model of visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 419–433. [PubMed]

Wolfson, S., & Landy, M. S. (1995). "Discrimination of orientation-defined texture edges." *Vision Research, 35,* 2863–2877. [PubMed]

Zhaoping, L. (2002). Pre-attentive segmentation and correspondence in stereo. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 357,* 1877–1883. [PubMed] [Article]

Zhaoping, L. (2005). Border ownership from intracortical interactions in visual area v2. *Neuron, 47,* 143–153. [PubMed]

Zhaoping, L. (2008a). Attention capture by eye of origin singletons even without awareness—A hallmark of 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. [PubMed] [Article]

Zhaoping, L. (2008b). *Eye of origin singletons outcompete the salient orientation singletons for gaze attraction despite their elusiveness to awareness*. Program 770.12, Annual meeting for Society for Neuroscience, November, 2008, Washington, DC. [Article]

Zhaoping, L., & May, K. A. (2007). Psychophysical tests of the hypothesis of a bottom–up saliency map in primary visual cortex. *PLoS Computational Biology, 3,* e62. [PubMed]

Zhaoping, L., & Snowden, R. J. (2006). A theory of a saliency map in primary visual cortex (V1) tested by psychophysics of color-orientation interference in texture segmentation. *Visual Cognition, 14,* 911–933.

Zhou, H., Friedman, H. S., & von der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *Journal of Neuroscience, 20,* 6594–6611. [PubMed]