5 The V1 hypothesis—creating a bottom-up saliency map for preattentive selection and segmentation

In this chapter, we focus on bottom-up visual selection, the second stage in the three-stage process of vision: encoding, selection, and decoding. We describe the hypothesis that V1 creates a bottom-up saliency map to guide visual selection, and we show that this hypothesis can solve certain V1 puzzles which elude the efficient coding principle underpinning the encoding stage. This hypothesis is motivated, formulated in detail, and applied to account for existing behavioral data on visual selection. Non-trivial and surprising predictions that result from the hypothesis are presented, along with their experimental confirmation. A circuit model describing how V1 mechanisms might implement this hypothesis is outlined. Finally, V1's role in selection is discussed in relation to selection by brain areas beyond V1.

5.1 Visual selection and visual saliency

5.1.1 Visual selection—top-down and bottom-up selections

Recall from Section 1.2 that selection and decoding are two visual stages after the visual encoding stage in the three stages of vision. The resources for processing visual input are limited, leading to the attentional bottleneck. Selection enables vision to focus the processing resource, i.e., focus the decoding, on just a fraction of this input. Therefore, selection often makes vision unable to properly decode or perceive the visual input outside our attentional spotlight. Accordingly, the effect of visual selection is measured by the *cueing effect*, which is the improvement in performance and/or speed of visual tasks on the selected input (Posner 1980), and the presence of an cueing effect is often used to demonstrate the availability of visual selection.

Selection is most obviously manifest in the fact that we shift our gaze, or saccade, to the visual locations we select. This overt form of selection is referred to as orienting. Meanwhile, selection can also be done covertly. Responses of neurons in the extrastriate cortex to an attended stimulus are often enhanced relative to their responses to unattended stimuli. This enhancement can result in the responses of neurons being the same as if the unattended stimulus was absent (see Section 2.6).

Reflecting upon our own subjective visual experience, we are more aware of our own goal-directed or voluntary selections, such as when we attend to a book when reading and ignore visual space outside the book page. Goal-directed selection can also be based on prior knowledge or expectation, such as in directing gaze to one's bookshelf when looking for a book, by the knowledge about the location of the bookshelf and an expectation that the book is likely on the shelf. Hence, it is not surprising that most theories or research frameworks have emphasized this goal-directed selection (Treisman and Gelade 1980, Duncan and Humphreys 1989, Tsotsos 1990, Desimone and Duncan 1995), which is also called top-down selection.

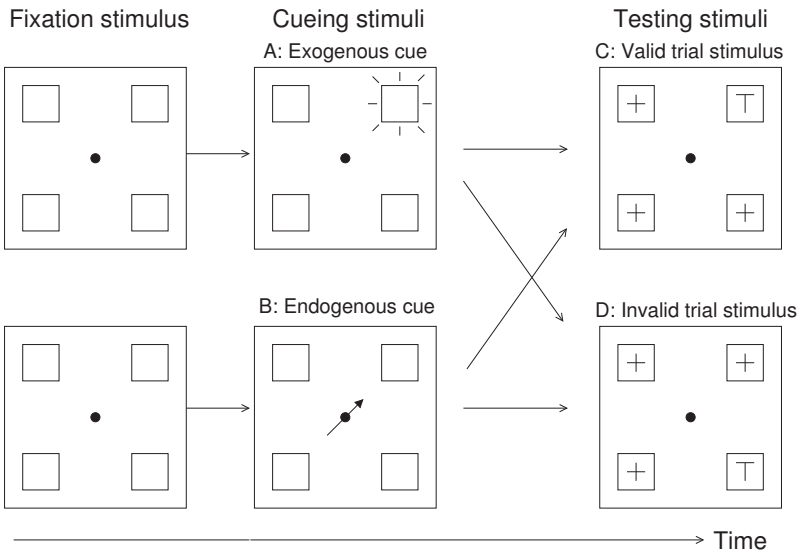


Fig. 5.1: A schematic of an example experimental design used to study top-down, or endogenous, guidance and bottom-up, or exogenous, guidance to attention. This schematic is representative of many other similar studies. Each of the six large square boxes contains a sketch of a stimulus on a fixed-location display. In a test trial, the fixation stimulus, a cueing stimulus, and a testing stimulus are shown consecutively. Observers have to discriminate an aspect of the brief test stimulus, e.g., the orientation of the letter "T" (whether its stem points downwards, upwards, to the left, or to the right). The location of this letter can be at any one of several (e.g., four in this figure) possible positions, unknown to the observer beforehand. This location is cued by a brief exogenous or endogenous cue in the cueing stimulus, which onsets at a time interval called stimulus onset asynchrony (SOA) before the test stimulus onsets. An exogenous cue indicates this location typically by a flash at or near this location; an endogenous cue does this by a symbol (e.g., an arrow pointing to the location) whose position (typically at the fixation point) is independent of the cued position. The actual location of the letter "T" in the test stimulus is at the cued location in a valid trial, or otherwise in an invalid trial. In typical experiments, observers have to keep their gaze fixed at the central fixation point in each stimulus throughout a trial. The exogenous cue has been found to be faster acting, more effective, and harder to ignore, and it can overwrite the endogenous cue (Müller and Rabbitt 1989). Adapted with permission from Müller, H. J. and Rabbitt, P. M., Reflexive and voluntary orienting of visual attention: time course of activation and resistance to interruption, *Journal of Experimental Psychology: Human Perception and Performance*, 15 (2): 315–330, Fig. 1, copyright © 1989, American Psychological Association.

We would be blind to unexpected things if selection was purely top down. There is thus a vital role for an alternative form of visual selection driven directly by visual inputs, or involuntary selection without the influence of task goals. These are called bottom-up selection. In some situations, such as during an emergency, bottom-up selection should be able to overwrite top-down selection, such that, e.g., we should direct our attention to a predator pouncing at us even while reading. In this sense, bottom-up selection serves an ultimate top-down goal of survival. However, this book follows the convention of referring to involuntary selections as being bottom-up. Top-down selection is also referred to as endogenous or reflective, since it is linked with internal goals or knowledge of the viewer. Bottom-up selection

is also referred to as exogenous, since it is driven by the external visual inputs, and is said to be reflexive.

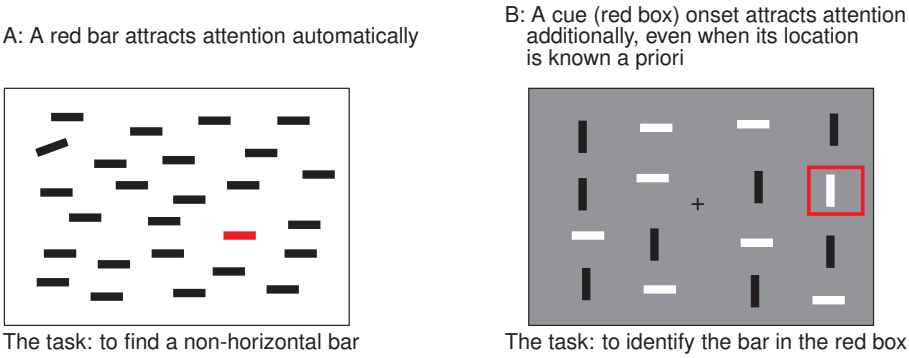


Fig. 5.2: Demonstration of the superior potency (A) and speed (B) of bottom-up over top-down selection. A: even if an observer's task is to find a non-horizontal bar in the image, the red non-target bar automatically distracts attention. B: The task for (human) observers is to keep gaze focused on the center cross, and to report whether the target bar in the red cue box (which is at the same, known, location in the image on each trial) is white vertical, black horizontal, or is like the non target bars (which are white horizontal and black vertical). Target and non-target bars are displayed simultaneously for a short time duration in each trial. Nakayama and Mackeben (1989) measured the shortest display duration necessary for the observers' report to be suitably accurate. This duration was longer when the red cue box remained on display throughout the experimental session to mark target location, compared to the case when the cue appeared about 50–150 ms ahead of the bars in each trial.

Bottom-up selection is often faster acting and more potent than top-down selection (Jonides 1981, Müller and Rabbitt 1989, Nakayama and Mackeben 1989), as one might expect from its role in emergencies. Figure 5.1 shows the kind of experimental setups used to investigate their respective characters. An exogenous cue, typically a brief flash, draws attention reflexively to the flashed location, such that discrimination of a test stimulus presented very soon after the cue is typically faster and more accurate at the flashed location than at another location. An endogenous cue, typically presented as symbols (e.g., an arrow to the northeast) to indicate a likely location of the upcoming test letter, is physically not at the same location as the test letter. It can also make discrimination better and faster at the cued rather than the uncued location. However, when instructed, observers can effectively ignore the endogenous cue, such that their performance is independent of whether the location is cued or uncued; but they are unable to ignore the exogenous cue (Jonides 1981). Furthermore, if the onsets of the cue and test stimuli differ by only a very brief stimulus onset asynchrony (SOA) such as 100–150 ms, the benefit of a valid cue, when the cued and tested locations agree, and the cost of an invalid cue, when the cued and tested locations disagree, are both larger with the exogenous than the endogenous cue. At longer SOAs, the difference between the performances at the cued and uncued locations decreases with SOA for exogenous cues but increases with SOA for endogenous cues, before this difference asymptotes at around SOA = 300–400 ms. Additionally, the performance benefit of a valid, endogenous, cue can be completely eliminated if, after the endogenous cue and about 100 ms before the test stimulus onset, an exogenous flash occurs at another location. This occurs even if this exogenous flash is completely uninformative about the location of the target letter (Müller and Rabbitt 1989).

Figure 5.2 A demonstrates that the red horizontal bar automatically attracts the attention of observers, even if they are intending to look for a non-horizontal bar. Such a distraction by a task-irrelevant salient color singleton is also hard to turn off voluntarily (Theeuwes 1992). Figure 5.2 B shows the stimulus patterns used by Nakayama and Mackeben (1989) to contrast top-down and bottom-up selection. Subjects were asked to discriminate a target bar in a briefly displayed, inhomogenous array of bars. They were unable to recognize the target bar reliably when the array was shown for only 33 ms, even though they knew the target location in the array long before the array appeared. This suggests that vision was too slow, or top-down selection was inadequate for this task, when the stimulus was displayed too briefly. However, task performance dramatically improved if a red box, the exogenous cue, surrounding the target location appeared about 50–150 ms before the array appeared. This bottom-up attraction by the red box, in addition to the top-down attraction due to the prior knowledge of the target location, caused a marked improvement in task performance. Without this bottom-up attraction, the array had to be shown for a much longer time in order to improve the task performance. This again suggests that, in this task, when bottom-up attention is available, vision is not too slow and the bottom-up attraction is faster acting.

Top-down and bottom-up attentional selection are often space-based, such that only visual input at or near a spatial location is selected. Selected locations are metaphorically referred to as being in the attentional spotlight. This also implies that the spotlight is spatially compact and has a finite size. Recall from Section 2.5 that spatial selection is closely linked with eye movements, such that the selected location typically coincides with the current gaze position or the destination of an impending saccade. Since our gaze can only be at one location at a time, it is not surprising that it is either impossible or very difficult to select two disjoint locations simultaneously (Cave and Bichot 1999), even for covert attention.

Selection can also be based on features of the input. Consider looking for a red cup. Selection can be based on a particular value of the feature of color, so that sensitivity to red-colored objects in the whole visual field is relatively enhanced. Obviously, feature-based selection is goal directed and is thus top down. Selection can also be object-based. For example, consider two visual locations in an image that are equally distant from the current gaze position, which is inside the image area of an object. If only one of the two locations is inside the image area of the same object, then sensitivity to inputs at this location is relatively higher. Since object-based selection relies on the perception of the object, knowledge about the object's shape or identity can influence the selection. Therefore, it is likely that object-based selection is not strictly bottom up.

To understand selection as a whole, we must clearly understand bottom-up selection, both because of its own potency and because of its competition and cooperation with top-down selection. This book focuses mostly on bottom-up (spatial) selection, mainly because more is known about its neural mechanisms. The book comparatively ignores feature-based or object-based attention.

5.1.1.1 Terminology: selection, attention, saliency, and priority

Voluntary or top-down visual selection is often colloquially referred to as "paying attention." When bottom-up selection overrides top-down selection, it is often said that attention has been distracted to a task-irrelevant input. In this colloquial sense, the word "attention" is understood as some sort of resource, which is applied to or spent on the selected input, enabling this input to be recognized or decoded. Hence, the term "preattentive" is understood as to refer to the processing stage before the resource of "attention" is applied, although exogenous selection process can be operative at the preattentive stage. Meanwhile, directing attention to somewhere or to something is also colloquially referred to as attending somewhere or something. In this sense, "attending" and "paying attention" refer to the act of selection,

regardless of whether this is by top-down or bottom-up means. In this book, we will use the term "selection," "selecting," or "to select" to mean the act rather than the resource, to avoid the confusion.

In this book, we define the *saliency* of a visual location as the degree to which this location attracts selection by bottom-up mechanisms only. A location with a large saliency value is said to be salient. Following Egeth and Yantis (1997), the term *priority* is used to describe the degree to which this location attracts selection as a result of combining top-down and bottom-up mechanisms. A saliency map is a map of saliency values of the visual field, while a priority map is a map of priority values. Behaviorally, selection follows the priority map, such that attention or gaze is more likely to be directed first to locations of higher priorities. The temporal order in which spatial locations are selected should follow the order of their priorities, deterministically or stochastically, such that, when a scene is viewed, a location having a higher priority is more likely selected before a location having a lower priority. Often, a location is often more likely selected, i.e., having a higher priority, because it is closer to the currently attended location. This is likely caused by top-down rather than bottom-up factors.

5.1.1.2 Probing bottom-up saliency behaviorally even though selection is controlled by priority

According to the definitions above, a location's priority can often be assessed behaviorally by the reaction time (RT) associated with finding or identifying a target at its location. Alternatively, it can be assessed by measuring how well observers discriminate or identify a visual target at the location, given a fixed viewing duration. This latter measurement is called the accuracy, expressed as the probability that the task is performed correctly. Accuracy should increase with the amount of time the target spends in the attentional spotlight. Therefore, given a fixed viewing duration, greater accuracies should be coupled to shorter RTs for selecting the target location. Although saliency is only one of the contributing factors to priority, it can be studied in terms of the difference between RTs or accuracies for different tasks. For example, studies of visual search often assume that a shorter RT indicates a larger saliency at the location of the search target. Below we explain why and when this assumption is approximately valid.

Let RT_{task} be the behaviorally measured RT in a task, e.g., to find a certain target. Imagine an ideal world in which we could measure $RT_{\text{selection}}$, which is the RT to select the task relevant location, e.g., the location of the target. Let us define

$$RT_{\text{other}} \equiv RT_{\text{task}} - RT_{\text{selection}}; \quad RT_{\text{other}} \text{ is understood as the RT required for all the non-selection processes to complete the task.} \tag{5.1}$$

For example, let visual task A be to look for a target bar tilted 20^o anticlockwise from horizontal, among many non-target bars uniformly tilted 20^o clockwise from horizontal. Experimentally, we measure RT_{task} as the time from the onset of the visual stimulus to the time when observers report the location of the target. This RT_{task} contains $RT_{\text{selection}}$ to put the target bar in the attentional spotlight, and RT_{other} to confirm that the bar in the spotlight is indeed the sought-after target and to execute the motor action to report the target's location. Note that when the absolute value of $RT_{\text{selection}}$ is difficult to measure, then so is RT_{other} since it is defined by $RT_{\text{task}} - RT_{\text{selection}}$. However, this should not affect our argument.

We can extract saliency from RT_{saliency}, defined as $RT_{\text{saliency}} \equiv RT_{\text{selection}}$ when the top-down contribution to selection is set to zero. However, eliminating all top-down contributions is impossible in typical behavioral experiments. In general, define

$$RT_{\text{top-down selection}} \equiv RT_{\text{selection}} - RT_{\text{saliency}}. \tag{5.2}$$

The term $RT_{\text{top-down selection}}$ could be negative or positive, depending on whether the top-down and bottom-up selection cooperate or compete. Putting the above together, we have

$$\begin{aligned} &RT_{\text{task}} = RT_{\text{saliency}} + RT_{\text{top-down selection}} + RT_{\text{other}}; \\ &\text{therefore, } RT_{\text{task}} \text{ can be a proxy for } RT_{\text{saliency}} \text{ when} \\ &\qquad RT_{\text{top-down selection}} + RT_{\text{other}} = \text{constant for different tasks.} \end{aligned} \tag{5.3}$$

One can design experiments to ensure that $RT_{\text{top-down selection}} + RT_{\text{other}}$ is a constant. For example, suppose that along with task A above, subjects execute task B, which requires finding the same target bar (tilted 20^o anticlockwise from horizontal) but among non-target bars that are horizontal rather than being tilted 20^o clockwise from horizontal. Provided that tasks A and B do not differ in other aspects (non-target numerosity, item sizes, etc.), it is reasonable to assume that the two tasks have approximately the same RT_{other}, since the target in the attentional spotlight is equally distinguishable from non-targets, and since the time it takes to report the target (after the target is recognized in the attentional spotlight) is unlikely to depend on the task. Since the target is the same, feature-based attention should also be set the same way, suggesting that $RT_{\text{top-down selection}}$ should also be the same. In total, if RT_{task} is shorter for tasks A than task B, then the target location can be considered to be more salient in task A, and vice versa.

Next, consider a slightly modified experiment in which the target feature is unknown ahead of time, e.g., when the task is simply to find a uniquely oriented bar in the image without specifying the orientation of this unique target bar. In this case, the top-down contribution will be more limited but will still be the same for the two tasks. In other experimental designs, the bottom-up saliency can be so strong that attention can be attracted to the target location automatically whether or not the target identity is known ahead of time, so that the top-down contribution to selection is negligible. Altogether, $RT_{\text{top-down selection}} + RT_{\text{other}}$ can be approximately independent of the task in many different situations. This can even be approximately true when the two tasks differ in both target and non-target identities.

Under the assumption that, among multiple tasks in a study, tasks with shorter RT_{task}s have more salient target locations, we can study how saliency is determined by input stimuli. This assumption is the basis for many behavioral and modeling studies into saliency, including those described in this book. One should nevertheless be wary of violations of the assumption, as sometimes (Zhaoping and Guyader 2007) a difference in RT_{task}s between tasks results from top-down rather than bottom-up effects.

Similar considerations and arguments apply when bottom-up saliency of a location is assessed by the accuracy of input discrimination at that location for sufficiently brief viewing durations.

Various features of experimental designs can minimize the impact of factors other than saliency in the measured RTs and accuracies. One is to minimize any a priori knowledge as to the possible positions and features of the visual inputs. Another is to make the visual input at a location task-irrelevant, to remove or minimize the top-down factors in selecting this location.

One can also make the input presentation very brief, such that there is insufficient time for more than one shift of attention, or one glimpse. Compared with this first shift, any second shift of visual attention is more influenced by the visual information gained during the initial glimpse, and such information can drive top-down influences. For example, take the case of asking an observer to find an apple in an image quickly but without revealing ahead of time what kind of image will be presented. When a picture of a kitchen scene is suddenly shown, the first gaze shift is likely to be governed by reflexes before the observer realizes what kind of scene is on show. However, the duration from the appearance of the image to the second gaze shift is often sufficient for the observer to realize the overall content or gist of the scene, and thence to exploit the top-down expectation that an apple is more likely to be on a kitchen counter than a kitchen floor. In turn, this can influence the second gaze shift. Conversely, a picture of a forest would direct the second gaze shift according to a different top-down factor.

Analogously, the effect of saliency is better manifested in shorter reaction times. A longer latency to respond after a brief display allows more top-down influences.

One should also design the task to minimize the contamination of the measured behavioral outcome by high level strategic factors. For example, in many conventional visual search tasks (Wolfe 1998), observers are asked to report the presence or absence of a target in an image. When observers do not find the target after an initial scan of the search display, they may decide to search further for the target, or report that the target is absent. Hence, the RT of the report is affected by the strategic decision of when to terminate the search, a strategic decision that may be task dependent. To avoid this, it helps to make the target present in all search trials and ask the observers to simply report an aspect of the target location, e.g., whether the target is in the left or right half of the visual display.

5.1.2 A brief overview of visual search and segmentation—behavioral studies of saliency

Studies of visual search are often used to examine how the saliency of a target location depends on the characteristics of the visual inputs, assuming that a shorter RT indicates a larger saliency at the location of the search target. In these experiments, human observers are asked to search for a target as quickly as possible, and their reaction time (RT) to report the target is recorded.

Visual search has been the target of extensive behavioral studies (Treisman and Gelade 1980, Duncan and Humphreys 1989, Julesz 1981, Wolfe, Cave and Franzel 1989, Wolfe 1998), and is introduced briefly in Fig. 5.3. It has been found, for instance, that if the target is characterized by a visual feature such as color or orientation that is sufficiently unique within a visual image or scene, the RT for finding it is often insensitive to the number of non-target items (distractors) in the scene. Visual search for which the target differs from all the non-targets in one unique feature is called *feature search*. Figure 5.3 E shows an example in which the target differs from the distractor not by a single feature but by a conjunction of two features: red and vertical. Each of these is present in the non-target items (which are blue-vertical or red-horizontal). Such a search is called a *conjunction search*, and it is usually more difficult than a feature search. RTs in conjunction searches usually grow quickly with the number of distractors. The total number of search items, target plus the distractors, is called the *set size* of the search. One can imagine that if a target were defined by a conjunction of more than two features, the search would be even more difficult.

Visual search in which the RT is almost independent of the set size is called *efficient*. By contrast, if RT increases with set size sufficiently quickly, the search is called *inefficient*. Efficient and inefficient searches have often been interpreted to depend on underlying neural processes that are, respectively, parallel and serial. Visual inputs in the scene are processed all at once for an efficient search and chunk by chunk for an inefficient search.

Efficiency in visual search has been used as an empirical definition of the notion of a basic feature dimension in visual input (Wolfe 1998). Accordingly, color, orientation, motion direction, stereoscopic depth, and sizes have been found to be among the basic feature dimensions, since a target that is sufficiently different from homogenous distractors in any one of these dimensions (e.g., Fig. 5.3 AD) can be found efficiently. Indeed, this has been used to define feature search in cases in which the target differs from the distractors by its feature value in a basic feature dimension. Feature searches are efficient if the non-targets have the same feature value in the basic feature dimension that distinguishes the target; they can be inefficient if their feature values in this dimension are sufficiently varied.

Often, a salient target in an efficient visual search is said to pop out of the scene *preattentively* or to require only preattentive mechanisms to be noticed by human observers. Here

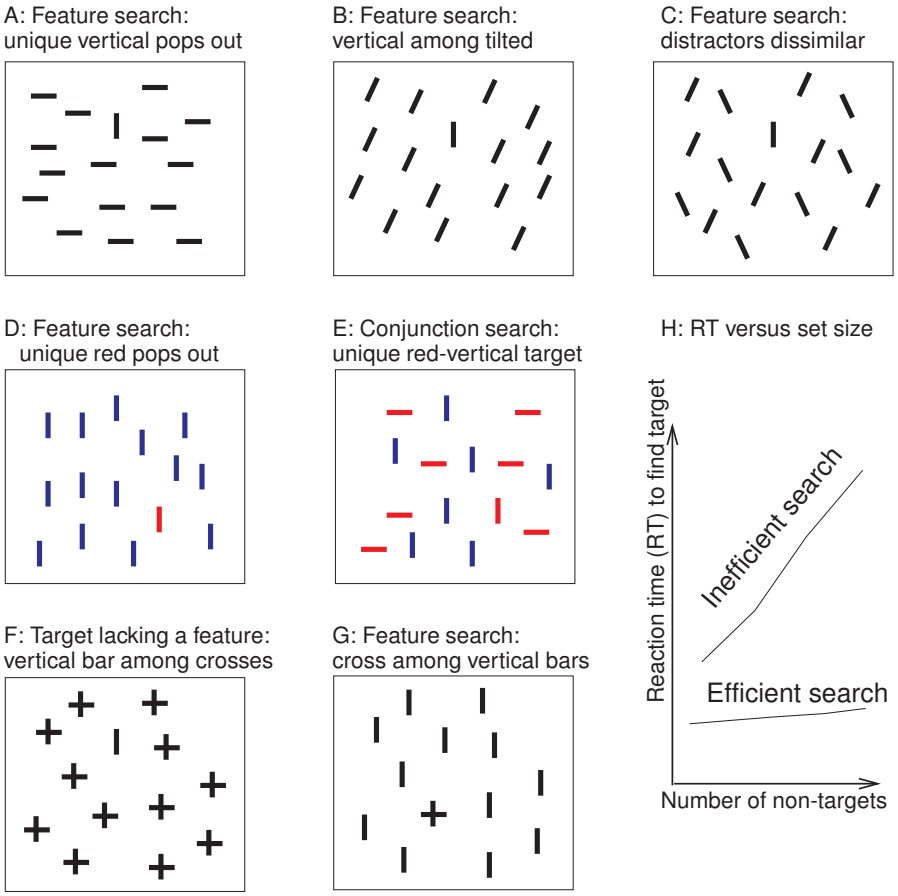


Fig. 5.3: A brief overview of visual search. A–G: Illustrative examples of visual search. The search target is a vertical bar in A–C and F, a red-vertical bar in D and E, and a cross in G. A–D and G are examples of feature search, when the target has a feature that is absent in the non-targets. E is an example of conjunction search when the target is defined by a unique conjunction of features, each of which is individually present in the non-targets. In F, the target is defined by the absence of a feature that is present in the non-targets. F and G together illustrate the asymmetry of visual search, when the ease of search changes by swapping identities of the targets and non-targets. H: Characteristics of efficient and inefficient searches in terms of how RTs depend on the number of non-target (distractor) items.

the term preattentive can be understood as without top-down attentional guidance. Hence, preattentive attentional guidance is by bottom-up attractions only. By contrast, an inefficient search requires something more than preattentive mechanisms. However, the meanings or definitions of these terms (e.g., pop-out, preattentive, etc.) may differ in the literature according to different research communities.

Efficiency in visual search can be affected by many factors, and there is a continuum rather than two discrete categories (efficient/parallel and inefficient/serial). Figure 5.3 AB demonstrate that searching for the vertical target is easier when the feature contrast (orientation contrast) between the target and distractors is larger; Fig. 5.3 BC demonstrate that search becomes more difficult when the distractors are not identical to each other (Duncan and

Humphreys 1989), even though the target's feature is unique in both examples. Figure 5.3 FG show a simple example of visual search asymmetry, in which the ease of search can change when the target and distractor swap identity. Figure 5.3 F is an example for which the target is more difficult to find when it is defined by the absence of a basic feature that is present in the distractors.

Further, efficiency, defined as the insensitivity of the search RT to the set size of the search, is insufficient to describe the ease of a search by itself. Some searches can require longish RTs even though these RTs are insensitive to the search set size.

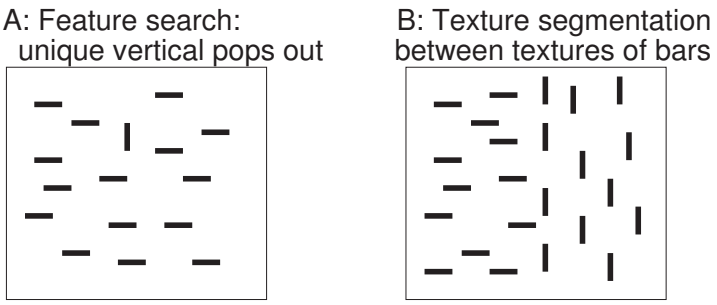


Fig. 5.4: Demonstration that visual search and segmentation are typically related. A: A vertical bar pops out among horizontal bars. B: A texture of vertical bars is readily segmented from a texture of horizontal bars.

Visual saliency is also manifest in texture segmentation behavior (Julesz 1981). Texture segmentation becomes easier when the border between two texture regions is more salient (Nothdurft 1991, Li 1999b, Li 2000b).

When a unique target pops out of non-targets in a visual search display, one texture region made of many of these target items is also typically easy to segment from another texture region made of the non-target items, as demonstrated in Fig. 5.4. This link between visual search and segmentation will be elaborated further in the chapter.

5.1.3 Saliency regardless of visual input features

One can compare the saliency induced by different input features such as color and orientation. For example, in Fig. 5.2 A, both the location of the red bar and the location of the non-horizontal bar are salient. If observers freely view this image in a task in which they do not have a preference for either feature, one can see which location attracts attention more strongly. Phenomenologically, it is as if there is a saliency map of the visual space, such that locations with higher saliency values in this map are more likely to attract attention, regardless of the visual input features that make those locations salient. In other words, once feature distinctions are converted into saliency values, or a raw visual image is turned into a saliency map, the original image feature values that caused these different saliencies are irrelevant as far as bottom-up attraction to attention is concerned. This is true even if the rules by which different features (or feature combinations) are converted into saliency values differ. In fact, this feature irrelevance is (implicitly) part of the definition of saliency, since the concept makes no reference to the feature dimensions or values that determine its values.

That saliency values are independent of input features may be a reason why traditional models (Koch and Ullman 1985, Wolfe et al. 1989, Itti and Koch 2000) compute saliency from visual inputs according to a framework which can be paraphrased as follows (Fig. 5.5 A).

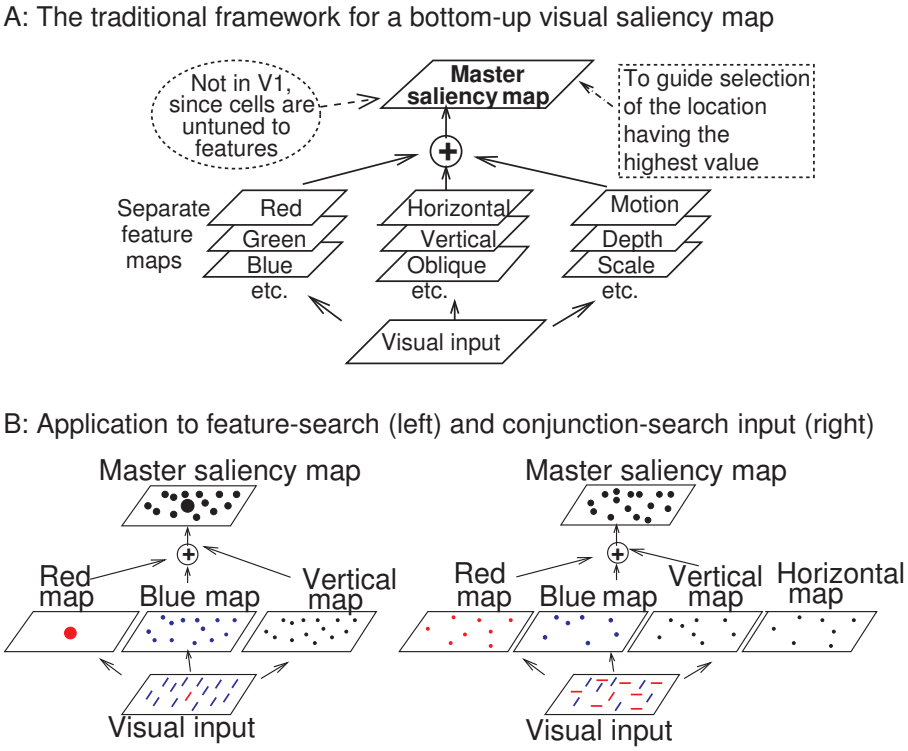


Fig. 5.5: A: Schematic of the framework for traditional models of visual saliency. This framework implies that a saliency map should be in a brain area (such as the lateral intraparietal area (LIP) (Gottlieb et al. 1998)) where cells are untuned to visual features. B: Application of this framework on feature-search (left) and conjunction-search (right) stimuli. Only the relevant feature maps are shown, and the activations in each feature map are higher when there are fewer items in that map. The master map has a hot spot at the location of the red bar in the image for the feature search to attract selection, but it has no hot spot for the conjunction search image. Adapted with permission from Zhaoping, L., Theoretical understanding of the early visual processes by data compression and data selection, *Network: Computation in Neural Systems*, 17(4): 301–334, Fig. 12, copyright © 2006, Informa Healthcare.

Visual inputs are analyzed by separate feature maps, e.g., red feature map, green feature map, vertical, horizontal, left tilt, and right tilt feature maps, etc., in several basic feature dimensions such as color, orientation, and motion direction. The activation of the unit representing an input at a particular location in its corresponding feature map decreases roughly with the number of the neighboring input items sharing the same feature value. Hence, in an image of a red bar among blue bars, as in the left example of Fig. 5.5 B, the red bar evokes a higher activation in the red map than that of each of the many blue bars in the blue feature map. Then, the saliency value for a location in the master map is the summation of the activations in all the separate feature maps associated with that location. The summation implies that the saliency values in the master map generalize across the actual input features. In the master saliency map for the left example of Fig. 5.5 B, the red bar evokes the highest activation at its location and attracts visual selection. By contrast, a unique red-vertical bar, among red-horizontal and blue-vertical bars, does not evoke a higher activation in any single feature map, red, blue, vertical, or horizontal, and thus not in the master map either.

The traditional framework provides a good phenomenological model of the saliency implied by behavior in feature and conjunctive searches. It has subsequently been made more explicit and implemented in computer algorithms (Itti and Koch 2000). It does not explicitly specify the neural mechanisms or the cortical area(s) underlying the feature and master maps. However, it implies that the master saliency map should be in a cortical area where neurons are not tuned to visual feature values, since combining the feature maps eliminates the feature selectivity in the master map. The LIP (lateral intraparietal area) or FEF (frontal eye field) could be candidates for this master map, since their neural activities are untuned to input features. This implication of feature irrelevance has had an obvious impact on the directions of experimental investigations—for many years, few experiments looked for the neural substrates of the saliency map in early visual areas, where neurons are feature selective.

Contrary to intuition, the fact that saliency is independent of particular input features does not mean that the cells reporting saliency *must* be untuned to input features. If saliencies are signaled by the activities or firing rates of neurons, then "signaling independent of input features" can simply mean that the neural firing rates associated with saliency are universal, independent of the neurons' feature preferences. For example, if one neuron prefers red color and another prefers vertical orientation, then the two neurons signal the same saliency value if they have the same firing rate, and the more active neuron signals a higher saliency than the less active neuron, *regardless* of their (different) feature preferences. This is just like the value of a pound sterling of English currency is independent of the race or gender of the currency holder.

In principle, according to this idea, V1 could contain the saliency map, since V1 neurons can use their activities to signal saliencies at the locations of their receptive fields, despite their respective feature preferences. This does not mean that the selectivities of the V1 neurons to features are useless for visual computation beyond saliency. For instance, these selectivities can be used for decoding visual inputs in a visual area downstream along the visual pathway. However, whether or not they are used as part of other computations should not be relevant for saliency signaling (Li 2002). Meanwhile, the saliency values represented in the firing rates of the saliency signaling neurons can be read out to execute visual selection, such as to execute a gaze shift to the most salient location. The neurons for saliency readout and execution, perhaps in the superior colliculus, can be untuned to features, but they are separate from the neurons computing and representing saliencies in the saliency map (we discuss this in more detail in Section 5.2.3).

The traditional framework, which uses separate feature maps to sum into a master saliency, also imposes an unnecessary and unjustified rule on the interaction between features for saliency. This is illustrated by the following example. Let there be two visual inputs containing bars, all of them are colored blue. One input contains a unique vertical (blue) bar among horizontal (blue) bars. The other input contains a unique rightward moving (blue) bar among leftward moving (blue) bars. Hence, there is an orientation singleton in the first input and a motion singleton in the other input. Let these two inputs be such that their respective master saliency maps by the traditional framework are identical to each other. In this saliency map, the singleton's location is the most salient location. Now let us change the color of the singleton in each input to red, with exactly the same red feature value for the two inputs. Hence, the original orientation and motion singletons are now, respectively, color-orientation and color-motion double-feature singletons. The traditional saliency framework predicts: (1) the saliency of each double-feature singleton is larger than that of the corresponding single-feature singleton; and, (2) the saliency increase from the single-feature to the corresponding double-feature singleton is the same in the two inputs, independent of the feature dimension which defines the original single-feature singleton. Prediction (1) arises from the feature summation rule; and prediction (2) arises from the separation of feature maps in making the

master saliency map. Although the summation rule and the separation of feature maps seem natural, and sufficient, for achieving the property of saliency regardless of features, they are both unnecessary for this property. It will be shown in Section 5.5.3 that both predictions (1) and (2), which are consequences of the summation rule and the separation of the feature maps, are inconsistent with experimental data.

In contrast, when saliencies are signaled by the activities or firing rates of feature-tuned neurons, and when some neurons are tuned to more than one feature dimension (such as those in V1), there is no separation of feature maps in making the saliency map. This allows rich interactions between various features for saliency, when these neural activities are read out for their universal saliency values and for the visual locations (but not features) they represent. In this sense, V1 could hold the saliency map, rather than many coexisting feature maps in separate subpopulations of neurons; and the SC could be a saliency read-out area, rather than a master saliency map to combine various feature maps. This will be detailed in Section 5.2.

5.1.4 A quick review of what we should expect about saliencies and a saliency map

Before we proceed further, it is worth reviewing and drawing some conclusions from the definition and expected manifestations of saliency. To recap, the saliency of a visual location is defined as the degree in which this location attracts visual selection in a purely bottom-up manner, such that a location having a higher saliency should be more likely to attract bottom-up selection before rather than after another location having a lower saliency. Below we elaborate this definition somewhat, list a few immediate consequences, and discuss some implications.

1. The saliency of a location increases with the probability with which this location is selected bottom-up before the other locations in the scene. It is inversely related, either deterministically or stochastically, to the order in which this location is selected by bottom-up manner mechanisms in the scene.
2. Saliency should be context dependent, since saliency at a visual location is associated with, and is defined in the context of, the whole visual scene. Hence, the location of a red apple may be the most salient in one scene full of green leaves, but the same location and the same apple in another scene made of many other red apples is not salient. In particular, the RT for gaze to reach this red apple should be much shorter in the first scene, unless there is a strong top-down influence.
3. In typical behavioral settings, visual selection depends on both the saliencies of visual locations and top-down factors. Thus, empirical selection approaches selection by saliency only in the asymptotic limit at which top-down factors are eliminated. This asymptotic limit is an ideal, which is difficult to reach experimentally. This is because observers who are behaving consciously inevitably have internal top-down goals that influence selection, and because viewing a visual input typically triggers awareness and internal knowledge of the scene, which also affect selection in a top-down, knowledge-driven, manner. Nevertheless, we can still compare saliencies between visual locations (or visual inputs) when the top-down factors that influence selection are made equal. Furthermore, we can minimize top-down contributions to selection by minimizing expectation, knowledge, or awareness of the visual inputs, and we can also shorten input viewing duration to avoid visual knowledge being triggered or having time to exert its top-down effects on selection. In particular (as we will see later), saliency can be well manifested when it works against the top-down factors.

4. Even if the transformation rule from visual inputs to saliency values depends on the visual input features concerned, the rule to transform the responses of the saliency signaling neurons to saliency values, by definition, cannot depend on visual input features.
5. Any visual location within the visual field should have a saliency value. Therefore, a saliency map should cover the whole visual field. Consequently, one expects that brain area(s) computing and reporting the saliency map must be able to respond to the whole visual field.
6. Since attention shifts typically change the selected location from the center of the visual field to an eccentric location and since the center of the visual field is often associated with the current top-down goal of the observer, visual saliency should be mainly, or at least more strongly, operative at more eccentric locations away from the center of the visual field.

The above points will be further elaborated in this chapter.

5.2 The V1 saliency hypothesis

The V1 saliency hypothesis was originally proposed in the 1990s, and was elaborated over several following years (Li 1997, Li 1999a, Li 1999b, Li 2002, Zhaoping 2005b). It states that V1 creates a bottom-up saliency map of visual space such that, first, the saliency of a location is represented by the highest of the responses of the V1 neurons whose receptive fields cover that location; and second, the receptive field location of the most active V1 cell in response to a visual scene is the most salient location in the scene.

Usually, a small image feature evokes responses from many V1 cells which have overlapping classical receptive fields (CRFs) and may have different feature preferences. For example, a short, red vertical bar can excite a neuron preferring the color red, another neuron preferring vertical orientation, a third neuron preferring both the color red and vertical orientation, and a fourth neuron whose most preferred orientation is 5 degrees from vertical, and so on. The receptive fields of all these neurons include the location of the bar. According to the V1 saliency hypothesis, the saliency at this bar's location is represented by the response of the fastest firing neuron (i.e., the neuron with the highest response) with a receptive field covering this location in response to the image containing the bar, regardless of the feature preference of the neuron having this response. Locating the cell that is most responsive to a scene overall locates the most salient location. Saliency does not depend on extraneous processing, such as whether input features are decoded beforehand, simultaneously, or never, from the responses of the same cell population (potentially in a complex and feature-specific manner from the population responses (Dayan and Abbott 2001); decoding will be discussed in Chapter 6).

It is not economical to use cortical areas beyond V1 along the visual pathway to realize a saliency map, whether or not the cells in those areas are feature independent. That V1 cells have small CRFs implies that the spatial resolution of a V1-based saliency map can be better than a map based anywhere downstream. Furthermore, since V1 is at an early stage on the visual pathway, saliency can be signaled quickly. High spatial resolution and alacrity are both desirable for bottom-up visual selection.

It may come as a surprise to many experienced vision researchers, who are familiar with the traditional framework of saliency (Fig. 5.5) and its implications, that V1's activities could signal saliency. It has been known since the 1960s that V1 neurons are tuned to local visual features like orientation, color, motion direction, binocular disparity (Hubel and Wiesel 1968), and input scales (see Chapter 2). It was not obvious that V1 neurons could signal salience,

which depends on global context—after all, a vertical bar is salient in the context of horizontal bars, but the same vertical bar is not salient among other vertical bars. Until recently, V1 had never been looked at as playing an essential role in computing saliency.

Figure 5.6 uses the metaphor of an auction to help explain this extended role of V1. An auction shop has the slogan "Attention auctioned here; no discrimination between your feature preferences; only spikes count;" three V1 neuron bidders are depicted, with one tuned to motion direction with one spike's worth of bidding "money," another tuned to red color with 3 spikes' worth, and the third one tuned to a tilted orientation with 2 spikes' worth. The auctioneer, although feature blind, can do his job perfectly provided that he can count the spike "money." Of course, a feature-blind auctioneer does not mean that the "attention" awarded to the highest bidder is feature blind—post-selectional decoding should recognize the features at the selected visual location. The superior colliculus (introduced in Section 2.5) could possibly play the role of the auctioneer—it receives monosynaptic inputs from V1; its neurons are poorly or not feature-tuned, but their receptive fields are retinotopically organized; and it directs eye movements, which can be seen as the ultimate manifestation of the attention that is awarded to the receptive field location of the highest bidder.

This metaphor also conveys an important aspect of the V1 saliency hypothesis: attention does not have a fixed price—it is just that the highest bidder wins. A given level of neural activity may signal the most salient location in one scene, when it is the highest among the responses of the population of V1 neurons, but the same activity level may signal only a mediocre saliency in another input scene, when it is only typical among the responses of the population. Hence, it is not sufficient to record the activity of a single V1 neuron to determine saliency; measurements across the neural population are required to determine whether one neuron signals the most salient location.

5.2.1 Detailed formulation of the V1 saliency hypothesis

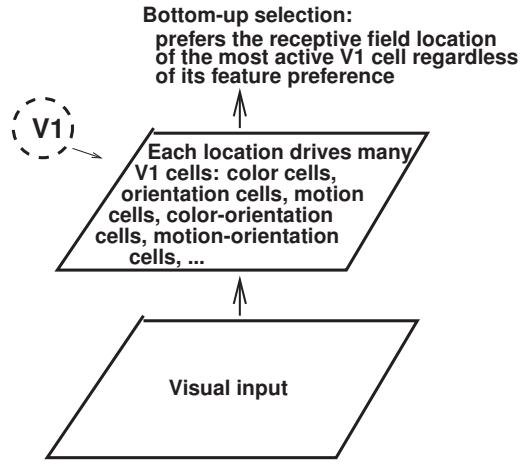
A location with a larger scalar saliency value in the saliency map is more likely to be selected by bottom-up attentional mechanisms for further visual processing. Here, by selection, we always mean selection by bottom-up mechanisms only. According to the V1 saliency hypothesis, saliency values are represented by the firing rates of V1 neurons. Let $(x_1, x_2, ..., x_n)$ denote the centers of the RF locations of the V1 cells with responses $(O_1, O_2, ..., O_n)$. Given a location x, let $x_i \approx x$ mean that the receptive field of the neuron with response O_i covers location x. Let

$$\begin{aligned} \text{SMAP}(x) \equiv \max_{x_i \approx x} O_i, \quad & \text{the highest response to } x; \\ & \text{then SMAP}(x) \text{ as a function of } x \text{ is the saliency proto-map.} \end{aligned} \tag{5.4}$$

The SMAP(x) value in this saliency proto-map is the value that location x bids for selection in the sense of the attentional auction described in Fig. 5.6. Therefore, this proto-map completely determines, and can be translated into, the saliency map through the read-out of the auction. This makes the following hold.

1. If two scenes generate identical saliency proto-maps, then their saliency maps are identical.
2. Within a scene, if SMAP(x) >SMAP(x'), then the saliency at location x is higher than that at location x'.
3. If the saliency proto-map values for two scenes are identical to each other at all locations except location x, then this location is more salient in the scene with the larger proto-saliency value SMAP(x).
4. The third point above is a special case of the following. Saliency at location x increases with the degree in which SMAP(x) is relatively higher than SMAP(x') at other locations

A: The theory of a bottom-up saliency map from V1



B: Its cartoon interpretation

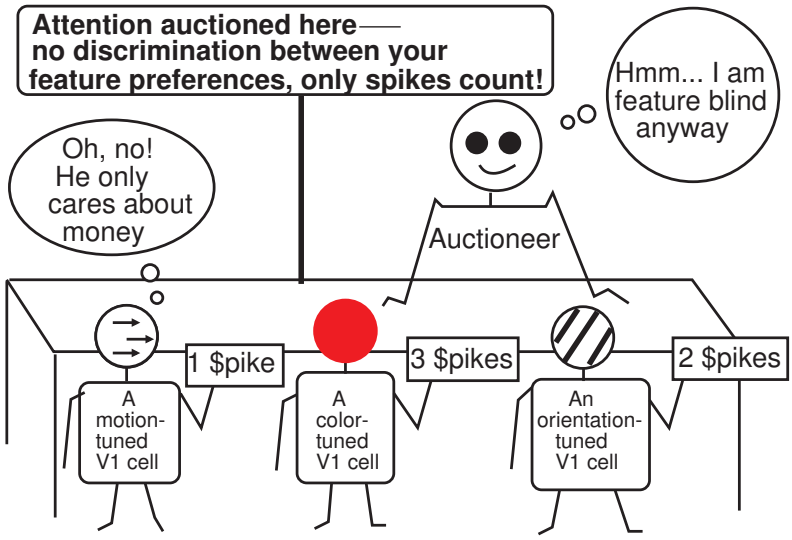


Fig. 5.6: A schematic summary and a cartoon interpretation of the V1 saliency hypothesis. In contrast with previous accounts, no separate feature maps, nor any summation of them, is needed in the V1 theory. V1 cells signal saliency despite their feature tuning. Adapted with permission from Zhaoping, L., Theoretical understanding of the early visual processes by data compression and data selection, *Network: Computation in Neural Systems*, 17(4): 301–334, Fig. 12, copyright © 2006, Informa Healthcare.

x'. Hence, for example, if $\bar{S}$ and σ_s are the mean and standard deviation of the proto-saliency SMAP(x) across space, a location x tends to be more salient when SMAP$(x)/\bar{S}$ and $(\text{SMAP}(x) - \bar{S})/\sigma_s$ are larger.

Since the saliency proto-map completely defines the saliency map through the attentional auction, from here on, we will refer to the saliency proto-map SMAP$(.)$ as the saliency map and SMAP(x) as the saliency value for location x, where it is not necessary to draw a

distinction between the saliency map and the saliency proto-map. We will see that this notion of a saliency map leads to non-trivial, and experimentally testable, qualitative and quantitative predictions.

The most salient location in the scene is the location with the highest saliency value in the saliency map:

$$\text{the most salient location } \hat{x} = \text{argmax}_x \left[\text{SMAP}(x)\right]. \tag{5.5}$$

The most salient location can also be identified as the RF location of the most active V1 cell, i.e.,

$$\text{the most salient location } \hat{x} = x_{\hat{i}}, \quad \text{where } \hat{i} = \text{argmax}_i O_i. \tag{5.6}$$

One might worry that the most salient locations defined by the two equations above are not precisely the same. For example, let the response $O_1 > O_{i\neq 1}$ of the first neuron be the highest among all V1 neural responses, and let the receptive field of this neuron cover a circle of one degree in diameter centered at location x_1. From equation (5.6), $\hat{x} = x_1$. Meanwhile, from equation (5.4), the saliency value SMAP(x) will be the same for all locations x within that one degree diameter circle centered at x_1. Then by equation (5.5), the most salient location includes all locations x within this circle, rather than just its center location x_1. This inconsistency however merely defines the spatial resolution of the saliency map. For the main functional role of saliency, which is to specify how attention should shift using saccades (Hoffman 1998), this resolution, as defined by the sizes of the V1 receptive fields, is adequate. In particular, it is no larger than the typical size of a saccadic error, which is the discrepancy between the target of a saccade and the actual gaze location brought by this saccade (Becker 1991). Note that the sizes of the V1 receptive fields scale with the eccentricity, and the saccadic error is about 10% of the eccentricity of the saccadic target (Becker 1991).

What we currently know does not determine whether "the V1 cells" that the V1 saliency hypothesis suggests to participate in the attentional auction include all cells in this area or just a subpopulation. Certainly, it is unlikely that these cells include the inhibitory interneurons in V1. However, "the V1 cells" should cover the whole visual field. For simplicity, in this book, we will not distinguish between a putative subpopulation and the other V1 cells.

5.2.2 Intracortical interactions in V1 as mechanisms to compute saliency

It has been suggested that intracortical interactions between V1 neurons are the neural mechanism by which saliency is computed. As seen in Section 2.3.9, the response of a V1 neuron to visual inputs in its classical receptive field (CRF) can be influenced by contextual inputs from outside this CRF. The intracortical neural connections, which link nearby neurons whose CRFs may or may not overlap, mediate these intracortical interactions between the neurons. These interactions have been suggested to underlie the contextual influences which underlie the context dependence of a neuron's response. This dependence is essential for computing saliency since, e.g., a vertical bar is salient in the context of horizontal bars but not other vertical bars.

The dominant contextual influence is iso-feature suppression, which is the suppression of the response to the input within a CRF when the context contains inputs with the same or similar features (Knierim and Van Essen 1992, Wachtler, Sejnowski and Albright 2003, Jones, Grieve, Wang and Sillito 2001). Iso-feature suppression is believed to be caused by the mutual antagonism between nearby V1 neurons tuned to similar features such as orientation and color. For the case of orientation, this suppression is known as iso-orientation suppression; see Section 2.3.9. Hence, e.g., a cell's response to its preferred orientation within its CRF is suppressed when the CRF is surrounded by stimuli sharing the same orientation.

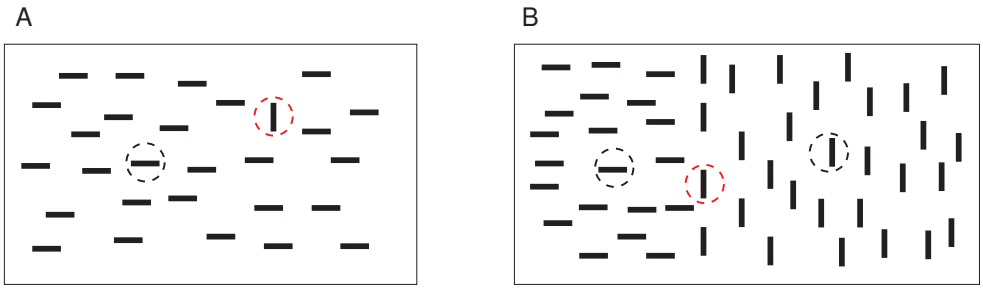


Fig. 5.7: The responses to an orientation singleton or a bar at a texture border will be higher because iso-orientation is absent or weaker, respectively. The dashed circles mark the CRFs of the neurons responding to the bars enclosed. In A, the vertical bar is unique in having no iso-oriented neighbors. Hence, a neuron tuned to a vertical orientation and responding to this bar is free from the iso-orientation suppression which affects neurons that are tuned to a horizontal orientation and respond to the horizontal bars. In B, a bar at the texture border, e.g., the one within the red circle, has fewer iso-oriented neighbors than a bar that is far from the border (e.g., the two bars in the black circles). Hence, neurons responding to the border bars are less affected by iso-orientation suppression.

Figure 5.7 illustrates how iso-orientation suppression makes V1 responses to a salient orientation singleton or an orientation texture border higher than responses to the background bars. For the case of the visual input in Fig. 5.7 A, a cell preferring vertical orientations and responding to the vertical bar escapes any iso-orientation suppression, because there is no neighboring vertical bar to evoke activity in neighboring cells that are also tuned to vertical. By contrast, a neuron preferring horizontal orientations and responding to one of the background horizontal bars experiences suppression from other horizontally tuned neurons responding to the neighboring horizontal bars. Consequently, when the contrast of all input bars is the same, the V1 cell that is activated most strongly by this image is the one responding to the vertical bar. According to the V1 saliency hypothesis, its location is then the most salient in this image.

Similarly, the bars at the border of an orientation texture have fewer iso-oriented neighbors than those away from the texture border. Thus, neurons responding to a texture border bar in Fig. 5.7 B experience a weaker iso-orientation suppression than that experienced by neurons responding to the other texture bars.

Figure 5.7 A can be seen as a special case of Fig. 5.7 B in that an orientation singleton is a texture region with just one texture element. Hence, this singleton itself can be viewed as its own texture border, and, by the reasoning above for Fig. 5.7 B, is more salient. This is why, as demonstrated in Fig. 5.4, the facilities of visual search and visual segmentation are typically related, when the visual features involved correspond.

In just the same way, iso-color suppression between neighboring V1 neurons that prefer similar colors, and iso-motion-direction suppression between neighboring V1 neurons that prefer similar motion directions, should both make for relatively higher V1 responses to a singleton in color or motion direction. More generally, iso-feature suppression should make V1 responses relatively higher at locations of higher input feature contrast. Thus, even though the CRFs are small and the intracortical connections that mediate contextual influences have a finite range, this mechanism allows V1 to perform a *global* computation

such that its neural responses reflect context beyond the range of the intracortical connections (Li 1997, Li 1999b, Li 2000b). By contrast, retinal neurons respond in a largely context independent manner, and so they could only adequately signal more specific and context independent forms of saliency, such as that caused by a bright image spot.

The neural mechanisms in V1 that mediate saliency have other properties. For instance, whether the saliency at the location of a red vertical bar is more likely signaled by a red-tuned cell or a vertically-tuned cell depends on the context. (For simplicity, we ignore neurons tuned both to the color red and to vertical orientation.) Both these cells respond to the red vertical bar; whichever one responds more vigorously should signal the saliency of this location (assuming this location has no other visual inputs). In the context of red horizontal bars in Fig. 5.8 A, it is the response of a vertical-tuned cell that determines its saliency; in the context of black vertical bars in Fig. 5.8 B, this is determined by a red-tuned cell. In either case, the most responsive neuron is determined by iso-feature suppression, and the saliency value depends only on the firing rate of the most responsive cell, regardless of whether it is color- or orientation-tuned.

Fig. 5.8: Contextual dependence of the neurons signaling the saliency of the red vertical bar. This bar evokes responses in cells preferring red and in cells preferring vertical orientations (ignoring the cells tuned to the color red and vertical orientations simultaneously for simplicity of argument). In A, iso-orientation suppression makes the vertical-tuned cell the most responsive to the red vertical bar; in B, iso-color suppression makes the red-tuned cell the most responsive.

5.2.3 Reading out the saliency map

The saliency map SMAP(x) is read out in order to execute an attentional shift. In principle, the saliency map in V1 could be ignored, i.e., read-out is unnecessary unless there is a need to shift attention. It is important to distinguish a brain area that contains the saliency map SMAP(x) from the brain areas that read out the saliency map for the purpose of shifting attention. V1's saliency output may be read by (at least) the superior colliculus (SC) (Tehovnik, Slocum and Schiller 2003), which receives inputs from V1 and directs gaze (and thus attention). In this case, the SC is not considered to compute saliency, but is merely a read-out area which selects the most salient location to execute an attentional shift.

Operationally, selecting the most salient location $\hat{x}$ does not require SMAP(x) to be calculated by means of the maximum operation in equation (5.4) to find the highest response to each location x. Rather, it only needs a single maximum operation $\hat{i} = \text{argmax}_i O_i$ over all neurons i, regardless of their RF locations or preferred input features. This is algorithmically perhaps the simplest possible operation to read a saliency map, and it can thus be performed very quickly. Being quick is essential for bottom-up selection.

If the read-out of saliency is deterministic, the most salient location $\hat{x}$ should be the first

one that bottom-up mechanisms select in the scene. If read-out is stochastic, this most salient location $\hat{x}$ is just most likely to be the first one selected.

Merely for the purpose of computing saliency, the maximum operation could be performed either in V1 or in the read-out area, or even both. The single maximum operation

$$\max_i O_i = \max_x(\text{SMAP}(x)) = \max_x(\max_{x_i \approx x} O_i) \tag{5.7}$$

over all responses O_i is equivalent to cascading two maximum operations, the first one locally $\max_{x_i \approx x}(O_i)$ to get SMAP(x) and then the second one globally $\max_x(\text{SMAP}(x))$. This is like selecting the national winner $\max_i O_i$ by having a two-round tournament: first, the local players near location x compete to get the local winner's score $\text{SMAP}(x) = \max_{x_i \approx x} O_i$; then, the local winners from different locations x compete globally to determine the overall winner $\max_x \text{SMAP}(x)$. If the local competition is performed in V1, and if the global competition is done in a read-out area such as the SC, then the explicit saliency map SMAP(x) should be found in the activities of the neurons projecting to the SC. This would license just a numerically small number of projecting neural fibers from V1 to the SC, consistent with anatomical findings (Finlay, Schiller and Volman 1976). If the competition is done in a single-round tournament in the SC or if both rounds of a two-round tournament are performed in the SC, then the SC needs to extract the saliency map from the whole population of V1 responses. The V1 saliency hypothesis is agnostic as to where and how the maximum operations are performed; these questions can be investigated separately.

Since V1's responses $\mathbf{O} = (O_1, O_2, ..., O_n)$ most likely play additional roles beyond saliency, it is necessary that the maximum operation or competition that selects the most salient location does not prevent the original responses $\mathbf{O}$ from being sent to other brain areas such as V2. Therefore, multiple copies of the signals $\mathbf{O}$ should be sent out of V1 via separate routes: one to the saliency read-out area and the others to other brain areas for other visual computations. For saliency, the maximum operation is only needed en route (perhaps in the layer 5 of V1) to, or in, the saliency read-out area. This need not distort the $\mathbf{O}$ values projecting to other brain areas.

5.2.4 Statistical and operational definitions of saliency

A salient location, such as that of the orientation singleton or the texture border in Fig. 5.7, is typically a place where visual input deviates from its context to a statistically significant degree. Consider covering the unique vertical bar in Fig. 5.7 A. One would naturally expect the bar at the covered image location also to be horizontal. In other words, conditional on the contextual input, the probability that this location contains a vertical bar is very small. Similarly, given the horizontal bars in the left half of the image in Fig. 5.7 B, without the knowledge of the presence of vertical bars in the right half of the image, the probability that the orientations of the bars are vertical in the middle of the image is quite low. Within textures of uniformly oriented bars as in Fig. 5.7, input statistics are translationally invariant, i.e., are identical at all locations. Hence, saliency could be linked to the degree to which the translation symmetry of one of a class of input statistics is broken (Li 1997, Li 1998b, Li 1999b, Li 2000b). This notion is related to the one that saliency is associated with surprise or novelty (Itti and Baldi 2006, Lewis and Zhaoping 2005). Other related notions of saliency include: a salient location is where an "interest point" detector (for a particular geometric image feature like a corner) signals a hit or where local (pixel or feature) entropy (i.e., information content) is high (Kadir and Brady 2001).

Meanwhile, we saw in Section 5.2.2 that iso-feature suppression in V1 allows the neurons in this area to detect and highlight the salient locations where local statistics (such as the average of the values of basic features such as color, orientation, and motion direction over

a neighborhood) change significantly. Ignoring any dependence on eccentricity for simplicity (or considering only a sufficiently small range of eccentricities), we assume that the properties of V1 CRFs and intracortical interactions are translation invariant. This implies that the input tunings and stimulus-bound responses to inputs within a neuron's CRF do not depend on the location of that CRF and that the interaction between two neurons depends only on the relative, but not the absolute, locations of their CRFs. In that case, the V1 responses should be translation invariant when the input is translation invariant (provided that there is no spontaneous symmetry breaking; discussed in Section 5.4.3). This input translation symmetry arises, for example, in an image comprising a regular texture of horizontal bars. It can also be generalized to cases such as the image of a slanted surface of a homogenous texture. However, when some statistics associated with these input features are not translation invariant, the responses of V1 neurons are expected to exhibit corresponding variabilities. Therefore, V1 mechanisms can often detect and highlight the locations where input symmetry breaks, and this will typically arise at the boundaries of objects. The V1-dependent process of highlighting salient object boundaries has also been termed as preattentive *segmentation without classification* (Li 1998b, Li 1999b), since the operation presumably occurs before object recognition or classification.

However, not all changes in visual input statistics make the corresponding input locations salient. Extensive studies have identified some of the kinds of input statistics whose change can make a location salient (Julesz 1981). For example, although there are exceptions (see Fig. 5.36 C), it has been observed that human observers can easily segment two textures which differ from each other according to their first and second order statistics, but not when the two textures differ in only higher order statistics. Hence, if one were to define saliency by the degree of change in some kind of input statistics, then this definition would need to include the individual sensitivity of saliency to changes in each kind of input statistics.

The basis of verifying whether a computational definition of saliency captures the reality should be our operational definition of saliency as the degree in which a visual location attracts attention by bottom-up mechanisms. This operational definition of saliency should also offer a basis to test any theory about saliency. In particular, the V1 saliency hypothesis should be tested against the behavioral data on saliency. For example, if a particular change in visual input statistics does not make the corresponding visual location behaviorally salient, then, if the hypothesis is correct, that location's saliency, as computed from V1 responses, should remain low, despite the change in the statistics of visual inputs.

As discussed in Section 5.1.1.2, saliency can be assessed by the reaction times and accuracies achieved in visual search and segmentation tasks. These tasks must be designed in such a way that the saliency of the visual location of interest is inversely related to the RT in the task (see Fig. 5.3), or monotonically related to the task performance accuracy using a brief visual display. To test the V1 saliency hypothesis, the predicted saliency from the V1 responses should be compared with that evident from the RTs and accuracies found in the behavioral experiments.

5.2.5 Overcomplete representation in V1 for the role of saliency

Chapter 4 discussed how the efficient coding principle could not readily explain the fact that V1 representation of visual input is highly overcomplete. This apparent over-representation greatly facilitates fast bottom-up selection by V1 outputs (Zhaoping 2006a). To see this more clearly, let us focus on orientation as a feature (ignoring other features such as color, motion, and scale). There is a large number of different cells in V1 tuned to many different orientations near the same location. This representation $\mathbf{O}$ helps to ensure that there is always a cell O_i at

each location that *explicitly* signals the saliency value of this location (at least in the case that the saliency arises from the orientation feature).

For example, let there be 18 neurons whose receptive fields cover a location x; each neuron prefers a different orientation, $\theta_i = i \cdot 10^o$ for $i = 1, 2, ..., 18$, spanning the whole 180^o range of orientation. The orientation tuning width of the neurons is sufficiently wide that presenting a bar of any orientation at x should excite some of the 18 neurons, with at least one being excited nearly optimally. This neuron will duly signal the saliency of the input. An input bar oriented at $\theta = 31^o$ would, for example, have its saliency signaled by the neuron tuned to 30^o.

Imagine instead an alternative representation which has only three neurons covering this location, preferring orientations 0^o, 60^o, and 120^o from vertical. To a 31^o tilted bar, the most responsive neuron prefers 60^o, the second most responsive one prefers 0^o, but the actual response of neither comes close to that of their preferred 60^o or 0^o bars, respectively. The maximum neural response to a 60^o bar would be higher than that to the 31^o bar, and if saliency is defined by the maximum neural response, the 31^o bar would appear less salient than the 60^o bar. To rescue the calculation, the saliency of the 31^o bar would have to be calculated as a function of responses from multiple underlying neurons (e.g., from the 0^o- and 60^o-preferring neurons, or from all three neurons), and this function would have to depend on the number of input bars at the same location x. The computational complexity of this calculation would increase dramatically when other feature dimensions and contextual inputs were also considered. It is completely unclear how such a saliency function could be computed and whether the computation would compromise the goal of fast saliency signaling along with adequate representation of the visual input.

It is likely that V1's overcomplete representation is also useful for other computational goals which could also be served by V1. Indeed, V1 also sends its outputs to higher visual areas for operations, such as recognition and learning, that go beyond selection. Within the scope of this chapter, I will not elaborate further upon our poor understanding of what constitutes the best V1 single representation for computing saliency as well as serving these other goals (although, as discussed, there can be different output channels for different goals).

5.3 A hallmark of the saliency map in V1—attention capture by an ocular singleton which is barely distinctive to perception

Everyday experience tells us that an item will only be salient if it is perceptually very distinct from its surroundings. This is the case, for example, for a red item among green ones or for a vertical bar among horizontal bars. It is thus surprising that the V1 saliency hypothesis predicts the following: an ocular singleton, which differs from surrounding items only by being shown to a different eye and is barely perceptually distinct, can be roughly as salient as a color or orientation singleton. For example, a horizontal bar shown to the left eye among surrounding horizontal bars shown to the right eye [14] is predicted to be highly salient.

This prediction arises because iso-feature suppression in V1, which is responsible for feature singleton pop-out, also applies for the feature that is the eye of origin of a visual input. This is underpinned by the many monocular neurons in this area, and indeed it is evident in

[14] Ocular singletons can be presented using stereo goggles. Some basic constraints need to be satisfied in such dichoptic presentations: vergence eye positions (which focus on specific locations in a three-dimensional scene) are anchored to elements of the display such as an image frame common to the two eyes; and the spacing of the elements should be such that neither binocular rivalry nor stereo matching occur for items in the two eyes (at least within a brief duration).

the observation that, when a V1 neuron responds to a monocular input, its response is more suppressed when the surrounding inputs are presented to the same, rather than to the other, eye (DeAngelis, Freeman and Ohzawa 1994).

However, an ocular singleton is not perceptually distinctive because few neurons downstream from V1 along the visual pathway are monocular. Various sources of evidence suggest that perception depends on the activity of extrastriate neurons; perception is thus blind to the eye of origin for monocular inputs. The lack of monocular cells in the extrastriate cortex is reflected in optical imaging of the ocular dominance columns. Figure 2.23 shows that these columns seen from the cortical surface stop abruptly at the border between V1 and V2, because V2 does not have enough monocular neurons. Therefore, the response of a binocular V2 neuron to a monocular input does not contain information regarding whether the input comes from the left or the right eye. This blindness at the neuron level is manifested behaviorally. For instance, in an experiment, observers were asked to report (and to guess, if necessary) whether there was a single item presented to the right eye among many background items presented to the left eye in a perceived image of these items (whose luminances are independently and randomly chosen). Their reports did not statistically differ from random guesses (Wolfe and Franzel 1988). Apparently, observers cannot distinguish an ocular singleton from the background items by its unique eye of origin. However, this does not mean that the singleton did not attract their attention. Indeed, this is an example for which the saliency of a visual input cannot be measured by the RT or accuracy of observers to find or identify this input.

In sum, the eye-of-origin signal is mainly, or exclusively, available in V1 among all the visual cortical areas. Furthermore, we know from Section 2.5 that upstream neurons are likely not involved in saliency. That is, in monkeys, the projection from the retina to the superior colliculus is normally not involved in visually guided eye movements; further, there is no projection from LGN to the superior colliculus. Therefore, the predicted high saliency of an ocular singleton (to guide attention or gaze shift) would be a clear fingerprint of the role of V1.

Despite the lack of reliable perception, one can probe the saliency of an ocular singleton by making it task irrelevant and testing if it interferes with visual search by distracting attention away from a true target. Figure 5.9 shows a visual input for observers who are asked to search for an orientation singleton among background bars. All bars are monocular, and one of the non-target bars is an ocular singleton. Observers perceive an image which is like the superposition of the image to the left eye and the image to the right eye. If gaze or attention is attracted in a bottom-up manner to the ocular singleton, it will interfere with the search, lengthening RTs. This was indeed observed. Take the case that the singletons are on opposite sides of the image, and 12^o from the center of the display, which is where gaze initially pointed before the search begins. The first gaze shift during the search was directed to the task-irrelevant ocular singleton on 75% of the trials. This was the case even though the orientation singleton target was very salient, since it was tilted 50^o away from 659 uniformly oriented non-target bars, and observers were told to search for it as quickly as possible (Zhaoping 2012). This is analogous to Fig. 5.2 A, in which the red non-target bar among black bars attracts attention automatically away from the target. However, unlike the ocular singleton, the red bar is highly perceptually distinctive.

We describe in more detail here the experiments showing that the ocular singleton could attract attention, even though observers could not perceive any visual difference between it and its neighboring bars. Three different dichoptic presentation conditions, monocular (M), dichoptic congruent (DC), and dichoptic incongruent (DI) are shown in Fig. 5.10. The superposition of the two monocular images is the same in these three conditions, and it resembles the perceived image, which has an orientation singleton bar in a background of uniformly oriented bars. The orientation singleton is the target of visual search. In the M

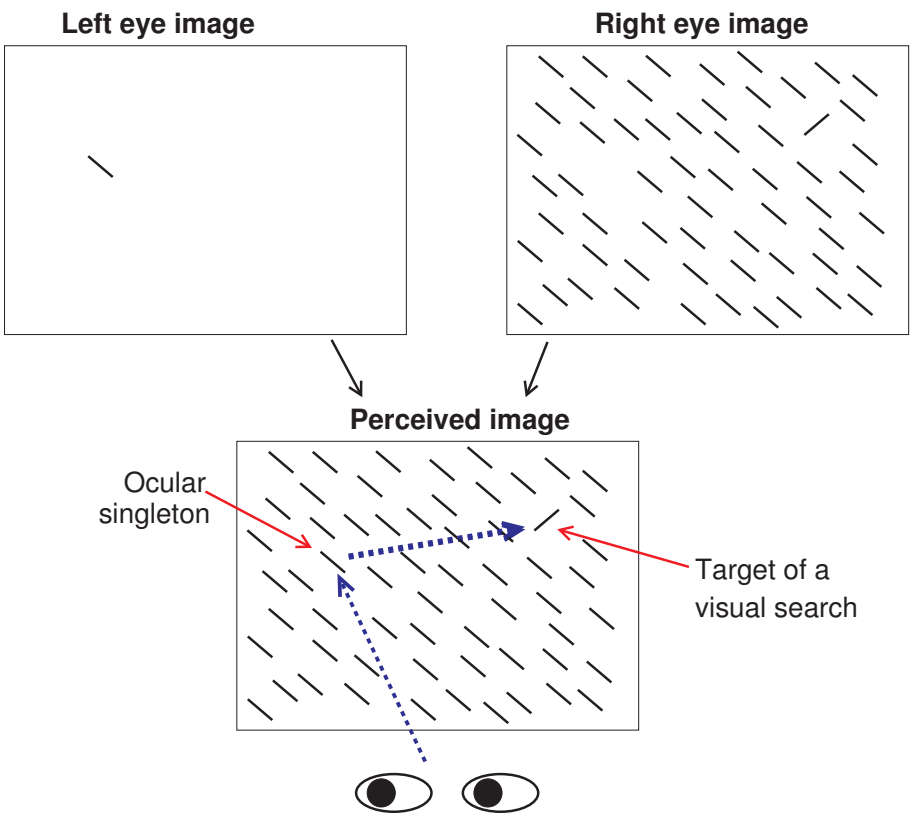


Fig. 5.9: An ocular singleton, though task-irrelevant and not perceptually distinct from background items, often attracts the first shift of gaze, before a subsequent shift to the target (the orientation singleton) of the visual search (Zhaoping 2012). The colored arrows are not part of the visual stimulus; they indicate the gaze shifts and point to the feature singletons.

condition, all bars are presented to the same single eye. In the DC condition, the target bar is an ocular singleton, since it is presented to a different eye than the other bars. In the DI condition, a non-target bar is an ocular singleton; it is presented to the opposite lateral side of the target from the center of the perceived image. In the search for the orientation singleton, the ocular feature is task irrelevant but could help or hinder the task in DC or DI conditions, respectively.

Figures 5.11 and 5.12 present experiments using such stimuli, together with their results (Zhaoping 2008). In Fig. 5.11, subjects had to report whether the orientation singleton, whose location the observers did not know ahead of each trial, was tilted clockwise or anticlockwise from horizontal. However, the images were presented so briefly that the task would be difficult unless attention was quickly guided to the location of the target. The degree of difficulty was measured by the error rate, which is the fraction of trials in which the observers performed the task erroneously. Figure 5.11 B shows that the error rate for this task was smaller in the DC trials than in the M and DI trials. This suggests that attention was guided to the target more effectively in the DC trials. One may see the DC or DI trials as ones in which an ocular singleton provides valid or invalid, respectively, guidance of attention to the target. In the M trials, there is no ocular singleton to guide attention.

Meanwhile, the second experiment in Fig. 5.11 revealed that the same observers were

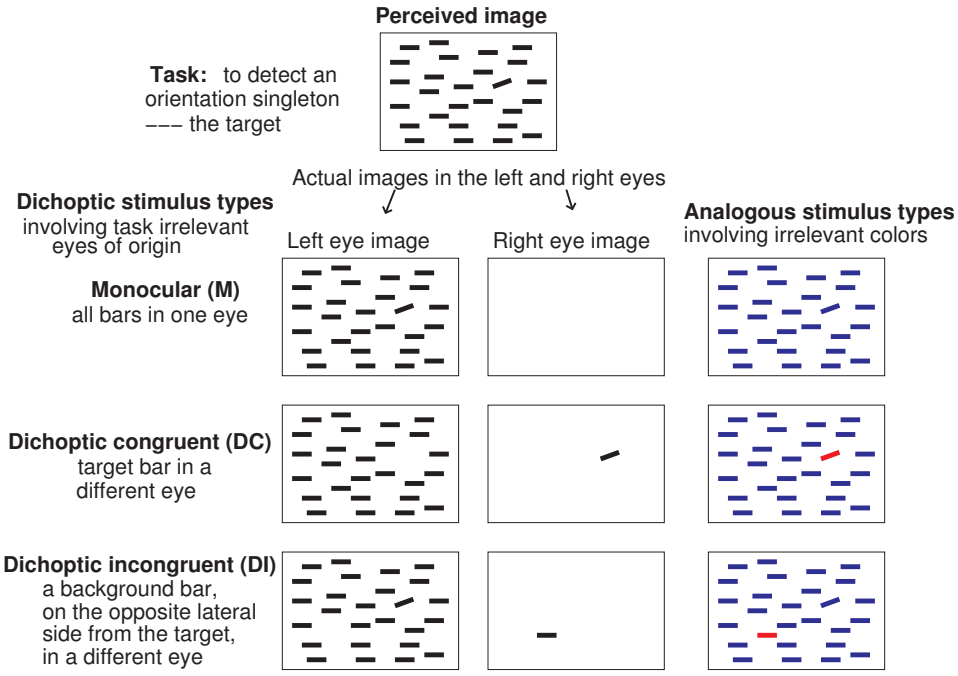


Fig. 5.10: Schematic of the stimulus used to test the automatic capture of attention by an eye-of-origin or ocular singleton, even though one can barely perceive any difference between inputs from different eyes. The ocular feature is irrelevant for the search for an orientation singleton, and the observers are not required to report it. They perceive an image with an orientation singleton target among background bars, but this perceived image could be made from three different dichoptic presentation conditions: monocular (M), dichoptic congruent (DC), and dichoptic incongruent (DI). The analogous case when color is the irrelevant feature disrupting the same task is shown on the right. If the ocular singleton is salient and attracts attention more strongly than the orientation singleton, it should help and hinder the task in the DC and DI conditions, respectively, by guiding attention to and away from the target.

not necessarily aware of these attention-guiding ocular singletons. When different bars in the display had randomly different luminances, observers could do no better than chance (error rate is 0.5) in reporting the presence or absence of the ocular singletons. They did better than chance when all the bars had the same luminance, because an ocular singleton can sometimes be identified by an illusory contrast different from the other bars. (Apparently, heterogenous luminances across the bars made this illusory contrast ineffective for identifying the ocular singleton by the observers.) Nevertheless, the ability of the task-irrelevant ocular singleton to guide attention in the first experiment did not depend on whether the luminance condition was such that, from the second experiment, we would expect subjects to have been able to identify the ocular singleton.

These findings suggest that RTs to locate the orientation singleton should be shorter in the DC trials and longer in the DI trials. This was indeed observed when observers were asked to report as quickly as possible whether the target was in the left or right half of the perceived image (which remained displayed until they made their choice); see Fig. 5.12. Let RT_M, RT_{DC}, and RT_{DI} denote the RTs for the monocular (M), dichoptic congruent (DC), and dichoptic incongruent (DI) stimulus conditions, respectively. The data show $RT_M > RT_{DC}$

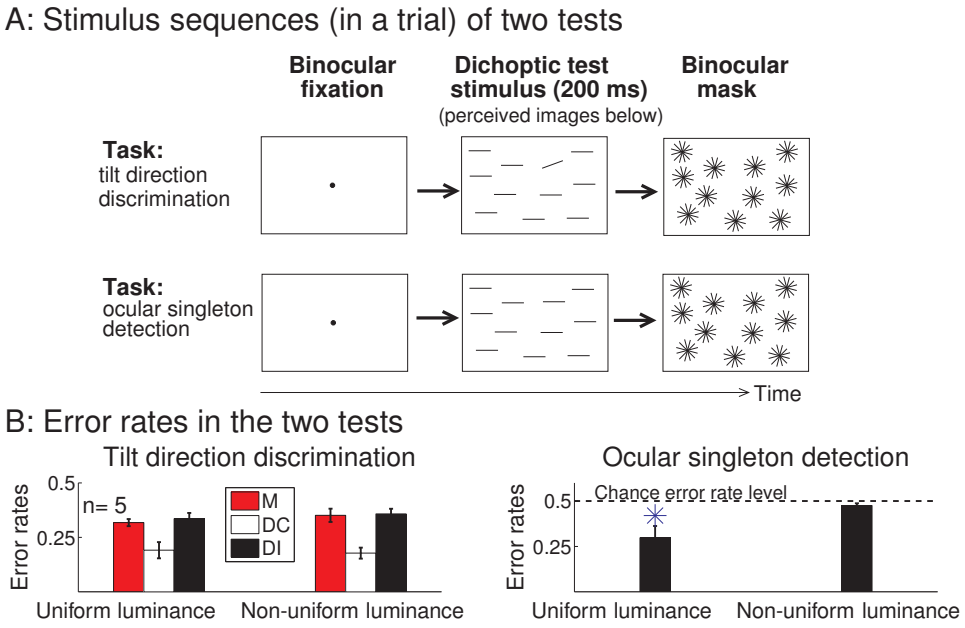
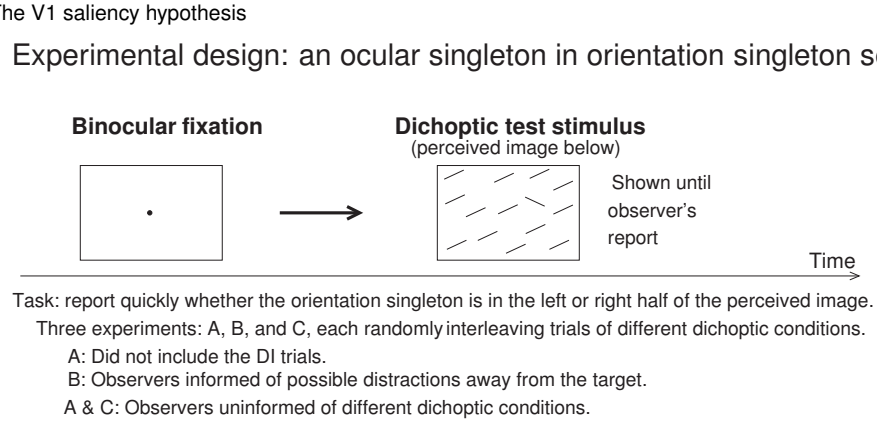


Fig. 5.11: Two experiments showing that an ocular singleton guides attention even when observers are unaware of its presence (Zhaoping 2008). A: Schematics of test trials in each of the two experiments. The dichoptic test stimulus, in which all bars are monocular, is binocularly masked after being displayed for only 200 ms. In one experiment (top), observers report whether an orientation singleton, tilted 20^o from 659 horizontal bars, was tilted clockwise or anticlockwise from horizontal. As in Fig. 5.10, all test stimulus bars are monocular, and a given trial can be randomly monocular (M), dichoptic congruent (DC), or dichoptic incongruent (DI). In the second experiment (bottom), the test stimulus is the same as in the first experiment except that all bars are horizontal, and the ocular singleton has an equal chance of being present or absent (if present, it is randomly at one of the locations for the ocular singleton in the first experiment). Observers report whether the ocular singleton is present. In each trial in both experiments, either all stimulus bars have the same luminance or different bars have different random luminances. B: Error rates in the two experiments, averaged across five observers (who participated in both experiments), are shown separately for the two luminance conditions. In the right plot in B, an error rate significantly different from the chance level (0.5) is indicated by a "*".

and $RT_{DI} > RT_M$, in which $RT_M \approx 0.6$ seconds for typical observers. These relationships remained true whether or not the subjects were informed that different dichoptic stimulus types could be randomly interleaved in the trials. Even when they were informed that a non-target bar might distract them in some of the trials and were explicitly told to ignore it (experiment B of Fig. 5.12), RT_{DI} was still greater than their RT_M. This suggests that the bottom-up attraction of the irrelevant ocular singleton could not be easily suppressed by top-down control. The RT difference $RT_{DI} - RT_M$ was around 0.2–0.3 seconds on average, comparable to typical time intervals between two saccades in a visual search. This suggests that, in typical DI trials, attention or gaze was first attracted to the task-irrelevant ocular singleton before being directed to the target. This was later confirmed by tracking the gaze of subjects who were doing this task; see Fig. 5.9.

As mentioned in Section 5.1.2, orientation is one of the basic feature dimensions. That is,



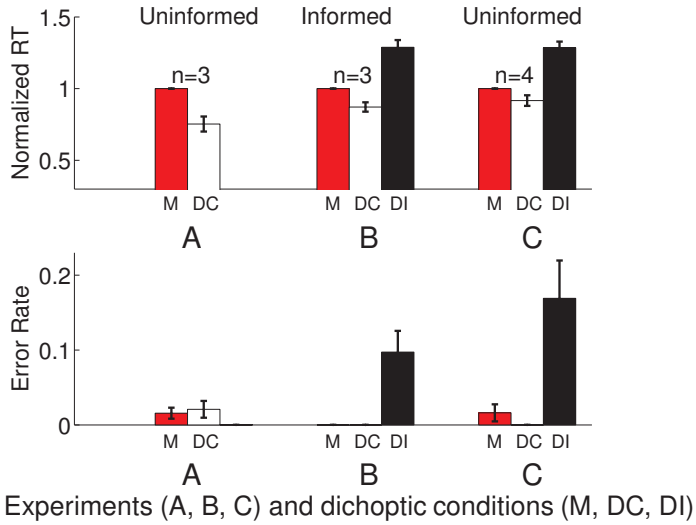


Fig. 5.12: An ocular singleton can speed up or slow down visual search for an orientation singleton. Each dichoptic search stimulus had 659 iso-oriented background bars and one orientation singleton bar, tilted 25^o from horizontal in opposite directions. Subjects reported as soon as possible whether the target was in the left or right half of the perceived image. There were three experiments, A, B, and C, each of which randomly interleaved trials of various dichoptic conditions: monocular (M), dichoptic congruent (DC), and dichoptic incongruent (DI), as in Fig. 5.10. As indicated in the bar charts, experiment A contained M and DC trials, and experiments B and C each contained M, DC, and DI trials. Observers were not informed of the different dichoptic conditions except in experiment B, in which they were informed that some trials might contain a distracting non-target. RTs (normalized by RT_M, which is around 600 ms, of individual observers, so that $RT_M = 1$) and error rates are averaged across $n = 3$, 3, and 4 observers, respectively, for experiments A, B, and C.

an orientation singleton is sufficiently salient such that visual search for it is efficient, with RT being independent of the search set size. However, the experiment depicted in Fig. 5.12 showed that an ocular singleton can attract gaze more strongly than an orientation singleton tilted 50^o from the background bars. Hence, the ocular singleton is more salient than the

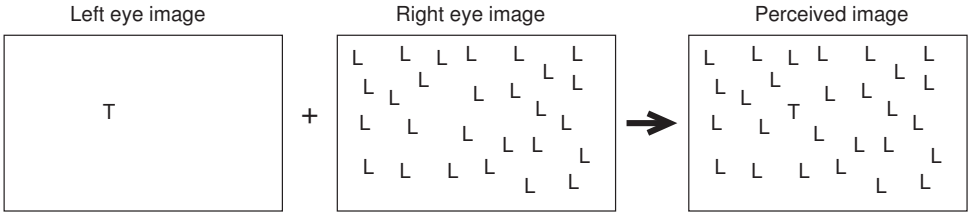


Fig. 5.13: Making the target "T" an ocular singleton renders efficient what is otherwise found to be an inefficient search for a "T" among "L"s (Zhaoping 2008).

orientation singleton, and so the eye-of-origin feature dimension must also be basic. Indeed, an inefficient search for a letter "T" among background letters "L" can be made efficient when "T" is an ocular singleton (Zhaoping 2008); see Fig. 5.13. This basic feature dimension of ocular origin was not recognized until the experimental findings described here, since this feature was not perceptually distinctive.

5.3.1 Food for thought: looking (acting) before or without seeing

At first, the prediction, that a visual item that is barely distinguishable from its neighbors can attract attention, like a red flower among green leaves, might seem surprising or even impossible. This reaction arises from our impression or belief, driven from experience, that seeing precedes looking, i.e., that we look at something after, or because, we have seen what it is. The confirmation of this counterintuitive prediction from the V1 saliency hypothesis invites us to ponder and revise our belief. Logically, one looks in order to see, and looking should be expected to precede seeing, at least for part of our visual behavior. This is analogous to the example in Fig. 1.4, when observers do the act of looking or shifting gaze before they know the identity of the visual input at the destination of their gaze shift. Looking should also be dissociable from seeing. Indeed, brain lesion patients who cannot recognize objects can still manipulate objects adequately (Goodale and Milner 1992). According to this analysis, it is likely that gaze is attracted to the location of the ocular singleton in the perceived image before the two monocular images have been combined to achieve the perception of the perceived image. Meanwhile, the perceived image, i.e., the image in observers' perception after combining the inputs from the two eyes, contains little or no information about the eye of origin that could influence gaze.

5.4 Testing and understanding the V1 saliency map in a V1 model

This section presents a model of V1. This model is intended to serve two main purposes: first, as a substitute for the real V1 to test the relationship between V1 activities and behavioral saliencies. The experiments on ocular singletons offer convincing support for the V1 saliency hypothesis. However, we should also examine whether the link between V1 responses and saliency also applies in general cases, including those for which the saliency effects are subtle. The literature (Wolfe 1998) contains a wide range of behavioral data on saliency in terms of reaction times or task difficulties in visual search and segmentation. However, physiological data based on stimuli used in the behavioral experiments are few and far between. Furthermore, according to the V1 saliency hypothesis, predicting the saliency of a location requires us to compare the V1 responses of neurons with that location as their classical RFs to the responses

of neurons favoring other locations. This would require simultaneous recordings of many V1 units responding to many locations, a very daunting task with current technology. Examining the responses of all neurons in a simulation of the model provides a simpler, though obviously inferior, alternative to recording in the actual V1. Figure 5.14 shows an outline of the V1 model and its function.

Second, examining the model neural circuit can help us understand how intracortical interactions in V1 lead to the computation of context-dependent saliency from local contrast inputs. Qualitative arguments such as those in Section 5.2.2 suffice for us to envisage how iso-feature suppression could explain the relative enhancement of V1 responses to very salient feature singletons and texture borders. However, they are insufficient for knowing whether or how V1 interactions could also account for subtle saliency effects and indeed whether the neural circuit dynamics are well-behaved. Obviously, the actual V1 neural dynamics *are* well behaved. However, testing whether a model of V1 mechanisms identified by us as responsible for saliency computation has well behaved dynamics enables us to test whether our understanding is correct. Simulating the intracortical interactions, showing how they produce iso-feature suppression as well as the less dominant interactions, can allow us to verify our intuitions and help to identify how various intracortical mechanisms shape visual saliency.

The material in this section is mostly adapted and extended from papers in the literature (Li 1998a, Li 1999a, Li 1999b, Li 2000b, Li 2001, Li 2002, Zhaoping 2003). These were all published before the confirmation of the ocular singleton effect and of other predictions of the V1 saliency hypothesis (described later this chapter). Therefore, this model actually served the purpose of assessing whether V1 mechanisms could feasibly subserve the computation of saliency. That is, it tested the V1 saliency hypothesis using behavioral data already known before the model was constructed, or using self-evident behavioral phenomena. In this section, we will show simulated V1 responses to representative visual inputs whose behavioral saliency profiles are well known. Furthermore, we will show examples in which model responses highlight locations where input statistics break translation symmetry.

5.4.1 The V1 model: its neural elements, connections, and desired behavior

V1 neurons can be tuned to orientation, color, scale, motion direction, eye of origin, disparity, and combinations of these. This is in addition to the selectivity to input spatial locations (and spatial phase/form) by their receptive fields. As an initial attempt to study whether it is feasible for V1 mechanisms to compute saliency, the model here focuses on only two features, spatial location and orientation. Thus, each model neuron is characterized by its preferred orientation and spatial location; all model receptive fields have the same size; and the centers of the receptive fields sit on a regular spatial grid. Hence, the model ignores other visual cues such as color, motion, and depth. Emulating and understanding the dependence of saliencies on the spatial configurations of oriented bars is arguably more difficult than emulating and understanding the dependence on luminance and color features. Once the feasibility of V1 mechanisms for saliency can be established in this simplified V1 model, the model can then be extended to include the other feature dimensions and neural selectivities (see Section 5.8.3).

Since the model focuses on the role of intracortical interactions in computing saliency, the model mainly includes orientation selective neurons in layers 2–3 of V1. These are coupled by intracortical connections, which are sometimes also called horizontal or lateral connections. The model ignores the mechanism by which the neural receptive fields are generated. Inputs to the model are images seen through the model classical receptive fields (CRFs) of V1 complex cells, which are modeled as edge or bar detectors (we use "edge" and "bar" interchangeably). (To avoid confusion, here the term "edge" refers only to local luminance contrast. Meanwhile,

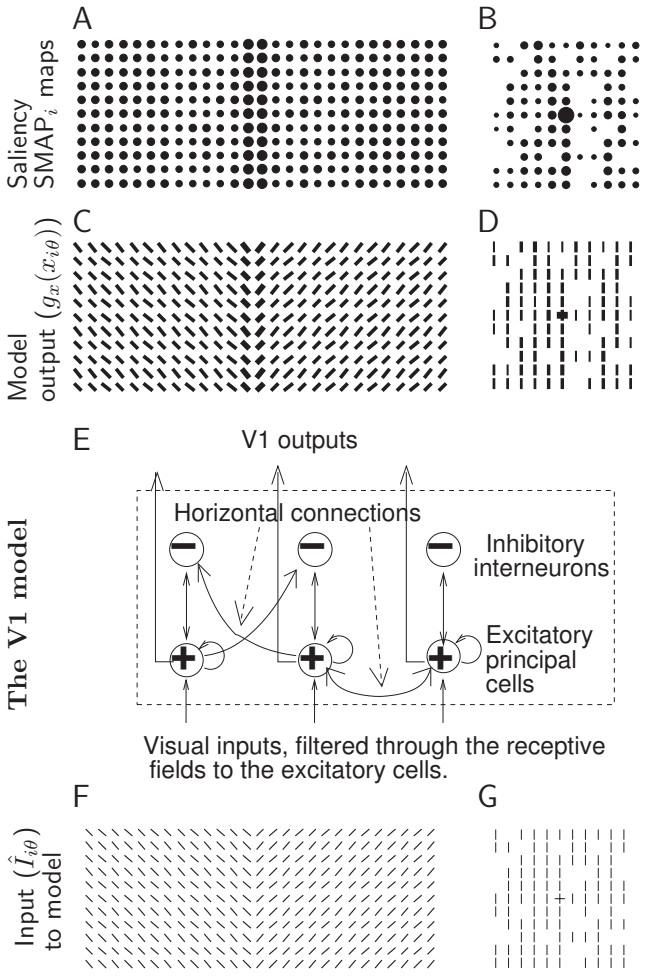


Fig. 5.14: The V1 model and its operation. The model (E) focuses on the part of V1 responsible for contextual influences: excitatory pyramidal (principal) cells in layers 2–3 of V1, interneurons, and intracortical (horizontal) connections. A pyramidal cell can excite another pyramidal cell monosynaptically, and/or inhibit it disynaptically via the inhibitory interneurons. The model also includes general and local normalization of activities. F and G are two example input images. Their evoked model responses, C and D, are those of the pyramidal cells preferring the corresponding positions and orientations. As in many figures in the rest of this chapter, the input contrast (strength) or output responses are visualized by (in proportion to) the thicknesses of the bars in the input or output images. A principal cell receives direct visual input only from the input bar within its CRF. Its response depends both on the contrast of the bar and the stimuli in the context, the latter via the intracortical connections. Each input/output/saliency image that is shown is only a small part of a larger, extended input/output/saliency image. At the top (A, B) are saliency maps, in which each location i is for a hypercolumn. The size of the disk at location i visualizes the highest response SMAP$_i$ among the pyramidal cells responding to this location. A location is highly salient if this disk is much larger (assessed by a z score) than the other disks in the map. The notations $\hat{I}_{i\theta}$ and $g_x(x_{i\theta})$ will be defined shortly. Adapted with permission from Zhaoping, L., Theoretical understanding of the early visual processes by data compression and data selection, *Network: Computation in Neural Systems*, 17(4): 301–334, Fig. 8, copyright © 2006, Informa Healthcare.

a boundary of a region is termed "boundary" or "border", which, especially in textures, may or may not correspond to any actual luminance edges in the image.) Intracortical connections (Rockland and Lund 1983, Gilbert and Wiesel 1983) mediate interactions between neurons such that patterns of direct inputs to the neurons via their CRFs are transformed into patterns of contextually modulated responses (firing rates) from these neurons.

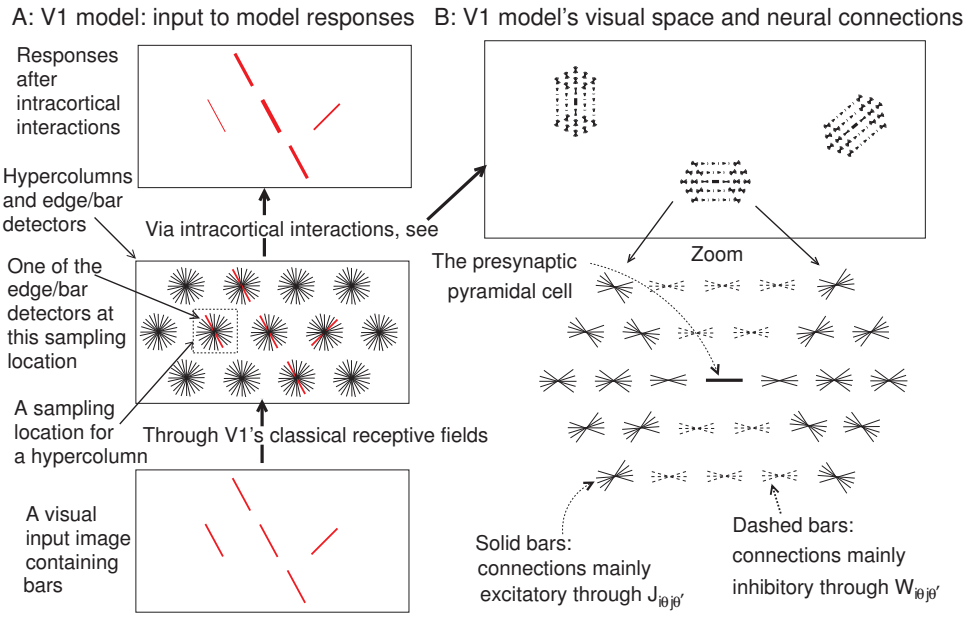


Fig. 5.15: Schematic of the V1 model. A: An example visual input contains five bars of equal contrast (marked by red color, to distinguish them from black bars visualizing neurons and the neural connection pattern); the (black) rectangle (not part of input image) frames the input image. In the middle is the V1 model, which contains many classical edge or bar detectors; each detector is visualized by a black bar and is modeled by a pair of mutually connected neurons: an excitatory pyramidal cell and an inhibitory interneuron (see Fig. 5.16). A single hypercolumn occupies a spatial sampling location and comprises many detectors preferring various orientations that span 180^o. Without intracortical interactions, five edge/bar detectors (shown in red in the middle frame) are equally excited by the five equal contrast input bars through their respective CRFs; no other detector is as substantially activated directly. Through intracortical interactions, the eventual responses from the five detectors are unequal, visualized by different thicknesses of the (red) bars in the top frame. B: A schematic of the lateral connections in the model. The rectangle frames the visual space. Three groups of neural connections (translated and rotated versions of each other) radiating from three presynaptic cells are shown. In the zoomed view of one group, the central horizontal bar marks the presynaptic pyramidal cell preferring horizontal orientations. The thin bars mark the locations and preferred orientations of the postsynaptic pyramidal cells: the solid ones are for cells mainly excited by the presynaptic cell through monosynaptic $J_{i\theta,j\theta'}$ connections; dashed ones are for cells mainly disynaptically inhibited by the presynaptic cell, via connections $W_{i\theta,j\theta'}$; see text.

Figures 5.15 and 5.16 show the elements of the model and the way they interact. Following original literature, we denote a spatial sampling location by i rather than x, which will

instead denote membrane potentials of pyramidal cells. At each spatial sampling location i, there is a model V1 hypercolumn composed of cells whose CRFs are centered at i. Each of these cells is tuned to one of $K = 12$ different orientations θ spanning 180^o. Based on experimental data (White 1989, Douglas and Martin 1990), each edge or bar detector at location i and preferring orientation θ is modeled by one pair of interconnected model neurons: one excitatory pyramidal cell and one inhibitory interneuron; detailed in Fig. 5.16. Hence, altogether, each hypercolumn consists of 24 model neurons. Each model pyramidal cell or interneuron is a simple rate-based neuron (see Section 2.1.2). It could model abstractly, say, 1000 spiking pyramidal cells or 200 spiking interneurons with similar CRF tuning (i.e., similar i and θ) in the real cortex. Therefore, a 1:1 ratio between the numbers of pyramidal cells and interneurons in the model does not imply such a ratio in the cortex. We often refer to the cells tuned to θ at location i as simply the edge or bar element $i\theta$. The image that is shown is represented as inputs $I_{i\theta}$ across various $i\theta$. Each $I_{i\theta}$ models the visual input image seen through the CRF of a complex (pyramidal) cell preferring location i and orientation θ.

Although readers can follow the rest of this section without any equations, the following equations summarize the neural interactions in the model (see Section 2.1.2 on neuron models):

$$\dot{x}_{i\theta} = -\alpha_x x_{i\theta} - g_y\left(y_{i,\theta}\right) - \sum_{\Delta\theta\neq 0} \psi\left(\Delta\theta\right) g_y\left(y_{i,\theta+\Delta\theta}\right)$$
$$+J_o g_x\left(x_{i\theta}\right) + \sum_{j\neq i,\theta'} \mathsf{J}_{i\theta,j\theta'} g_x\left(x_{j\theta'}\right) + I_{i\theta} + I_o + I_{\text{noise}}, \tag{5.8}$$

$$\dot{y}_{i\theta} = -\alpha_y y_{i\theta} + g_x\left(x_{i\theta}\right) + \sum_{j\neq i,\theta'} \mathsf{W}_{i\theta,j\theta'} g_x\left(x_{j\theta'}\right) + I_c + I_{\text{noise}}. \tag{5.9}$$

In the above equations, $x_{i\theta}$ and $y_{i\theta}$ model the membrane potentials of the pyramidal cell and the interneuron, respectively, for edge or bar element $i\theta$; $g_x(x)$ and $g_y(y)$ are sigmoid-like functions modeling cells' firing rates or responses given membrane potentials x and y for the pyramidals and interneurons; $-\alpha_x x_{i\theta}$ and $-\alpha_y y_{i\theta}$ model the decay to resting potentials with time constants $1/\alpha_x$ and $1/\alpha_y$; $\psi(\Delta\theta)$ models the spread of inhibition within a hypercolumn; $J_o g_x(x_{i\theta})$ models self-excitation; $\mathsf{J}_{i\theta,j\theta'}$ and $\mathsf{W}_{i\theta,j\theta'}$ are neural projections from pyramidal cell $j\theta'$ to excitatory and inhibitory postsynaptic cell $i\theta$; I_c and I_o are background inputs modeling the general and local normalization of activities; and I_{noise} is input noise which is independent between different neurons. The pyramidal cell outputs $g_x(x_{i\theta})$ (or temporal averages over these) represent the V1 responses. Equations (5.8) and (5.9) specify how the pyramidal activities $g_x(x_{i\theta})$, which are initialized by external inputs $I_{i\theta}$, are modified by the contextual influences via the neural connections. This model can be reproduced using the complete details in the appendix of this chapter (see Section 5.9).

Note that notations in this chapter often have different semantics from those in other chapters. For example, K means the number of preferred orientations in a hypercolumn, and should not be confused with the kernels or filters in the previous chapters. This book tries to balance between self-consistency within its own notation, and consistency with the notation used in the original literature.

The pyramidal responses or output activities $g_x(x_{i\theta})$, which are sent to higher visual areas as well as subcortical areas such as the superior colliculus, will be used to quantify the saliencies of their associated locations and edge elements. The inhibitory cells are treated as interneurons. The input $I_{i\theta}$ to pyramidal cell $i\theta$ is obtained by filtering the input image through the CRF associated with $i\theta$. Hence, when the input image contains a bar of contrast $\hat{I}_{i\gamma}$ at location i and oriented at angle γ, this bar contributes to $I_{i\theta}$ by the amount

$$\hat{I}_{i\gamma}\phi(\theta-\gamma), \quad \text{where } \phi(\theta-\gamma) \text{ is the orientation tuning curve of the neurons. (See Section 5.9 for the actual } \phi(x) \text{ used.)} \tag{5.10}$$

To visualize the strength of the input (contrast) and the model responses, the widths of the bars plotted in each figure are made to be larger for stronger input strength $I_{i\theta}$, or greater pyramidal responses $g_x(x_{i\theta})$ (or their temporal averages).

In the absence of intracortical interactions between different edge elements $i\theta$, the reciprocal connections between each pyramidal cell and its partner inhibitory interneuron would mainly provide a form of gain control for the direct input $I_{i\theta}$ (and make the response transiently oscillatory). The response $g_x(x_{i\theta})$ from the pyramidal cell $i\theta$ would only be a function of this direct input, in a context independent manner. With intracortical interactions, the influence of one pyramidal cell on the response of its neighboring pyramidal cell is excitatory via monosynaptic connections and inhibitory via disynaptic connections through the interneurons. Consequently, a pyramidal cell's response depends on inputs outside its CRF, and the pattern of pyramidal responses $\{g_x(x_{i\theta})\}$ is typically not just a scaled version of the input pattern $\{I_{i\theta}\}$ (see Fig. 5.15 A).

Figure 5.15 B shows the structure of the lateral connections in the model (Li 1999b). Connection $\mathsf{J}_{i\theta,j\theta'}$ from pyramidal cell $j\theta'$ to pyramidal cell $i\theta$ mediates monosynaptic excitation. It is present if these two segments are tuned to similar orientations $\theta \approx \theta'$ and the centers i and j of their CRFs are displaced from each other roughly along their preferred orientations θ and θ'. Connection $\mathsf{W}_{i\theta,j\theta'}$ from pyramidal cell $j\theta'$ to the inhibitory interneuron $i\theta$ mediates disynaptic inhibition from pyramidal cell $j\theta'$ to pyramidal cell $i\theta$. It tends to be present when the preferred orientations of the two cells are similar $\theta \approx \theta'$, but the centers i and j of their CRFs are displaced from each other along a direction roughly orthogonal to their preferred orientations. This V1 model has a translation invariant structure, such that all neurons of the same type have the same properties, and the neural connections $\mathsf{J}_{i\theta,j\theta'}$ (or $\mathsf{W}_{i\theta,j\theta'}$) have the same structure from all the presynaptic neurons $j\theta'$ except for translation and rotation to suit the position and orientation of the presynaptic receptive field $j\theta'$ (Bressloff, Cowan, Golubitsky, Thomas and Wiener 2002). The structure of the connections from a single pyramidal cell resembles a bow-tie.

Figure 5.16 illustrates the intracortical connections and their functions in further detail. The input image in Fig. 5.16 contains just horizontal bars. Hence, neurons preferring non-horizontal orientations are not strongly excited directly and are omitted from the figure. Here, the monosynaptic connections J link neighboring horizontal bars displaced from each other roughly horizontally, and the disynaptic connections W link those bars displaced from each other more or less vertically in the visual input image plane. The full lateral connection structure from a cell preferring a horizontal bar to cells preferring other bars (including bars that are not horizontal) is shown in Fig. 5.15 B.

In the input image, the five horizontal bars have the same input contrast, giving equal strength input $I_{i\theta}$ to the five corresponding pyramidal cells. Nevertheless, the output responses from these five pyramidals are different from each other, illustrated in the top plate of Fig. 5.16 by the different widths of the bars. The three horizontally aligned bars evoke higher output responses because the corresponding neurons facilitate each other's activities via the monosynaptic connections $\mathsf{J}_{i\theta,j\theta'}$. The other two horizontal bars evoke lower output responses because the corresponding neurons receive no monosynaptic lateral excitation but receive disynaptic lateral inhibition from (neurons responding to) the neighboring horizontal bars displaced vertically from, and not co-aligned with, them. (To avoid excessive words, we sometimes use the term "bars" to refer to "neurons receiving direct inputs from the bars" when the meaning is clear from context). The three horizontally aligned bars, especially the middle one, also receive disynaptic inhibitions from the two vertically displaced bars.

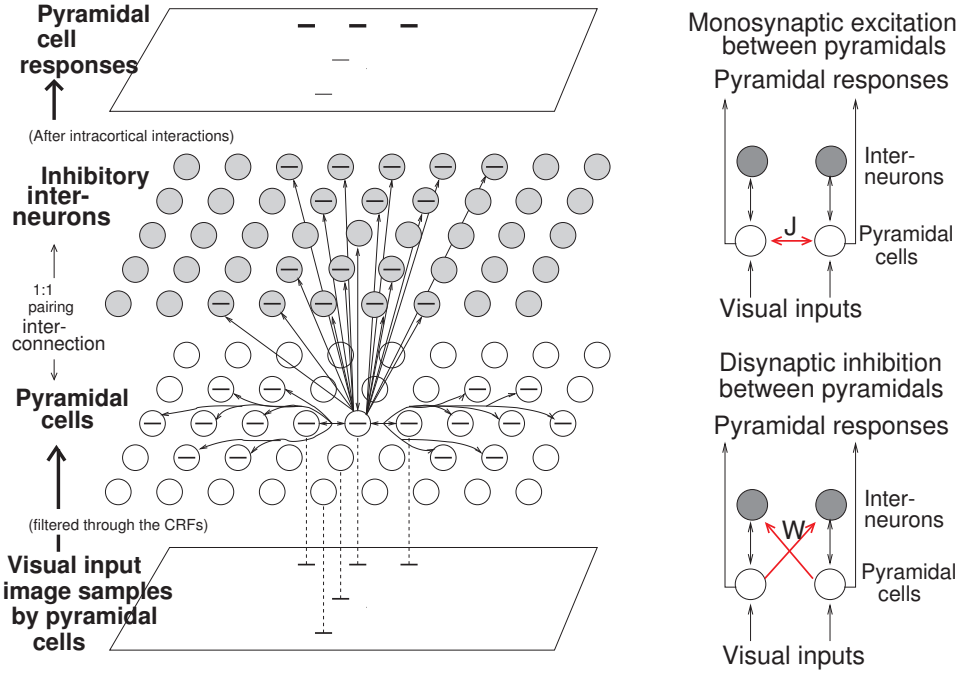


Fig. 5.16: Model elements. To avoid excessive clutter, only cells tuned to horizontal orientations are shown; and only connections to and from the central pyramidal cell are drawn. A horizontal bar, marking the preferred orientation, is drawn on the central pyramidal cell and the postsynaptic cells to which it is linked via lateral connections. In the input image plane, the central pyramidal neuron sends axons to other pyramidal cells displaced from it locally in a roughly horizontal direction, and to the interneurons which are also displaced locally, but in a roughly vertical direction. These axons are, respectively, for the monosynaptic excitation and disynaptic inhibition between the pyramidal cells (illustrated in the plots on the right). Five horizontal bars of equal contrast are shown in the input image in the bottom plane; each excites a pyramidal cell with the corresponding CRF (the correspondences are indicated by the dashed lines). The three aligned bars evoke higher responses, while two bars displaced vertically from them evoke lower responses (shown in the top plate). These differential responses are caused by facilitation between the three aligned bars via the monosynaptic connections J and the suppression between the vertically displaced bars by the disynaptic inhibition mediated by W. Adapted with permission from Li, Z., Pre-attentive segmentation in the primary visual cortex, *Spatial Vision*, 13(1): 25–50, Fig. 2C, copyright © 2000, Koninklijke Brill NV.

When the input image is a homogenous texture of horizontal bars, each bar receives monosynaptic lateral excitation from its (roughly) left and right neighbors but disynaptic lateral inhibition from its (roughly) top and bottom neighbors. The intracortical connections in the model are designed so that the sum of the disynaptic inhibition overwhelms the sum of the monosynaptic excitation in an iso-orientation texture. Hence, the net contextual influence on any bar in an iso-oriented and homogenous texture will be suppressive—this is iso-orientation suppression. Therefore, it is possible for the same neural circuit to exhibit iso-orientation suppression when the input image is a uniform texture, and to exhibit colinear facilitation, or contour enhancement, when the input image is an isolated contour made of multiple co-aligned bar segments. This is what has been observed in physiological experiments (Knierim

and Van Essen 1992, Kapadia et al. 1995); see Section 2.3.9. Note that an iso-orientation texture can be seen as an array of parallel contours or lines.

Figure 5.17 illustrates how a smooth contour in a noisy background or along a texture border should evoke higher V1 responses than the same contour lying within a texture. The contextual influence depends on both the orientation and the spatial configuration of the context due to the following reasons: first, lateral connections tend to link bars having similar orientations; and second, the interaction between these similarly oriented bars tend to be monosynaptic and excitatory when they are co-aligned but disynaptic and inhibitory when they are not co-aligned. Each of the three vertical bars in dashed circles in Fig. 5.17 is part of a vertical contour. However, the contours are either along a texture border, within an iso-oriented texture, or embedded in a random background. Each enjoys the monosynaptic excitation from its co-aligned neighbors. However, iso-orientation suppression in V1 implies that this monosynaptic excitation is overwhelmed by the disynaptic inhibition when the contour is in the center of an iso-oriented texture. Meanwhile, when the contour is along a texture border such that it has fewer parallel contours as neighbors, this disynaptic inhibition should be reduced. The disynaptic inhibition should be minimal when there is no parallel contour neighbor, such as when the contour is isolated or embedded in a random background, as in Fig. 5.17 B. These intuitions will be confirmed by model simulations later in this chapter.

In most, or all figures of this chapter, we only show a small segment of the actual visual inputs and model responses, and the actual spatial extent of the input and response patterns should be understood to extend spatially well beyond the boundaries of the plotted regions. The model has a periodic or wrap-around boundary condition to simulate an infinitely large visual space; this is a conventional idealization of reality.

5.4.2 Calibration of the V1 model to biological reality

We intend to use the V1 model as a substitute for the real V1 to test whether saliency computations can feasibly be carried out by V1 mechanisms. Thus, we need to ensure that the relevant behaviors of the model resemble those of real V1 as much as possible. This is just like calibrating an experimental instrument in order to be able to trust subsequent measurements taken with this instrument. This does not mean that the model should include parts to model neural spikes and ionic channels on the neural membrane. (Later on in this chapter, in Section 5.8, it will be argued that equations (5.8) and (5.9) give a minimal model for V1's saliency computation.) However, when we use the visual inputs for which the firing rate responses from the real V1 are known, the model neuron's firing rate response, which will be used to predict saliency, should qualitatively resemble the real V1 responses.

More specifically, we examine representative visual input cases (see Fig. 2.24), in which contextual influences in real V1 have been studied. We simulate V1 model responses to these inputs, and compare the average firing rates of model and real V1 units to assess the qualitative resemblance (see Fig. 5.18 and Fig. 5.19). Figure 5.18 A–D model the contextual suppression that was seen physiologically by Knierim and Van Essen (1992). Figure 5.18 E–H model the contextual facilitation that Kapadia et al. (1995) recorded. To make model responses and the real V1 responses agree with each other, a neural circuit containing separate excitatory and inhibitory neurons is employed for the model V1, and a bow-tie pattern of the neural connections has been designed (see Fig. 5.15 B).

The model neurons' responses, in particular their dependence on the input context, varies with the strength or contrast of the input bar on which the contextual influence is being examined. As in physiological data, stronger and weaker input contrast are associated with stronger suppression and facilitation, respectively (see Section 5.4.6).

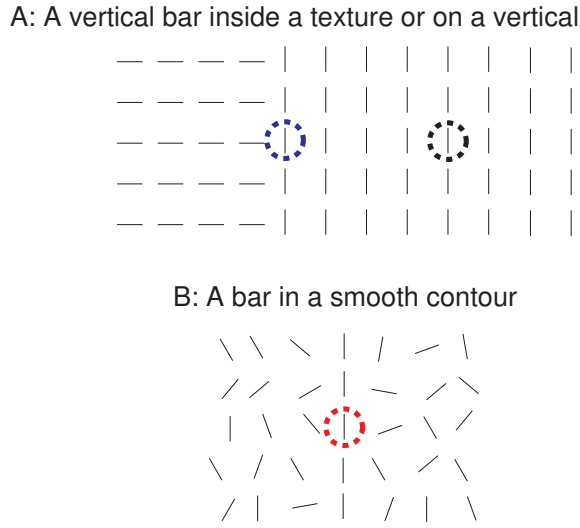


Fig. 5.17: Colinear facilitation and iso-orientation suppression arising from excitation and inhibition between V1 neurons. All three vertical bars, in blue, black, and red dashed circles, respectively (the circles are for illustration; they are not present in the visual input), receive strong monosynaptic excitation, because each is co-aligned with its top and bottom neighboring vertical bars. Meanwhile, these three vertical bars receive different degrees of disynaptic inhibition. Inhibition increases with the number of neighboring bars parallel to, but not co-aligned with, each of them. The bar in the red circle is minimally affected by disynaptic inhibition; the bar in the black circle is maximally affected, and inhibition can overwhelm the monosynaptic excitation.

5.4.2.1 Some conventions in displaying the model behavior

Figure 5.18 also illustrates some conventions used to display the model behavior in many figures of this chapter. To display model input and responses, only a limited spatial range of the locations i of model units $i\theta$'s is shown for illustration. This limited region should be understood as being only a part of an infinitely large image, and the plotted image content should extrapolate beyond the plotted region. (Otherwise, translation invariance of inputs breaks at the outer boundary of the plotted images, and this break should also manifest in substantial non-homogeneities in the response levels.)

In addition, unless otherwise stated explicitly, the model is always simulated in a two-dimensional visual space in a wrap-around or periodic boundary condition. In particular, let location $i = (i_x, i_y)$ of the model neural units $i\theta$ have the horizontal and vertical components i_x and i_y, respectively, in a Manhattan grid, such that $i_x = 1, 2, ..., N_x$ and $i_y = 1, 2, ..., N_y$; then location $i = (i_x = 1, i_y)$ is the horizontal neighbor of location $i' = (i_x = N_x, i_y)$, and location $i = (i_x, i_y = 1)$ is the vertical neighbor of location $i' = (i_x, i_y = N_y)$. Analogous conditions apply if the visual inputs are sampled in a hexagonal grid. Furthermore, N_x and N_y are much larger than the maximum length $|i - j|$ of the lateral connections $\mathsf{J}_{i\theta,j\theta'}$ and $\mathsf{W}_{i\theta,j\theta'}$.

Furthermore, to avoid clutter in plots to visualize model responses, we only show bars whose output responses $g_x(x_{i\theta})$ exceed a threshold. For example, due to the finite width of orientation tuning curves (see equation (5.10)), a bar $\hat{I}_{i\beta}$ at location i in the input image actually provides direct inputs $I_{i\theta}$ to multiple model neurons with similar, but not identical, preferred orientations θ. When input $\hat{I}_{i\beta}$ and contextual facilitations are sufficiently strong,

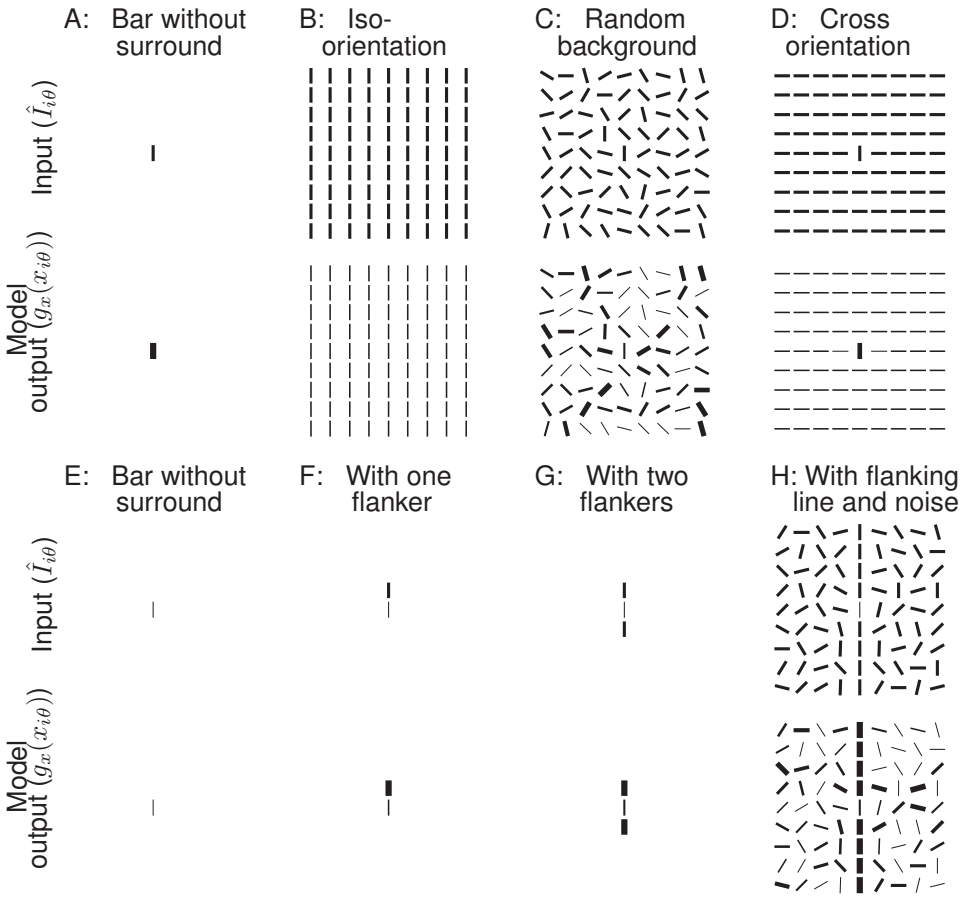


Fig. 5.18: The V1 model qualitatively reproduces representative observations of contextual influences in V1. Each model input pattern has a central vertical (target) bar with or without contextual stimuli. All visible bars are presented at the same high contrast ($\hat{I}_{i\theta} = 3.5$) except for the target bar in E, F, G, H where $\hat{I}_{i\theta} = 1.05$ is near threshold. Input and output strengths are visualized by the widths of the bars, using the same scale in all plots. Isolated high and low contrast bars are presented in A and E. B, C, and D simulate various forms of contextual suppression of the response to the high contrast target. F, G, and H simulate various forms of contextual facilitation of the response to the low contrast target. Note that the response to the near threshold target bar in H is stronger than that to the high contrast target bar in B. Output responses weaker than a threshold are not plotted to avoid clutter. Adapted with permission from Li, Z., Pre-attentive segmentation in the primary visual cortex, *Spatial Vision* 13(1): 25–50, Fig. 3A–3H, copyright © 2000, Koninklijke Brill NV.

more than one model neuron at location i can be activated (making $g_x(x_{i\theta}) > 0$). However, the responses of the less activated bars at this location are often below the threshold we use for visualization, and so these bars do not appear in the plots. Similarly, model input plots are typically plotted according to the values of $\hat{I}_{i\theta}$ (i.e., the actual input image) rather than $I_{i\theta}$ (the direct inputs to individual model neurons).

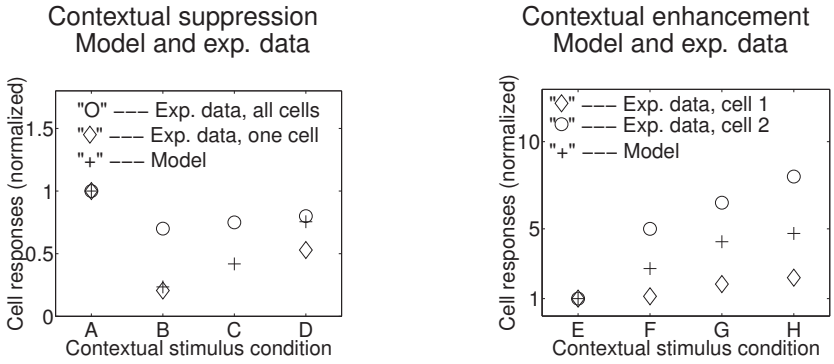


Fig. 5.19: Comparison between the output of the model in Fig. 5.18 and physiological observations. The labels A, B, C, D, E, F, G, and H on the horizontal axes mark the various contextual configurations in the subplots of Fig. 5.18. Responses are normalized relative to the response to the isolated bar. In the left plot, data points "o" and "◇" are taken from Knierim, J.J. and Van Essen, D.C. Neuronal responses to static texture patterns in area V1 of the alert macaque monkey, *Journal of Neurophysiology*, 67(4):961–980, figures 4b and 11, 1992. In the right plot, data points "o" and "◇" are taken from the two cell examples in figures 12B and 12C of Kapadia, M.K., Ito, M., Gilbert, C.D., and Westheimer, G. Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys, *Neuron*, 15(4):843–56, 1995. Adapted with permission from Li, Z., Pre-attentive segmentation in the primary visual cortex, *Spatial Vision*, 13(1): 25–50, Fig. 3I–3J, copyright © 2000, Koninklijke Brill NV.

5.4.3 Computational requirements on the dynamic behavior of the model

The V1 model should be applied to visual inputs which have not been used in physiological experiments. Hence, in addition to calibrating the model to the existing physiological observations, the model should also be designed such that it is well behaved in a manner expected for a visual system that computes saliency appropriately. This imposes the following requirements on the model; some of them also help to ensure that the model is properly calibrated to existing physiological data.

The first requirement is that when the input is not translation invariant, and if the location where the input changes is conspicuous, the model should give relatively higher responses to this location. Figure 5.17 A presents an example for which the orientations of the bars change at the texture border. Elevated responses to the texture border bars highlight the conspicuous input locations, consistent with their higher saliency. As we have argued, this can be achieved by mutual suppression between neurons responding to neighboring iso-oriented bars. Border bars have fewer iso-oriented neighbors and so experience less suppression and have relatively higher responses. Hence, iso-orientation suppression should be sufficiently strong to make the degree of highlights sufficient; indeed as strong as that observed physiologically.

The second requirement is that, when the model is exposed to an homogenous texture, the population response should also be homogenous. In particular, this means that if inputs $I_{i\theta}$ to the model are independent of the spatial location i, then the outputs $g_x(x_{i\theta})$ should also be (neglecting the response noise, which should be such that they do not cause qualitative differences). If this requirement was not satisfied by real V1, then we would hallucinate inhomogenous patterns even when the input image did not contain them, or we would hallucinate

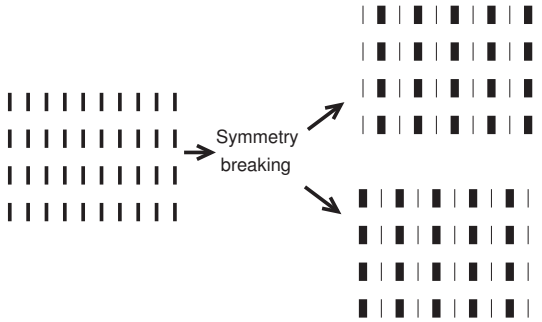


Fig. 5.20: Spontaneous symmetry breaking. Given sufficient mutual suppression between vertical arrays of bars, the output in response to the homogenous texture input (on the left) can evolve to one of the two inhomogenous response patterns on the right. Which pattern will emerge depends on how the initial activities deviate from the homogeneity—an initial deviation (caused by noise) toward one of the final patterns will be amplified to increase the chance of the emergence of the corresponding final pattern. The real V1 avoids such symmetry breaking (in normal conditions); it should therefore also be avoided in the model.

salient locations when there are none. This requirement has to be satisfied in order to obtain the model behavior demonstrated in Fig. 5.18 B.

It may seem that this requirement should be satisfied automatically, since translation invariant (i.e., homogenous) inputs might seem obviously to give rise to translation invariant outputs when the intracortical connections $J_{i\theta,j\theta'}$ and $W_{i\theta,j\theta'}$ are all translation invariant. However, translation invariant dynamical systems are subject to spontaneous symmetry breaking, and so they could generate non-homogenous responses even when fed with homogenous inputs. For instance, a thin stick standing vertically has a strong tendency to fall sideways to one side or another. The symmetric equilibrium position of upright standing is dynamically unstable—a small perturbation of the stick to one side will be amplified further by the dynamics under gravity.

In the V1 model, just as for the case of the stick, a homogenous response to a homogenous texture input pattern (such as the regular texture of vertical bars in Fig. 5.20) is also an equilibrium point in a dynamic system. Again, as for the stick, this equilibrium point can be unstable if the neural dynamics are incorrectly modeled. In particular, iso-orientation suppression between neighboring vertical bars in Fig. 5.20 makes neighboring vertical arrays of bars suppress each other. Consider the case that one array has slightly higher response than the other because of noise included in the dynamics. Then this array will suppress neighboring arrays more; those arrays could then suppress the first array less, making the first array's responses higher still. Thus, the perturbation could be amplified by a form of positive feedback in the dynamics. If this positive feedback is too strong, spontaneous pattern formation occurs, as schematized in Fig. 5.20. The mutual suppression between the arrays is caused by iso-orientation suppression. Intuitively, reducing the strength of this suppression should help reduce the instability.

However, reducing the strength of the iso-orientation suppression will compromise the first requirement to highlight conspicuous input locations where input changes. Hence, there is a conflict between the need to have strong iso-orientation suppression to highlight conspicuous input locations, e.g., at a texture border or a feature singleton, and the need to have a weak iso-orientation suppression in order to prevent spontaneous symmetry breaking to homogenous inputs. Mathematical analysis of the dynamic system of neural circuits, explained in detail

in Section 5.8, shows that resolving this conflict imposes the following requirement on the model's neural circuit: mutual suppression between principal neurons should be mediated disynaptically by inhibitory interneurons, as in the real V1. This circuit requirement precludes implementing iso-orientation suppression by direct inhibition between the principal units, as is often the case in artificial neural networks or computer vision algorithms (such as the Markov random field model).

Thirdly, the strength of mutual excitation between neurons should be limited, in order to prevent ubiquitous non-zero responses of pyramidal neurons to zero direct input given contextual inputs. In particular, the colinear facilitation implied by Fig. 5.18 FGH should not be so strong as to activate a neuron whose most preferred stimulus bar is absent in the input but is an extrapolation of a straight line present in the input image. Otherwise, the visual system would hallucinate the eternal growth of short, unchanging input lines.

If V1 does not create a saliency map in the form proposed by the V1 saliency hypothesis, then the above requirements for a well-behaved model for saliency computation is not expected to be consistent with the requirement that the model being calibrated to sufficiently resemble the real V1 (as in Fig. 5.18 and Fig. 5.19). Nevertheless, a single set of model parameters (presented in the appendix to this chapter; see Section 5.9) has been found that satisfies both sets of requirements, reinforcing the plausibility of the hypothesis that V1 creates a bottom-up saliency map. The design and analysis of the recurrent neural circuit are mathematically somewhat challenging. Hence, I separate the mathematical details into a separate section (Section 5.8) for readers interested in the nonlinear neural dynamics (Li 1999b, Li 2001, Li and Dayan 1999). However, the challenging mathematics is far less formidable than simultaneous *in vivo* recordings from hundreds of primate V1 neurons using visual search stimuli and the current technology in physiological experiments.

5.4.4 Applying the V1 model to visual search and visual segmentation

The model parameters include the neural connections $\mathsf{J}_{i\theta,j\theta'}$ and $\mathsf{W}_{i\theta,j\theta'}$, the activation functions $g_x(.)$ and $g_y(,)$, the neurons' decay constants, the way the model activities are normalized, the local interactions within a hypercolumn, and characteristics of the input noise. Following the design and calibration, all these parameters were fixed (to the values presented in the appendix to this chapter; see Section 5.9), and the model's response to a variety of input stimuli (including stimuli not used for calibration) can be tested.

In particular, we examine representative visual inputs for which the saliency properties, e.g., which locations are salient and how saliency depends on input characteristics, are known from visual experience or behavioral experiments. We compare these saliency properties with those predicted from the responses of the V1 model.[15] These representative inputs and saliency properties are:

1. Images containing orientation singletons, or borders between iso-orientation textures;
2. images contrasting saliencies of visual search targets in feature and conjunction searches;
3. images demonstrating visual search asymmetry;
4. images demonstrating how saliencies depend on input feature contrasts, spatial densities of input items, or regularities of texture elements;
5. images demonstrating the conspicuousness of a hole in the visual input pattern, or of a missing input;

[15] In principle, one could design the model such that the model's predicted saliency behavior agrees with those observed in visual behavior or experience. If so, this agreement should be included as one of the computational requirements for the model in Section 5.4.3. In practice, the model was designed, and its parameters were fixed, without first ensuring this agreement.

6. images containing more complex textures whose boundaries are conspicuous to varying degrees.

Because the model parameters are fixed, the differences in model responses arise solely from the differences in the input stimuli $\hat{I}_{i\theta}$ (and, sometimes, the difference between Manhattan and hexagonal input grids, which we use to sample the input more proficiently).

To illustrate the function of the intracortical interactions, many model simulations use input patterns in which all visible bars $i\theta$ have the same underlying input contrast $\hat{I}_{i\theta}$, such that differential responses to different visible bars can only arise systematically from the intracortical interactions. For each bar element $i\theta$, the initial model response $g_x(x_{i\theta})$ is dictated only by the external inputs $I_{i\theta}$ to this bar. However, due to intracortical interactions, the response $g_x(x_{i\theta})$ is significantly affected by inputs $I_{j\theta'}$ to other bar elements $j\theta'$ within about one membrane time constant $1/\alpha_x$ after the initial neural response. (The current model implementation has the parameter $\alpha_y = \alpha_x$.) This agrees with physiological observations (Knierim and Van Essen 1992, Kapadia et al. 1995, Gallant, Van Essen and Nothdurft 1995), if this time constant is assumed to be of the order of 10 milliseconds (ms).

5.4.4.1 Model behavior, and additional conventions in its presentation, in an example: two neighboring textures

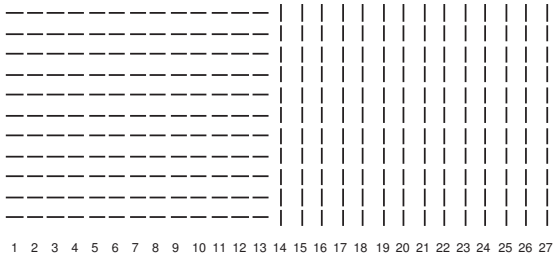
Figure 5.21 shows an example of the temporal evolution of the model responses. The activities of units in each texture column initially rise quickly to an initial response peak and then decrease. The initial responses at time $t = 0.7$ (in units of the membrane time constant, and excluding latency from retinal input to LGN output) after stimulus onset are roughly the same across the columns, since they are mainly determined by the direct, rather than the contextual, input to the receptive fields. By time $t = 0.9$, responses to the horizontal bars near the vertical texture border are relatively weaker than responses elsewhere, because these bars enjoy less monosynaptic colinear facilitation. Neural responses reach their initial peak at around $t = 1.2$. Then, iso-orientation suppression starts to manifest itself. This suppression lags the colinear facilitation since it is mediated disynaptically. The suppression is most obvious away from the texture border, where each bar has more iso-orientation neighbors. The black curve plots the mean responses after input onset as a function of the column, averaging over many cycles of the oscillating neural responses. This curve is another version of the plot in Fig. 5.22 C.

In the rest of this chapter, we generally omit the temporal details of the responses, and so we report just the temporal averages of the neural activities $g_x(x_{i\theta})$ after the model has evolved for several time constants after the onset of the visual input $I_{i\theta}$. (For simplicity, we often use "outputs," "responses," or "$g_x(x_{i\theta})$" to mean the temporal averages of the model pyramidal responses $g_x(x_{i\theta})$.) Further, inputs $I_{i\theta}$ are typically presented to the model at time 0 and persist, unless stated otherwise.

This focus on static inputs and temporally averaged model responses is motivated by the following considerations: (1) most of the behavioral data on saliency are from experiments using static visual images presented for a much longer time duration than the time constant of neurons; (2) even though the initial presentation of the image will lead to a strong impulse of saliency at locations of all image items, the behavioral effects that are typically measured depend on the differences between saliencies at locations of different input items, and these should be most pronounced *after* this impulse has subsided. (Of course, the model can also be applied to temporally varying inputs or asynchronously presented inputs. For example, if all except one item in an homogenous array are presented simultaneously, the temporally unique item, if presented with a sufficiently long delay, should make its location very salient by the saliency impulse associated with its onset.)

For each model simulation, the input contrasts, which are represented by $\hat{I}_{i\theta}$, are adjusted

A: Input image ($\hat{I}_{i\theta}$) to model (texture column numbers at bottom)



B: Responses $g_x(x_{i\theta})$ versus texture columns above at various time since visual input onset, or temporal average responses (black).

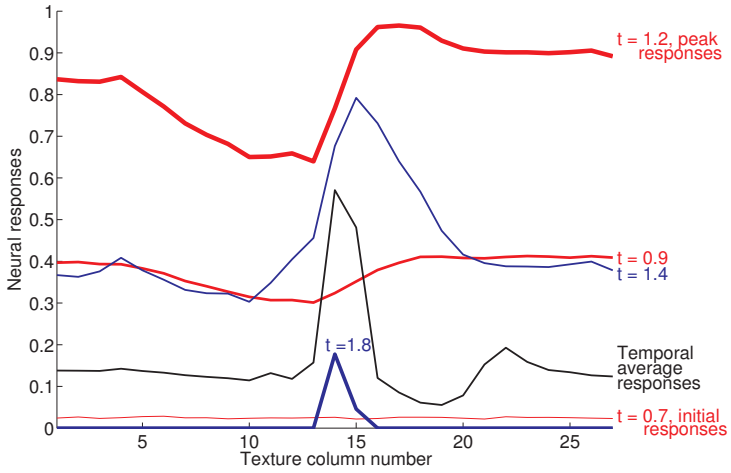


Fig. 5.21: The temporal evolution of the model responses to an input pattern. A: The input pattern contains two regions (excluding the texture column numbers indicated at the bottom); each visible bar has the same input strength $\hat{I}_{i\theta} = 2.0$. The input pattern is presented at time $t = 0$ and remains presented thereafter. Only 11 rows by 27 columns of the input bars are plotted out of a larger image (of 22 rows by 60 columns, using wrap-around boundary conditions). B: The response traces are the average of $g_x(x_{i\theta})$ across positions i within the same texture column, where θ is the orientation of the input bars in the column. Red and blue curves plot responses at various time points t (indicated to the right of each curve) during the rising and decaying phases, respectively, of the initial phase of the responses. Time t is in the units of membrane time constant, and it excludes the latency from retina input to LGN output. The black curve plots the responses averaged over a duration from $t = 0$ to $t = 12$. The initial responses (at $t = 0.7$) are not context dependent; but contextual influences are apparent within half a time constant after the initial response.

to mimic the corresponding conditions in physiological and psychophysical experiments. In the model, the input dynamic range is $\hat{I}_{i\theta} = (1.0, 4.0)$, which will allow an isolated bar to drive an excitatory neuron from threshold activation to saturation. Hence, low contrast input bars, which are typically used to demonstrate colinear facilitation in physiological experiments, are represented by $\hat{I}_{i\theta} = 1.05$ to 1.2. Intermediate or high contrast inputs (e.g., $\hat{I}_{i\theta} = 2 - 4$)

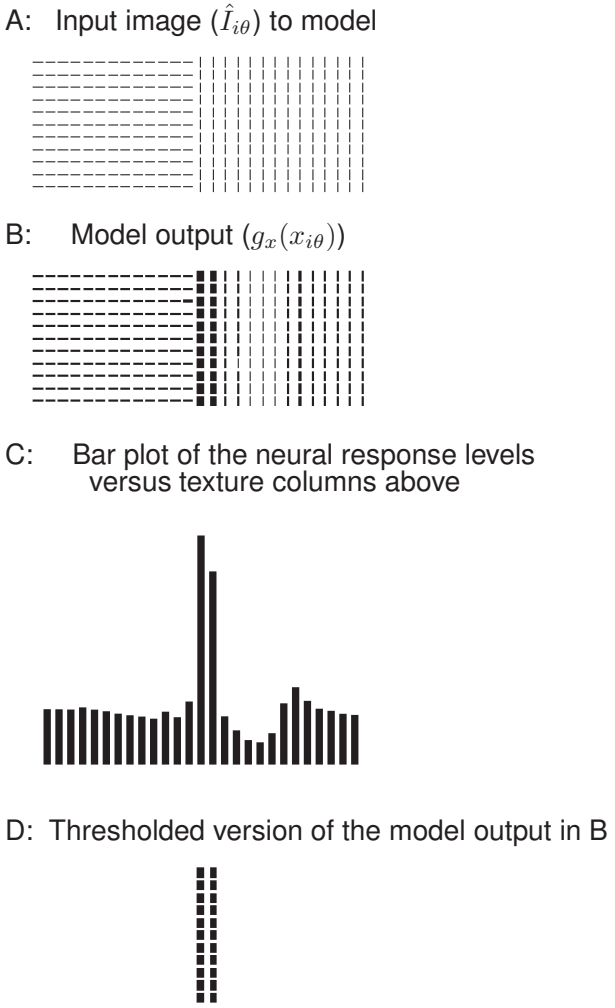


Fig. 5.22: Texture segmentation. A is the same as the input pattern in Fig. 5.21 A. B: Model output responses to A, i.e., temporal averages of $g_x(x_{i\theta})$ for the bars. C: The average model response in each column in B (considering only the most responsive neuron at each texture element location) is represented by the height of the bar for each column location. This plot shows the same information as the black curve in Fig. 5.21 B. D: The result of applying a threshold of half of the maximum response among all bars to the responses $g_x(x_{i\theta})$ in B. Adapted with permission from Li, Z., Visual segmentation by contextual influences via intracortical interactions in primary visual cortex, *Network: Computation in Neural Systems*, 10(2): 187–212, Fig. 3, copyright © 1999, Informa Healthcare.

are used for all the visible bars in other input images, including those illustrating texture segmentation and feature pop-out. Meanwhile, the neural output $g_x(x_{i\theta})$ ranges from 0 to 1.

Figure 5.22 B further illustrates the model response to the same input as in Fig. 5.21 by showing the average responses $g_x(x_{i\theta})$ for a substantial patch of the input texture. Figure 5.22 C plots the (temporal averaged) responses $g_x(x_{i\theta})$ to the bars averaged in each column in Fig. 5.22 B. It shows that the most salient bars are indeed near the region boundary. Figure

5.22 D confirms that the boundary can be identified by thresholding the output responses using a threshold parameter, $thresh = 0.5$, set to be a proportion of the maximum response to the image, to eliminate weak outputs that would otherwise clutter the figure. Thresholding is not performed by V1 but is only used for visualization.

According to the V1 saliency hypothesis, the visual locations surviving the thresholding are more likely to be selected first by bottom-up mechanisms. If the diameter of the attentional spotlight is smaller than the length of the texture border, then only a part of the border can be selected first. We might reasonably assume that the reaction time for an observer to complete the overall task of segmenting two neighboring textures decreases with the time it takes until any part of the border is first selected. Thus, we can use this reaction time to probe the saliency of a texture border, without addressing how the full task of segmentation is completed after the selection of only a small part of the texture border.

Here we also briefly point out something beyond the scope of investigating saliency. In Fig. 5.22 B, the response highlights are not distributed symmetrically around the texture border. This could make the viewers perceive the location of the texture border as being biased slightly to the right of the border. This has indeed been observed psychophysically (Popple 2003), although there may be additional causes for such biases beyond V1, such as the perception of figure and ground. This is a demonstration that the V1 saliency mechanisms make V1 responses distort the visual input image. The ultimate percept is likely the outcome of additional processing based on the V1 responses.

5.4.4.2 Assessing saliency from model V1 responses: illustrated by the effect of the orientation contrast at a texture border

According to equation (5.4), the saliency value at each location is the highest (pyramidal) response to inputs at that location. A location in the V1 model is denoted by i, and various neurons $i\theta$ give responses $g_x(x_{i\theta})$. Therefore, saliency value at location i is

$$\text{SMAP}_i \equiv \max_\theta \left[g_x(x_{i\theta}) \right]. \tag{5.11}$$

As discussed in Section 5.2.1, these pseudo-saliency values at various locations need to be compared with each other in order to determine the most salient location in an image. The actual saliency value of a location should reflect this comparison. For this purpose, let

$$\begin{aligned} \bar{S} &\equiv \text{the average of SMAP}_i \text{ over } i \quad \text{and} \\ \sigma_s &\equiv \text{the standard deviation of SMAP}_i \text{ over } i \end{aligned} \tag{5.12}$$

be the mean and standard deviation of the SMAP_i values at all locations i, or alternatively, at all locations i with non-zero neural responses. The salience of a location i can then be assessed by

$$r_i \equiv \frac{\text{SMAP}_i}{\bar{S}} \quad \text{and} \quad z_i \equiv \frac{\text{SMAP}_i - \bar{S}}{\sigma_s}. \tag{5.13}$$

In our plots of the model responses, quantities r can be visualized by the thickness of the plotted output bars. Meanwhile, z models the psychological z score.

The quantities $\bar{S}$ and σ_s in equation (5.12) could alternatively be defined as

$$\begin{aligned} \bar{S} &\equiv \text{average of } g_x(x_{i\theta}) \text{ over } (i,\theta) \quad \text{and} \\ \sigma_s &\equiv \text{standard deviation of } \ g_x(x_{i\theta}) \ \text{ over } (i,\theta). \end{aligned} \tag{5.14}$$

This alternative is conceptually and algorithmically simpler, since it omits the intermediate step of obtaining $\text{SMAP}_i = \max_\theta \left[g_x(x_{i\theta}) \right]$, in which neurons are grouped according to their receptive field location i. Using the alternative should only make quantitative rather than

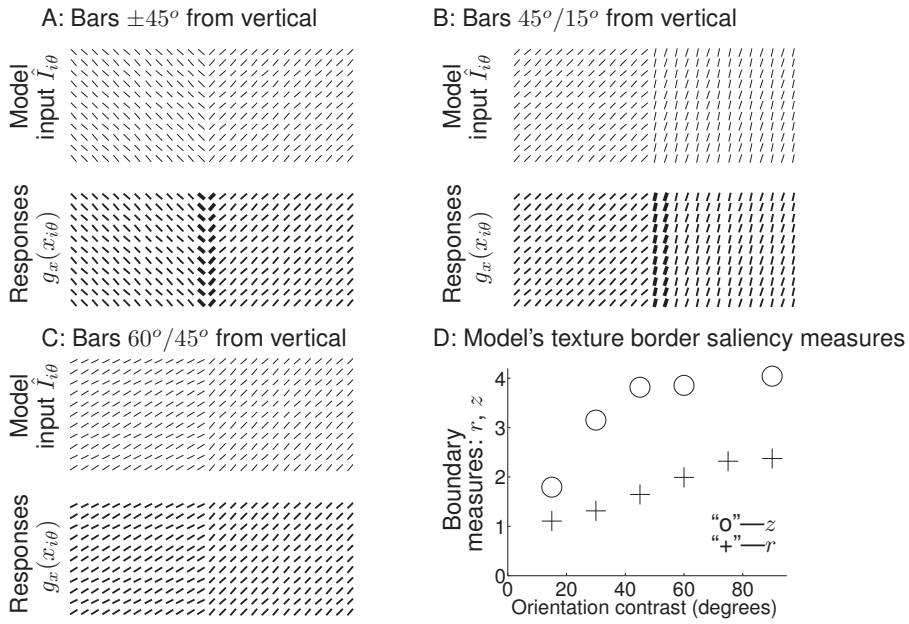


Fig. 5.23: A, B, C: Additional examples of model behavior at orientation texture borders. Each example contains two neighboring textures, in which texture bars have orientation θ_1 and θ_2, respectively, meeting in the middle at a vertical border. In A, B, and C, the saliency measures for the borders are $(r, z) = (1.4, 3.4)$, $(r, z) = (1.7, 3.7)$, and $(r, z) = (1.03, 0.78)$. D: Texture border saliency measures r, z (indicated by "+" and "o", respectively) from the model as a function of the orientation contrast at the border. Each data point is the averaged measure from borders of all possible pairs of θ_1 and θ_2 for a given $|\theta_1 - \theta_2|$. The most salient column in B is in fact the second left column in the texture region on the right. In C, the texture border is barely detectable without close scrutiny. Although the texture border bars are among the most salient ones, their evoked responses are only slightly ($\sim 10\%$) higher than those of the other bars (this is imperceptible in the line widths shown in the output). Adapted with permission from Li, Z., Visual segmentation by contextual influences via intracortical interactions in primary visual cortex, *Network: Computation in Neural Systems*, 10(2): 187–212, Fig. 4, copyright © 1999, Informa Healthcare.

qualitative difference to r and z. In this book, the r and z values are obtained by using $\bar{S}$ and σ_s in equation (5.12), with the locations i used to obtain the mean $\bar{S}$ and σ_s only including the locations which have non-zero responses $g_x(x_{i\theta})$ for at least one θ.

To assess the saliency of a texture border, we replace the SMAP_i in equation (5.13) by the average SMAP_i in the most salient grid column parallel to, and near, the texture boundary. A salient texture border should give large values for (r, z). For instance, in Fig. 5.22, $(r, z) = (3.7, 4.0)$ at the texture border.

V1 does not (and does not need to) calculate r and z. These two values just help us characterize the saliencies of visual locations in order to compare them with our visual experience or behavior, e.g., to see whether locations with high r and z values indeed correspond to the locations that are more conspicuous. In particular, locations with smaller z scores are expected to take longer to select, due to the competition between multiple locations for selection. A z score larger than three makes a location quite salient and indeed, likely to be the most salient

in the scene. An example is the texture border in Fig. 5.22. Meanwhile a location with $z \sim 1$ is not so salient, even if it has the largest z score in the scene.

Consider applying these tools to the examples of orientation textures shown in Fig. 5.23. One can see for oneself how conspicuous is each texture border. Texture borders with orientation contrasts of 90^o (Fig. 5.23 A) or 30^o (Fig. 5.23 B) are quite conspicuous, i.e., salient. However, a border with an orientation contrast of only 15^o (Fig. 5.23 C) is rather difficult to notice without scrutiny. These observations agree with the model's z scores. The z score for this 15^o contrast border is indeed only $z = 0.78$. Other 15^o contrast borders will lead to higher z scores—for instance, if one texture comprises vertical bars, and the other texture comprises bars that are 15^o clockwise from vertical.

Psychologically, the just-noticeable orientation contrast for a texture border to be detected quickly is indeed about 15^o. In this model, a border with a 15^o orientation contrast has an average $z \approx 1.8$ (averaged over all possible orientations θ_1 and θ_2 for the bars in the two textures, given $|\theta_1 - \theta_2| = 15^o$); see Fig. 5.23 D. This is expected for a border with only a moderate saliency psychophysically.

The dependence of the border saliency on the orientation contrast is mainly caused by the decrease in the suppression between two neighboring bars as the orientation difference between them increases. This suppression is strongest between parallel, but not co-aligned, bars, it remains substantial when the two bars are similarly but not identically oriented, and it is much reduced when the two bars are orthogonal to each other. This is reflected in the bow-tie connection pattern between the V1 neurons shown in Fig. 5.15 B, and it is manifest in the contextual influences that are observed physiologically; see Fig. 5.18 ABD.

Henceforth, the model saliency of a visual location i is assessed by the z score only. In particular, the z score for the location of a target of a visual search will be assessed this way, to link with psychophysical data on visual search tasks (Li 1997, Li 1999b, Li 1999a, Li 2002).

5.4.4.3 Feature search and conjunction search by the V1 model

Figure 5.24 demonstrates the model's behavior for a feature search and a conjunction search. The same target "$\neq$" is presented in two different contexts in Fig. 5.24 A and Fig. 5.24 B. Against a texture of "$\not|$", it is highly salient because its horizontal bar is unique. Against a texture of "$\not|$" and "$\diagdown$", it is much less salient because only the conjunction of "—" and "$\diagup$" distinguishes it. This is consistent with psychophysical findings (Wolfe et al. 1989, Treisman and Gelade 1980). In the V1 model, the unique horizontal target bar in Fig. 5.24 A leads to the response of a V1 neuron that is not subject to iso-orientation suppression. All the other input bars are suppressed in this way. Thus, the horizontal bar evokes the highest response among all V1 neurons and makes the target location salient. Meanwhile, in Fig. 5.24 B, the V1 responses to both bars in the target suffer from iso-orientation suppression, just like all the other bars in the image. Hence, neither of the bars in the target evokes a response that is significantly greater than typical responses to the other bars, and so the location of the target is not salient.

Therefore, V1 mechanisms can be the neural substrate underlying the psychological "rule" that feature searches are typically easy and conjunction searches are difficult (Treisman and Gelade 1980).

Two kinds of feature tunings

Our observations suggest the following: consider a visual characteristic such as orientation or color that psychophysical rules deem to be a "feature" dimension, supporting such phenomena as easy or efficient search. Then, we can expect two neural properties to be tuned to feature values in this feature dimension. The first is that (the responses of) some V1 neurons should be

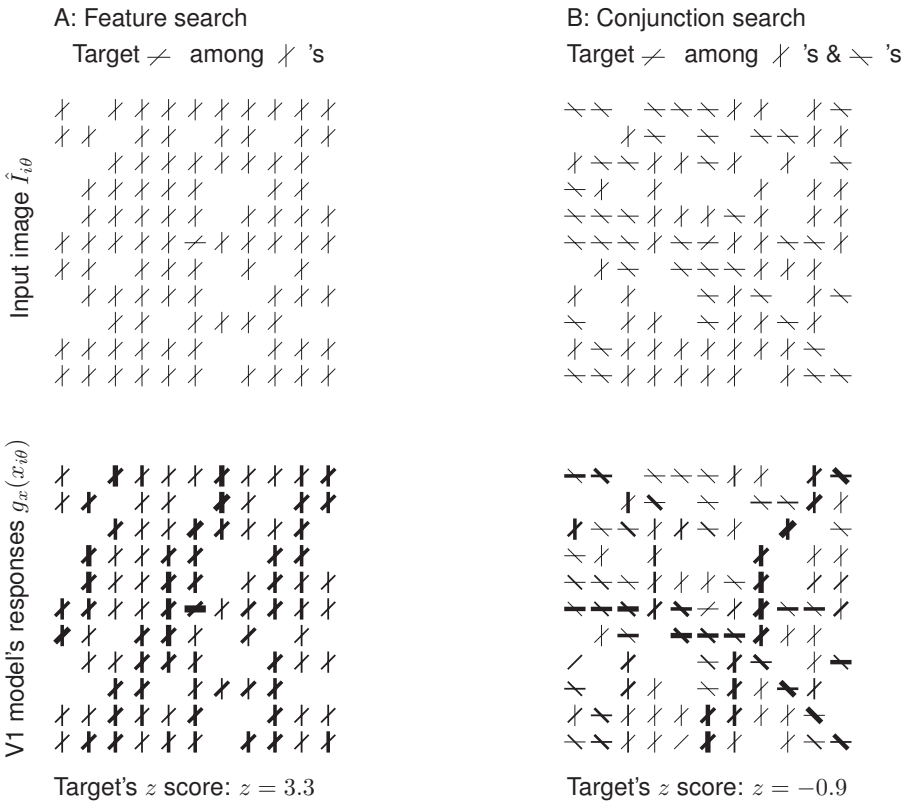


Fig. 5.24: The behavior of the model in feature (A) and conjunction (B) searches. Stimulus (top), model responses (bottom), and the z scores for the targets (displayed in the center of each pattern for convenience) are shown for the two examples. The target in both A and B is made of a horizontal bar and a 45^o (tilted clockwise from vertical) oblique bar intersecting each other. A: The target is unique in having a horizontal bar, making it a case of orientation feature search, and leading to a high z score, $z = 3.3$. B: Each target feature, i.e., the horizontal or the oblique bar, is present in the distractors; these differ from the target only in the conjunctions of the two orientations. This leads to a low z score for the target, $z = -0.9$.

tuned to this dimension—this is of course a classical concept. Orientation tuning is an example; it enables some V1 neurons to signal the saliencies caused by their preferred features.

The other neural property that should be tuned is the intracortical connection pattern between V1 neurons, such that the strength of the intracortical connection between neurons roughly decays with the difference between the preferred features of the two neurons. In other words, the intracortical connections are tuned to the preferred features of the linked V1 cells. The most critical aspect of this tuning should be iso-feature suppression. Tuned intracortical suppression makes a feature singleton salient. There can also be a feature tuning width analogous to that in the neural response tuning to features; this tuning width should be compatible to the minimum feature difference necessary between a feature singleton and the background feature values to make the singleton sufficiently salient. For example, an orientation singleton can be viewed as having a sufficiently unique orientation for salient pop-out if the intracortical connections between neurons most activated by the singleton and background features are absent or insignificant.

The V1 model explains the results of the feature and conjunction search tasks in Fig. 5.24 without any explicit representation of the conjunctions between features. According to the argument above, a lack of explicit representation of the conjunction, i.e., a lack of tuning to the conjunction feature, prevents the conjunction feature from behaving like a basic feature in terms of saliency. Therefore, a target whose uniqueness is only defined by a feature conjunction cannot be salient. On the other hand, the target in Fig. 5.24 A is salient not because the whole object item "$\not\prec$" is recognized or signaled by a single neuron; instead, it is salient because one of its component features, namely the horizontal bar, is a unique basic feature and is sufficiently salient to attract attention strongly by itself. As far as saliency is concerned, the oblique bar in the target is not visible to the saliency system, which only looks at the highest response at each location.

In Fig. 5.24, the background items are not spatially uniform or regular, and so responses to the background bars are not uniformly low. The response to each bar is determined by its particular contextual surround. An accidental alignment of a given bar with its local context facilitates (or at least reduces the suppression of) the final response. On the other hand, if the bar has more iso-orientation neighbors with which it is not aligned, then the response will be more greatly suppressed. Despite the heterogeneity in the population responses, the response to the target horizontal bar in Fig. 5.24 A is still substantially higher than most of the background responses, making the ultimate z score high.

5.4.4.4 A trivial example of visual search asymmetry through the presence or the absence of a feature in the target

Given the observation that a feature search is easier than other searches, it is straightforward to understand the simple example of visual search asymmetry in Fig. 5.25. Search asymmetry is the phenomenon that the ease of a visual search can change when the target and distractors are swapped—for instance, searching for a cross among vertical bars is easier than vice versa. The target cross is easier to find in Fig. 5.25 A because it can be found by feature search. The horizontal bar in the cross is the unique feature, and it evokes the highest V1 response since it is the only one which lacks iso-orientation neighbors. Meanwhile, in Fig. 5.25 B, the target fails to possess any unique feature lacking in the non-targets; hence, it cannot be found by feature search. The target vertical bar and the vertical bars in the background crosses are almost equally suppressed; thus the target's z score is too low for it to pop out.

As in Fig. 5.24 A, the target cross in Fig. 5.25 A is easier to find not because the whole cross is recognized; instead, it is because one of its components, namely the horizontal bar, evokes the highest overall V1 response. Its other component, the vertical bar, does not contribute to the z score for the target.

Note that the search asymmetry between the cross and the vertical bar cannot be predicted from the idea that the ease of finding a target depends on how different it is from the distractors, since this difference does not change when target and distractors swap identity. A long-standing psychological rule (Treisman and Gelade 1980) is that a target having an unique (basic) feature which is lacking in the non-targets (as in Fig. 5.25 A) is easier to find than a target defined by lacking a (basic) feature which is present in the non-targets (as in Fig. 5.25 B). We suggest V1 saliency mechanisms provide the neural substrate of this rule.

5.4.4.5 The ease of a visual search decreases with increasing background variability

The formula for a search target's z score, $z = (\mathrm{SMAP}_i - \bar{S})/\sigma_s$, suggests that increasing σ_s, by increasing the heterogeneity of the responses to non-targets, should decrease the target's z score when the target is at least minimally salient, i.e., when its highest evoked response SMAP_i is above the average response $\bar{S}$ to the scene. This is demonstrated in Fig. 5.26. A

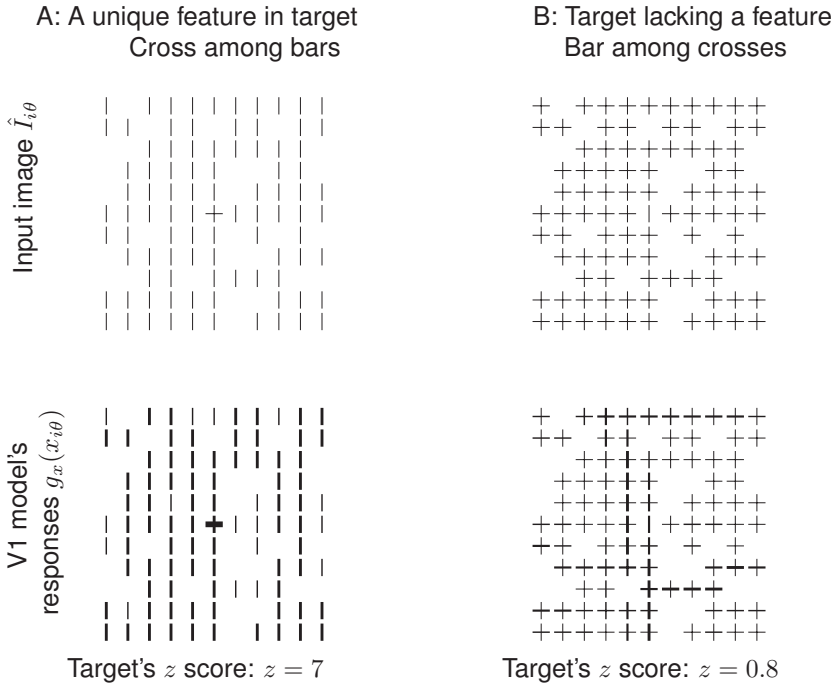


Fig. 5.25: A simple example (Li 1999b, Li 1999a) of search asymmetry in the V1 model. Searching for a cross among vertical bars (A) is easier than searching for a vertical bar among crosses (B). This figure is shown using the same format as that in Fig. 5.24. These examples also demonstrate that a target is easier (or more difficult) to find when it is defined by having (or lacking) a feature (e.g., the horizontal bar) that is absent (or present) in the distractors. The horizontal bar in the target in A is the only one in the image to evoke a V1 response that is not suppressed by iso-orientation suppression; the target vertical bar in B, however, suffers the same iso-orientation suppression experienced by other vertical bars.

target's saliency according to the model decreases when the non-targets are more variable, either because the non-targets are irregularly positioned in space, as in Fig. 5.26 A, or because the non-target feature values are heterogeneous, as in Fig. 5.26 B. Psychological observations have previously led to the rule that a target is more difficult to find when the background variabilities increase in these ways (Duncan and Humphreys 1989); and it has been suggested that random background variability acts as noise and limits the performance of visual search (Rubenstein and Sagi 1990).

Contextual influences can arrange for two identical visual items to evoke different V1 responses when in different contexts. This effect underlies the heterogeneous responses to non-targets in Fig. 5.26 A. Meanwhile, heterogeneous non-targets placed in a regular grid, as in Fig. 5.26 B, also evoke heterogeneous responses, since the contextual influences depend on the feature similarity between neighboring input items.

The model responses in Fig. 5.26 AB are more heterogeneous than those in Fig. 5.26 C. Therefore σ_s is larger in Fig. 5.26 AB. For example, if the maximum response SMAP_i to the target at location i is 10% above the average response $\bar{S}$, it will still stand out, making the target very salient if no other item in the scene evokes a response more than 5% above $\bar{S}$. However, if the background responses vary between 50% to 150% of the average $\bar{S}$, the target

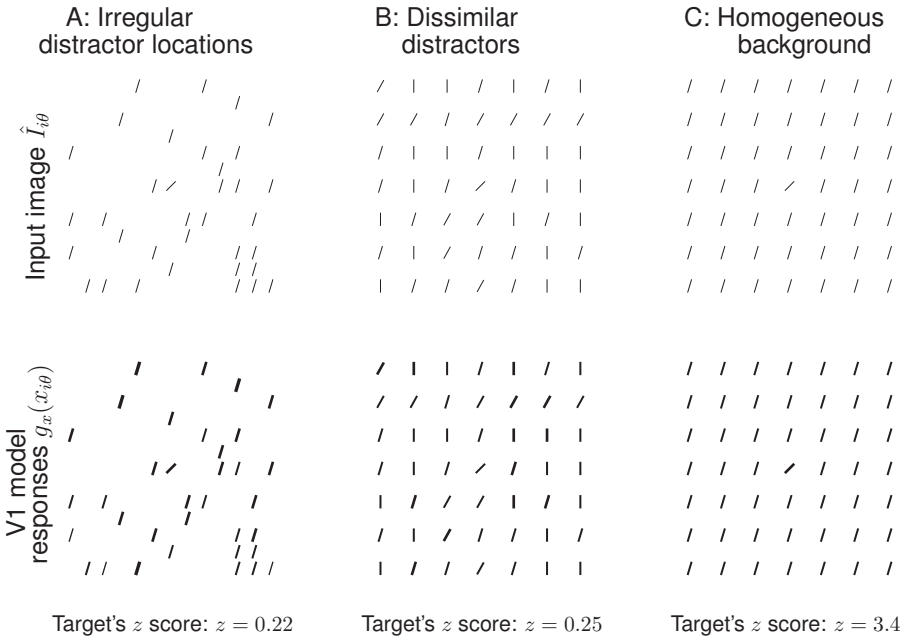


Fig. 5.26: The effect in the model of background variability on the saliency of a target (Li 2002). A, B, and C show visual search images and model responses. The target bar, tilted 45^o clockwise from vertical (and shown in the center of each image for convenience), is among distractors, which are irregularly placed identical bars tilted 15^o clockwise from vertical (A), or regularly placed bars randomly drawn from a selection of those tilted 0^o, 15^o, or 30^o clockwise from vertical (B), or regularly placed identical bars tilted 15^o clockwise from vertical (C). The z scores for the targets are listed immediately below each example.

would not be salient, since a response of only 10% above the average would be comparatively mediocre.

Of course, if the $\text{SMAP}_i < \bar{S}$, the target is not at all salient anyway, regardless of the variability in the background responses.

5.4.4.6 Saliency by feature contrast decreases with a decreasing density of input items

Contextual influences are mediated by intracortical connections, which are known to extend over only a finite range. These influences are thus reduced when the visual input density decreases, since this reduces the number of contextual neighbors within the reach of each visual input item via the intracortical connections. In turn, this reduces many saliency effects. For instance, it is apparent in the images in Fig. 5.27 that it is more difficult to segment two neighboring textures when the texture density is lower. This has also been observed in more rigorous behavioral experiments (Nothdurft 1985). The V1 model shows the same behavior (the right column of Fig. 5.27). The ease of the segmentation is reflected in the highest z score among the texture columns near the texture border. This z score is $z = 4.0$ in the densest example in Fig. 5.27 A, and it is $z = 0.57$ in the sparsest example in Fig. 5.27 D, which is quite difficult to segment without scrutiny.

To be concrete, iso-orientation suppression is weaker in sparser textures. Therefore, the dependence of V1-evoked responses on contextual inputs is weaker in sparser textures, and

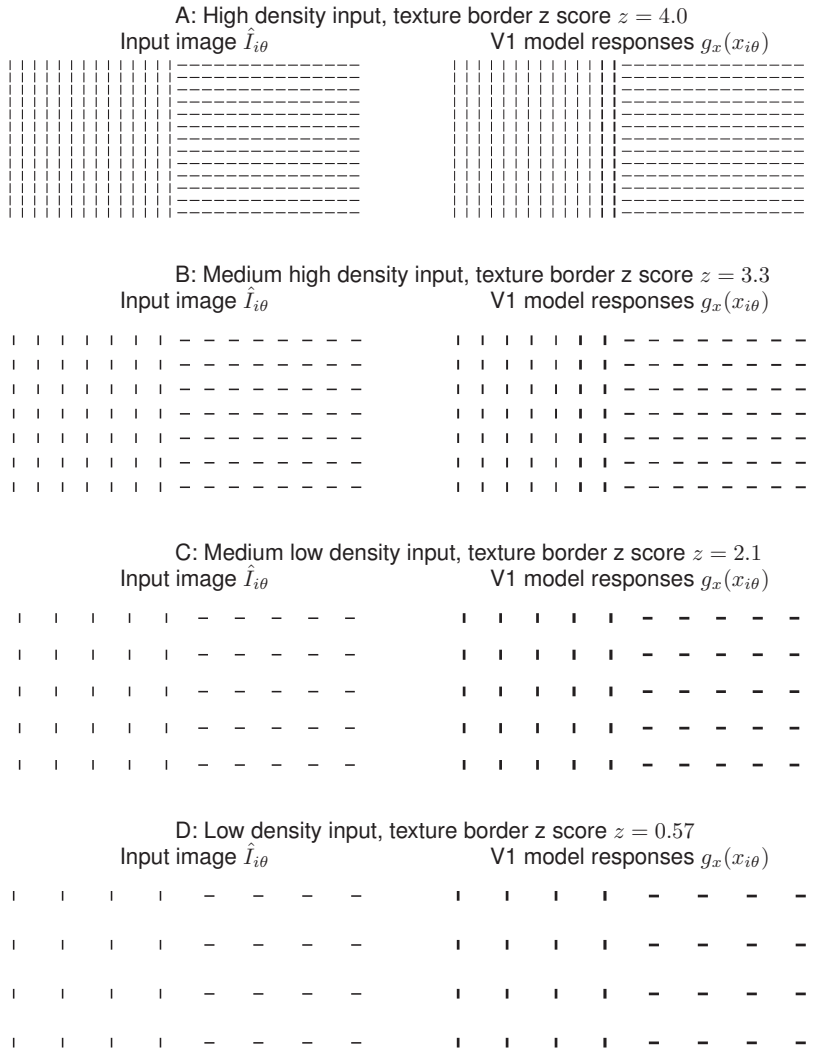


Fig. 5.27: Texture segmentation is more difficult in sparser textures. This is evident from examining the input images, and from the z scores of the texture columns at the borders that are obtained from the V1 model's responses (shown in the right column). All texture bars have input value $\hat{I}_{i\theta} = 2.0$. The average responses $g_x(x_{i\theta})$ to all texture bars are 0.15 (A), 0.38(B), 0.56(C), and 0.54(D).

so the response to a texture bar will be less sensitive to the proximity of this bar to a texture border. More specifically, the highlight at a texture border is caused by the difference between the contextual suppression of the border bars and that of the non-border bars (as explained in Fig. 5.17 A). When the distance between any two texture bars is longer than the longest intracortical connection, there should be zero iso-orientation suppression and so no saliency highlight at the texture border. For each background texture bar, the strength of iso-orientation suppression is largely determined by the number of iso-orientation neighbors that are within reach of the intracortical connections responsible for the suppression. Denser textures provide more iso-orientation neighbors to make this suppression stronger, making the texture border

more salient. Indeed, in Fig. 5.27, the average response to all the texture bars is lowest in the densest texture and higher in sparser textures. This argument also applies to the saliency of a feature singleton in a homogenous background texture. Indeed, such a singleton is easier to find in denser textures (Nothdurft 2000).

5.4.4.7 How does a hole in a texture attract attention?

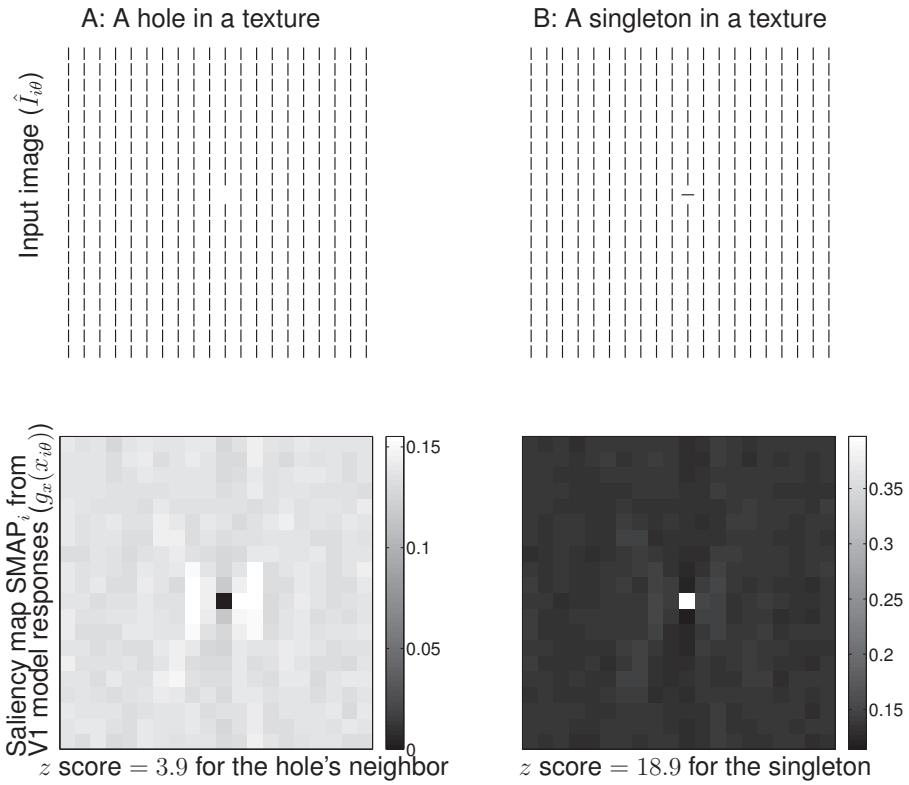


Fig. 5.28: Comparison between the conspicuousness of a hole (A) and a singleton (B). This figure uses the same format as in previous figures, except that the model responses are visualized by grayscale images, in which the gray level at each pixel i represents the maximum response magnitude SMAP$_i$ according to the scale bar on the right of the plot. (Gray scales rather than widths of the bars are used to visualize model responses, since otherwise the small but significant differences in the responses in A would be difficult to manifest as differences in the widths of the bars.) The two grayscale plots have different scale bars, although the average SMAP$_i$ values across i are similar around SMAP$_i \sim 0.136$. Much of the fluctuations in the responses further away from the hole or the singleton are caused by the input noise. In A, attention can be guided to the hole by first being attracted to its most salient neighbor.

It is apparent from Fig. 5.28 A that a hole in a texture is also conspicuous when the background is homogenous. Since a hole, or a missing bar in a texture, does not evoke any V1 response, how can its location attract attention? This can be understood from the observation that the hole still destroys the homogeneity of the texture. In particular, the bars near the hole are subject to weaker iso-orientation suppression because they have one fewer iso-orientation neighbor

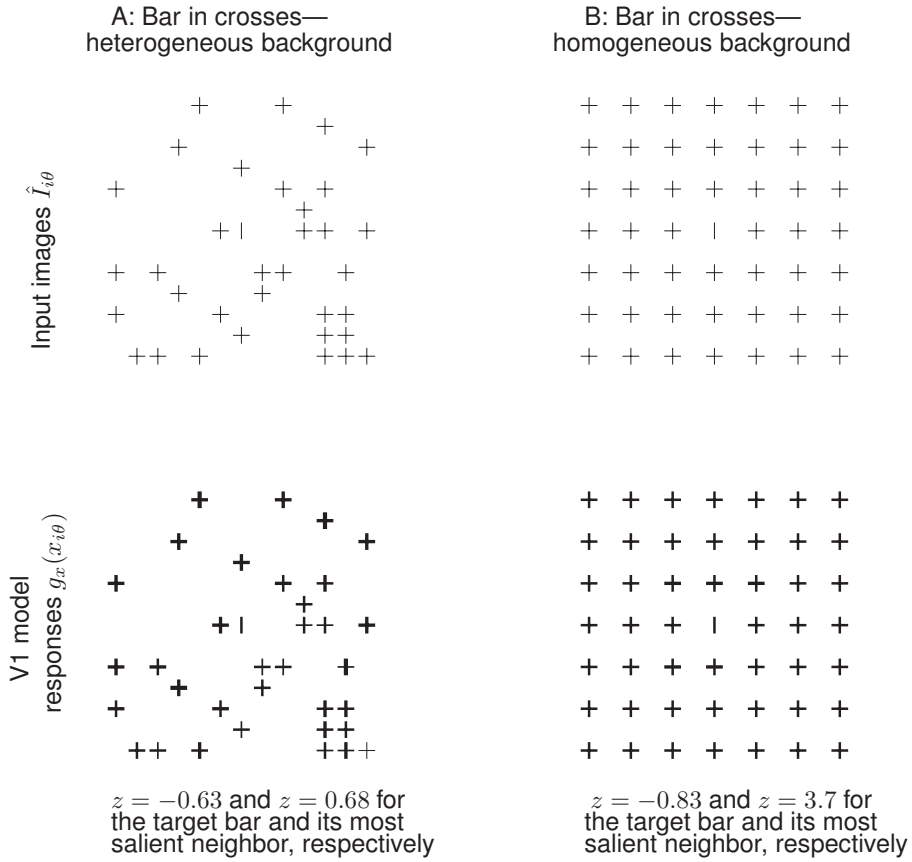


Fig. 5.29: Two additional examples of a target bar in distractor crosses (Li 2002), which are analogous to a hole in a texture as in Fig. 5.28 A. The distractor crosses are more regularly placed in B than A. Although the z score of the target vertical bar is higher in A than B, the most salient neighbor of the target bar has a higher z score in B than A. This underpins the observation that the target is more conspicuous in B than A, guided by the salient neighbor.

due to the hole. Although the suppression is reduced by only a small fraction, this fraction can generate a sizable z score when the background responses are sufficiently homogenous. In the example of Fig. 5.28 A, the mean and standard deviation of the responses over all the texture bars are 0.136 and 0.005 respectively. Meanwhile the response to the most salient neighbor of the hole is 0.155, giving this neighbor a z score of $z = (0.155 - 0.136)/0.005 = 3.9$. This salient neighbor attracts attention; although this attraction is weaker than that of an orientation singleton in the same background texture (Fig. 5.28 B). If the size of the attentional window is sufficiently large (as is suggested by experimental data (Motter and Belky 1998)), the hole can be contained within this window centered on the salient neighbor. Consequently, it may appear to awareness that our attention is attracted by the hole.

From the above interpretation, one prediction is that, in a visual search for a hole, gaze might land on a neighbor of the hole before making a corrective saccade to land on the target. Another prediction is that the conspicuousness of the hole can be manipulated by manipulating the input strength of its neighbors. In particular, the hole would be less conspicuous if its neighbors have slightly weaker input strength than those of the background texture elements.

This prediction has been supported by some preliminary observations (Zhaoping 2004, Zhou and Zhaoping 2010).

If the background texture is not so homogenous, as in the case of Fig. 5.64 B in which the non-homogeneity is caused by multiple holes randomly distributed in the texture, then the z score would be lower and the hole would be less conspicuous. In such cases, the missing input at the hole may be viewed as having been filled-in because it escapes attention. Note that this form of filling-in is not caused by a response to the hole, as would happen if there was a texture element at the location of the hole. This will be discussed more when analyzing Fig. 5.64.

Looking for a hole in a texture can be viewed as a special case of searching for a target lacking a feature that is present in the non-targets. Therefore it is natural that searching for a hole is more difficult than searching for a singleton target defined by the presence of a feature. This is seen in Fig. 5.28: the singleton target in the same texture generates a much higher z score. In the example of a target bar among crosses in Fig. 5.25 B, the target bar's z score $z = 0.8$ is in fact lower than the z score $z = 1.4$ of its left neighbor, although this more salient neighbor is not as salient as the horizontal bar in the target cross in Fig. 5.25 A. In general, the neighbors of a target lacking a feature present in the non-targets are not necessarily more salient than the target, because the actual responses depend on the contextual configurations of the visual input.

Figure 5.29 shows two additional examples of a bar among background crosses. In both examples, the z scores of the target location are negative, indicating that the responses to the target location are below the average responses (maximized at each location) at the locations of other visual items. Comparing Fig. 5.29 A and Fig. 5.29 B, the target has a higher z score in the former but appears to attract attention more strongly in the latter. This is because the most salient neighbor of the target has a higher z score $z = 3.7$ in the latter. The responses to the horizontal bars above and below the target vertical bar in Fig. 5.29 B are slightly higher than most of the other responses, because the missing horizontal bar in the target reduces the iso-orientation suppression on these neighboring horizontal bars by a small but significant fraction.

So far, all the examples of behavior of the model can be more or less intuitively and qualitatively understood from iso-feature suppression, which is the dominant intracortical interaction in V1. This intuition has been used to understand feature versus conjunction searches, search asymmetry between cross and bar, and the saliency effects by texture density, input heterogeneity, a hole, and the orientation contrast between textures. The model simulations merely confirm our intuitive understanding. However, it is desirable to test whether V1 mechanisms can also explain more complex and subtler saliency effects that cannot be intuitively or qualitatively understood from only the effects of iso-orientation suppression. Therefore, we next apply the V1 model to some complex examples, and we will see that these subtler saliency effects are often the net outcome from multiple balancing factors.

5.4.4.8 Segmenting two identical abutting textures from each other

Figure 5.30 A shows that the V1 model responses can even highlight a texture border between two identical textures. Perceptually, the texture border in Fig. 5.30 B seems more salient than that in Fig. 5.30 A, as if there were an illusory vertical border cutting between the two textures. However, the V1 model provides a z score that is somewhat larger for the texture border in Fig. 5.30 A. The reason for this may be that the perception of the illusory contour, rather than saliency, is more likely to arise in V2 rather than V1, as suggested by experimental data (von der Heydt et al. 1984, Ramsden, Hung and Roe 2001). The perception of the illusory contour could be mistaken as the saliency effect.

In each of these examples, all texture bars have about the same number of iso-orientation

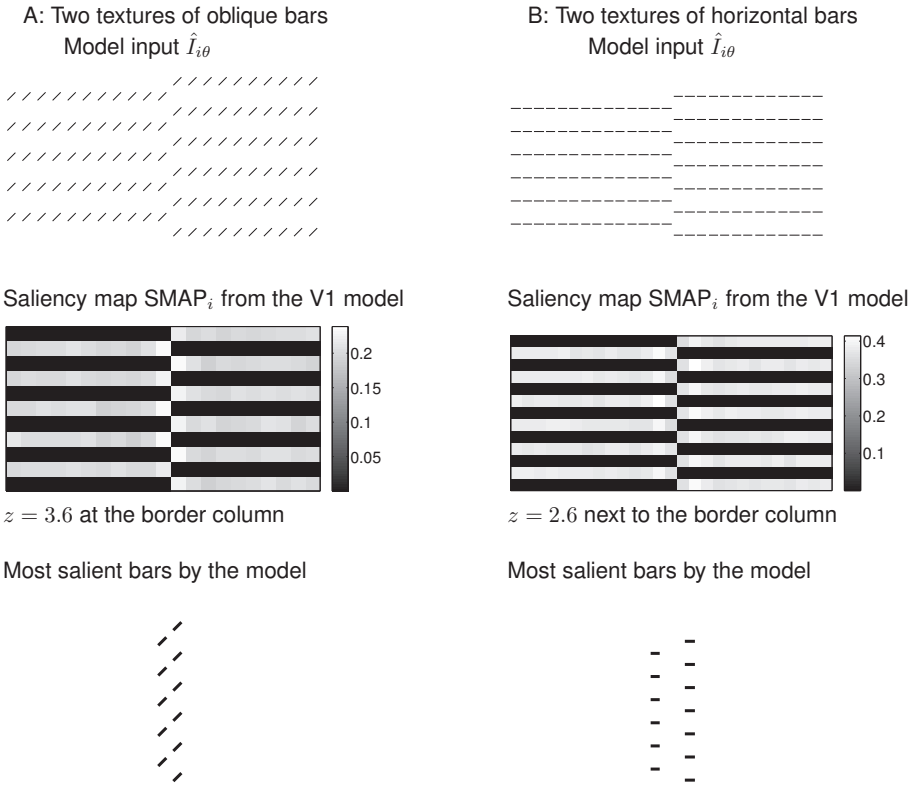


Fig. 5.30: Segmenting two identical textures by detecting the salient border where input statistics change. In both A and B, the two neighboring textures are identical but are displaced from each other vertically. The top two rows of the figure use the format in Fig. 5.28, with the grayscale at each pixel i in the middle row representing the SMAP$_i$ value. The bottom row visualizes the most salient bars. All visible bars have $\hat{I}_{i\theta} = 2$ and $\hat{I}_{i\theta} = 3.5$ in A and B, respectively. In A, the most responsive locations are at the texture border; the bars there have SMAP$_i = 0.23$ against a background $\bar{S} = 0.203$. In B, the most responsive locations are one column away from the border, with SMAP$_i = 0.4$, against a background $\bar{S} = 0.377$.

neighbors regardless of their positions relative to the texture border. It is no longer obvious whether the border bars should be less subject to iso-orientation suppression. Nevertheless, the spatial configuration of the context of each texture bar depends on whether this bar is close to the texture border. This configuration is an aspect of the input statistics and determines the contextual influence. Although the net influence from all the iso-orientation neighbors is typically suppressive, some iso-orientation neighbors can give rise to colinear facilitation when they are co-aligned with the central bar. Apparently, the configurations of the contextual surrounds are such that the net suppression is weaker for a texture bar at or near the texture border.

In both of the examples in Fig. 5.30, the subtle changes in the spatial configuration of the surround are such that the model V1 responses to the locations near the borders are relatively higher. This is consistent with the experience of conspicuous borders.

5.4.4.9 More subtle examples of visual search asymmetry

Some example visual search asymmetries, shown in Fig. 5.31, are much more subtle than that in Fig. 5.25. In each example, the ease of the visual search changes slightly upon swapping the target and the distractor. For example, in Fig. 5.31 E, it is slightly easier to find an ellipse among circles than a circle among ellipses. Readers can examine them to see which target-distractor condition seems easier for finding the target.

The asymmetry between bars and crosses in Fig. 5.25 involves a clear case of the absence versus presence of a basic feature, namely orientation, in the target. Both the neurons and intracortical connections in V1 are tuned to this feature dimension. Hence, via V1 mechanisms, this orientation feature drives a strong asymmetry in an obvious manner. By contrast, there is not a clear V1 feature that distinguishes a circle and an ellipse. If the sizes of the circle and ellipse are comparable to those of the CRFs of the V1 neurons which are not tuned to orientation (see Fig. 3.32), then the circle and the ellipse should evoke similar response levels, if anything perhaps slightly favoring the circle (i.e., opposite to the direction of the asymmetry). Most individual V1 neurons only respond to the line or curve segments in the circles and ellipses, according to their own oriented receptive fields. The V1 model treats the circle as eight line segments, oriented in four different orientations, in a particular spatial arrangement; and the ellipse as ten line segments in five different orientations. None of the ten line segments in the ellipse is oriented sufficiently differently from all the line segments in the circle, and vice versa. So the asymmetry between circle and ellipse cannot be realized in the model in terms of an obvious differential presence of a feature in one versus the other target.

The V1 model indeed generates the asymmetry. The largest z score for the bars in the target ellipse among circles is larger than that for the bars in the target circle among ellipses. The asymmetry arises as the net result of many sources of contextual influences, including iso-orientation suppression, colinear facilitation, and general surround suppression, which is independent of the orientations concerned. None of these contextual influences obviously weighs for or against the direction of the asymmetry. This is similar to the two examples in Fig. 5.30, where the relatively higher saliencies at the texture borders arise not from an obvious change in iso-orientation suppression but from a net result of subtle changes in both contextual suppression and contextual facilitation.

The V1 model was applied to all the search images in Fig. 5.31 and their random variations (such as the random changes in the spatial arrangements of the visual items). As in all the examples of the model application, the model parameters had already been fixed beforehand by the requirements from model calibration and dynamic behavior (described in Sections 5.4.2 and 5.4.3). The z score of the target is calculated as the maximum z score among the line segments which make up the target. In all the five examples of visual search asymmetry, the directions of the asymmetry predicted by the V1 model agree with those observed behaviorally (Treisman and Gormican 1988, Li 1999a).

Note that if V1 responses are not responsible for these subtle examples of asymmetry, then a prediction from the V1 saliency hypothesis on the direction of the asymmetry would only match the behavioral direction by chance. Whether the predicted directions match the behavioral ones in all the five examples provides a stringent test of the V1 saliency hypothesis.

Conventional psychological theories (Treisman and Gormican 1988, Wolfe 2001) presume that each target-distractor pair that exhibits search asymmetry implies the presence and absence of a preattentive basic feature in the easier and the more difficult, respectively, search of the pair. For example, since the ellipse is easier to find among circles than vice-versa, one should conclude that the ellipse has an "ovoid" feature that is absent in a circle (i.e., the ellipse is seen as a departure from the circle). This, of course, leads to feature proliferation. That the

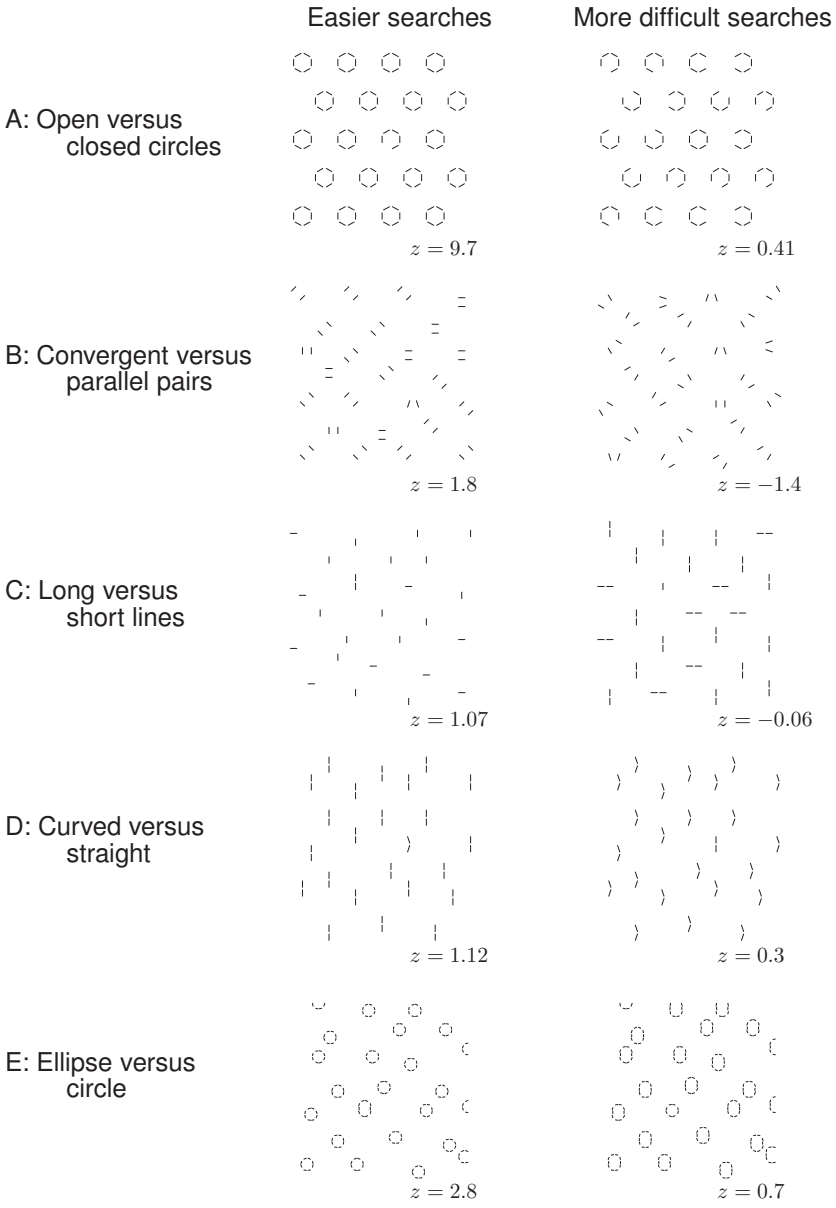


Fig. 5.31: Five pairs of images for the subtle examples of visual search asymmetry. They resemble those studied by Treisman and Gormican (1988). The V1 model can account for the directions of all these asymmetries. Stimulus patterns ($\hat{I}_{i\theta}$) are shown with the targets' z scores (as the largest z score for the bar segments which comprise the target) from the model marked underneath. Adapted with permission from Zhaoping, L., Theoretical understanding of the early visual processes by data compression and data selection, *Network: Computation in Neural Systems*, 17(4): 301–334, Fig. 10, copyright © 2006, Informa Healthcare.

Box 5.1: **Some examples of visual search asymmetries are due to higher level mechanisms**

Another example of search asymmetry is shown in Fig. 5.32: a target letter "N" is more difficult to find among mirror images of "N"s than the reverse (Frith 1974). The letter "N" and its mirror image differ only in the direction of the oblique bar in their shape, and there are no known mechanisms in V1 to break this mirror-reflection symmetry. To explain this asymmetry, conventional psychological theories suggest that a more familiar letter "N" lacks a novelty feature which is present in its mirror image. It seems difficult to envision that V1 mechanisms might account for any such feature based on object familiarity or novelty.

However, later observations (Zhaoping and Frith 2011) indicate that there is little asymmetry between the reaction times for gaze to reach the respective targets, the letter "N" and its mirror image, during the visual searches. This suggests that the search asymmetry does not result from the initial visual selection by bottom-up saliency of the targets. (Note that either target is very salient, having an uniquely oriented oblique bar in the image.) Instead, the asymmetry would originate from confusing the target as a non-target, since all items in the search image have the same viewpoint-invariant shape. This confusion is of the kind we saw in Fig. 1.4, occurring at the shape recognition stage *after* a visual input location is selected. Apparently, this confusion is more effective when the target is "N" in its more familiar, rather than the unfamiliar, view; the familiarity makes the shape recognition faster, allowing an earlier onset of confusion during the task execution (Zhaoping and Frith 2011).

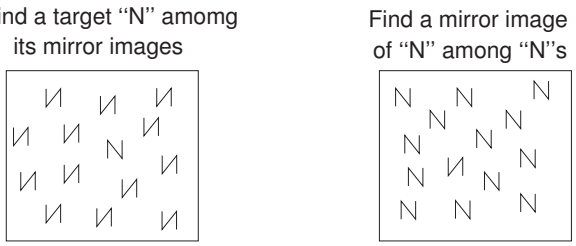


Fig. 5.32: Object shape confusion, not saliency, makes the search on the left more difficult (Zhaoping and Frith 2011).

The asymmetry between the "N" and its mirror image as the target-distractor pair is an example in which the reaction time $RT_{\text{task}} = RT_{\text{saliency}} + RT_{\text{top-down selection}} + RT_{\text{other}}$ (see equation (5.3)) to report the search target is not indicative of the relative degree of bottom-up saliency of the targets in the two searches. This is because this RT's non-saliency component, $RT_{\text{top-down selection}} + RT_{\text{other}}$, is not a constant between the two different searches. Among all the known examples of visual search asymmetry, it has yet to be worked out which examples are mainly caused by bottom-up saliency processes to test the V1 saliency hypothesis more extensively.

V1 model can successfully predict all these asymmetries suggests that it is unnecessary to introduce a feature for each such target-distractor, since an asymmetry can also be caused by the complexity of V1 circuit dynamics in response to the spatial configurations of primitive bar/edge segments in visual inputs.

The V1 model also suggests that it is not necessary to have custom neural detectors for a circle, ellipse, cross (see Fig. 5.25), curvedness, parallelness, closure, or perhaps even a face, in order to exhibit saliency effects associated with these input shapes. V1 detectors for primitive bars and edges, and the associated intracortical interactions, enable V1 responses to exhibit response properties which can be specific to spatial configurations of bar/edge

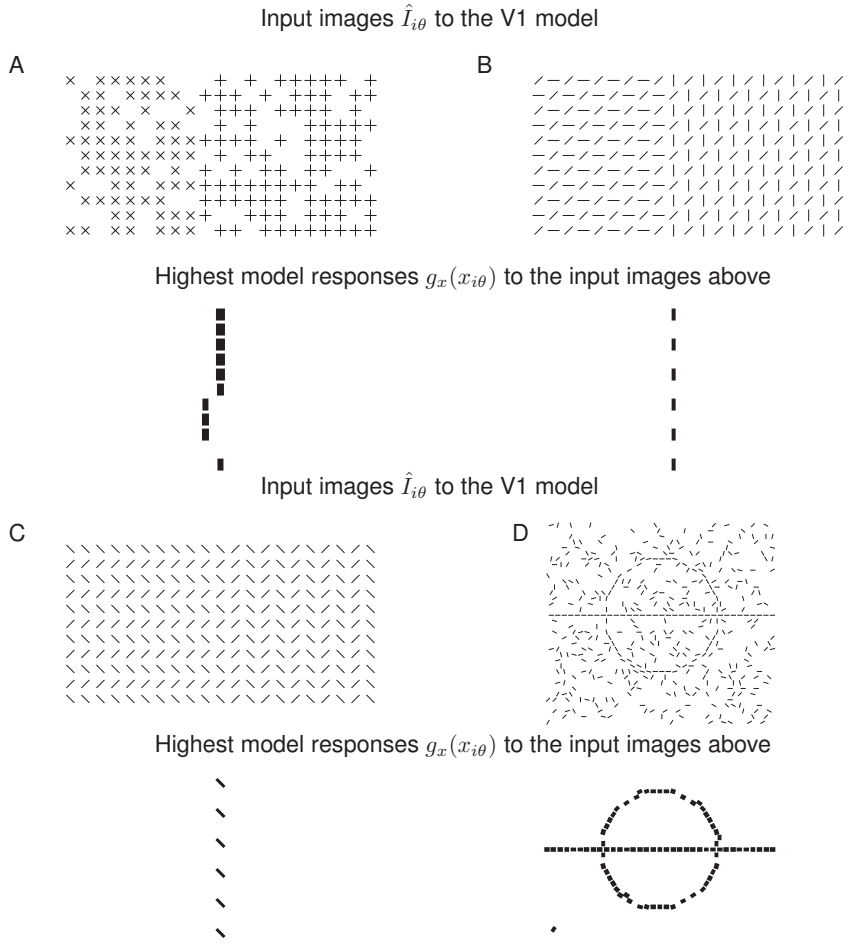


Fig. 5.33: Four examples of V1 model's response to highlight the input locations where input statistics deviate from the statistics of the context.

segments. In principle, these configurations could include many meaningful object shapes such as those of crosses and ellipses.

However, since the V1 model is a poor imitation of the real V1, the z scores of the search targets predicted by the V1 model in the stimuli in Fig. 5.31 can be quantitatively quite different from what is suggested by the behavioral data. A better test of whether V1 mechanisms can account for the asymmetries is to examine the response of the real V1 to the stimuli concerned while preventing top-down interference.

5.4.4.10 Complex examples where V1 responses highlight input locations where input statistics deviate from that of the context

We have argued that places where visual input statistics deviate significantly from those of the context are often predominant examples of salient locations. These locations are often at boundaries of objects, such as the border between two textures. Figures 5.23 and 5.27 show that the V1 model works well to highlight these input deviations at the borders between simple textures, each made of iso-oriented bars. Figure 5.30 shows that this also works in

two examples when the borders are between identical textures. Figure 5.33 shows that it also works in more complex examples.

The V1 model can highlight borders between two textures that are stochastic (Fig. 5.33 A), that involve checkerboard patterns of elements (Fig. 5.33 BC), or that have identical individual elements, but different second order correlations between texture elements (Fig. 5.33 C). Like the real V1 (Li, Piëch and Gilbert 2006), the V1 model can also highlight a contour in a noisy background (Fig. 5.33 D). The V1 saliency hypothesis and the behavior of the V1 model in Fig. 5.33 ABC suggest that the real V1 should also detect such complex input deviations from surrounding statistics. This suggestion is consistent with subsequent observations by functional magnetic resonance imaging of the cortex (Joo, Boynton and Murray 2012) in response to complex arrays of Gabor patterns.

5.4.5 Other effects of the saliency mechanisms—figure-ground segmentation and the medial axis effect

The foreground (figure) of a visual scene typically attracts attention more strongly than the background (often called just the "ground"). When both are textures made from iso-oriented bars, the figure has been observed to evoke higher V1 responses than the ground (Lamme 1995, Lee, Mumford, Romero and Lamme 1998), a phenomenon known as the *figure-ground effect*. Equally, V1 responses to a figure grating can sometimes be higher when it is presented against a background grating having a different (e.g., orthogonal) orientation, versus when a blank background is used (Sillito et al. 1995); this is termed *cross-orientation enhancement*. Finally, V1 responses to the central or medial axis of a figure texture can sometimes be higher than its responses to other regions of the figure that are not borders (Lee et al. 1998). This is called the *medial axis effect*.

Medial axes can be useful for characterizing deformable object shapes, e.g., to represent a human body as a stick figure. Hence, the figure-ground and medial axis effects appear to provide tantalizing hints that V1 could play a role in figure-ground segmentation and higher-order object representation—operations that go beyond highlighting salient border regions (which is the *border effect*). In this section, we show that these effects can be explained (Zhaoping 2003) as side effects of V1 saliency mechanisms that stress image locations where input translation invariance breaks down. This analysis explains why such effects are weaker than the border effect, and, furthermore, predicts that these side effects occur only for particular sizes of figures.

For example, when the figure is sufficiently small, as in Fig. 5.34 B, the responses its bars evoke should all be higher than those of the larger background texture, since each bar is part of the border. This is analogous to the pop-out of an orientation singleton. This "figure-ground" effect is observed electrophysiologically when the RF of the recorded neuron is in the figure region. However, it was predicted (Li 2000a) that this effect should disappear when the figure is large enough such that the RF of the recorded neuron is no longer a part of the texture border. This prediction was subsequently confirmed physiologically (Rossi, Desimone and Ungerleider 2001), and is illustrated in the V1 model simulations shown in Fig. 5.34 C–E.

In fact, the higher responses to the texture border enhance the iso-orientation suppression suffered by the figure regions immediately next to the border; this region is thus termed the *border suppression region*. We may refer to the suppression of the border suppression region by the salient border as the *border's neighbor effect*. Hence, when the figure size is such that the center of the figure is also within the border suppression region flanked by two or more border sides, as in Fig. 5.34 C, the response to the center of the figure should even be weaker than the typical responses to the background texture.

However, when the figure is large enough, as in Fig. 5.34 E, the distance between the center

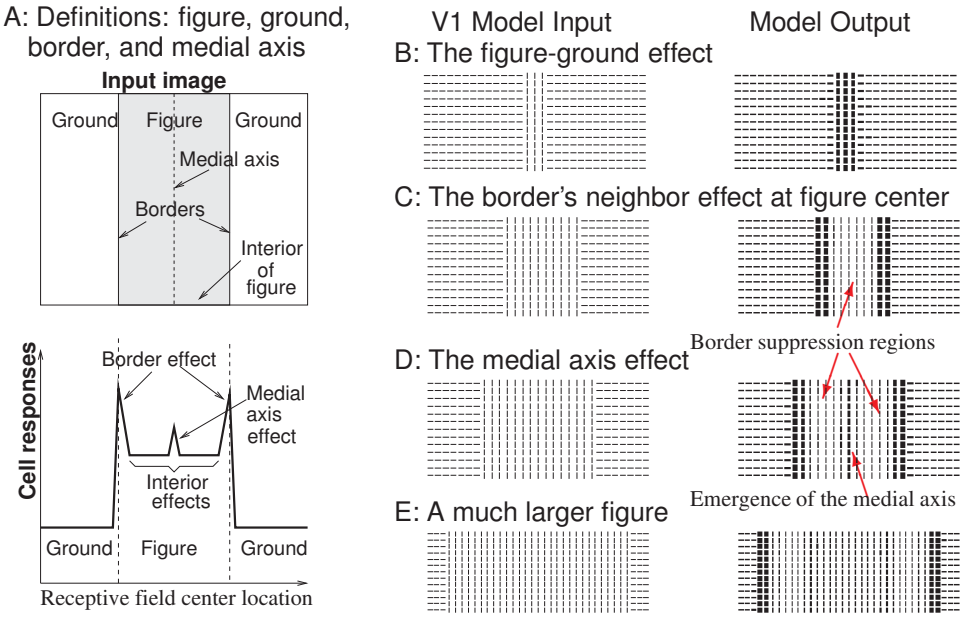


Fig. 5.34: Appropriate sizes of the figures evoke figure-ground effect, border's neighbor effect, and medial axis effect as side effects of the *border effect*, which is the relatively higher response to texture borders caused by V1 saliency mechanisms. A defines the terms. B, C, D, and E show V1 model responses to the figure texture for various figure sizes. The figure-ground effect, defined as higher responses to the figure, emerges in B when the figure size is small enough, making the whole figure its own border. In C–E, responses to the border suppression regions—the figure texture region next to the figure border—are low due to the *border's neighbor effect*, the stronger iso-orientation suppression from the salient figure border. D manifests the medial axis effect, since the axis escapes the suppression from the borders by virtue of (1) being sufficiently far from both borders and (2) being subject to weaker iso-orientation suppression from the two flanking border suppression regions that have themselves been suppressed by the salient borders. E shows the V1 model responses when the figure size is much larger. Adapted with permission from L. Zhaoping, V1 mechanisms and some figure-ground and border effects, *Journal of Physiology-Paris*, 97(4–6): 503–515, figure 1 and figure 4, copyright © 2003, Elsevier.

of the figure and either border is much longer than the typical length of the intracortical V1 connections responsible for iso-orientation suppression. In this case, the response to the figure center becomes indistinguishable from typical responses to the background texture. Feedback from higher visual areas could subsequently enhance the responses to the figure center, as suggested physiologically (Lamme, Rodriguez-Rodriguez and Spekreijse 1999, Scholte, Jolij, Fahrenfort and Lamme 2008), and could partly explain behavioral aspects of the figure-ground effect. However, modulation of V1 responses by the immediate context, which is responsible for the border effect, remains intact after V2 inactivation (Hupé et al. 2001) and is present whether the animal is awake or under anaesthesia (Knierim and Van Essen 1992, Nothdurft et al. 1999).

Figure 5.34 D shows how the medial axis effect can arise as a further consequence of the border effect. The response to the medial axis will be enhanced when the figure is just the right size such that the following two conditions are satisfied: first, the figure center is out of

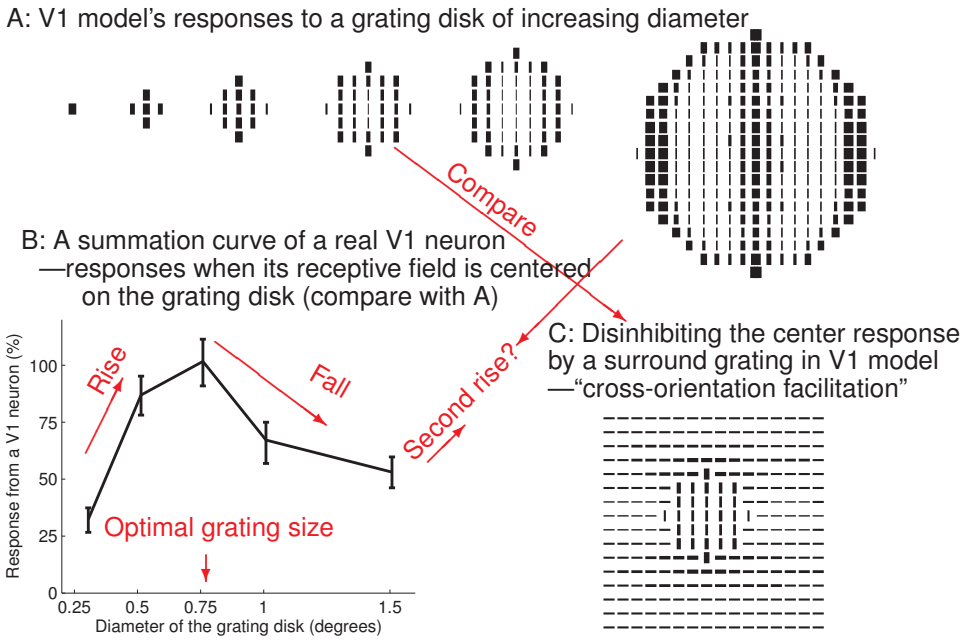


Fig. 5.35: V1 responses to a disk grating, and cross-orientation facilitation (Zhaoping 2003). A: V1 model's responses to coarse-sampled disk gratings. As the disk size increases, the center of the grating changes from being part of the border (with a high response), to being part of the border suppression region (with a suppressed response), to being included in the emerging medial axis (with an enhanced response). B: Responses of a real V1 neuron to a disk grating as a function of the diameter of the disk (this curve is called a summation curve). The model's response to the largest disk grating in A predicts a second rise in this summation curve associated with the medial axis effect. C: When the central disk grating is larger than the optimal size, such that the center of the disk is in the border suppression region (the fourth grating in A), a surrounding grating can suppress the responses to the border of the central grating by general (orientation-unspecific) surround suppression, thereby releasing the central response from the border suppression. This may explain some physiological observations (Sillito et al. 1995) of cross-orientation facilitation. Data in B from Jones, H. E., Grieve, K. L., Wang, W., and Sillito, A. M. Surround suppression in primate V1, *Journal of Neurophysiology*, 86(4): 2011–28, 2001.

reach of iso-orientation suppression caused by both lateral borders of the figure; and second, the figure center is within reach of the relatively weaker suppression occasioned by the border suppression regions associated with these borders. The resulting suppression of the media axis could then be weaker than that suffered by typical background texture bars. Therefore, the medial axis effect should only be observed for certain figure sizes, as is indeed the case physiologically (Lee et al. 1998).

The dependence of the neural response on the size of the figure is also manifest in the way that a V1 neuron's response to a disk grating varies with the disk's diameter (Zhaoping 2003); see Fig. 5.35. When the orientation of the grating is that preferred by the neuron whose RF is centered on the disk, the neural response increases with the diameter for small diameters when the disk is smaller than the classical receptive field of the neuron, and then the response decreases with the diameter when the center of the disk moves out of the disk border and

into the border suppression zone. The overall relationship between response level and disk diameter is called a *summation curve*.

When the disk grating is even larger, its center moves out of the border suppression zones, and the response it evokes should therefore rise again. This is the medial axis effect. It should lead to a second rise in the summation curve; see Fig. 5.35 B. This prediction is supported by recent physiological data.[16]

Consider the case that the disk grating is somewhat larger than the optimal size (where the summation curve peaks), such that its center is in the border suppression region associated with the disk border. Then, adding a surrounding grating of a different (e.g., orthogonal) orientation should suppress the responses to the border of the grating disk, via general, orientation-unspecific, surround suppression. In turn, this should lessen the iso-orientation suppression by the disk grating's border onto the disk center. In other words, the surround grating disinhibits the response to the center of the figure grating; see Fig. 5.35 AC. This could explain physiological observations of cross-orientation enhancement (Sillito et al. 1995), which is indeed often observed when the figure grating is somewhat larger than the optimal size where the summation curve peaks.

5.4.6 Input contrast dependence of the contextual influences

Contextual influences are dependent on the contrast or strength of visual inputs, with suppression decreasing as the input contrast decreases. As will be explained later in equation (5.94), this is because the inhibitory interneurons, which mediate the suppression, are less sensitive to inputs from the excitatory cells at lower input contrast. Since salience depends on iso-orientation suppression, this implies that orientation singletons and texture borders are less salient at low input contrast. This is why most simulations of the V1 model use medium or high input contrast, as indeed is also true of most behavioral experiments into saliency.

Since figure-ground and medial axis effects arise from border effects, they are also weaker at lower input contrast. Consequently, the radius of the grating where a neuron's summation curve (Fig. 5.35 B) peaks tends to be larger when the input contrast is weaker. This is true in the model and is also observed physiologically.

5.4.7 Reflections from the V1 model

In total, building and applying the V1 model has led to the following conclusions.

1. It is possible to build a V1 model which can simultaneously satisfy two requirements: (1) reproducing the contextual influences that are observed physiologically; and (2) being able to amplify selective deviations from homogeneity in the input, without hallucinating heterogeneous responses to homogenous visual inputs. Therefore, V1 mechanisms are plausible neural substrates for saliency.
2. Under the V1 saliency hypothesis, the responses of the V1 model to representative visual inputs produce saliency maps that are consistent with subjective visual experience and previous behavioral observations.
3. The V1 model confirms the intuition that iso-orientation suppression is the dominant mechanism underlying various saliency effects. Such effects include the qualitative distinction between feature and conjunction searches, the greater saliency of locations where feature contrast is greater, and the dependence of saliency on visual input density and heterogeneity. The V1 model also demonstrates that other intracortical interactions,

[16] Private communication from Kenneth D. Miller (2013), who collaborated with Dan Rubin and Stephen Van Hooser on an investigation which revealed these data.

including colinear facilitation and general, feature-unspecific, contextual suppression, also play essential roles in shaping saliency. This is especially the case for visual inputs that are more similar to typical visual inputs and so are more complex than those used in feature searches.

4. The V1 model can signal saliency at locations of complex shapes such as ellipses and crosses, even though there is no V1 cell tuned to such shapes. This reaffirms our understanding that selection of a visual location can occur before recognition of inputs or objects at this location. One may even ask whether V1 mechanisms can also contribute to attentional attraction of, e.g., a face, which, like a cross, is a particular spatial configuration of image elements (like bars and patches of luminance and color) that activate V1 neurons.

5. Saliency mechanisms have side effects, and these can be understood.

Recall from Fig. 1.1 that an important role for a model is to be an intermediary between a theoretical hypothesis and experimental data. This role can be fulfilled, for example, by demonstrating the theory in particular instances or by fitting data to a particular manifestation of the theory. In our current example, the theory is the V1 saliency hypothesis, the data are observations of bottom-up visual selection, and the V1 model played a role of verifying the theory by testing the ability of a restricted set of V1 mechanisms to account for some behavioral observations. The restrictions include: (1) that the model contains only neurons tuned to spatial locations and orientations; (2) that the model ignores many physiological details; and, (3) that all the model neurons have the same receptive field size. These restrictions imply that the model can only be tested against a restricted set of saliency data. For example, the model is not expected to account well for the saliency of a scale singleton because it omits the multiscale property of V1. Nevertheless, we can ask whether the V1 model can be successful when applied to an appropriately restricted set of data.

The success of the V1 model with the restricted data suggests that one can extrapolate and generalize beyond the current model. For example, the model can be extended to include model V1 neurons tuned to feature dimensions other than orientation, such as color, motion direction, scale, disparity, and ocularity (tuning to ocularity can be defined as a relative sensitivity to inputs from the two eyes). One can expect, and verify, that iso-feature suppression should work in the same way for these feature dimensions as it does for orientation. Indeed, one such extension has been carried out for the case of color (Li 2002, Zhaoping and Snowden 2006). Similarly, although the model has mostly been applied to synthetic images (with a few exceptions (Li 1998a, Li 1999b, Li 2000b)), one can expect, and test, that the theory also applies to more realistic visual inputs such as those from natural scenes.

V1 neurons that have large and unoriented receptive fields (see Fig. 3.32) can also be included in the model. Extending iso-feature suppression to the feature of the round shapes of these receptive fields (of a given scale), one would expect mutual suppression between all nearby neurons of this class. These neurons are likely to play an important role in saliency for round shapes or patches, perhaps contributing to the attentional attraction of a face (of a similar size to the receptive fields).

While a model can be used to build confidence in a theory, the theory should be able to stand despite a model's inaccuracies or fail despite a model's fit to many details. Furthermore, a theory should be ultimately tested against experiment data rather than just against model simulations (Fig. 1.1). The test in Section 5.3 of the predicted high saliency of an ocular singleton is an example of a direct test of the theory without the aid of the V1 model. We next turn to more such tests.

5.5 Additional psychophysical tests of the V1 saliency hypothesis

This section presents additional non-trivial predictions and their behavioral tests. Each of these predictions exploits either a distinctive characteristic of V1 physiology, to test the specifically V1 nature of the hypothesis, or a qualitative difference between the V1 saliency hypothesis and conventional ideas about saliency. The V1 model is not necessary as an intermediary between the hypothesis and the link between physiology and saliency behavior. This is because the hypothesis is so explicit, and because the knowledge about V1's physiology is extensive, that it is easy to predict from physiology to behavior via the medium of the hypothesis. It is also typically easier to test behavioral predictions using psychophysical experiments than to test physiological predictions by electrophysiological experiments.

One prediction is based on the feature-blind, "auction" nature of selection by saliency that is depicted in Fig. 5.6. It states that texture segmentation should be more severely impaired than traditional theories would imply if a task-irrelevant texture is superposed. This prediction arises because the saliency value at a location can be hijacked by the irrelevant features whose evoked V1 responses are higher than those of the task-relevant features. This prediction cannot be derived from the traditional saliency frameworks that are depicted in Fig. 5.5, so it allows us to test them against the V1 saliency hypothesis.

The second prediction arises from colinear facilitation, which is a characteristic of V1 physiology. Via the saliency hypothesis, this implies how the ease of texture segmentation can be influenced by the degree of spatial alignment between the texture bars.

The third prediction arises from the observation that whereas some V1 neurons are tuned simultaneously to color and orientation (see Section 3.6.6.3), and some V1 neurons are tuned simultaneously to orientation and motion direction, very few V1 neurons are tuned simultaneously to color and motion direction (Horwitz and Albright 2005); see Section 3.6.9. Based on this, according to the V1 saliency hypothesis, we can predict whether the RTs for finding a feature singleton that is unique in two feature dimensions should be shorter than the statistically appropriate combinations of the RTs for finding feature singletons that are unique in just one of the two feature dimensions.

These three predictions are qualitative, in that they anticipate that the RT in one situation (RT_1) should be shorter than another RT in a different situation (RT_2). They do not predict a precise value for the difference $RT_2 - RT_1$. The fourth prediction is quantitative, based on the assumption that there are no V1 neurons (or an insignificant number of V1 neurons) tuned simultaneously to the three feature dimensions: color, orientation, and motion direction. It derives a precise relationship among the RTs for a single observer for finding feature singletons that differ from a uniform background in one, two, or three of these dimensions, and it uses this relationship to predict the whole distribution of one of these RTs from the distributions of the other RTs. This is a quantitative prediction that is derived without any free parameters. Therefore, the V1 saliency hypothesis could be easily falsified if it is incorrect, since there is no freedom to fit data to the prediction. We will show an experimental confirmation of this prediction.

5.5.1 The feature-blind "auction"—maximum rather than summation over features

According to the V1 saliency hypothesis, the saliency of a location is signaled by the highest response to this location, regardless of the feature preference of the neurons concerned. For instance, the cross among bars in Fig. 5.25 A is salient due to the response of the neuron tuned to the horizontal bar, with the weaker response of a different neuron tuned to the vertical bar

being ignored. Therefore, the "less salient features" at any location are invisible to bottom-up saliency or selection, even though they are visible to attention attracted to the location due to the response to another feature at the same location. This leads to the following prediction: a visual search or a segmentation task can be severely interfered with by task-irrelevant stimuli that evoke higher V1 responses than the task-relevant stimulus does at the same location. This is because the task-irrelevant stimulus, which makes the task-relevant stimulus invisible to saliency, will determine the saliency values of the stimulus and thereby control bottom-up selection. This attention control by task-irrelevant stimuli makes the task performance inefficient.

Figure 5.36 shows texture patterns that illustrate and test the prediction. Pattern A has two iso-orientation textures, activating two populations of neurons, one tuned to left tilts and one to right tilts. Pattern B is a checkerboard, evoking responses from another two groups of neurons tuned to horizontal and vertical orientations.

With iso-orientation suppression, neurons responding to the texture border bars in pattern A are more active than those responding to the background bars, since each border bar has fewer iso-orientation neighbors to exert contextual iso-orientation suppression (as explained in Fig. 5.7). For ease of explanation, let us say that the responses from the most active neurons to a border bar and a background bar are 10 and 5 spikes/second respectively. This response pattern renders the border location more salient, making texture segmentation easy. Each bar in pattern B has as many iso-orientation neighbors as a texture border bar in pattern A, and so it also evokes a response of (roughly) 10 spikes/second.

The composite pattern C, which is made by superposing patterns A and B, activates all neurons responding to patterns A and B. For simplicity (and without changing the conclusions), we ignore interactions between neurons tuned to different orientations. Therefore, the neurons tuned to oblique orientations respond roughly to the same degree as they do to A alone (we call these relevant responses); and the neurons tuned to horizontal or vertical orientation respond roughly to the same degree as they do to B alone (irrelevant responses). This implies that all texture element locations evoke the same maximum response of 10 spikes/second, which is the largest of the relevant and irrelevant responses to each location.

According to the feature-blind auction framework of the V1 hypothesis, it is this maximum response to a location x, $\text{SMAP}(x) = \max_{x_i \approx x} O_i$ (from equation (5.4)) that determines the saliency $\text{SMAP}(x)$ at that location, where x_i is the center of the receptive field (which covers location x) of neuron i giving the response O_i. Thus, by the V1 hypothesis, all locations are equally salient (or non-salient), without a saliency highlight at the texture border. Therefore texture segmentation is predicted to be much more difficult in C than A, as indeed is apparent in Fig. 5.36. Any saliency signal associated with the task-relevant, oblique, bars is swamped by the uniform responses to the task-irrelevant horizontal and vertical bars.

If saliency was instead determined by the summation rule $\text{SMAP}(x) \propto \text{sum}_{x_i \approx x} O_i$ (this is a modification of equation (5.4)), responses to the various orientations at each texture element location in pattern C could sum to preserve the border highlight as $20 = 10 + 10$ (spikes/second) against a background of $15 = 10 + 5$ (spikes/second). This predicts that texture segmentation should be easy (Zhaoping and May 2007). This summation rule is the basis of traditional saliency models (Itti and Koch 2001, Wolfe et al. 1989) (depicted in Fig. 5.5). By the maximum rule, it may seem a waste not to include the contributions of "less salient features" to obtain a "more informative" saliency measure of locations, as in the summation rule. However, reaction times for locating the texture border[17] confirmed the prediction of the maximum rather than the summation rule; see Fig. 5.36 D.

[17] In the experiment (Zhaoping and May 2007), each stimulus display consisted of 22 rows $\times$ 30 columns of items (of single or double bars) on a regular grid with unit distance 1.6^o of visual angle. Observers were instructed to press

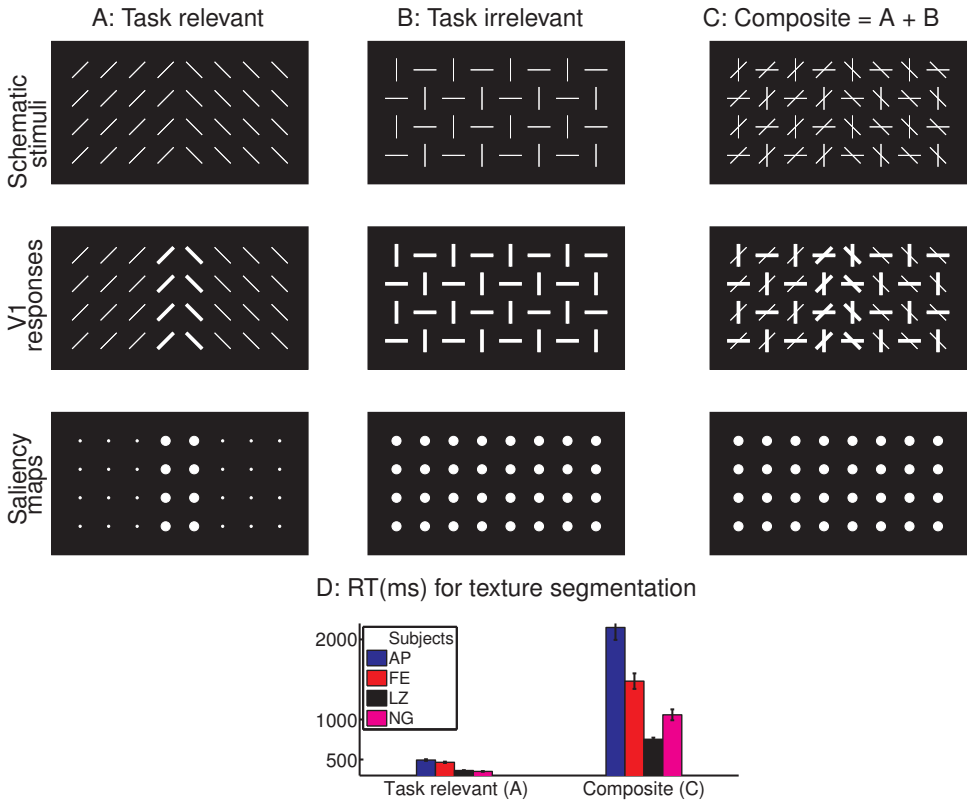


Fig. 5.36: Psychophysical confirmation of the maximum rule used by the V1 saliency hypothesis, instead of the summation rule used by traditional models of saliency. A, B, C: Schematics of texture stimuli (extending continuously in all directions beyond the portions shown), each followed by schematic illustrations of V1's responses and a saliency map, formulated as in Fig. 5.14. Each dot in the saliency map scales with the maximum V1 response to the corresponding location, rather than the sum of all V1 responses there. Every bar in B, or every texture border bar in A, experiences less iso-orientation suppression. The composite stimulus C, made by superposing A and B, is predicted to be difficult to segment, since the task-irrelevant features from B interfere with the task-relevant features from A, giving no saliency highlights to the texture border. D: Reaction times of four observers (subjects) for the texture segmentation task using stimuli similar to A and C. Adapted from Zhaoping, L. and May, K. A., Psychophysical tests of the hypothesis of a bottom-up saliency map in primary visual cortex, *PLoS Computational Biology*, 3(4):e62, Fig. 1, copyright © 2007, Zhaoping, L. and May, K. A.

The two halves of Fig. 5.36 C have very different second order statistics of visual inputs.[18] This is an example for which the breakdown in the translation symmetry of the input statistics, even though it involves low (i.e., second) order statistics, does not lead to high saliency.

a left or right button as soon as possible to indicate whether the texture border was in the left or right half of the display.

[18] These two halves can be easily distinguished by a standard texture segregation model (Bergen and Landy 1991), which works by examining whether two textures have identical visual inputs in matching orientation channels.

5.5.1.1 Further discussion and exploration of interference from task-irrelevant features

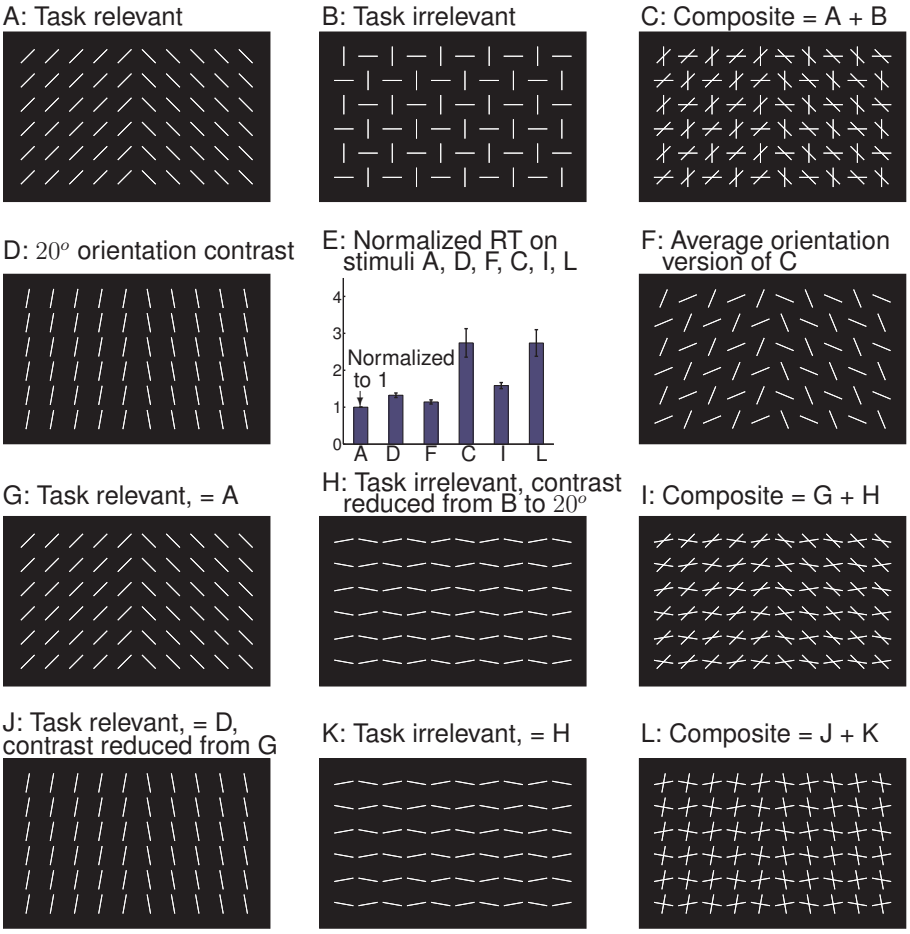


Fig. 5.37: Further illustrations of the interference wrought by task-irrelevant features. A, B, and C are the schematic stimuli from Fig. 5.36. D is a version of A, with bars being 10^o from vertical, thus reducing the orientation contrast at the texture border to 20^o. F is derived from C by replacing each texture element of two intersecting bars by one bar whose orientation is the average of the two intersecting bars. G, H, and I are derived from A, B, and C by reducing the orientation contrast (to 20^o) in the interfering bars; each is 10^o from horizontal. J, K, and L are derived from G, H, and I by reducing the task-relevant contrast to 20^o. E plots the average of the normalized reaction times for three subjects, on stimuli A, D, F, C, I, and L (which were randomly interleaved within a session). Each normalized RT is obtained by dividing the actual RT by that of the same subject for stimulus A. Error bars denote standard error of the mean. Adapted from Zhaoping, L. and May, K. A., Psychophysical tests of the hypothesis of a bottom-up saliency map in primary visual cortex, *PLoS Computational Biology*, 3(4):e62, Fig. 2, copyright © 2007, Zhaoping, L. and May, K. A.

One might wonder whether the composite texture elements in Fig. 5.36 C (each of which comprises two intersecting bars) are acting (for saliency) as single bars having the average

orientation at each location; see Fig. 5.37 F. This would make the relevant orientation feature noisy and impair performance. The control experiment reported in Fig. 5.37 E demonstrates that this would not have caused so large an impairment. The RT for this "orientation-averaged" stimulus (Fig. 5.37 F) is at least 37% shorter than that for the composite stimulus (Fig. 5.36 C).

Box 5.2: **Alternative accounts for the interference by task-irrelevant features**

One may seek alternative explanations for the observed interference by task-irrelevant features that is predicted by the V1 saliency hypothesis. For instance, in Fig. 5.36 C, one may assign a new feature type, let us call it feature "X", to "two bars crossing each other at 45^o." Then, each texture element is this "X" at a particular orientation, and each texture region in Fig. 5.36 C is a checkerboard of two orientations of "X". So the segmentation could be more difficult in Fig. 5.36 C, in the same way that it is more difficult to segment the texture of "ABABAB" from "CDCDCD" in a stimulus pattern "ABABABABABCDCDCDCDCD" than it is to segment "AAA" from "CCC" in "AAAAAACCCCCC." This approach of creating new feature types to explain hitherto unexplained data could of course be extended to accommodate other cases, such as double-feature conjunctions (e.g., color-orientation conjunction), triple, quadruple, and other multiple feature conjunctions, or even complex stimuli like faces. It is not clear how long this list of new feature types would have to be. By contrast, the V1 saliency hypothesis is a parsimonious account, since it explains all these data without invoking additional free parameters or mechanisms. It was also used in Section 5.4 to explain visual searches for, e.g., a cross among bars or an ellipse among circles without any detectors for crosses or circles/ellipses. Our aim should be to explain the most data with the fewest necessary assumptions or parameters. Additionally, the V1 saliency hypothesis is based on substantial physiological findings. When additional data reveal the limitation of V1 for bottom-up saliency, the search for additional mechanisms for bottom-up saliency can be guided by following conclusions suggested by visual pathways and cortical circuits in the brain (Shipp 2004).

From the analysis above, one can see that the V1 saliency hypothesis also predicts a decrease in the interference if the irrelevant feature contrast is reduced. This is evident from comparing Fig. 5.37 GHI with Fig. 5.37 ABC, and it is confirmed by the RT data (Fig. 5.37 E). The neighboring irrelevant bars in Fig. 5.37 I have more similar orientations, inducing stronger iso-feature suppression. Consequently, their evoked responses are decreased, say, from 10 to 7 spikes/second. (Colinear facilitation is also greater for this stimulus; however, iso-orientation suppression dominates colinear facilitation physiologically, so the net effect is a decreased response to each task-irrelevant bar.) Consequently, the relevant responses to the texture border, i.e., the border highlights, are no longer submerged by the irrelevant responses. The irrelevant responses interfere less with the relevant responses, although the fact that the former (at 7 spikes/second) still dominate the latter (5 spikes/second) in the background implies that there would still be some interference (with the border highlight being reduced from 5 to 3 spikes/second).

Analogously, interference can be increased by decreasing the task-relevant orientation contrast at the texture border. This is demonstrated by comparing Fig. 5.37 JKL and Fig. 5.37 GHI, and it is also confirmed in experimental data (Fig. 5.37 E). Reducing the relevant feature contrast makes the relevant responses to the texture border weaker, say from 10 to 7 spikes/second. Consequently, these relevant responses are even more vulnerable to being submerged by the irrelevant responses. Therefore, interference is stronger in Fig. 5.37 L than in Fig. 5.37 I.

In sum, the existence and strength of the interference depend on the relative levels of responses to task-relevant and task-irrelevant features, with these responses depending on the

corresponding feature contrasts and direct input strengths. When the relevant responses dictate saliency everywhere, and when their values are barely affected by the presence or absence of the irrelevant stimuli, there should be little interference. Conversely, when the irrelevant responses dictate saliency everywhere, interference with the visual selection required for the task is strongest. When the relevant responses dictate the saliency value at the location of the texture border but not in the texture background, the degree of interference is intermediate. In both Fig. 5.37 C and Fig. 5.37 L, the irrelevant responses (approximately) dictate the saliency everywhere, so the texture borders are predicted to be equally non-salient. This is confirmed in the data (Fig. 5.37 E). However, the RT performance of subjects for Fig. 5.37 CL varies widely, perhaps because the bottom-up saliency is so weak for these two stimuli that subject-specific top-down factors contribute significantly to the RTs.

Additional data (Zhaoping and May 2007) confirmed analogous predictions from the V1 theory, such as predictions of interference by irrelevant color with orientation-based tasks.

5.5.1.2 Contrasting the V1 saliency hypothesis and traditional frameworks for saliency

As mentioned, in traditional models, the saliency values in the master map come from $\text{SMAP}(x) \propto \text{sum}_{x_i \approx x} O_i$, i.e., summing the activations in various feature maps, each based on one visual feature such as a particular color or orientation; see Fig. 5.5.

The V1 saliency theory differs from traditional theories, partly because one of its motivations was to understand V1. It also aims for fast computation, and thus it calculates saliency without requiring separate feature maps or decoding of input features. Indeed, many V1 neurons are tuned to more than one feature dimension (Livingstone and Hubel 1984, Lennie 1998) (e.g., to orientation and motion direction), making it impossible that separate groups of V1 cells represent separate feature dimensions or separate feature maps.

In contrast, the traditional theories were motivated by explaining behavioral data. They do not specify the cortical location of the feature maps or the master saliency map, or aim for algorithmic simplicity. For example, although the summation rule seems natural for the feature-blind saliency, it is in practice more complex to implement. The value of $\text{SMAP}(x) \propto \text{sum}_{x_i \approx x} O_i$ is sensitive to exactly which activations O_i should be included in the sum, considering that the receptive fields of different neurons have different sizes, shapes, center locations, and sharpness of their boundaries. If the boundary of the receptive field of neuron i is vague, and if this vague boundary barely covers location x, should the neural activation O_i be included in the summation for $\text{SMAP}(x)$? Should the summation rule be implemented as a weighted summation of neural activations, and if so, what weights should be used? The summation step has to be carried out in order to find the most salient location to direct attention to. In comparison, the most salient location by maximum rule can be easily found by finding the neuron with the highest neural response.

From the perspective of the feature-blind auction process, feature maps, and thus a master map, are unnecessary. The observations in Fig. 5.36 thus motivate the framework of visual selection without separate feature maps.

5.5.2 The fingerprints of colinear facilitation in V1

Two nearby V1 neurons can facilitate each other's responses if their preferred bars or edges are aligned with each other such that these bars or edges could be parts of a single smooth contour (Nelson, and Frost 1985, Kapadia et al. 1995). Although such colinear facilitation is much weaker than the iso-feature suppression which is mainly responsible for singleton pop-out in bottom-up saliency, it also has consequences for saliency behavior.

Figure 5.38 shows the first such consequence. Figures 5.38 A and 5.38 B both have

two orientation textures with a 90^o contrast between them. The texture borders pop out automatically. However, in Fig. 5.38 B, the vertical texture border bars in addition enjoy full colinear facilitation, since each has more colinear neighbors than the other texture border bars in either Fig. 5.38 A or Fig. 5.38 B. The vertical texture border bars are thus more salient than other border bars. We call a texture border made of bars that are parallel to the border a colinear border. In general, for a given orientation contrast at a border, a colinear border is more salient than other borders (Li 1999b, Li 2000b). This is also seen in the output of the V1 model—compare Fig. 5.22 and Fig. 5.23 A.

Hence, one can predict that it takes longer to detect the border in Fig. 5.38 A than in Fig. 5.38 B. This prediction was indeed confirmed (Fig. 5.38 E, in the same experiment reported in Fig. 5.36 D). A related observation (Wolfson and Landy 1995) is that it is easier to discriminate the curvature of a colinear than a non-colinear texture border.

Since both texture borders in Fig. 5.38 A and Fig. 5.38 B are so salient that they require very short RTs, and since RTs cannot be shorter than a certain minimum for each subject, even a large difference between the saliencies of these borders will only be manifest as a small difference in the RTs to detect them. However, the saliency difference can be unmasked by the interference caused by task-irrelevant bars. This is shown in Fig. 5.38 CD, involving the superposition of a checkerboard pattern of task-irrelevant bars tilted 45^o away from the task-relevant bars. This manipulation is the same as that to induce interference in Fig. 5.36. Again, for convenience, let us refer to relevant bars as leading to relevant responses from relevant neurons, and similarly for the irrelevant bars. As argued in Fig. 5.36, the irrelevant responses in the background texture region are higher than the relevant responses, and so they dictate the saliency of the background in both Fig. 5.38 C and Fig. 5.38 D. Meanwhile, the RT for detecting the texture border in Fig. 5.38 D is much shorter than that for Fig. 5.38 C, since the interference is much weaker in Fig. 5.38 D.

For concreteness, let us say, as we did when analyzing Fig. 5.36, that the relevant responses in Fig. 5.38 C are 10 spikes/second at the border and 5 spikes/second in the background, and that they are 15 spikes/second and 5 spikes/second, respectively, in Fig. 5.38 D. Meanwhile, the irrelevant responses are roughly 10 spikes/second at all locations in both Fig. 5.38 CD (as in Fig. 5.36). At the colinear vertical border bars in Fig. 5.38 D, the relevant responses (15 spikes/second) are much higher than the irrelevant responses (10 spikes/second), and so are less vulnerable to being submerged. However, because the irrelevant responses dictate and raise the background saliency, the irrelevant texture still causes interference by reducing the ratio between the maximum responses to the border and background, from a ratio of $15/5 = 3$ to $15/10 = 1.5$. This interference is much weaker than that in Fig. 5.38 C, whose border-to-background response ratio is reduced from 10/5 to 10/10.

Figure 5.39 demonstrates another, subtler, fingerprint of colinear facilitation. The task-relevant stimulus component is as in Fig. 5.38 A. The task-irrelevant stimulus consists of horizontal bars in Fig. 5.39 A and vertical bars in Fig. 5.39 B. Away from the border, both relevant and irrelevant bars lack orientation contrast. Thus, they have comparable iso-orientation suppressions and comparable final responses there. However, some readers might notice that the border in Fig. 5.39 A is slightly easier to notice than in Fig. 5.39 B. This can be understood by considering three types of intracortical interactions: iso-orientation suppression between the relevant responses, a general contextual suppression between the relevant and irrelevant responses regardless of the orientation preferences of the neurons, and colinear facilitation between irrelevant responses. The effects of the first two interactions are the same in Fig. 5.39 A and Fig. 5.39 B, but the effect of the third differs between the two stimuli.

Consider iso-orientation suppression from the relevant responses to the texture border to the relevant responses to the border suppression region next to the border (see Fig. 5.34 BC for an illustration of the border suppression region). Because the relevant responses to the

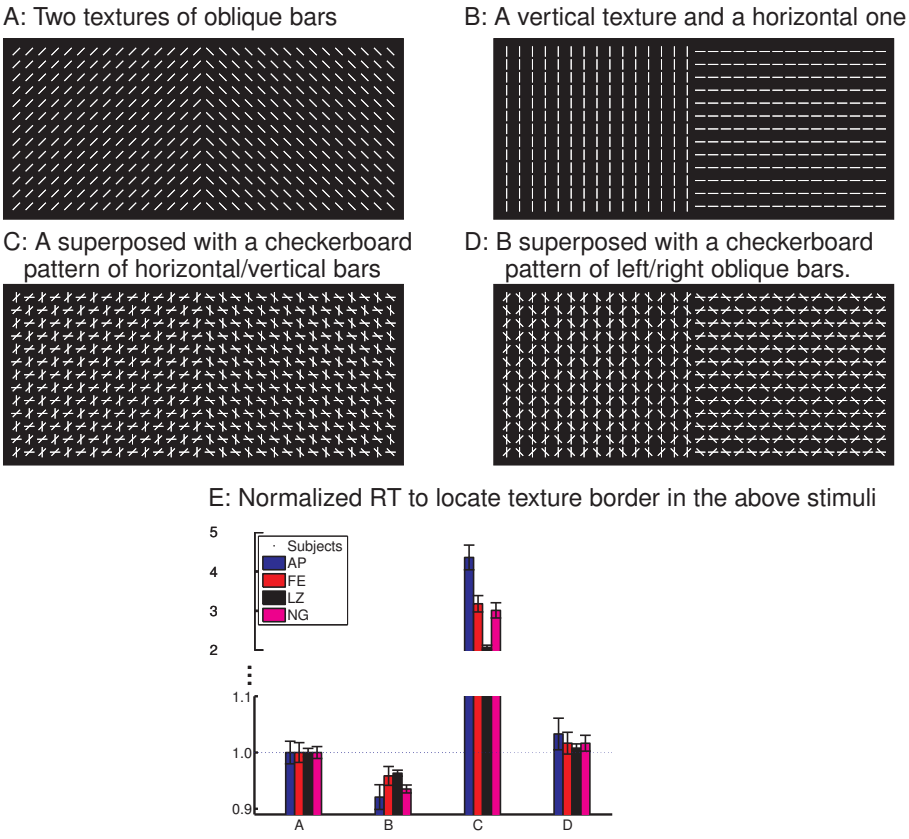


Fig. 5.38: Fingerprint of colinear facilitation in V1: a texture border with texture bars parallel to the border (called a colinear border) is more salient (Zhaoping and May 2007). A and B: Stimulus patterns for texture segmentation; each contains two neighboring orientation textures with 90^o orientation contrast at the texture border. The texture border in B appears more salient. The interference by task-irrelevant bars in C (as schematized in Fig. 5.36) is analogous to that in D. Nevertheless, the interference is much less effective in D since the more salient, task-relevant, colinear border bars are less vulnerable to interference. E: Normalized RT of subjects to localize the texture borders, given by the ratio of actual RT to each subject's (trial averaged) RT for stimulus condition A (493, 465, 363, 351 ms for AP, FE, LZ, and NG, respectively).

border bars are the strongest among the relevant responses, the iso-orientation suppression that they exert is also strongest, making the relevant responses in the border suppression region weaker than those further away from the border. In turn, these weaker relevant responses in the border suppression region generate less general suppression on the local irrelevant neurons, making the local irrelevant responses slightly higher than the other irrelevant responses. Hence, in the border suppression region, the relevant responses are slightly weaker, and the irrelevant responses slightly stronger, than their respective values in the homogenous region further away from the border. In this region, the irrelevant responses therefore dictate the local saliencies; furthermore, because these saliencies are slightly higher than those in the background, they induce interference for the task by reducing the relative saliency of the texture border. Figures 5.39 A and 5.39 B differ in the direction of the colinear facilitation

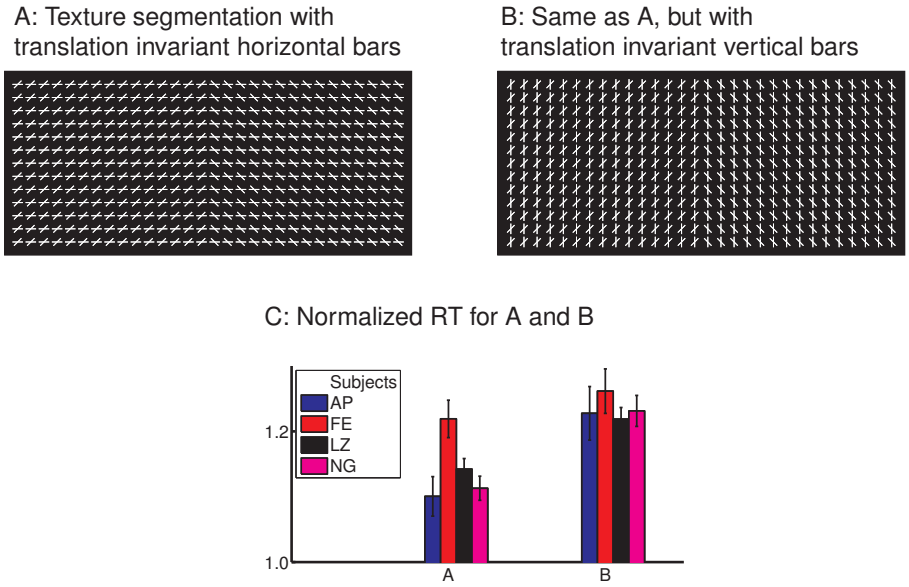


Fig. 5.39: Differential interference by irrelevant bars due to colinear facilitation (Zhaoping and May 2007). Stimuli A and B are made by superposing task-irrelevant horizontal (A) or vertical (B) bars on top of the relevant stimulus from Fig. 5.38 A. C: Normalized reaction times to locate the texture border in A and B (using the same subjects and presented the same way as in Fig. 5.38). The RT for B is significantly longer than that for A ($p < 0.01$) in three out of four subjects. By matched sample t-test across subjects, the RT for B is significantly longer than that in A ($p < 0.01$). For each subject, RTs for both A and B are significantly longer ($p < 0.0005$) than that for Fig. 5.38 A.

among the irrelevant bars. This direction is perpendicular to the border in Fig. 5.39 A and parallel with it in Fig. 5.39 B. Mutual facilitation between neurons tends to equalize their response levels, thereby smoothing away the response variations in the direction of colinear facilitation. Consequently, the local peaks in the irrelevant responses in the border suppression region should be somewhat smoothed away in Fig. 5.39 A but not in Fig. 5.39 B. This predicts stronger interference in Fig. 5.39 B than in Fig. 5.39 A, as indeed is confirmed by the segmentation RTs; see Fig. 5.39 C.

5.5.3 The fingerprint of V1's conjunctive cells

Figure 5.40 shows that a bar that is unique in color, orientation, or in both color and orientation can pop out of (at least statistically) identical backgrounds made of bars with uniform orientation and color. We call the first two cases single-feature singletons and single-feature pop-outs, and the third case a double-feature singleton in a double-feature pop-out. If it takes a subject a reaction time of $RT_C = 500$ ms to find the color singleton, and another reaction time of $RT_O = 600$ ms to find the orientation singleton, one may wonder whether the reaction time RT_{CO} for finding the double-feature singleton should be 500 ms or less.

Let us consider an extremely ideal case, when $RT_C = 500$ ms and $RT_O = 600$ ms always, without any stochasticity or trial to trial fluctuations. Then, if

$$RT_{CO} = \min(RT_C, RT_O) = 500 \text{ ms},$$

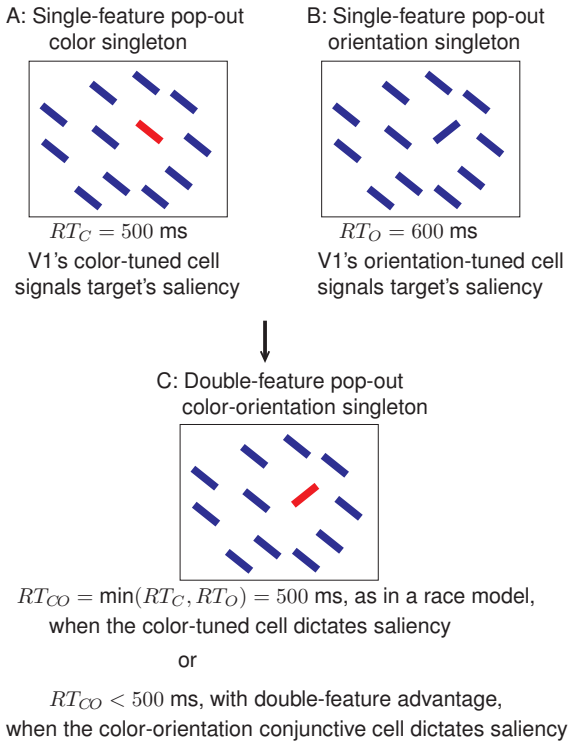


Fig. 5.40: Schematic of single- and double-feature pop-out in color and/or orientation. A: The saliency of the color (C) singleton is dictated by the response of the cell tuned to a red color, which is the only cell free from iso-feature suppression for this input stimulus. B: Similarly, the saliency of the orientation (O) singleton is dictated by the response of the cell tuned to its orientation. C: The color-orientation (CO) double-feature singleton highly activates all three cell types: color-tuned, orientation-tuned, and conjunctive color-orientation–tuned cells; the most activated among them should dictate the singleton's saliency. Consider the simplest case when A, B, and C do not differ in the neural responses to the background bars; furthermore, let the color (only)-tuned cell respond identically to the singletons in A and C, and let the orientation (only)-tuned cell respond identically to the singletons in B and C. When the RTs to find the singletons are, for example, $RT_C = 500$ ms for the color singleton and $RT_O = 600$ ms for the orientation singleton, whether RT_{CO} for the color-orientation singleton is less than or equal to $\min(RT_C, RT_O) = 500$ ms depends on whether the conjunctive cell is the most active cell responding to the CO singleton and is more active than its activation in A and B.

we call RT_{CO} an outcome of a *race model*, as if RT_{CO} is the result of a race between two racers with times RT_C and RT_O, respectively. If $RT_{CO} < \min(RT_C, RT_O)$, we say that there is a double-feature advantage. The idealization to treat RTs as deterministic, rather than stochastic, will be removed later when we work with real behavioral RTs. Meanwhile, for the ease of explanation, we use this idealization without changing the conclusions. We will explain below that the V1 saliency hypothesis predicts a double-feature advantage when V1 has cells tuned conjunctively (or simultaneously) to features in both feature dimensions—in this example, color and orientation.

V1 has conjunctive neurons tuned to color (C) *and* orientation (O), or to orientation *and* motion direction (M). However, experiments have observed few V1 neurons tuned to color *and*

motion direction (Horwitz and Albright 2005). Therefore, the V1 saliency hypothesis predicts that a double-feature advantage should exist for a color-orientation (CO) double feature and a motion-orientation (MO) double feature, but this double-feature advantage should be absent for a color-motion (CM) double feature. It is known that V2, receiving inputs from V1, has neurons selective to all three types of feature conjunctions: CO, MO, and CM (Gegenfurtner, Kiper and Fenstemaker 1996). Thus, if V2, or visual areas that are further downstream, are responsible for bottom-up saliency, then one would predict double-feature advantage for all three types of double-feature singletons. Therefore, we refer to the prediction of a double-feature advantage for CO and MO singletons but not for CM singleton as a V1 fingerprint.

Below, we provide a rigorous argument for the prediction, starting with the example of CO singleton. For intuition, though, consider the activity of neurons whose relevant tuning is solely to C or O, or conjunctively to CO. Due to iso-feature suppression, a C (only)-tuned neuron should, by definition, respond identically to the CO singleton and a C singleton, but it should be less activated by an O singleton. Similarly, an O (only)-tuned neuron should respond identically to the CO singleton and an O singleton, but it should be less activated by a C singleton. Finally, the response from a CO-tuned neuron to a CO singleton should be no less than its response to a C or O single-feature singleton. Thus, among all neurons, whether they are tuned to C, O, or CO, the highest response to the CO singleton should be no less than the highest response to the C singleton or the O singleton. Provided that, for different singletons, the statistical properties, e.g., the average and standard deviation, of the V1 neural responses to the background bars are sufficiently similar, the V1 saliency hypothesis predicts that the CO singleton will be no less salient than the C and O singletons. Since a singleton's saliency should relate inversely to the RT for finding it, $RT_{CO} \leq \min(RT_C, RT_O)$ follows.

For ease of notation, in this section we eschew the usual notation O_i for the output or response of a V1 neuron indexed by i. Instead, let α denote an input bar, and let C_α, O_α, or CO_α, respectively, denote the highest response to this bar from a population of neurons tuned solely to C, or O, or conjunctively to CO (and these neurons have their RFs cover the location of this input bar). The value α can be $\alpha = C, O,$ or CO for a C, O, or CO singleton or $\alpha = B$ for a background non-singleton bar. Hence (C_C, O_C, CO_C) is the triplet of responses to a color singleton, (C_O, O_O, CO_O) to an orientation singleton, $(C_{CO}, O_{CO}, CO_{CO})$ to a CO double-feature singleton, and (C_B, O_B, CO_B) to one of the many bars in the background. The maximum rule states that the saliency of the bar indexed by $\alpha = C, O, CO,$ or B is

$$\text{SMAP}_\alpha \equiv \max(C_\alpha, O_\alpha, CO_\alpha)\,. \tag{5.15}$$

Note that, among the neurons responding to the bar α, the number of neurons tuned to C may not be the same as the number of neurons tuned to O (or CO). However, this does not matter in our formulation since C_α, O_α, or CO_α marks the highest response to the bar from a subpopulation of neurons having a particular tuning property regardless of the number of neurons in this subpopulation.

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

$$C_{CO} = C_C, \quad O_{CO} = O_O, \tag{5.16}$$

$$C_O = C_B, \quad O_C = O_B. \tag{5.17}$$

(Note that, although, e.g., $C_O = C_B$, this C neuron is still tuned to color.) Furthermore, iso-color and iso-orientation suppression, and the strong saliency of the singletons, imply

$$C_C > C_B \quad \text{and} \quad O_O > O_B. \tag{5.18}$$

Generalizing iso-feature suppression to the conjunctive cells, we expect

$$CO_{CO} \geq CO_O, \quad CO_{CO} \geq CO_C, \tag{5.19}$$
$$CO_O \geq CO_B, \quad CO_C \geq CO_B. \tag{5.20}$$

Since the singletons $\alpha = C, O$, or CO pop out, we have

$$\text{SMAP}_\alpha \gg \text{SMAP}_B \quad \text{for } \alpha = C, O, \text{ or } CO. \tag{5.21}$$

Since $O_C = O_B$ (by equation (5.17)), then

$$\text{SMAP}_C = \max(C_C, O_C, CO_C) = \max(C_C, O_B, CO_C).$$

This, combined with $\text{SMAP}_C \gg \text{SMAP}_B$ and $\text{SMAP}_B \geq O_B$, leads to

$$\text{SMAP}_C = \max(C_C, CO_C), \quad \text{and analogously,} \quad \text{SMAP}_O = \max(O_O, CO_O). \tag{5.22}$$

Then we can derive

$$\begin{aligned}
\text{SMAP}_{CO} &= \max(C_{CO}, O_{CO}, CO_{CO}) && && (5.23)\\
&= \max(C_C, O_O, CO_{CO}) && \{\text{by equation (5.16)}\} && (5.24)\\
&= \max(C_C, O_O, \max(CO_{CO}, CO_C, CO_O)) && \{\text{by equation (5.19)}\} && (5.25)\\
&= \max(\max(C_C, CO_C), \max(O_O, CO_O), CO_{CO}) && && (5.26)\\
&= \max(\text{SMAP}_C, \text{SMAP}_O, CO_{CO}) && \{\text{by equation (5.22)}\} && (5.27)\\
&\geq \max(\text{SMAP}_C, \text{SMAP}_O)\,. && && (5.28)
\end{aligned}$$

In the above, each $\{...\}$ is not part of the equation, but it contains text pointing out the equation used to arrive at the mathematical expression to its left. Equations (5.27) and (5.28) mean that the double-feature singleton CO can be more salient than both the single-feature singletons C and O if there are conjunctive cells whose response CO_{CO} has a non-zero chance of being larger than both SMAP_C and SMAP_O to dictate the saliency of the CO singleton (this is achieved when CO_{CO} is larger than O_O, C_C, CO_C, and CO_O). When there is no conjunctive cell CO, we can simply make $CO_\alpha = 0$ in the above equations, eliminating its ability to dictate the saliency value. Then, inequality (5.28) becomes an equality:

$$\text{SMAP}_{CO} = \max(\text{SMAP}_C, \text{SMAP}_O) \quad \text{when there is no conjunctive CO neuron.} \tag{5.29}$$

The saliency SMAP_α is taken as determining the RT_α for detecting the singleton α via a monotonic function $f(.)$:

$$RT_\alpha = f(\text{SMAP}_\alpha), \quad \text{such that } f(x_1) > f(x_2) \text{ when } x_1 < x_2, \tag{5.30}$$

by the definition of saliency. Equations (5.28–5.30) then lead to

$$\begin{aligned}
RT_{CO} &= \min\left[RT_C, RT_O\right], \\
&\quad \text{the race model, when there is no conjunctive CO cell.}
\end{aligned} \tag{5.31}$$

$$\begin{aligned}
RT_{CO} &= \min\left[RT_C, RT_O, f\left(CO_{CO}\right)\right] \\
&\leq \min\left[RT_C, RT_O\right], \\
&\quad \text{double-feature advantage, with conjunctive CO cells.}
\end{aligned} \tag{5.32}$$

Hence, without conjunctive CO cells, RT_{CO} to detect a CO double-feature singleton can be predicted by a race model between two racers SMAP_C and SMAP_O with their respective racing times as RT_C and RT_O for detecting the corresponding single-feature singletons. With

conjunctive cells, there may be a double-feature advantage. The RT_{CO} can be shorter than predicted by the race model between the two racers SMAP_C and SMAP_O, since there is now a third racer, CO_{CO}, with its RT as $f(CO_{CO})$; see equation (5.32). Note that, when we say the race model for the RT of a double-feature singleton, we mean a race between *only two* racers whose racing times are the RTs for the two corresponding single-feature singletons, *without* any additional racers.

Now let us remove the deterministic idealization and treat the RTs and the V1 neural responses as stochastic, as they are in reality. The neural responses in single trials can be seen as being drawn from a probability distribution function (pdf). Thus, SMAP_C, SMAP_O, and CO_{CO} are really all random variables drawn from their respective pdfs, making SMAP_{CO} another random variable. Accordingly, the RTs are also random variables by their respective pdfs. In particular, when the race model holds, i.e., when there is no CO conjunctive cell, Monte Carlo simulation methods based on equation (5.31) can be used to predict RTs for the double-feature singleton as follows. Let us denote RT_{CO}(race) as the RT_{CO} by the race model. We randomly sample one RT each from the distribution of RT_C and that of RT_O, respectively, and call these samples RT_C(sample) and RT_O(sample). This gives a simulated sample, RT_{CO}(race sample), of RT_{CO}(race) according to the race model as

$$RT_{CO}(\text{race sample}) \equiv \min[RT_C(\text{sample}), RT_O(\text{sample})] \tag{5.33}$$

by equation (5.31). Using a sufficient number of such samples, one can generate a distribution of RT_{CO}(race). We can then test whether human RTs to detect a CO singleton is statistically shorter than RT_{CO}(race) predicted by the race model.

The response CO_{CO} of the CO neuron to the CO singleton is also stochastic, and its corresponding (would-be) RT $f(CO_{CO})$ also follows a pdf. Averaged over trials, according to equations (5.27) and (5.32), as long as this additional racer CO_{CO} has a non-zero chance of being larger than both SMAP_C and SMAP_O, the trial-averaged RT_{CO} should be shorter than the one predicted by the race model. Note that this double-feature advantage can happen even when the average response of the CO neurons are no larger than those of the C and O neurons.

We also note that, even when there is no CO cell, the race-model predicted RT_{CO}(race) can be on average (over the trials) shorter than both the average RT_C and the average RT_O (unless the distributions of RT_C and RT_O do not overlap), since RT_{CO}(race) is always the shorter one of the two single-feature RT samples.

The derivation and analysis above can be analogously applied to the double-feature singletons MO and CM, involving the motion-direction feature. Hence, the fingerprints of V1's conjunctive cells are predicted to be as follows: compared to the RT predicted by the race model from the RTs for the corresponding single-feature singletons, RTs for the CO and MO double-feature singletons should be shorter, but the RT for the CM double-feature singleton should be the same as predicted.

This fingerprint was tested (Koene and Zhaoping 2007) in a visual search task for a singleton bar (among 659 background bars) regardless of the features of the singleton, using stimuli as schematized in Fig. 5.40. Each bar is about 1×0.2^o in visual angle, takes one of the two possible isoluminant colors (green and purple) against a black background, is tilted from vertical in either direction by a constant amount, and moves left or right at a constant speed. All background bars are identical to each other in color, tilt, and motion direction, and the singleton pops out by virtue of its unique color, tilt, or motion direction, or any combination of these features. The singleton has an eccentricity 12.8^o from the initial fixation point at the center of the display in the beginning of each search trial. Subjects have to press a button as soon as possible to indicate whether the singleton is in the left or right half of the display, regardless of the singleton conditions, which are randomly interleaved and unpredictable.

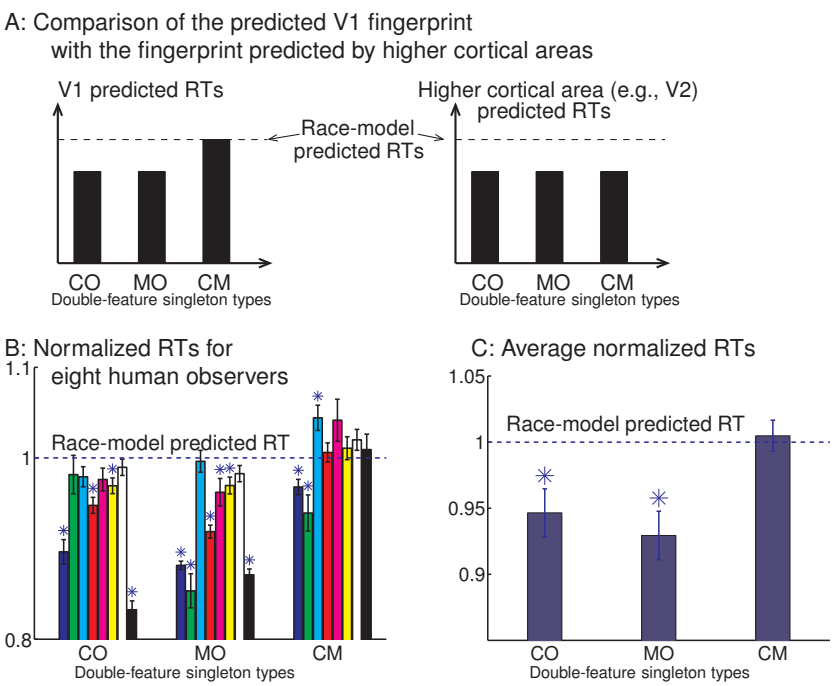


Fig. 5.41: Testing the fingerprint of V1 conjunctive cells in bottom-up saliency. A: The predicted V1 fingerprint, depicted in the left plot, compared with the prediction by V2/higher cortical areas (right plot). The dashed lines indicate the value of the predicted RT for the double-feature singletons by the race model. If bottom-up saliency in these tasks were computed by higher cortical areas, double-feature advantage, by an RT shorter than predicted from the race model, should occur in all double-feature singletons CO, MO, and CM. By contrast, V1 predicts a double-feature advantage for CO and MO singletons but not for the CM singleton. This is because V1 has conjunctive CO and MO cells but no CM cells, but V2/higher areas have all the three cell types. B,C: Experimental findings by Koene and Zhaoping (2007). The plotted bars show normalized mean RTs across trials for each subject (in B) or the average of these means across subjects (in C). The normalization factor comes from the predictions of the race model. Error bars indicate the standard errors of the means. In C, by matched sample two-tailed t-tests, the observed RT_{CO} and RT_{MO} for the double-feature singletons CO and MO are significantly shorter than those predicted by the race model, whereas the observed RT_{CM} for the double-feature singleton CM is not significantly different from the race-model prediction. In B and C, a "*" above a data bar indicates a significant difference between the RT data and that predicted from a race model.

Trials with incorrect button presses or with RTs shorter than 0.2 seconds or longer than three standard deviations above the average RTs were excluded from data analysis.

The experiment by Koene and Zhaoping (2007) was designed such that there was a symmetry between the two possible feature values in each feature dimension, i.e., between the isoluminant green and purple colors, between left-tilt and right-tilt orientations, and between the leftward and rightward movements. Hence, for our derivation, the highest responses from the V1 neurons to a bar is assumed to be regardless of whether the bar takes one or the other of the two possible feature values, e.g., C_C (the highest response of V1 neurons tuned to color only to a color singleton bar) is regardless of whether a color singleton is the unique green

bar among purple bars or the other way around (even though the highest response to a green singleton is from a cell tuned to green and that to a purple singleton is from a cell tuned to purple). This feature symmetry allows us to pool together the RT data for symmetry-related singletons, e.g., a green singleton and a purple singleton, in the data analysis.

Figure 5.41 BC plot the observed RTs for the double-feature singletons, normalized by the RTs predicted by the race model. For example, for each observer, a distribution of RT_{CO}(race) can be predicted from the histograms of the behavioral RT_C and RT_O data from this observer, using the Monte Carlo method above. The normalized RT of this observer for the CO singleton is his/her behavioral RT to detect the CO singleton divided by the average RT_{CO}(race) predicted by the race model for the same observer. Therefore, a double-feature advantage is manifest in a normalized RT smaller than unity, and a race-model predicted RT gives a normalized RT equal to unity. The results confirm the predicted V1 fingerprint. A double-feature advantage for the CO singleton has been previously observed (Krummenacher, Müller and Heller 2001). Similarly, a lack of double-feature advantage has also been observed when both features are in the orientation dimension (Zhaoping and May 2007), consistent with the V1 saliency hypothesis, since there is no V1 cell conjunctively tuned to two different orientations.

Note that traditional models of saliency would predict that a double-feature singleton should, if anything, be more salient than the single-feature singletons. In particular, recall from Section 5.1.3 that the traditional models should predict that, in the experiment by Koene and Zhaoping, a CO singleton should be more salient than an O singleton in the same way, and by the same amount, as a CM singleton is more salient than an M singleton. These predictions arise from the separation between feature maps and from the summation rule. The observations shown in Fig. 5.41 refute these predictions.

5.5.4 A zero-parameter quantitative prediction and its experimental test

Equation (5.31) shows that if there were no V1 neuron tuned simultaneously to both C and O, then one could quantitatively predict RT_{CO} from RT_C and RT_O. Hence, for example, an $RT_C = 500$ ms and an $RT_O = 600$ ms could together predict $RT_{CO} = \min(RT_C, RT_O) = 500$ ms from the race model, without any free parameters (see Fig. 5.40). As both RT_O and RT_C follow their respective probability distributions, the probability distribution of RT_{CO} could also be quantitatively derived without any parameters, by drawing random samples of RT_{CO} as $RT_{CO} = \min(RT_O, RT_C)$ from sample pairs (RT_O, RT_C).

However, because V1 has neurons tuned conjunctively to C and O, the measured probability distribution of RT_{CO} is different from this race-model prediction derived just from the distributions of RT_C and RT_O. This is shown in Fig. 5.42 B.

We mentioned in the previous section that few CM V1 neurons have been found that are tuned conjunctively to color (C) and motion direction (M). Hence, when RT_{CM} is the RT to find a double-feature singleton unique in both color (C) and motion direction (M) and RT_M is the RT to find a singleton unique in motion direction, the distribution of RT_{CM} is the same as that of $\min(RT_C, RT_M)$. This is consistent with the observation that, across observers, the average RT_{CM} is indeed the same as that predicted by the race model (by drawing random samples of RT_{CM} as $RT_{CM} = \min(RT_C, RT_M)$) from sample pairs (RT_C, RT_M). However, a closer observation of Fig. 5.41 B suggests that $RT_{CM} < \min(RT_C, RT_M)$ (averaged across trials) for two out of the eight observers. Indeed, findings by different researchers differ as to the existence of CM cells in V1 (Horwitz and Albright 2005, Michael 1978)—perhaps they exist in some observers but are just less numerous than CO and MO double-feature conjunctive cells.

Even if there are some CM neurons in V1, there has yet to be a report of triple feature conjunctive cells, CMO, which are simultaneously tuned to color (C), motion direction (M), and orientation (O) features. (Note that the CMO cells should be a subset of the CM cells, and hence they cannot be more numerous than the CM cells.) Indeed, this dearth is as expected from the input signal-to-noise considerations discussed in Section 3.6.9 that preclude substantial numbers of V1 neurons from being tuned to multiple feature dimensions. Hence, we can use exactly the same idea as the one that led to $RT_{CO} = \min[RT_O, RT_C]$ in equation (5.31) to derive the following parameter-free equation (Zhaoping and Zhe 2012b) from the V1 saliency hypothesis:

$$\min(RT_C, RT_M, RT_O, RT_{CMO}) = \min(RT_{CM}, RT_{CO}, RT_{MO}). \tag{5.34}$$

In this expression, CO, MO, and CM are the conjunctions of two features indicated by the respective letters (C, M, and O), and CMO is the triple feature conjunction of C, M, and O, and RT_α is the RT to find a single-, double-, or triple-feature singleton denoted by α, which can take values $\alpha = C$, M, O, CM, MO, CO, or CMO.

To derive equation (5.34), we proceed as in the last section. Let C, M, O, CM, CO, and MO denote V1 neurons tuned to a single or double (conjunctive) feature(s) indicated by the respective letters, and let C_α, M_α, O_α, CM_α, CO_α, and MO_α be the responses of these neurons to singleton α or a background item $\alpha = B$. Then, as in equation (5.15), the saliency at the location of item α is

$$\text{SMAP}_\alpha = \max(C_\alpha, M_\alpha, O_\alpha, CM_\alpha, CO_\alpha, MO_\alpha). \tag{5.35}$$

Just as in equations (5.16–5.20), the neurons should respond more vigorously to a singleton whose feature uniqueness matches more of its preferred tuning, and its response should be indifferent to feature contrast in a dimension to which it is not tuned. Hence, statistically,

$$\begin{gathered}
C_C = C_{CO} = C_{CM} = C_{CMO} > C_B = C_O = C_M = C_{MO}, \\
O_O = O_{CO} = O_{MO} = O_{CMO} > O_B = O_C = O_M = O_{CM}, \\
M_M = M_{CM} = M_{MO} = M_{CMO} > M_B = M_C = M_O = M_{CO}, \\
CM_{CM} = CM_{CMO}, \quad CM_M = CM_{MO}, \quad CM_C = CM_{CO}, \quad CM_B = CM_O, \\
CO_{CO} = CO_{CMO}, \quad CO_C = CO_{CM}, \quad CO_O = CO_{MO}, \quad CO_B = CO_M, \\
MO_{MO} = MO_{CMO}, \quad MO_M = MO_{CM}, \quad MO_O = MO_{CO}, \quad MO_B = MO_C.
\end{gathered} \tag{5.36}$$

Furthermore, since the singletons are very salient, we have, for $\alpha = C, M, O, CM, CO, MO$ and CMO,

$$\text{SMAP}_\alpha > \text{SMAP}_B = \max(C_B, M_B, O_B, CM_B, CO_B, MO_B). \tag{5.37}$$

From these equations, one can derive

$$\begin{aligned}
&\max(\text{SMAP}_C, \text{SMAP}_M, \text{SMAP}_O, \text{SMAP}_{CMO}) = \\
&\qquad \max(\text{SMAP}_{CM}, \text{SMAP}_{CO}, \text{SMAP}_{MO}),
\end{aligned} \tag{5.38}$$

which is the saliency equivalent of the RT equation (5.34) due to the monotonically inverse relationship between SMAP_α and RT_α. The above equation can be verified by substituting equation (5.35) for each SMAP_α in equation (5.38), using properties in equation (5.36) and noting that

$$\max[\max(a, b, ...), \max(a', b'...), ...] = \max[a, b, ..., a', b', ...] \tag{5.39}$$

for various quantities a, b, a', and b', etc.

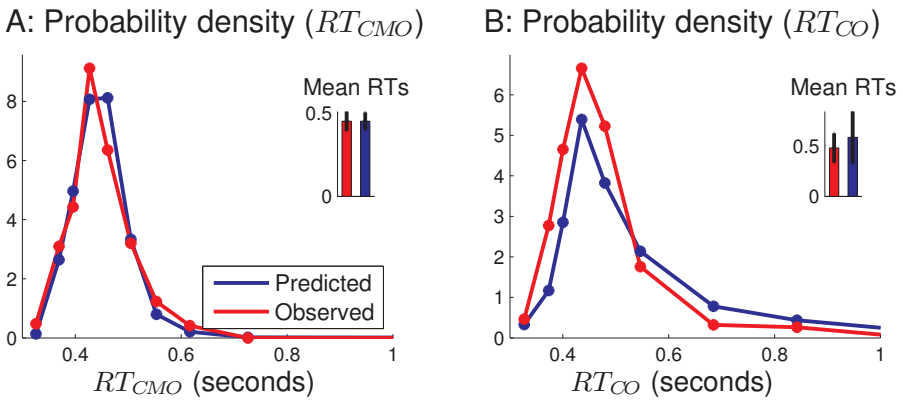


Fig. 5.42: Testing a quantitative prediction from the V1 saliency hypothesis using data collected by Koene and Zhaoping in the same experiment as in Fig. 5.41. Only six out of the eight observers in the experiment had the complete set of data on all feature singletons. A: The distribution of RT_{CMO} for one of the six observers is predicted from the other RTs of the same observer according to equation (5.34). The predicted and observed quantities are plotted in blue and red, respectively. In comparison, B shows the disconfirmation (using data from the same observer) of the incorrect prediction, $RT_{CO} = \min(RT_C, RT_O)$ (i.e., predicting RT_{CO} from RT_C and RT_O by a race model), which does not arise from the hypothesis, because of the presence of CO neurons in V1. In A, but not in B, the predicted distribution is not significantly different from the observed distribution.

Let $RT_1 \equiv \min(RT_C, RT_M, RT_O, RT_{CMO})$ and $RT_2 \equiv \min(RT_{CM}, RT_{CO}, RT_{MO})$. Then equation (5.34) states that $RT_1 = RT_2$. Because neural responses are stochastic, the actual equality is between the probability distributions of RT_1 and RT_2, respectively. Given the observed distributions of RT_C, RT_O, RT_M, RT_{CM}, RT_{CO}, and RT_{MO}, one can derive the distribution of RT_{CMO} by finding the one which minimizes the difference (quantified by some appropriate measure) between the probability distributions of RT_1 and RT_2, respectively.

Figure 5.42 A shows that the predicted and observed distributions of RT_{CMO} agree with each other quantitatively, up to the noise in estimating these distributions that comes from the finite number of search trials. Statistical analysis confirms that the difference between the predicted and observed distributions is not significant for any of the six observers. By contrast, a test of the incorrect prediction $RT_{CO} = \min(RT_C, RT_O)$ of RT_{CO} from a race model that assumes no CO neuron fails, since there is a significant difference between the predicted and the observed distributions of RT_{CO} for most of the six observers. Figure 5.42 shows an example. Two other (examples of) incorrect predictions (which cannot be predicted by the V1 theory without additional requirements on V1 physiology) of RT_{CMO} by the race equations $\min(RT_C, RT_{MO}, RT_{CMO}) = \min(RT_{CM}, RT_{CO}, RT_M, RT_O)$ and $RT_{CMO} = \min(RT_C, RT_M, RT_O)$, respectively, also fail to agree with data for at least some observers.

The agreement between the quantitative prediction and experimental data further supports the idea that V1 is the substrate for saliency computation. This is because higher cortical areas (such as V2) downstream along the visual pathway do have neurons tuned to the triple or more conjunctions of simple features,[19] as expected from the general observation that neural selectivities become more complex in higher visual cortical areas. Since our prediction

[19] Private communication from Stewart Shipp, 2011.

requires an absence of these triple conjunctive cells, it is unlikely that the higher cortical areas, instead of V1, used the maximum rule (from equation (5.4)) to compute saliency for the feature singletons in our search stimuli. Otherwise, the predicted RT_{CMO} would be statistically longer than the observed RT_{CMO}, just like the race-model predicted RT_{CO} is longer than the RT_{CO} in reality.

5.5.5 Reflections—from behavior back to physiology via the V1 saliency hypothesis

Recall from Chapter 1 that one of the aims of a theory is to link physiological and behavioral observations. So far, we have used the V1 saliency hypothesis to predict behavior in visual search and segmentation tasks from the physiological properties of V1. From a surprisingly "impossible" prediction of a salient ocular singleton to a quantitative prediction derived without any free parameters, experimental confirmations of these predictions build confidence in the theory.

These successes encourage us to reverse the direction of prediction by applying this theory to predict V1 neural properties from behavioral data. For example, a significant difference between the distribution of the behavioral RT_{CO} and that of the race model $RT_{CO} = \min(RT_C, RT_O)$, seen in Fig. 5.42 B, predicts a non-trivial contribution to the saliency of the CO singleton from CO neurons in V1; see equations (5.27) and (5.32). This predicted contribution is the portion that is beyond that by the same CO cells to the saliency of the single-feature (C and O) singletons, and it can be quantitatively assessed from the behavioral RT data (Zhaoping and Zhe 2012a). We predict that the V1 CO neurons should respond to their preferred conjunction of C and O features more vigorously when this conjunction is a double-feature rather than a single-feature singleton. Consequently, the contextual suppression on a CO cell responding to this conjunction is predicted to be weaker when the contextual inputs differ from this conjunction in both, rather than just one of, the C and O dimensions. Analogous predictions hold for the contextual influences on the MO neurons in V1.

5.6 The roles of V1 and other cortical areas in visual selection

According to Fig. 2.3 and Fig. 2.29, V1 is just one of the visual areas that send signals to the superior colliculus (SC) to control gaze. The SC also receives inputs from the retina, extrastriate areas, lateral intraparietal cortex (LIP), and the frontal eye field (FEF). Figure 5.43 shows a simplified schematic of the brain areas involved in gaze control. If we identify gaze control with the control of selection (ignoring covert selection), then it is likely that other brain areas must also contribute to selection, i.e., the guidance of attention.

It is instructive to imagine that the decision as to where, or to what, to direct attention is by a committee. Various brain areas, including V1, are the committee members which send their contributions to the decision; and the SC transforms the decision by the committee to motor responses. The impact of each brain area on the decision is determined by various factors, including the strength and timeliness of its contribution. Hence, some decisions are dominated by top-down effects, while others by bottom-up ones.

It is generally believed that the frontal and parietal brain areas are involved in many top-down aspects of attentional selection (Desimone and Duncan 1995, Corbetta and Shulman 2002). By contrast, observations in Section 5.3 and Section 5.5.4 suggest that cortical areas beyond V1 play little role in the bottom-up control of selection mediated by the saliency of the

very salient singletons in the eye of origin, color, orientation, and motion direction. However, the V1 saliency hypothesis does not preclude additional influences from other cortical areas to bottom-up selection for more complex stimuli. It is therefore important to understand the extent of their contribution.

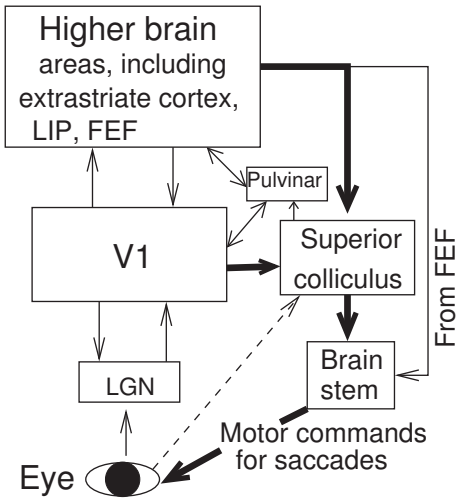


Fig. 5.43: Brain areas governing gaze control (Schiller 1998). V1 is only one of a number of areas contributing to the control of gaze. In monkeys and cats, the retina plays a very limited role in controlling saccades driven by visual inputs (Schiller et al. 1974, Schiller 1998). The role of higher visual areas can be assessed by investigating the influence on gaze control of aspects of visual perception that are not processed by V1.

First, various observations suggest that the retina plays little role in visually guided saccades in normal (non-lesioned) monkeys and cats, even though it can play a role in stabilizing retinal images during viewer or scene motion via their projection to the accessory optic system (Schiller 1998). In monkeys, a very small and relatively poorly understood fraction of retinal ganglion cells, called W cells, projects to the superficial layers of the SC (Schiller 1998). The axons of these neurons conduct spikes more slowly than parvo- and magnocellular ganglion cells (Schiller and Malpeli 1977). When V1 in monkeys or cats is removed or cooled, neurons in the intermediate and deep layers of the SC, and in particular, those eye-movement cells which activate to evoke saccades, can no longer be driven by visual stimuli. This is the case even though the animals can still make saccades in the dark (but not in response to visual stimulation) and even though these cells still fire before non-visually guided saccades (Schiller et al. 1974, Schiller 1998), which can be controlled by FEF, which can bypass the SC to control saccades. Also, monkeys suffering V1 lesions do not have proper visually guided saccades for up to two months after the lesion (Isa and Yoshida 2009).

The LGN lacks direct input to the SC. Thus, apart from the retina, only V1 and brain areas downstream along the visual pathway can be responsible for visually guided saccades or selection for normal primates. One key difference between V1 and downstream areas is latency—the latter typically have longer latencies than V1 in response to visual input, and so their contributions to bottom-up selection, or to top-down selection contingent on visual input, are likely to lag behind that of V1 (Bullier and Nowak 1995, Schmolesky, Wang, Hanes, Thompson, Leutgeb, Schall and Leventhal 1998). One can imagine situations in which V1's contribution is so strong and fast that the non-V1 contributions could be too slow to have an

impact. The non-V1 contribution could also be ignored if it is too weak, or if it and V1's contributions are redundant. Conversely, it could be substantial when V1's contribution is too weak to reach a quick decision.

To investigate the respective contributions by V1 and other brain areas in selection, we focus on those selections which are contingent on external visual inputs, assuming that V1 plays little role in the other selections. Note that top-down and task-driven factors can influence input contingent selections. This is because, when selection has a sufficiently long latency after visual input, the gist of the scene obtained by observers during the latency could exert influence associated with the knowledge of the scene or ongoing tasks.

5.6.1 Using visual depth feature to probe contributions of extrastriate cortex to attentional control

To explore contributions to bottom-up selection beyond V1, it helps to identify visual processes that are carried out in higher visual areas but not in V1, and to investigate how these visual processes guide selection. A good candidate is stereo vision, which analyzes surfaces and their depth orders to achieve the perception of three-dimensional (3D) surfaces. Even though V1 cells are tuned to binocular disparities, 3D perception requires stereo processes to suppress false matches, which occur between visual inputs to the two eyes arising from two different object features in the scene (see Fig. 6.7 C). It is known that these stereo matching processes aimed at surface perception are centered outside V1, notably in V2 (Cumming and Parker 2000, Bakin et al. 2000, von der Heydt et al. 1984, von der Heydt et al. 2000, Qiu and von der Heydt 2005, Janssen, Vogels, Liu and Orban 2003). Hence, attentional guidance by depth or 3D cues should reflect contributions coming from beyond V1.

It has been shown (Nakayama and Silverman 1986, He and Nakayama 1995) that searching for a target defined by a unique conjunction of depth and another feature is much easier than typical conjunction searches that lack the depth feature (e.g., the color-orientation conjunction in Fig. 5.3 E). This suggests that 3D cues can help direct attention to task-relevant locations. We can measure and compare selection with and without 3D cues while the 2D cues are held constant. The speed-up of attentional guidance by the 3D cues is identified as a contribution from beyond V1.

One such study (Zhaoping, Guyader and Lewis 2009) is an extension to the experiment shown in Fig. 5.36, which was used to test the maximum rule in V1 saliency computations. In that experiment, the segmentation of a task-relevant texture was subject to interference from a superposed task-irrelevant texture surface. Denote the task-relevant image (texture A in Fig. 5.36) by $I_{\rm rel}$, the task-irrelevant image (texture B in Fig. 5.36) by $I_{\rm ir}$, and the composite image (texture C in Fig. 5.36) as $I_{\rm com} = I_{\rm rel} + I_{\rm ir}$. The interference from $I_{\rm ir}$ can be reduced when $I_{\rm ir}$'s position is slightly shifted horizontally from $I_{\rm rel}$ by a displacement x. Let us denote this shifted version of $I_{\rm ir}$ as $I_{\rm ir}(x)$, and the resulting composite image as $I_{\rm com}(x) = I_{\rm rel} + I_{\rm ir}(x)$; see Fig. 5.44. The RT for segmenting $I_{\rm com}(x)$ is less than that for segmenting the original composite $I_{\rm com}$. (One can also simulate the V1 model in Section 5.4 and confirm that the V1 saliency value at the texture border is higher in the 2D offset images $I_{\rm com}(\pm x)$ than in the original $I_{\rm com}$.) A horizontal shift $-x$ of $I_{\rm ir}$ in the opposite direction, giving a composite image $I_{\rm com}(-x)$, would reduce the RT just as effectively. These RT reductions are not caused by any 3D cues, since exactly identical textures $I_{\rm com}(\pm x)$ are presented to the two eyes.

If $I_{\rm com}(x)$ and $I_{\rm com}(-x)$ are viewed dichoptically by the two eyes, the percept is 3D: the two texture surfaces $I_{\rm rel}$ and $I_{\rm ir}$ separate in depth (see Fig. 5.44). Whether $I_{\rm rel}$ appears in front of or behind $I_{\rm ir}$ depends on whether the right or left eye sees $I_{\rm com}(x)$. If the separation in

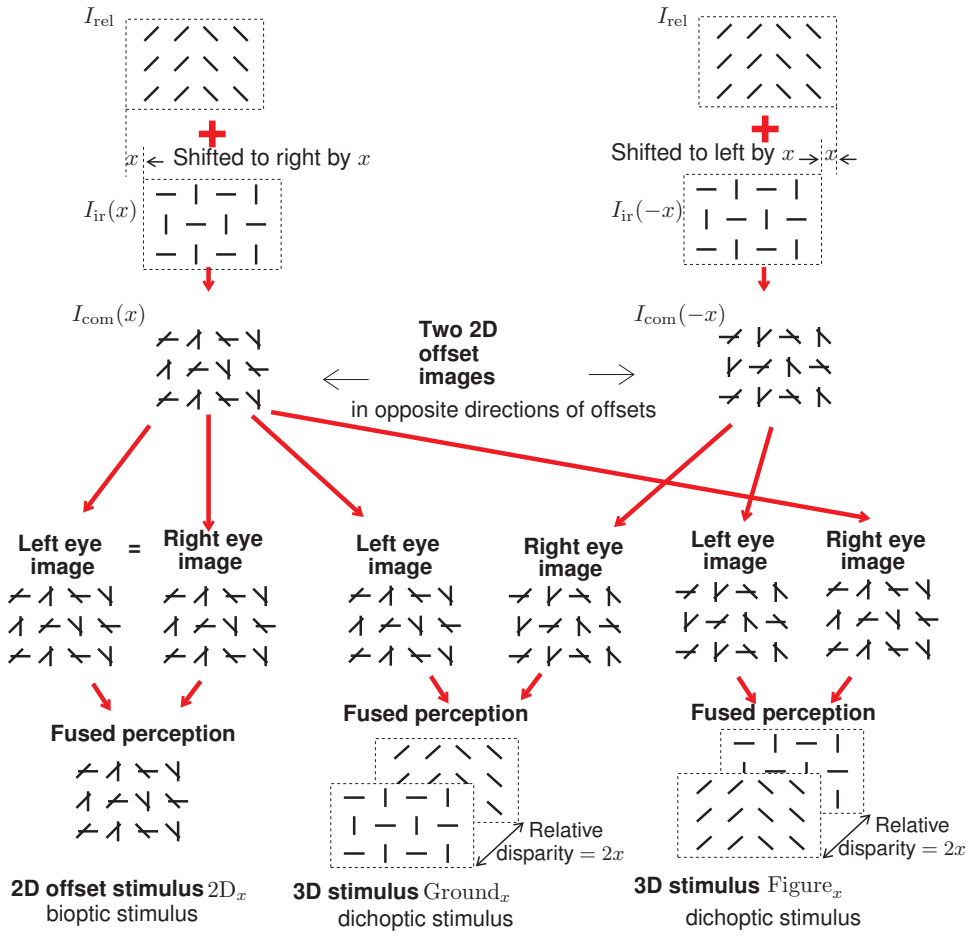


Fig. 5.44: Construction of 2D and 3D stimuli used to assess the contribution to selection of 3D processes in brain areas beyond V1. The texture images I_{rel} are as textures A in Fig. 5.36, and texture images $I_{\text{ir}}(\pm x)$ are spatially shifted (horizontally by $\pm x$) versions of texture B in Fig. 5.36. Superposing I_{rel} and $I_{\text{ir}}(\pm x)$ makes $I_{\text{com}}(\pm x)$, and $I_{\text{com}}(x = 0)$ is as texture C in Fig. 5.36. The bottom row shows the 2D offset stimulus 2D_x, created by presenting the 2D offset image $I_{\text{com}}(x)$ (or $I_{\text{com}}(-x)$) identically to both eyes, and the 3D stimuli Ground_x and Figure_x, created by presenting $I_{\text{com}}(x)$ to one eye and $I_{\text{com}}(-x)$ to the other. The relative disparity between I_{rel} and I_{ir} in the 3D stimuli is $2x$. Adapted with permission from Zhaoping, L., Guyader, N., and Lewis, A., 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), article 20, doi: 10.1167/9.11.20, Fig. 2, copyright © 2009, ARVO.

depth makes segmentation faster than for the 2D percept arising from binocular (sometimes called bioptic) viewing of $I_{\text{com}}(x)$, the post-V1 3D effect should be credited.

Denote the 2D bioptic stimulus when the 2D offset image $I_{\text{com}}(x)$ or $I_{\text{com}}(-x)$ is presented identically to both eyes as 2D_x, and the 3D dichoptic stimuli when $I_{\text{com}}(x)$ and $I_{\text{com}}(-x)$ are presented to different eyes as Figure_x and Ground_x, when I_{rel} is in the foreground or background, respectively. These stimuli share the same 2D cues, notably the same

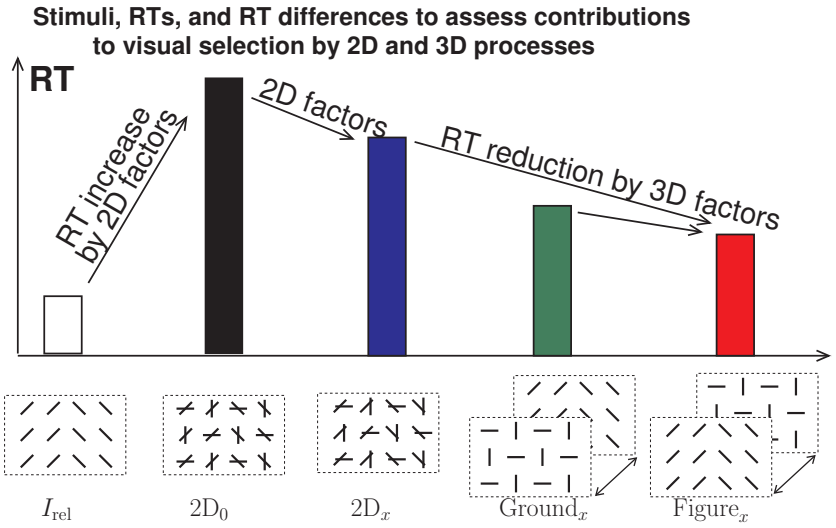


Fig. 5.45: The contributions of 2D and 3D processes to selection are manifest in the differences between RTs for texture segmentation using five different types of texture stimuli: I_{rel}, $2\mathrm{D}_0$, $2\mathrm{D}_x$, Figure_x, and Ground_x. Each of the last four stimuli (created as in Fig. 5.44) contains two texture surfaces, I_{rel}, which is task relevant, and I_{ir}, which is task irrelevant. These two surfaces are placed at the same depth, as in $2\mathrm{D}_0$ and $2\mathrm{D}_x$, or at different depths, as in Figure_x and Ground_x, in which I_{rel} is in the foreground and background respectively. The 2D offset stimulus $2\mathrm{D}_x$ has a spatial offset $\pm x$ between textures I_{rel} and I_{ir}; this offset is zero in $2\mathrm{D}_0$. The contribution of 3D processes to selection should be manifested in the RT difference $RT(2\mathrm{D}_x) - RT(\mathrm{Figure}_x)$, and it is perhaps also manifested in the RT difference $RT(\mathrm{Ground}_x) - RT(\mathrm{Figure}_x)$ regardless of the eye dominance. Adapted with permission from Zhaoping, L., Guyader, N., and Lewis, A., 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), article 20, doi: 10.1167/9.11.20, Fig. 3, copyright © 2009, ARVO.

2D positional offset between the task-relevant and task-irrelevant textures. However, the 3D stimulus has an additional 3D cue, the depth separation between the two textures, to which 3D perception is sensitive. If $RT(2\mathrm{D}_x)$, $RT(\mathrm{Figure}_x)$, and $RT(\mathrm{Ground}_x)$ are the RTs for the segmentation task for the corresponding stimuli (see Fig. 5.45), then any 3D contribution to selection is likely to manifest as the following two differences between the RTs,

$$\text{the first RT difference} \equiv RT(2\mathrm{D}_x) - RT(\mathrm{Figure}_x) \quad \text{and}$$
$$\text{the second RT difference} \equiv RT(\mathrm{Ground}_x) - RT(\mathrm{Figure}_x),$$

being positive. The second RT difference may be positive if the task-relevant surface in the foreground helps steer attention.

The result was that these differences were only significantly positive, indicating a contribution from 3D processes, for RTs that were at least 1 second long; see Fig. 5.46. This RT is the time it takes for observers to press one of the two buttons to report whether the texture border in I_{rel} component is in the left or right half of the visual display. Assuming that it takes subjects around 300–400 ms after making the decision actually to press the button, this one second RT implies a roughly 600–700 ms RT for the task decision. The data in Fig. 5.46 thus suggest that if the saliency signal from V1 is sufficiently fast and adequate for the task, then a

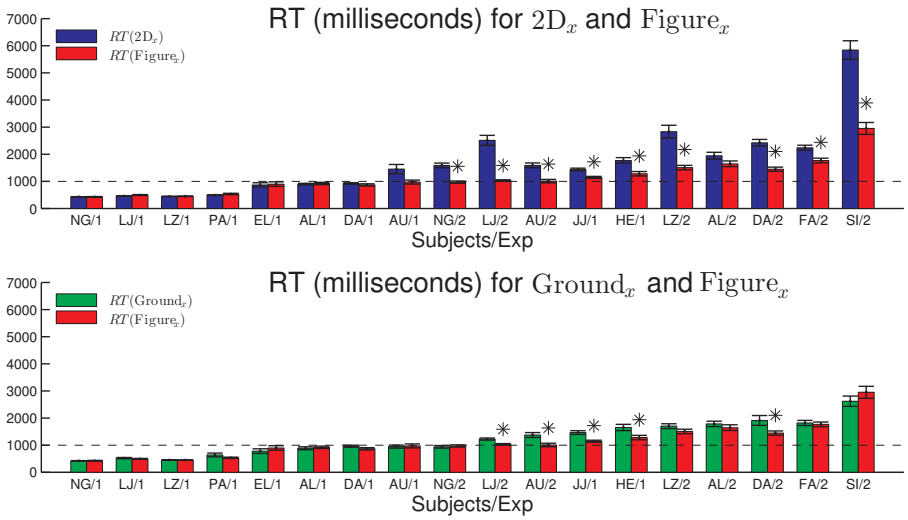


Fig. 5.46: Contributions from 3D visual processes to selection in a segmentation task are absent unless observers take at least 1000 ms to register their decision. The task is depicted in Fig. 5.45. RTs are plotted for various subjects in experiments 1 and 2. These two experiments differ in that the orientation contrast at the texture border in I_{rel} is 90^o in experiment 1 (with I_{rel} as in Fig. 5.45), and 14^o in experiment 2 (I_{rel}, not shown, has roughly oblique texture bars). Experiment 2 is designed to reduce the saliency of the texture border so that a longer RT is required. The horizontal axes label the subjects and the experiments in which these RTs were obtained, e.g., Subject/Exp = "LZ/2" means subject LZ in experiment 2. A "*" on top of the red data bar (for $RT(\mathrm{Figure}_x)$) indicates a significantly larger (two-tailed t-test) $RT(2D)$ or $RT(\mathrm{Ground}_x)$ than $RT(\mathrm{Figure}_x)$. Data from Zhaoping, L., Guyader, N., and Lewis, A., 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), article 20, 2009, doi: 10.1167/9.11.20.

decision can be made quickly without waiting for contributions from higher brain areas. This situation should apply for cases in which subjects can respond manually within 1 second for the task, regardless of whether the stimuli are 2D or 3D.

However, if the visual input is such that the V1 saliency signal is inadequate for a quick and confident task decision, contributions from higher brain areas can be important. This situation should apply to 3D stimuli like Figure_x and Ground_x, when their monocular 2D component images do not give rise to adequate V1 saliency signals for the task within a short time. Then, depth perception by higher brain areas aids attentional guidance. If the additional contribution from higher visual areas is absent, or it and the contribution from V1 are redundant, and the V1 contribution is weak, then the task RT can be long. This should apply to the situation in which observers take longer than 1 second to respond to the stimuli $2D_x$ or $2D_0$ without 3D cues. In other words, the findings suggest that, at least for depth processing, extrastriate areas do not contribute to input contingent selection immediately upon visual input, but they do contribute around several hundred milliseconds after the input onset. Meanwhile, V1 dominates in controlling selection within this initial time window after visual input onset or after a sudden unpredicted change to the visual scene.

It has been observed (Einhäuser, Spain and Perona 2008) that human saccades on static photographs are better predicted by visual objects (i.e., recognizable objects which are likely

meaningful to the viewers) than by saliency. However, the first few saccades are very similar to those made by observers who have visual object agnosia (Mannan, Kennard and Husain 2009), suggesting that the early saccades are primarily controlled by bottom-up saliency rather than object processes occurring outside V1. These findings are consistent with our observations using the depth feature.

In everyday life, we may divide our visual experience into separate episodes, each being defined by our visual exposure to a constant environment, such as a room or an outdoor field. A typical visual episode may last many seconds, minutes, or hours. Many of the eye movements in an episode, such as directing gaze to a kitchen counter to look for a kettle, are controlled by our knowledge of the environment. This knowledge can be obtained very quickly, within perhaps a second or so after we enter a new environment (and aided by our stored knowledge of what typical kitchens, or other environments, look like). Viewed from the perspective of temporal durations of selection control, V1's contribution to attentional guidance is confined to just the first second, and will only be a tiny fraction of the total contributions from all brain areas in typical daily situations. However, by exerting initial control, V1's role in selection is special.

5.6.2 Salient but indistinguishable inputs activate early visual cortical areas but not the parietal and frontal areas

Neural activities correlated or associated with bottom-up saliency (which we call saliency-like signals) have been observed in LIP (Gottlieb et al. 1998, Bislay and Goldberg 2011, Arcizet, Mirpour and Bisley 2011) and FEF (Thompson and Bichot 2005). Areas such as the pulvinar and V4 have also been observed to be linked with attentional guidance or maintenance (Schiller and Lee 1991, Robinson and Petersen 1992, Mazer and Gallant 2003). These areas, and particularly a fronto-parietal network, are often considered to organize the guidance of attention (Corbetta and Shulman 2002) or the maintenance of information associated with attention, especially in task-dependent manners. Thus, it is important to know whether their computational role is confined to controlling top-down and task-dependent attention, and whether V1 is the origin of the saliency-like signals which are received in order to be combined with top-down signals in the computation. Importantly, saliency-like signals have always been evoked in experiments using salient inputs which are highly distinctive perceptually. Hence, the relevant activities in these regions might come from recognizing or perceiving the input, rather than saliency itself.

Attentional capture by an eye-of-origin singleton, shown in Fig. 5.9, indicates that some inputs can be quite salient without evoking any perceptual awareness of how they are distinct from background items. If there is a difference between the neural response to such a salient input and the response to a perceptually identical, yet non-salient, background input, this difference should represent a relatively pure signal for saliency, with minimal contamination by non-saliency factors. Identifying brain areas that exhibit such a pure saliency signal can help us to identify brain areas involved in computing saliency. Assessing this was the intent of an experiment (Zhang, Zhaoping, Zhou and Fang 2011) in which functional magnetic resonance imaging (fMRI) and event-related electroencephalography potentials (ERPs) were used to probe neural responses to salient inputs that are not perceptually distinctive.

In the experiment, we used an input texture containing a regular array of iso-oriented bars except for a foreground region of 2×2 bars tilted differently from the background bars; see Fig. 5.47. This whole texture was shown for only 50 ms and was quickly replaced by a high-contrast mask texture (which itself lasted for just 100 ms). Meanwhile, the observers directed their gaze to an ever-present fixation point. The foreground texture could only be at one of the two possible locations, which were 7.2^o to the lower left or lower right of the fixation point.

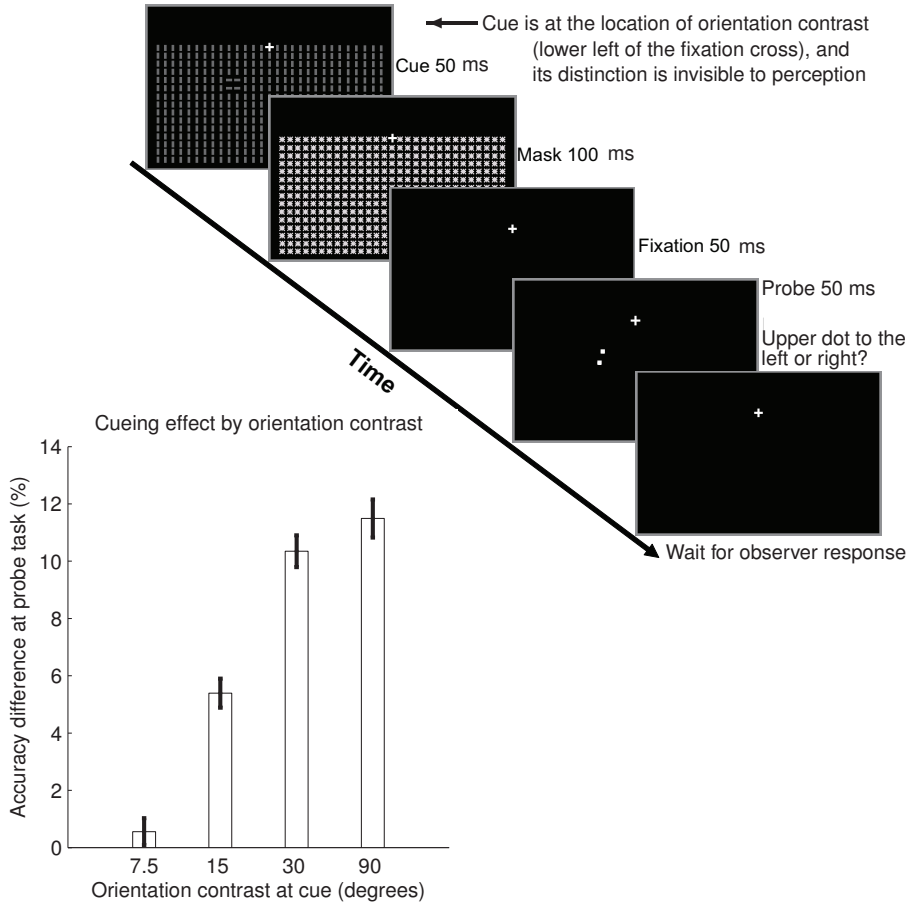


Fig. 5.47: A salient cue, which improves discrimination at the cued location but whose distinction from background inputs is invisible to perception. Upper: stimuli for the cue and probe. The cue, at the orientation contrast (which may be 0^o, i.e., no cue, or 7.5^o, 15^o, 30^o, or 90^o), and the probe, the two dots, appeared at the same (in this figure) or different locations when the probe was cued or uncued, respectively. The cross in each image was the fixation point. Observers reported whether the upper dot was to the left or right of the lower dot in the probe. Lower: the cueing effect was the increase in the percentage of the correct reports by observers in the cued relative to the uncued condition. Adapted from *Neuron*, 73 (1), Xilin Zhang, Li Zhaoping, Tiangang Zhou, and Fang Fang, Neural Activities in V1 Create a Bottom-Up Saliency Map, pp. 183–92, figures 1 and 2, copyright © (2012), with permission from Elsevier.

However, since the presentation was so brief and the mask so powerful, observers could not tell whether the foreground region was at one or the other location, even if forced to guess. Nevertheless, given sufficient orientation contrast, the foreground was salient in that it could serve as an effective exogenous cue to influence the discrimination of an input probe stimulus that was shown for 50 ms, starting 50 ms after the mask.

The probe consisted of two small dots; observers had to report whether the upper dot was to the left or right of the lower dot. This task was difficult, such that the accuracy of the task performance, i.e., the percentage of the reports that were correct, was typically only about

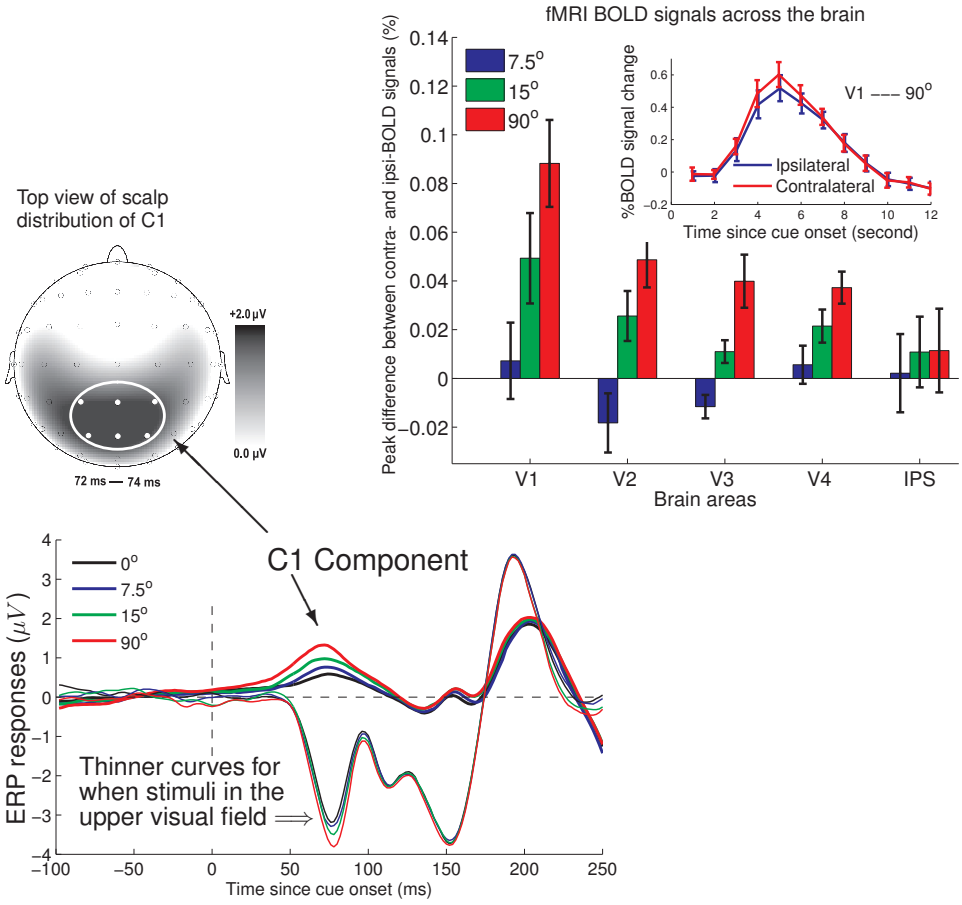


Fig. 5.48: Brain responses to the salient cue in Fig. 5.47. Brain activations averaged over observers were seen in the ERPs (lower left). The thinner curves were ERP responses (initially negative) to visual stimuli shown in the upper visual field for comparison. The C1 component in ERP, emerging around 55 ms and peaking around 73 ms from the cue onset, was mainly caused by V1 activities in the occipital region at the back of the scalp. Its polarity depended on whether the evoking stimulus was in the upper or lower visual field. The upper right plot shows cue-evoked brain activations probed by fMRI in V1–V4 and the intraparietal sulcus (IPS). They (plotted as colored bars) were significant in V1–V4 but not in the IPS. Each bar marks the difference between the peak BOLD (blood-oxygenation-level-dependent) responses to the cue at cortical locations (in a particular retinotopic brain area) contralateral and ipsilateral to the cue. The inset shows the time courses of the BOLD responses in the region of the retinotopic V1 for the cue location (contralateral) and in the corresponding region in the opposite hemisphere (ipsilateral), when the orientation contrast at the cue was 90^o. All data in this figure are for when the visual stimuli were in the lower visual field, except for the thinner curves in the ERP responses (lower left). Data from Zhang, X., Zhaoping, L., Zhou, T., and Fang, F., Neural activities in V1 create a bottom-up saliency map, *Neuron*, 73(1): 183–192, 2012.

70%. However, performance on cued trials, when the probe was shown at the same location as the cue, was up to more than 10% better than on uncued trials, when the probe was shown at the other location where the cue could have been. This increase in the accuracy is called the cueing effect, and it makes manifest the saliency of the cue, despite the fact that the cue's distinction was invisible to the observers.

The cueing effect was significant when the orientation contrast between the bars at the cue and those in the background was more than 7.5^o. The basic cueing effect, and its increase with orientation contrast, was also observed using other probe tasks, e.g., discriminating the motion direction of moving dots or the orientation of a Gabor patch.

We used fMRI to compare the blood-oxygenation-level-dependent (BOLD) signals in regions of the brain processing the visual stimuli for the two possible locations of the salient but non-distinctive cue. A difference between these signals at these two locations is seen as a saliency signal. Areas V1, V2, V3, and V4 exhibited significant differences, but the intraparietal sulcus (IPS, which is thought to contain the human homologue of LIP) did not.

Briefly, increased neural processing in any brain region is thought to increase its demand for blood and thus the local BOLD signal (which has a slow time constant of a few seconds). The difference between the BOLD signals with and without any sensory input may be seen as the BOLD response to this sensory input. In a retinotopic cortical area contralateral to the cue location, one can identify the region of interest (ROI) as the cortical surface patch that responds to visual inputs at the cue location. Then, one can quantify a form of saliency signal by comparing BOLD signals in the ROI to those in the equivalent cortical surface patch in the ipsilateral hemisphere. We define this to be the maximum difference between the BOLD responses in these two surface patches over the time course of the fMRI responses for a trial. This saliency signal is averaged over all trials for a given orientation contrast at the cue location, with the uniform orientation of the background texture bars randomly chosen for each trial. These signals are plotted in the upper right panel in Fig. 5.48 for different retinotopic brain regions (V1, V2, V3, V4, and the IPS), and different orientation contrasts of the cue. In FEF and LGN, where retinotopy and/or the spatial resolution of the fMRI are too poor, the lack of a saliency signal was determined by a lack of a difference between the BOLD responses to cues of different saliencies.

We measured ERPs in a separate experiment. These showed that the earliest scalp potential response evoked by the invisible cue was a C1 component. This component emerged around 55 ms after the onset of the cue, reaching its peak response around 73 ms after the onset, and crucially, had an amplitude that increased significantly with the orientation contrast at the cue. Its polarity depended on whether the cue (with the whole accompanying texture) was in the upper or lower visual field (see the lower left panel in Fig. 5.48). This C1 component is believed to be generated mainly by neural activities in V1 in the occipital lobe (at the back of the brain), because it has a short latency and because of the location-dependence of its polarity.[20]

The fMRI activations in V1–V4 evoked by saliency also increased significantly with the orientation contrast of the cue, but to a degree that decreased from V1 to V4. This suggests that the saliency activation in V1 is unlikely to be caused by that in V4. Further evidence for this is that lesioning V4 in a monkey impairs its visual selection of non-salient objects but does not impair the selection of salient objects (Schiller and Lee 1991); see Fig. 2.28 B.

Furthermore, for a given orientation contrast, the fMRI activation evoked by saliency also decreased from V1 to V4. This finding contrasts with that of another study (Melloni,

[20] The lower and upper visual fields are mapped to the retinotopic V1 at the upper and lower banks, respectively, of the calcarine fissure. They activate neurons with geometrically opposite orientations in their spatial layout (Jeffreys and Axford 1972).

Van Leeuwen, Alink and Müller 2012), in which the stimulus contained four oriented grating patches and observers had to find a perceptually distinct orientation-singleton patch (which was oriented orthogonally to the non-target patches). In that study, the difference between the fMRI responses to the target and non-target patches *increased* from V1 to V4. The contrast between the findings suggests that recognition, and perhaps also the task-dependency, of the orientation contrast are the causes for higher saliency signals in higher rather than lower visual areas. Furthermore, in that study, fMRI responses in IPS were stronger when there was a need to suppress the distraction of a salient color singleton which was not a target; while fMRI responses in FEF were stronger to enhance the less salient targets. These observations are consistent with task-oriented functions in the IPS and FEF regions.

5.7 V1's role beyond saliency—selection versus decoding, periphery versus central vision

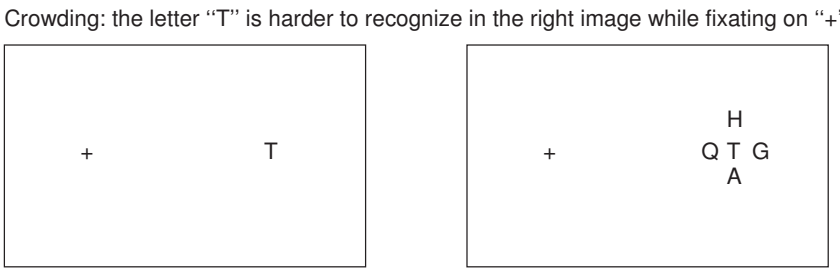


Fig. 5.49: Crowding is an impairment of the ability to recognize a stimulus in visual periphery that is caused by surrounding stimuli. In both the left and right images, the letter "T" does not overlap with any other image elements. However, fixating on the "+," you will find the "T" in the right image much harder to recognize.

Our overall framework sees vision in terms of encoding, selection, and decoding. In Chapters 3 and 4, we discussed extensively the possible role of V1 in encoding. Here, we have discussed how V1 influences selection via its output to the superior colliculus. V1 also projects to higher cortical areas, receives feedbacks from them, and sends feedback to the LGN (Fig. 5.43), so it is important to consider how it contributes to post-selectional decoding.

Visual spatial selection, and in particular saliency, shifts attention from the current focus to a new location. Since the current focus generally lies within the fovea, saliency mechanisms should primarily operate outside it. Take iso-orientation suppression, which is the essential contextual influence underpinning how orientation contrast leads to saliency. Behavioral studies suggest that this is mostly absent in the fovea (Petrov, Carandini and McKee 2005). Following this argument, iso-feature suppression for other feature dimensions should also be absent in the fovea. This absence of contextual influence at the fovea is beneficial for decoding stimuli shown there, since contextual influences would distort the relevant V1 responses. Just such distortion makes it hard to decode peripheral inputs that are surrounded by context. This might be part of the cause for crowding, which is an impairment of the ability to recognize or discriminate peripheral inputs that is caused by having contextual inputs nearby (Levi 2008); for a demonstration, see Fig. 5.49. Crowding makes it very difficult to read text more than a few characters down the line from the current fixation.

Hence, V1 saliency mechanisms operate in the periphery to help select the next focus of attention and bring it to the fovea by a saccade. Meanwhile, the lack of saliency mechanisms in the fovea allows the visual representation there to be faithful to the input to facilitate post-selectional visual decoding; see Fig. 5.50.

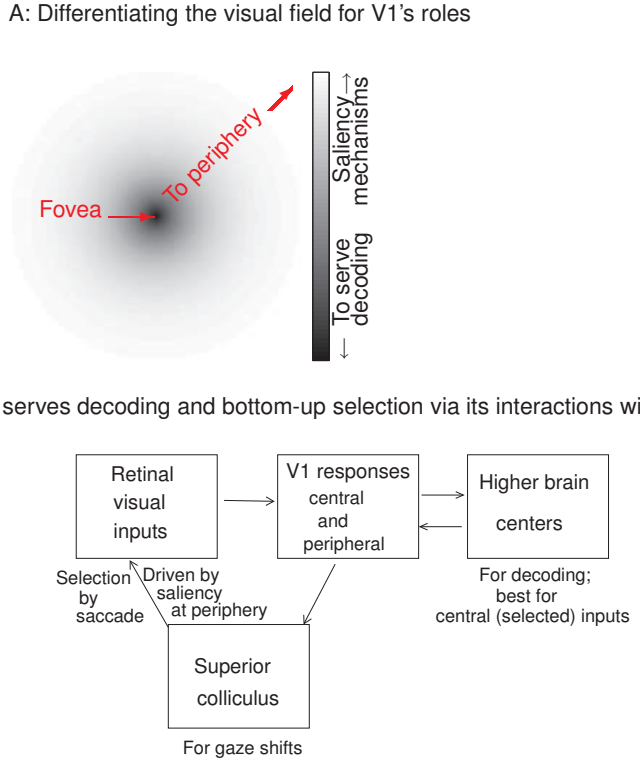


Fig. 5.50: The distinct roles of V1 in peripheral and central vision. A: The visual field is shown in light and dark shades to represent preferential involvement of V1 in saliency and decoding, respectively. B: V1's contribution to bottom-up selection is mediated through the superior colliculus; its contribution to decoding involves higher visual areas.

The next target of visual selection is typically influenced by the stimulus that is currently being decoded. Consequently, along with bottom-up saliency, selection is affected by other factors including the knowledge of decoded visual objects and the ongoing task. Since decoding is better for foveal inputs, the relative impact of saliency in selection increases with increasing eccentricity. In the example depicted in Fig. 5.9, both the ocular and orientation singletons are highly salient because of the contrast between the singleton features and the background features; however, the ongoing task of finding an orientation singleton makes top-down selection favor the latter. As the singletons become more peripheral, recognizing them (i.e., decoding) becomes more difficult, diminishing the task-dependent advantage of the orientation-singleton target in the competition for selection. Consequently, the non-target ocular singleton should become an increasingly effective distractor, damaging performance at the outset of the search. This was indeed observed (Zhaoping 2012)—75% of the first saccade during search were directed to the lateral side of the ocular singleton when both singletons

had an eccentricity of 12 degrees from the initial gaze position (at the center of the displays). By contrast, only 50% were inappropriately directed when the eccentricity was 7.3 degrees.[21]

As shown in Fig. 5.50 B and argued in Chapter 6, feedback from higher visual areas to early visual areas such as V1 is expected to help with decoding, i.e., visual recognition. Given that decoding is favored in the central visual field, one might expect that feedback, particularly from ventral visual areas associated with "what" vision, is more extensive at and near the fovea to help recognizing object features. This suggestion can be tested empirically. It is consistent with recent observations[22] that feedbacks to central and peripheral regions of primate V1 are predominantly from the cortical areas in the ventral and dorsal streams, respectively. It is also supported by observations in a recent behavioral study (Zhaoping 2013a). The stimuli in this behavioral study were adapted from the dichoptic stimuli in the experiment shown in Fig. 3.15 designed to test efficient stereo coding. The percepts induced by such stimuli are ambiguous, such that two likely percepts correspond respectively to the sum ($\mathcal{S}_+$) and difference ($\mathcal{S}_-$) of the visual inputs to the two eyes. Top-down expectations would favor the sum over the difference, in view of the normal correlations between binocular inputs in daily visual experience, and is known to influence perception (see Chapter 6). In the fovea, there was indeed a bias to perceive the sum rather than the difference signal, consistent with this top-down influence; the bias was weak or absent in the periphery (at about 10^o eccentricity), even though the stimulus was adequately enlarged at periphery to compensate for the drop in visual acuity.

5.7.1 Implications for the functional roles of visual cortical areas based on their representations of the visual field

We have argued above that, at least to a first approximation, visual selection by saccades brings the selected visual inputs to the central visual field to be decoded. Consequently, the neural representation of the central and peripheral fields in all visual cortical areas should depend on whether the area is mainly concerned with selection, decoding, or both (Zhaoping 2011). If an area is concerned with bottom-up selection, then the whole visual field should be represented in its neurons, since salient locations can arise unexpectedly anywhere in the visual field. However, areas that are mainly concerned with decoding or top-down or task-dependent selection (which is typically dependent on the current attended object or on knowledge and memory rather than external sensory stimulus) should have neurons that respond mainly or solely to central visual inputs.

In particular, if a cortical area contains a bottom-up saliency map, then it should represent the whole visual space, including the full ranges of eccentricity and polar angles. Conversely, cortical areas further downstream along the visual pathway are likely to devote their resources to the attended, i.e., near foveal, regions, since they are more likely involved in post-selectional processing.

These arguments allow the experimental literature to be used to discriminate among brain regions as to which are likely to contain a saliency map. V1 and V2 respond to the whole visual space (Gattass, Sousa and Rosa 1987, Rosa, Sousa and Gattass 1988) up to at least 80^o eccentricity. Nevertheless, a recent fMRI study showed that V2 (and V3) devotes more cortical area than V1 to the central 0.75 degree of the visual field (Schira, Tyler, Breakspear and Spehar 2009). However V3 and V4 represent only the central 35–40 degrees (Gattass, Sousa and Gross 1988). Toward the culmination of the ventral visual pathway, which is

[21] The stimulus pattern was also different for the two different eccentricity cases, but the densities of the background bars (when the spatial dimension is measured in the unit of bar length) were comparable (Zhaoping 2012).

[22] Private communication from Henry Kennedy (2013).

devoted to processing object features or "what" processing, neurons in the IT cortex (area TE) have very large receptive fields, typically covering the central gaze region and extending to both the left and right half of the visual field. However, they are devoted to inputs within 40^o eccentricity (Boussaoud, Desimone and Ungerleider 1991). Along the dorsal visual pathway, which is more concerned with "where" visual processing, the visual field representation may be expected to reflect visual selection better. However, area MT has no neural receptive field beyond 60^o (Fiorani, Gattass, Rosa and Sousa 1989). Experiments differ as to the maximum eccentricities of the RFs of neurons in area LIP, with one (involving a central fixation task) finding few neurons with receptive fields centered beyond 30^o eccentricity (Ben Hamed, Duhamel, Bremmer and Graf 2001), and another (Blatt, Andersen and Stoner 1990), using anesthetized monkeys, finding few neurons with receptive fields beyond 50^o (albeit with those having receptive-field radii of up to 20^o). V6 (the parieto-occipital area PO), which has many non-visual neurons and neurons influenced by eye positions (Galletti, Battaglini and Fattori 1995), is substantially devoted to peripheral vision. Its neurons can respond to visual inputs up to 80^o in eccentricity (Galletti, Fattori, Gamberini and Kutz 1999). The FEF receives inputs from both the ventral and dorsal streams; however, the spatial coverage of neurons in FEF is poorly studied. Few receptive fields beyond 35^o eccentricity have so far been mapped; the receptive fields concerned have been seen as being open-ended in the periphery since their true extent is unclear (Mohler, Goldberg and Wurtz 1973).

These observations are collectively consistent with the idea that V1 creates a bottom-up saliency map to guide attention. They also imply that some higher visual cortical areas such as V4 and IT along the ventral pathway are less likely to be involved in bottom-up selection than in decoding. Since some cortical areas in the dorsal visual stream cover a large extent of the peripheral visual field, their role in bottom-up visual selection cannot be excluded, although it is also likely that the peripheral coverage serve the purpose of visually guided action (such as grasping and reaching). V2's coverage of the whole visual field suggests that it may also be involved in bottom-up selection. This motivates future investigations. In sum, applying the perspective of inferring functional roles from visual field representations (Zhaoping 2011), V1 is the most likely candidate to compute a bottom-up saliency map.

The functional role of V1 should have direct implications on the role of V2 and other downstream cortical areas along the visual pathway. If V1 creates a saliency map to guide attention in a bottom-up manner, then the downstream areas might be better understood in terms of computations in light of the exogenous selection, and these computations include endogenous selection and post-selectional decoding (Zhaoping 2013b). This is consistent with observations (see Section 2.6) that top-down attention associated with an ongoing task can typically modulate the neural responses more substantially in the extrastriate cortices than in V1.

5.7.2 Saliency, visual segmentation, and visual recognition

Selection by (bottom-up) saliency may be the initial step in visual segmentation, which is the problem of separating out from the rest of the scene those image locations that are associated with visual objects that need to be recognized; see Fig. 5.51 A. Most non-trivial visual functions involve object recognition and localization to enable motor responses and memory storage; segmentation is a fundamental issue for recognition because it needs to be carried out before and during this operation. Computer vision approaches have been tried to solve the problem of image segmentation for decades without a satisfactory solution in the general input situation. The crux of the problem is the following dilemma: to segment the image area containing an object, it helps to recognize it first; while to recognize the object requires having

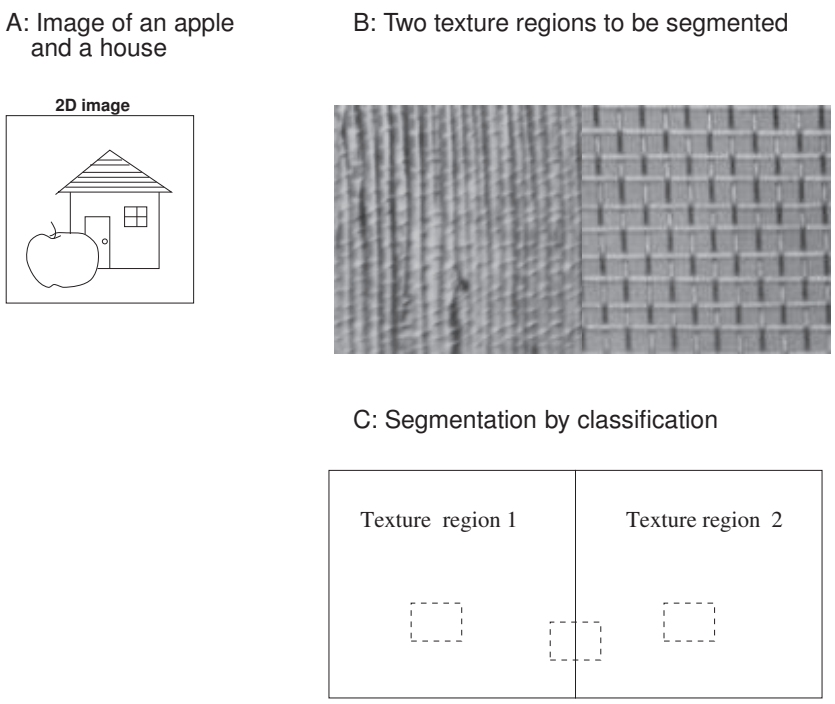


Fig. 5.51: Demonstration of the segmentation-classification dilemma. A: To recognize the apple, it helps to segment the image area associated with it; however, to segment this image area, it helps to recognize the apple. B: Segmenting the two texture regions from each other is hard, since the two regions do not obviously differ by mean luminance or another simple measure. Characterizing local image areas by measures such as smoothness, regularity, orientation, spectrum of spatial frequencies, etc., could help to distinguish different texture regions. C: To segment the image into a priori unknown regions, each local image area, denoted by dashed boxes, needs to be characterized by some such measures.

first segmented the image area that contains it. (Here, we exclude recognizing the gist of a scene without recognizing individual objects.)

Thus, the many computer vision algorithms that have been developed for image segmentation can all be viewed as performing "segmentation by recognition" or "segmentation by classification." Consider segmenting the two texture regions in Fig. 5.51 B; this is not trivial since the two texture regions do not differ in some obvious measure (such as the mean luminance or color), and it is not even known a priori whether the image contains one or two or more regions. Conventional algorithms start by taking any image area, e.g., one of the dashed boxes in Fig. 5.51 C, and trying to characterize it by some image-feature measures. These measures might quantify the mean pixel luminance, regularity, smoothness, dominant spatial frequency, dominant orientations, characteristics of the histogram of the image pixel values, or other aspects of the local image area. Each measure is called a "feature" of the image area, and the image area can be described by a feature vector whose components are the various feature measurements. When two image areas differ sufficiently by their feature vectors, they are presumed to belong to different surface regions. Hence, such algorithms perform "segmentation by classification," i.e., they segment by classifying the feature vectors.

However, these algorithms operate under the assumption that each image area chosen

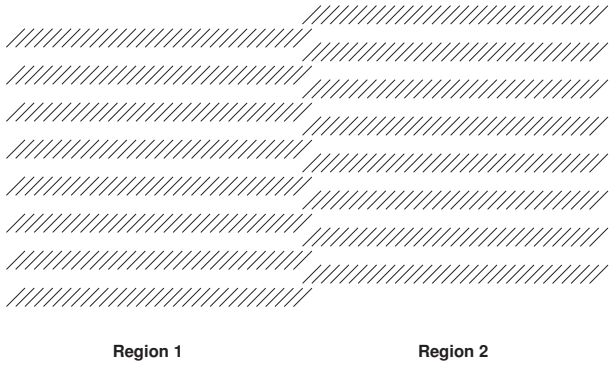


Fig. 5.52: An example demonstrating that biological vision can operate without performing segmentation by classification. We can readily see two regions in this image, even though these regions share all the same feature values. Thus, feature classification is neither sufficient nor necessary to segment the two regions. There is also no vertical contrast edge at the vertical region border, so algorithms using edge-based approaches for segmentation would also fail. Reproduced with permission from Li, Z., Visual segmentation by contextual influences via intracortical interactions in primary visual cortex, *Network: Computation in Neural Systems*, 10(2): 187–212, Fig. 1, copyright © 1999, Informa Healthcare.

to be classified happens to fall into a single surface region to be segmented. This is not guaranteed since we do not know a priori where the region boundaries are. If a chosen area, e.g., the central image area bounded by the central dashed box in Fig. 5.51 C, falls on the border between two regions, it would be hard to characterize its features. The chance of such an event can be reduced by making the individually inspected image areas smaller. This inevitably makes the feature vector values less precise, since many feature values, such as the value of the dominant spatial frequency, require the image area to be large enough for them to be quantified with sufficient precision. This problem stems ultimately from the dilemma that segmentation requires classification and classification requires segmentation.

The "classification" of the image patches in the above example is not the same as recognizing an object, although it can provide clues to the underlying object (e.g., inferring a leopard by its skin) or at least a surface of the object. Nevertheless, the fundamental interdependence between recognition and segmentation remains.

Figure 5.52 demonstrates that biological vision can operate without employing segmentation-by-classification, since classifying the two identical texture regions flanking the texture border is neither necessary nor sufficient for the segmentation. One may argue that special image processing operators could be constructed to detect the border between these two textures. However, such image processing operators would almost certainly be bespoke for this particular image example. Different examples analogous to this one would require different tailored operators to achieve segmentation. It is not desirable to build a big bag of many tricks to tackle this problem, since one can build many special examples that require special purpose operators and so make the bag infeasibly large. Apparently, human vision can carry out segmentation without classification (Li 1999b). This is analogous to making a saccade to a visual location before recognizing the object at that location (see in Fig. 1.4).

Selection by saliency can underpin segmentation without classification. If the border between the two texture regions in Fig. 5.52 is salient, it attracts selection. Locating the border between two objects might be the first step to segmenting them. This first step can be coarse but can nevertheless provide an initial condition in what could be an iterative process,

alternating between segmentation and recognition. In other words, the initial segmentation, by the selection of the border due to its high saliency, can lead to preliminary recognition which can refine segmentation. In turn, this can refine the recognition, and so on. This iterative process is likely to involve both V1 and other cortical areas. Understanding the underlying process is a challenge.

5.8 Nonlinear V1 neural dynamics for saliency and preattentive segmentation

The credibility of the hypothesis that V1 creates a saliency map is significantly bolstered by the demonstration in Section 5.4 that a model using plausible V1 mechanisms could realize the computation concerned. In this section, we show how this model was designed through the analysis of neural circuit dynamics. Readers not interested in these details can skip this section.

The computation of saliency transforms one representation of visual inputs based largely on image contrast to another representation based instead on saliencies. We identify V1 with this transformation, suggesting that its input, the visual stimulus filtered through the classical receptive fields of the V1 neurons, is transformed to an output represented by the activities from the V1 output cells, such that the output can be read out for saliency through the maximum rule in equation (5.4); and the mechanisms it employs are the intracortical interactions mediated by its nonlinear recurrent neural circuit.

There are two characteristics of this V1 saliency transform. First, we focus on cases in which top-down feedback from higher visual areas does not change during the course of the saliency transform but merely sets a background or operating point for V1. In such cases, V1's computation is autonomous, consistent with its being bottom-up or preattentive. Of course, more extensive computations can doubtlessly be performed when V1 interacts dynamically with other visual areas.

Second, the saliency of a location should depend on the global context. Making the output of a V1 neuron depend non-locally on its inputs would be hard to achieve in a purely feed-forward network with retinotopically organized connections and local receptive fields. Rather, the recurrent dynamics enable computations to occur at a *global* scale despite the local neural connectivity.

Nonlinear dynamics involving many recurrently connected neurons is typically difficult to understand and control. As we have seen from Fig. 5.15 B, V1 pyramidal neurons are generally engaged in mutual excitation or mutual inhibition (via interneurons). Since mutual excitation or mutual inhibition involves a positive feedback loop, a recurrent neural network with both interactions is typically unstable against random fluctuations unless the interactions are very weak. The difficulty of understanding such nonlinear recurrent networks in order to properly control and design them is apparent in many previous works (Grossberg and Mingolla 1985, Zucker, Dobbins and Iverson 1989, Yen and Finkel 1998). Nevertheless, harnessing this dynamics is essential to realize the saliency computation.

In this section, we summarize analysis from various research papers (Li 1997, Li 1998a, Li 1999b, Li and Dayan 1999, Li 2001) which addressed the following central issues: (1) computational considerations regarding how a saliency model should behave; (2) a minimal model of the recurrent dynamics for computing saliency, i.e., to achieve (1); (3) the specific constraints on the recurrent neural connections; and (4) how recurrent dynamics give rise to phenomena such as region segmentation, figure-ground segregation, contour enhancement, and filling-in. In addressing these issues, we perform a stability analysis of nonlinear dynamics to examine the conditions governing neural oscillations, illusory contours, and (the absence

of) visual hallucinations. By contrast, single neural properties such as orientation tuning that are less relevant to computations at a global scale will not be our focus. Some of the analytical techniques, e.g., the analysis of the cortical microcircuit and the stability analysis of the translation-invariant networks, can be applied to study other cortical areas that share similar neural elements and neural connection structures with V1's canonical microcircuit (Shepherd 1990).

5.8.1 A minimal model of the primary visual cortex for saliency computation

A minimal model is the one which has just barely enough components to execute the necessary computations without anything extra. This criterion is inevitably subjective, since there is no fixed recipe for a minimalist design. However, as a candidate, I present a model that performs all the desired computations but for which simplified versions fail. Since the minimal model depends on the desired computation to be carried out by the model, I will also articulate this saliency computation as to what this model should do and what this model should not do.

Throughout the section, we try to keep our analysis of the characteristics of the recurrent dynamics general. However, to illustrate particular analytical results, approximations, and simplification techniques, I often use a model of V1 whose specifics and numerical parameters are presented in the appendix to this chapter, so that the readers can try out the simulations.

We use notation such as $\{x_{i\theta}\}$ to denote a vector containing components $x_{i\theta}$ for all $i\theta$. Hence, $\{I_{i\theta}\}$ is the input, and $\{g_x(x_{i\theta})\}$ is the response. The V1 model should transform $\{I_{i\theta}\}$ to $\{g_x(x_{i\theta})\}$, with higher responses $g_x(x_{i\theta})$ to input bars $i\theta$ which have higher perceptual saliency. This is achieved through recurrent interactions between neurons. What kind of recurrent model is needed?

5.8.1.1 A less-than-minimal recurrent model of V1

A very simple recurrent model of the cortex can be described by this equation

$$\dot{x}_{i\theta} = -x_{i\theta} + \sum_{j\theta'} \mathsf{T}_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_{i\theta} + I_o, \tag{5.40}$$

where $-x_{i\theta}$ models the decay (with a time constant of unity) in membrane potential, and I_o is the background input. The recurrent connection $\mathsf{T}_{i\theta,j\theta'}$ links cells $i\theta$ and $j\theta'$. Visual input $I_{i\theta}$ (taken as being static for illustration) initializes the activity levels $g_x(x_{i\theta})$ and also persists after onset. The activities are then modified by the network interaction, making $g_x(x_{i\theta})$ dependent on input $I_{j\theta'}$ for $(j\theta') \neq (i\theta)$. The connections are translation invariant in that $\mathsf{T}_{i\theta,j\theta'}$ depends only on the vector $i - j$ and on the angles of this vector (in 2D space) to the orientations θ and θ'. Reflection symmetry (e.g., when a horizontal bar facilitates another horizontal bar to its right, so should the latter facilitate the former, with the same facilitation strength) gives the constraint $\mathsf{T}_{i\theta,j\theta'} = \mathsf{T}_{j\theta',i\theta}$.

Many previous models of the primary visual cortex (e.g., Grossberg and Mingolla (1985), Zucker, Dobbins, and Iverson (1989), and Braun, Niebur, Schuster, Koch (1994)) can be seen as more complex versions of the one described above. The added complexities include stronger nonlinearities, global normalization (e.g., by adding a global normalizing input to the background I_o), and shunting inhibition. However, they are all characterized by reciprocal or symmetric interactions between model units, i.e., $\mathsf{T}_{i\theta,j\theta'} = \mathsf{T}_{j\theta',i\theta}$. It is well known (Hopfield 1984, Cohen and Grossberg 1983) that in such a symmetric recurrent network, the dynamic trajectory $x_{i\theta}(t)$ (given a static input pattern $\{I_{i\theta}\}$) will converge in time t to a fixed point. This fixed point is a local minimum (attractor) in an energy landscape

$$E(\{x_{i\theta}\}) = -\frac{1}{2}\sum_{i\theta,j\theta'} \mathsf{T}_{i\theta,j\theta'} g_x(x_{i\theta}) g_x(x_{j\theta'}) - \sum_{i\theta} I_{i\theta} g_x(x_{i\theta}) + \sum_{i\theta} \int_0^{g_x(x_{i\theta})} g_x^{-1}(x) dx, \tag{5.41}$$

where $g_x^{-1}(x)$ means the inverse function of $g_x(x)$. Empirically, convergence to attractors typically occurs even when the complexities in the previous models mentioned above are included.

The fixed point $\bar{x}_{i\theta}$ of the motion trajectory, or the minimum energy state $E(\{x_{i\theta}\})$ where $\partial E/\partial g_x(x_{i\theta}) = 0$ for all $i\theta$, is given by (when $I_o = 0$)

$$\bar{x}_{i\theta} = I_{i\theta} + \sum_{j\theta'} \mathsf{T}_{i\theta,j\theta'} g_x(\bar{x}_{j\theta'}). \tag{5.42}$$

Without recurrent interactions ($\mathsf{T} = 0$), this fixed point $\bar{x}_{i\theta} = I_{i\theta}$ is a faithful copy of the input $I_{i\theta}$. Weak but non-zero T makes pattern $\{\bar{x}_{i\theta}\}$ a slightly modified version of the input pattern $\{I_{i\theta}\}$. However, sufficiently strong interactions T can make $\bar{x}_{i\theta}$ dramatically unfaithful to the input. This happens when T is so strong that one of the eigenvalues $\lambda^{\mathbb{T}}$ of the matrix $\mathbb{T}$ with elements $\mathbb{T}_{i\theta,j\theta'} \equiv \mathsf{T}_{i\theta,j\theta'} g'_x(\bar{x}_{j\theta'})$ satisfies $Re(\lambda^{\mathbb{T}}) > 1$ (here, g'_x is the slope of $g_x(.)$ and $Re(.)$ means the real part of a complex number). For instance, when the input $I_{i\theta}$ is translation invariant such that $I_{i\theta} = I_{j\theta}$ for all $i \neq j$, there is a translation-invariant fixed point $\bar{x}_{i\theta} = \bar{x}_{j\theta}$ for all $i \neq j$. Strong interactions $\mathbb{T}$ could destabilize this fixed point, such that it is no longer a local minimum of the energy landscape $E(\{x_{i\theta}\})$. Consequently, the recurrent dynamics pulls $\{x_{i\theta}\}$ into an attractor in the direction of an eigenvector of $\mathbb{T}$ that is not translation invariant, i.e., $x_{i\theta} \neq x_{j\theta}$ for $i \neq j$ at the attractor.

Computationally, a certain unfaithfulness to the input, i.e., making $g_x(x_{i\theta})$ not to be a function of $I_{i\theta}$ alone, is actually desirable. This is exactly what is required for unequal responses $g_x(x_{i\theta})$ to be given to input bars of equal contrast $I_{i\theta}$ but different saliencies (e.g., a vertical bar among horizontal bars when all bars have the same input contrast). However, this unfaithfulness should be driven by the nature of the input pattern $\{I_{i\theta}\}$ and in particular, driven by how the input pattern deviates from homogeneity (e.g., smooth contours or figures against a background). If, instead, spontaneous or non-input-driven network behavior—*spontaneous pattern formation or symmetry breaking*—occurs, then visual hallucinations (in this case we mean saliency outcomes which are drastically different from the saliency values of the input) would result (Ermentrout and Cowan 1979). Such hallucinations, whose patterns are not meaningfully determined by external inputs, should be avoided.

For example, given homogenous input $I_{i\theta} = I_{j\theta}$, if $\{x_{i\theta}\}$ is an attractor, then so is a translated state $\{x'_{i\theta}\}$ such that $x'_{i\theta} = x_{i+a,\theta}$ for any translation a. This is because $\{x_{i\theta}\}$ and $\{x'_{i\theta}\}$ have the same energy value E. The two possible patterns after symmetry breaking on the right part of Fig. 5.20 are instances of this, being translations of each other. (When the translation a is one-dimensional, such a continuum of attractors has been called a "line attractor" (Zhang 1996). For two- or more dimensional patterns, the continuum is a "surface attractor.") That the absolute positions of the hallucinated patterns are random, can even shift, and are not determined by the sensory input $\{I_{i\theta}\}$ implies a degree of unfaithfulness that is undesirable for saliency.

To illustrate, consider the case that $\mathsf{T}_{i\theta,j\theta'}$ is non-zero only when $\theta = \theta'$, i.e., the connections only link cells that prefer the same orientation. (This is a limit of the observations (Gilbert and Wiesel 1983, Rockland and Lund 1983) that the lateral interactions tend to link cells preferring similar orientations.) The network then contains multiple, independent, subnetworks, one for each θ. Take the $\theta = 90^o$ (vertical orientation) subnet, and for convenience, drop the subindex θ. We have

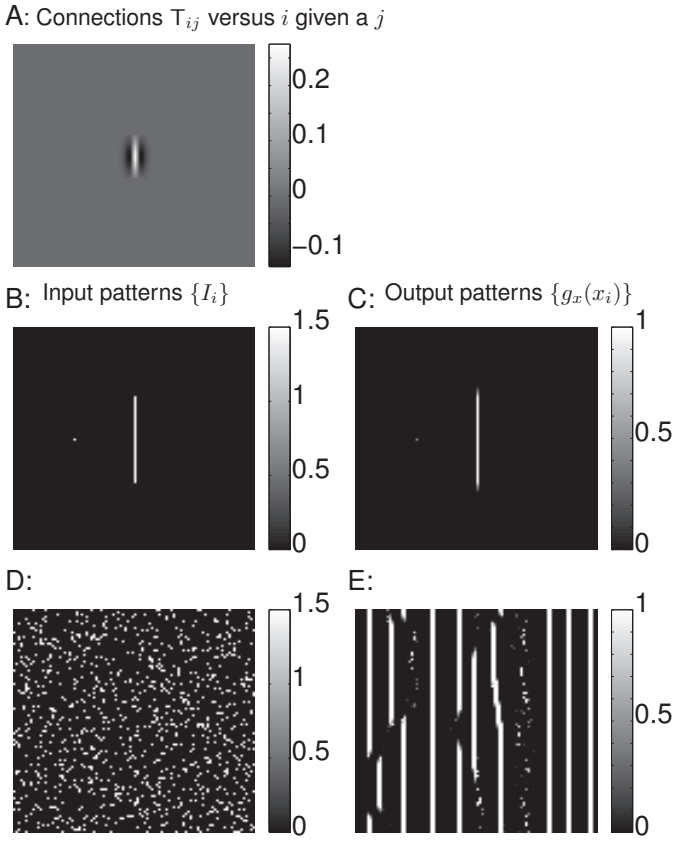


Fig. 5.53: A reduced model consisting of symmetrically coupled cells tuned to vertical orientation ($\theta = 90^o$), as in equation (5.43). Five grayscale images are shown; each has a scale bar on the right. The network has 100×100 cells arranged in a two-dimensional (2D) array, with wrap-around boundary conditions. Each cell models a cortical neuron tuned to vertical orientation, arranged retinotopically. The function $g_x(x)$ gives $g_x(x) = 0$ when $x < 1$, $g_x(x) = x-1$ when $1 \leq x < 2$, and $g_x(x) = 1$ when $x > 2$. A: The connection pattern T between the center cell j and the other cells i. This pattern is local and translation invariant, with excitation or inhibition between i and j which are, respectively, roughly vertically or horizontally displaced from each other. B: An input pattern $\{I_i\}$, consisting of an input line and a noise spot. C: Output response $\{g_x(x_i)\}$ to the input in B. The line induces a response that is $\sim 100\%$ higher than the response to the noise spot. D: A sample noisy input pattern $\{I_i\}$. E: Output response $\{g_x(x_i)\}$ to the input in D, showing hallucinated vertical streaks. Adapted with permission from Li, Z., Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex, *Neural Computation*, 13(8): 1749–1780, Fig. 1, copyright © 2001, MIT Press.

$$\dot{x}_i = -x_i + \sum_j \mathsf{T}_{ij} g_x(x_j) + I_i, \tag{5.43}$$

in which T_{ij} is still symmetric, $\mathsf{T}_{ij} = \mathsf{T}_{ji}$, and translation invariant. As an example, let T be a simple, center-surround pattern of connections in a Manhattan grid (for which grid location $i = (m_i, n_i)$ has horizontal and vertical coordinates m_i and n_i, respectively). Let

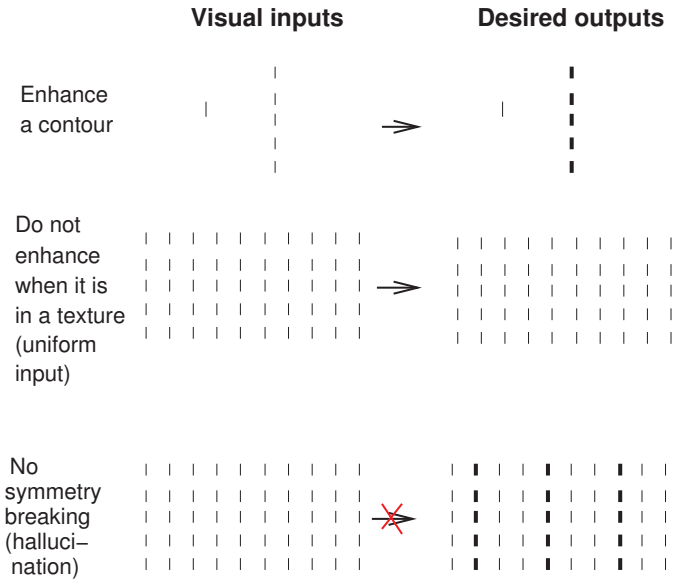


Fig. 5.54: Desired input-output mapping for saliency computation in three special input cases. Adapted with permission from Li, Z. and Dayan, P., Computational differences between asymmetrical and symmetrical networks, *Network: Computation in Neural Systems*, 10(1): 59–77, Fig. 1, copyright © 1999, Informa Healthcare.

$$\mathsf{T}_{ij} \propto \begin{cases} T & \text{if } i = j, \\ -T & \text{if } (m_j, n_j) = (m_i \pm 1, n_i) \text{ or } (m_i, n_i \pm 1), \\ 0 & \text{otherwise.} \end{cases} \tag{5.44}$$

If T is sufficiently strong, then even with an homogenous input $I_i = I_j$ for all i, j, the network can settle into an "antiferromagnetic" state in which neighboring units x_i exhibit one of the two different activities $x_{m_i,n_i} = x_{m_i+1,n_i+1} \neq x_{m_i+1,n_i} = x_{m_i,n_i+1}$, arranged in a checkerboard pattern. This pattern $\{x_i\}$ is just a spatial array of the replicas of the center-surround interaction pattern T. Note that the patterns after the spontaneous symmetry breaking in Fig. 5.20 are simply one-dimensional checkerboard patterns.

Intracortical interaction that are more faithful to V1 (Kapadia et al. 1995, Polat and Sagi 1993, Field, Hayes and Hess 1993) have T_{ij} depend on the orientation of $i - j$. However, this T_{ij} will still be translation invariant, i.e., independent of the absolute value of i and depending only on the magnitude and orientation of $i - j$. In the subnet of vertical bars, such V1-like interactions specify that two nearby bars i and j excite each other when they are co-aligned and inhibit each other otherwise. More directly, $\mathsf{T}_{ij} > 0$ between i and j which are close and roughly vertically displaced from each other, and $\mathsf{T}_{ij} < 0$ between i and j which are close and more horizontally displaced. Figure 5.53 shows the behavior of such a subnet. Although the network enhances an input (vertical) line relative to an isolated (short) bar, it also hallucinates other vertical lines when exposed to noisy inputs.

Instead, the recurrent network should have the properties illustrated in Fig. 5.54. First, its response to a smooth contour should be higher than to a bar segment that is either isolated or is an element in a homogenous texture. Second, it should not respond inhomogenously to a homogenous texture. In other words, the network should selectively amplify certain inputs against some other inputs. The ability of the network to achieve this property can be measured

by the gain (or sensitivity) to a contour relative to a homogenous texture. We call this the *selective amplification ratio* (Li and Dayan 1999):

$$\text{selective amplification ratio} = \frac{\text{gain to contour input}}{\text{gain to texture input}}. \tag{5.45}$$

A higher selective amplification ratio makes it easier to distinguish salient input (such as a contour) from the less salient inputs (such as homogenous textures). For instance, if the level of noise in the neural responses is comparable to the mean response to the homogenous texture, a selective amplification ratio comfortably larger than two is desirable to make the response to a contour stand out relative to the responses to a background texture. Physiological data (Nelson et al. 1985, Knierim and Van Essen 1992, Kapadia et al. 1995) shown in Fig. 5.19 suggest that the selective amplification ratio is up to at least four to five.

The competition between internal interactions T and the external inputs $\{I_i\}$ to shape $\{x_i\}$ makes it impossible to achieve a high selective amplification ratio. For analysis, consider the following simple pattern of interaction in the vertical bar subnet:

$$\begin{cases} \mathsf{T}_{ij} > 0, \text{when } i \text{ and } j \text{ are nearby and in the same vertical column,} \\ \mathsf{T}_{ij} < 0, \text{when } m_j = m_i \pm 1, \text{i.e., } i \text{ and } j \text{ are in the neighboring columns,} \\ \mathsf{T}_{ij} = 0, \text{otherwise.} \end{cases}$$

Furthermore, denote the total excitatory connection to a neuron from its contour as

$$T_0 \equiv \sum_{j, m_j = m_i} \mathsf{T}_{ij} > 0, \tag{5.46}$$

and denote the total suppressive connection to a neuron from neighboring contours as

$$T' \equiv - \sum_{j, m_j = m_i \pm 1} \mathsf{T}_{ij}. \tag{5.47}$$

In addition, for simplicity, take a piece-wise linear function for $g_x(x)$:

$$g_x(x) = \begin{cases} x - x_{\text{th}} & \text{if } x_{\text{th}} \le x \le x_{\text{sat}}, \text{ where } x_{\text{th}} \text{ is the threshold,} \\ & \quad \text{and } x_{\text{sat}} > x_{\text{th}} \text{ is the point of saturation,} \\ x_{\text{sat}} - x_{\text{th}} & \text{if } x > x_{\text{sat}}, \\ 0 & \text{otherwise.} \end{cases} \tag{5.48}$$

A vertical contour input has $I_i = I > x_{\text{th}}$ for i with $m_i = 1$, and $I_i = 0$ otherwise. Call the neurons i with $m_i = 1$ on the vertical line the "line units." We can ignore all other neurons since they will be at most suppressed by the line unit, and so none can be activated beyond threshold. By symmetry, at the fixed point, all the line units i have the same state $\bar{x}_i = \bar{x}$, where

$$\bar{x} = I + \sum_{j, m_j = m_i} \mathsf{T}_{ij} g_x(\bar{x}) = I + T_0 g_x(\bar{x}), \tag{5.49}$$

$$\rightarrow \quad g_x(\bar{x}) = \frac{I - x_{\text{th}}}{1 - T_0}, \tag{5.50}$$

when $I > x_{\text{th}}$ is not too large. Thus, a large $T_0 < 1$ helps to give the following high gain

$$\frac{\delta g_x(\bar{x})}{\delta I} = \frac{1}{1 - T_0} \quad \text{to an isolated input contour, and } 1 > T_0 \text{ is required for stability.} \tag{5.51}$$

By contrast, for a homogenous texture with $I_i = I > x_{\rm th}$ for all units i, the fixed point $\bar{x}_i = \bar{x}$ of the response is

$$\bar{x} = I + \left(\sum_{j,|m_j - m_i| \leq 1} \mathsf{T}_{ij} \right) g_x(\bar{x}) = I + (T_0 - T') g_x(\bar{x}) \tag{5.52}$$

$$\rightarrow \quad g_x(\bar{x}) = \frac{I - x_{\rm th}}{1 + (T' - T_0)}. \tag{5.53}$$

This means that

$$\text{the gain } \frac{\delta g_x(\bar{x})}{\delta I} = \frac{1}{1 + (T' - T_0)} \quad \text{to a homogenous input texture} \tag{5.54}$$

can be made small when the net suppression $T' - T_0$ is made large. Note that $T' - T_0$ quantifies the net iso-orientation suppression in a homogenous texture. Then,

$$\text{the selective amplification ratio} = \frac{\text{gain to contour input}}{\text{gain to texture input}} = \frac{1 + (T' - T_0)}{1 - T_0} \tag{5.55}$$

increases with increasing net iso-orientation suppression $T' - T_0$ and increasing contour facilitation T_0.

However, in a homogenous texture, a large net suppression destabilizes the homogenous fixed point $x_i = \bar{x}$ to fluctuations. Consider the fluctuation $x'_i = x_i - \bar{x}$, and assume that this lies within the linear range of $g_x(.)$, i.e., $x_{\rm th} < x'_i + \bar{x} < x_{\rm sat}$. Then

$$\begin{aligned} \dot{x}'_i &= -x'_i + \sum_j \mathsf{T}_{ij} \left[g_x(\bar{x} + x'_j) - g_x(\bar{x}) \right] \\ &= -x'_i + \sum_j \mathsf{T}_{ij} x'_j. \end{aligned} \tag{5.56}$$

This linear equation has an inhomogenous eigenmode, $x'_i = x' \cdot (-1)^{m_i}$ for all i, which is a spatially oscillating pattern (with amplitude x') of activities like those in the symmetry-breaking solution of Fig. 5.20. To see this, substitute $x'_i = x' \cdot (-1)^{m_i}$ into the above equation to obtain the equation of motion for the amplitude x':

$$\dot{x}' = -x' + \left(\sum_{j,m_j = m_i} T_{ij} - \sum_{j,m_j = m_i \pm 1} T_{ij} \right) x' \tag{5.57}$$

$$= -x' + (T' + T_0)\, x' \tag{5.58}$$

$$= (T' + T_0 - 1)\, x'. \tag{5.59}$$

This has a solution in which x' evolves with time t as

$$x'(t) \propto \exp[(T' + T_0 - 1)\, t]\,. \tag{5.60}$$

When $T' + T_0 > 1$, the amplitude x' of the inhomogenous eigenmode grows exponentially with time t. Given an initial deviation $\{x'_i(0)\}$ at $t = 0$ from the homogenous equilibrium state $x_i = \bar{x}$, the projection of this deviation pattern on this inhomogenous eigenmode grows exponentially, driving the activity pattern $x_i = \bar{x} + x'_i$ inhomogenous very quickly after the onset of visual input. The growth will eventually saturate because of the nonlinearity in $g_x(x)$. Hence, given homogenous input, the network spontaneously breaks symmetry from

the homogenous fixed point and hallucinates a saliency wave whose period is two columns. As shown in Fig. 5.20, there are two such waves—one has $x'_i = x' \cdot (-1)^{m_i}$ and the other has $x'_i = x' \cdot (-1)^{m_i+1}$—and they are a 180^o phase apart. Both of these inhomogenous states are also equilibrium points of the network even for homogenous input; however, they are stable to fluctuations, and they arise when $T' + T_0 > 1$ to make the homogenous fixed point unstable. Whether the network state will approach one or the other stable fixed point depends on the direction of the initial fluctuation pattern.

In sum, contour enhancement makes the network prone to "see" ghost contours whose orientations and widths match the interaction structure in T. Avoiding such hallucinations requires $T' + T_0 < 1$. This, together with $1 > T_0$ required by equation (5.51), implies that the selective amplification ratio is limited to the following:

$$\text{selective amplification ratio} = \frac{1 + (T' - T_0)}{1 - T_0} < 2. \tag{5.61}$$

Consequently, enhancement of contours relative to the background is insufficient (Li and Dayan 1999). Similar numerical limits on the selective amplification ratio apply to the cases of a general $g_x(x)$.[23] Although symmetric recurrent networks are useful for associative memory (Hopfield 1984), which requires significant input errors or omissions to be corrected or filled-in, they imply too much distortion for early visual tasks that require the output to be more faithful to the input.

5.8.1.2 A minimal recurrent model with hidden units

The strong tendency to hallucinate input in the symmetrically connected model of equation (5.40) is largely dictated by the symmetry of the neural connections. Hence, this tendency cannot be removed by introducing more complex cellular and network mechanisms without removing the symmetry of the neural connections. (These cellular and network mechanisms include, for instance, ion channels, spiking rather than firing rate neurons, multiplicative inhibition, global activity normalization, and input gating (Grossberg and Mingolla 1985, Zucker et al. 1989, Braun, Niebur, Schuster and Koch 1994), which are used by many neural network models.) Furthermore, attractor dynamics are untenable in the face of the well-established fact of Dale's law, namely that real neurons are overwhelmingly either exclusively excitatory or exclusively inhibitory. It is obviously impossible to have symmetric connections between excitatory and inhibitory neurons.

Mathematical analysis (Li and Dayan 1999) showed that asymmetric recurrent EI networks with separate excitatory (E) and inhibitory (I) cells can perform computations that are inaccessible to symmetric networks. In particular, EI networks support much larger selective amplification ratios without degenerating into hallucination. To illustrate this, start again with the simplification of a separated subnet (equations (5.43)) and piece-wise linear $g_x(x)$, as in equation (5.48). Then, replace neural units and connections (as in the example in Fig. 5.55):

$$\begin{aligned} \text{neural unit } \ x_i &\rightarrow \text{an EI pair [excitatory } x_i\text{, inhibitory } y_i \text{ with time constant } \tau_y\text{]}, \\ \text{connection } \ \mathsf{T}_{ij} &\rightarrow \mathsf{J}_{ij} \text{ from } x_j \text{ to } x_i\text{, and } \mathsf{W}_{ij} \text{ from } x_j \text{ to } y_i, \end{aligned}$$

such that the circuit's equation of motion becomes

[23] In such cases, let $\bar{x}_{\text{contour}}$ and $\bar{x}_{\text{texture}}$ be the fixed points for the contour and texture inputs, respectively. The response gains for a contour and a homogenous texture are $\delta g_x(\bar{x})/\delta I = g'_x(\bar{x}_{\text{contour}})/\left[1 - T_0 g'_x(\bar{x}_{\text{contour}})\right]$ and $\delta g_x(\bar{x})/\delta I = g'_x(\bar{x}_{\text{texture}})/\left[1 + (T' - T_0) g'_x(\bar{x}_{\text{texture}})\right]$, respectively, and the requirement for avoiding hallucinations becomes $(T' + T_0) g'_x(\bar{x}_{\text{texture}}) < 1$. Consequently, the selective amplification ratio is limited by an upper bound $2 \frac{g'_x(\bar{x}_{\text{contour}})}{g'_x(\bar{x}_{\text{texture}})} \frac{\left[1 - T_0 g'_x(\bar{x}_{\text{texture}})\right]}{\left[1 - T_0 g'_x(\bar{x}_{\text{contour}})\right]}$. If $g'_x(\bar{x}_{\text{contour}}) = g'_x(\bar{x}_{\text{texture}})$, this upper bound is again 2.

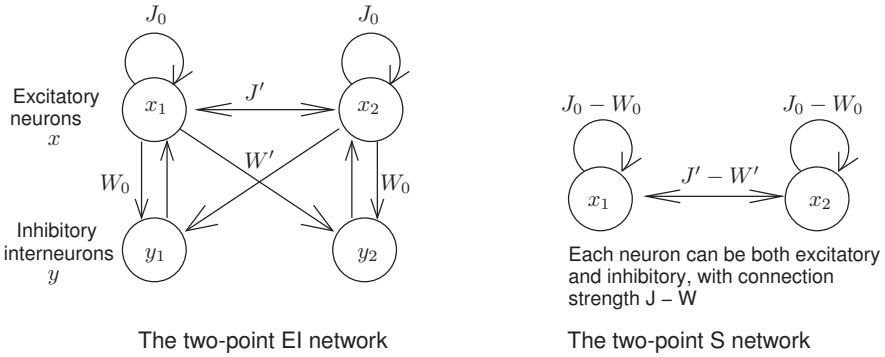


Fig. 5.55: Two-point EI (as in equations (5.66) and (5.67)) and S networks. There are austere models to elucidate the essential computation in a recurrent V1 sub-network involving only neurons tuned to a single orientation. The two networks are exact counterparts when the interneurons y_1 and y_2 are linear, with $g_y(y) = y$. The fixed points of the dynamics in one network are also the fixed points in the other, but the stabilities of the fixed points, and thus the computational power, differ in the two networks.

$$\dot{x}_i = -x_i - g_y(y_i) + \sum_j \mathsf{J}_{ij} g_x(x_j) + I_i \tag{5.62}$$

$$\tau_y \dot{y}_i = -y_i + \sum_j \mathsf{W}_{ij} g_x(x_j) \tag{5.63}$$

In this circuit, x_i is the excitatory unit and conveys the output of network, and y_i is the inhibitory interneuron (with output $g_y(y_i)$), which acts as an auxiliary or hidden unit of the network.

The fixed point $\{\bar{x}_i, \bar{y}_i\}$ of this EI network satisfies $\dot{x}_i = \dot{y}_i = 0$. The network can be designed such that these fixed points are identical (ignoring the y dimension) to the fixed points $\{\bar{x}_i\}$ of the original symmetric network in equation (5.43). This EI network is then a formal counterpart of the symmetric network (which we call the S network). This is particularly simple in the case when $g_y(y) = y$ is linear. Then, as the time constant τ_y of the interneurons approaches zero, such that $y_i = \sum_j W_{ij} g_x(x_j)$, equation (5.62) becomes

$$\dot{x}_i = -x_i + \sum_j (\mathsf{J}_{ij} - \mathsf{W}_{ij}) g_x(x_j) + I_i. \tag{5.64}$$

Hence, an EI network with very fast interneurons is equivalent to the S network when

$$\mathsf{J}_{ij} - \mathsf{W}_{ij} = \mathsf{T}_{ij}, \quad \tau_y = 0, \quad \text{and} \quad g_y(y) = y. \tag{5.65}$$

If $\tau_y > 0$, these two networks are counterparts of each other, with the same fixed points but different dynamics for the motion trajectories. From now on, for simplicity, we always take $\tau_y = 1$, and use this simple model to compare EI and S networks.

Consider an EI network that is the counterpart of the particular S sub-network that only involves vertical bars (described in equations (5.43–5.48)), and with translation-invariant J and W, where $J_0 \equiv \sum_{j,m_j=m_i} \mathsf{J}_{ij}$, $J' \equiv \sum_{j,m_j=m_i\pm1} \mathsf{J}_{ij}$, and similarly for W_0 and W'. We have $T_0 = J_0 - W_0$ and $T' = W' - J'$. Since the EI network has the same fixed points as the S network, the selective amplification ratio, which is evaluated at fixed points, is identical for the two networks. A high value for this ratio cannot be realized in the S network

because of the tendency for hallucination, resulting from the instability of the homogenous fixed point. However, in the EI network, under homogenous input, all the three fixed points, one homogenous and two inhomogenous, are unstable. As a result, the primary mode of instability of the homogenous solution in the EI network (to the homogenous input) is a spatially homogenous, temporal oscillation, because the network state cannot approach the non-homogenous, unstable, fixed points.

These conclusions can be understood by considering a highly simplified problem. In this simplification, $I_i \equiv I_1$ in all the odd columns have the same strength, and $I_i \equiv I_2$ in all the even columns also have the same strength. To quantify the selective amplification ratio, we need to consider responses to two input cases: a homogenous input pattern and a single contour. The former can be straightforwardly modeled by setting $I_1 = I_2$. For the latter, since the connections T_{ij} do not span more than a single column, we can set $I_1 > I_2 = 0$ or $0 = I_1 < I_2$, and thus consider non-interacting contours on either the odd or even columns.

Taking advantage of this simplification and assuming (as is arranged by the dynamics of the network) that all excitatory and inhibitory units in each column have the same activities, the input and state variables can be described by two-dimensional vectors $(I_1, I_2)^T$, $(x_1, x_2)^T$ and $(y_1, y_2)^T$ for the various quantities associated with the odd and even columns. We call this simplified system the two-point system (see Fig. 5.55) which captures the essence of our problem. The equations of motion of the two-point system are

$$\dot{x}_a = -x_a - y_a + J_0 g_x(x_a) + J' g_x(x_{a'}) + I_a, \tag{5.66}$$

$$\dot{y}_a = -y_a + W_0 g_x(x_a) + W' g_x(x_{a'}), \tag{5.67}$$

where $a = 1, 2$ and $a' \neq a$. Thus, the EI network has been reduced to two pairs of EI units, one for the odd columns and the other for the even columns of units. The 2×2 connection matrices for this reduced EI network are

$$\mathsf{J} = \begin{pmatrix} J_0 & J' \\ J' & J_0 \end{pmatrix} \quad \text{and} \qquad \mathsf{W} = \begin{pmatrix} W_0 & W' \\ W' & W_0 \end{pmatrix}. \tag{5.68}$$

The S network that is the counterpart of this EI network has just two neurons (rather than four), and the connection matrix $\mathsf{T} = \mathsf{J} - \mathsf{W}$.

From Fig. 5.54, we require relatively higher responses to the one-point input, $(I_1, I_2) \propto (1, 0)$, which corresponds to a contour input, and lower responses to the (uniform) two-point input, $(I_1, I_2) \propto (1, 1)$, which corresponds to a homogenous texture. The symmetry of the responses must be preserved for the two-point input $(I_1, I_2) \propto (1, 1)$.

In the two-point S system, the input response function to the one-point and two-point inputs are the same as those in equations (5.50) and (5.53) respectively, with a selective amplification ratio as in equation (5.55).

We can linearize the EI network to examine the approximate evolution of the deviations

$$(x'_a, y'_a) \equiv (x_a, y_a) - (\bar{x}_a, \bar{y}_a)$$

from the homogenous fixed point in response to homogenous input $(I_1, I_2) \propto (1, 1)$. The deviation (x'_a, y'_a) follows the equations (for $a \neq a'$)

$$\dot{x}'_a = -x'_a - y'_a + J_0 x'_a + J' x'_{a'}, \tag{5.69}$$

$$\dot{y}'_a = -y'_a + W_0 x'_a + W' x'_{a'}. \tag{5.70}$$

In comparison, in the two-point S network, the deviations x'_a follow

$$\dot{x}'_a = -x'_a + (J_0 - W_0) x'_a + (J' - W') x'_{a'}. \tag{5.71}$$

Note that matrices J, W, and T commute with each other, with common eigenvectors

$$V^{(+)} \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad V^{(-)} \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \tag{5.72}$$

which are called plus and minus modes, respectively, (note that we used analogous modes for the stereo summation and opponency signals in the stereo encoding in Section 3.5) or spatial synchronous and anti-phase modes, respectively. The corresponding eigenvalues of J, W, and T are $\lambda_{\pm}^{\mathsf{J}} = J_0 \pm J'$, $\lambda_{\pm}^{\mathsf{W}} = W_0 \pm W'$, and $\lambda_{\pm}^{\mathsf{T}} = \lambda_{\pm}^{\mathsf{J}} - \lambda_{\pm}^{\mathsf{W}}$, respectively. Then states (x_1', x_2') and (y_1', y_2') can be represented by their projections $x_{\pm}$ and $y_{\pm}$ onto these eigenmodes

$$\begin{pmatrix} x_1' \\ x_2' \end{pmatrix} = x_+ V^{(+)} + x_- V^{(-)}, \qquad \begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = y_+ V^{(+)} + y_- V^{(-)}. \tag{5.73}$$

Equations (5.69) and (5.70) can then be transformed to

$$\dot{x}_{\pm}' = -x_{\pm}' - y_{\pm}' + \lambda_{\pm}^{\mathsf{J}} x_{\pm}', \tag{5.74}$$

$$\dot{y}_{\pm}' = -y_{\pm}' + \lambda_{\pm}^{\mathsf{W}} x_{\pm}'. \tag{5.75}$$

Eliminating $y_{\pm}'$ from these equations, the EI dynamics follow

$$\ddot{x}_{\pm}' + (2 - \lambda_{\pm}^{\mathsf{J}})\dot{x}_{\pm}' + (\lambda_{\pm}^{\mathsf{W}} - \lambda_{\pm}^{\mathsf{J}} + 1)x_{\pm}' = 0. \tag{5.76}$$

Similarly, the S network dynamics is

$$\dot{x}_{\pm}' = -x_{\pm}' + (\lambda_{\pm}^{\mathsf{J}} - \lambda_{\pm}^{\mathsf{W}})x_{\pm}'. \tag{5.77}$$

The solutions to the linear equations are

$$x_{\pm}'(t) \propto \exp\left(\gamma_{\pm}^{EI} t\right) \qquad \text{for the EI network,} \tag{5.78}$$

$$x_{\pm}'(t) \propto \exp\left(\gamma_{\pm}^{S} t\right) \qquad \text{for the S network,} \tag{5.79}$$

where $\gamma_{\pm}^{EI}$ and $\gamma_{\pm}^{S}$ are the eigenvalues of the linear system in equation (5.76) for the EI network and equation (5.77) for the S network, respectively,

$$\gamma_{\pm}^{EI} = -1 + \frac{1}{2}\lambda_{\pm}^{\mathsf{J}} \pm \left(\frac{1}{4}(\lambda_{\pm}^{\mathsf{J}})^2 - \lambda_{\pm}^{\mathsf{W}}\right)^{1/2}, \qquad \text{and} \tag{5.80}$$

$$\gamma_{\pm}^{S} = -1 - \lambda_{\pm}^{\mathsf{W}} + \lambda_{\pm}^{\mathsf{J}}. \tag{5.81}$$

The fixed point for the EI or S network is unstable if $Re(\gamma_{\pm}^{EI})$ or $Re(\gamma_{\pm}^{S})$ is positive, as then fluctuations will grow. Note that, since $\lambda_{-}^{\mathsf{J}} = J_0 - J'$, $\lambda_{-}^{\mathsf{W}} = W_0 - W'$, $T_0 = J_0 - W_0$, and $T' = W' - J'$, it follows that

$$\gamma_{-}^{S} = -1 - \lambda_{-}^{\mathsf{W}} + \lambda_{-}^{\mathsf{J}} = -1 + W' - J' + (J_0 - W_0) = -1 + T_0 + T'.$$

The above equation and equation (5.60) are showing the same thing: when $T_0 + T' > 1$, the S network is unstable against fluctuations in the x_- mode, which is opposing fluctuations in the odd and even columns (or in x_1 and x_2).

Although the fixed points for the two networks are the same, their eigenvalues $\gamma_{\pm}^{S}$ and $\gamma_{\pm}^{EI}$ are different, and so their stabilities can also differ. Since $\lambda_{\pm}^{\mathsf{J}}$ and $\lambda_{\pm}^{\mathsf{W}}$ are real, $\gamma_{\pm}^{S}$ is always

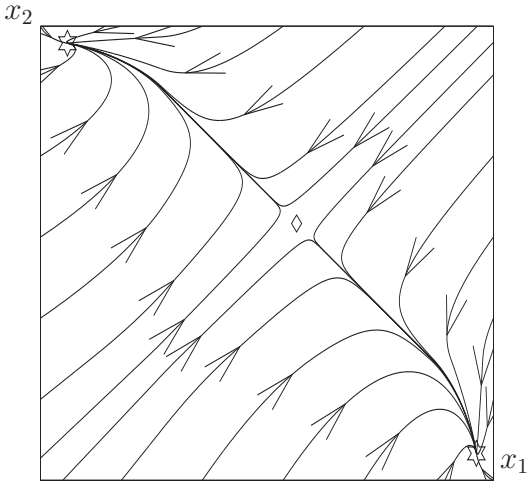


Fig. 5.56: The motion trajectory of the two-point S network under input $I \propto (1,1)$. The symmetric fixed point (marked by $\diamond$) becomes unstable when the two asymmetric fixed points (marked by ☆'s) appear. The symmetric fixed point is a saddle point, and the asymmetric ones are energy minima, of an energy landscape; the former repels and the latter attract nearby trajectories. The counterpart EI network has the same three fixed points but different dynamics (and no energy landscape). When the asymmetric fixed points in the EI networks are also unstable (and thus unapproachable), the network state can oscillate along the diagonal line $x_1 = x_2$ around the symmetric fixed point into the y dimensions without breaking symmetry. Adapted with permission from Li, Z. and Dayan, P., Computational differences between asymmetrical and symmetrical networks, *Network: Computation in Neural Systems*, 10(1): 59–77, Fig. 4, copyright © 1999, Informa Healthcare.

real. However, $\gamma_{\pm}^{EI}$ can be a complex number, leading to oscillatory behavior if its imaginary part $Im\left(\gamma_{\pm}^{EI}\right)$ is non-zero. One can derive that, for $k = +$ or $k = -$,

$$\text{when } \gamma_k^S > 0, \text{then } Re\left(\gamma_k^{EI}\right) > 0,$$
$$\text{i.e., the EI net is no less stable than the S net;} \tag{5.82}$$
$$\text{when } Im\left(\gamma_k^{EI}\right) \neq 0, \text{then } \gamma_k^S < 0,$$
$$\text{i.e., the S net is stable when the EI net is oscillatory (stable or not).} \tag{5.83}$$

These conclusions hold for any fixed point. Equation (5.82) can be proven by noting that $\gamma_k^S = -1 - \lambda_k^{\mathsf{W}} + \lambda_k^{\mathsf{J}} > 0$ gives $\lambda_k^{\mathsf{W}} < -1 + \lambda_k^{\mathsf{J}}$; hence $\left[\frac{1}{4}(\lambda_k^{\mathsf{J}})^2 - \lambda_k^{\mathsf{W}}\right]^{1/2} > \left|-1 + \frac{1}{2}\lambda_k^{\mathsf{J}}\right|$, and thus $Re\left(\gamma_k^{EI}\right) > 0$ (for one of the roots). Equation (5.83) can be proven by noting that $\frac{1}{4}\left(\lambda_k^{\mathsf{J}}\right)^2 < \lambda_k^{\mathsf{W}}$ leads to $\gamma_k^S < -1 - \lambda_k^{\mathsf{W}} + 2\sqrt{\lambda_k^{\mathsf{W}}} = -\left(1 - \sqrt{\lambda_k^{\mathsf{W}}}\right)^2 \leq 0$.

Now we can understand how the EI network maintains spatial symmetry under homogenous input $(I_1, I_2) \propto (1,1)$, even when $T' + T_0$ is large enough for the S network to break symmetry. Figure 5.56 shows the energy landscape and motion trajectory for the two-point S network under the homogenous input, with symmetry breaking. As analyzed above, the symmetry breaking is accompanied by three fixed points: one symmetric $\bar{x}_1 = \bar{x}_2$ and two asymmetric $\bar{x}_1 \neq \bar{x}_2$. The symmetric one is a saddle point, stable against fluctuations of x_+ (i.e., synchronous fluctuations in x_1 and x_2) from it but unstable against fluctuations of x_-

(i.e., opposing fluctuations in x_1 and x_2). The x_- fluctuation grows, with its initial value determining to which of the two asymmetric fixed points (x_1, x_2) will converge.

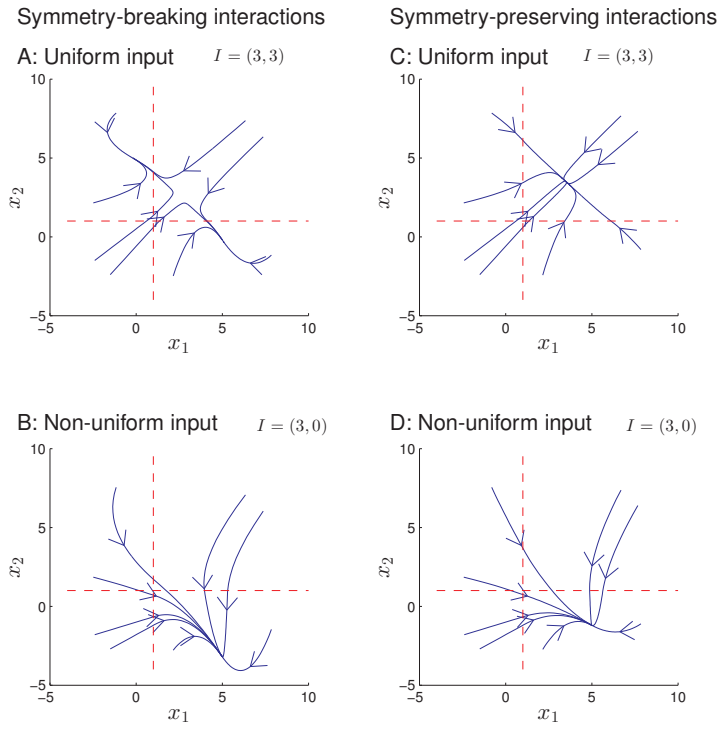


Fig. 5.57: Motion trajectories of the two-point S network. The interactions in A and B are symmetry breaking, with $T_0 = 0.5$ and $T' = 0.8$, so that the responses to uniform inputs converge to asymmetric fixed points (A). C, D: Lowering the inter-unit suppression to $T' = 0.3$ allows the network to preserve symmetry; however, the selective amplification ratio is now quite small. The function $g_y(y) = y$, and $g_x(x)$ is a threshold linear function with $x_{\text{th}} = 1$ and no saturation. The red dashed lines mark the threshold (x_{th}). Adapted with permission from Li, Z. and Dayan, P., Computational differences between asymmetrical and symmetrical networks, *Network: Computation in Neural Systems*, 10(1): 59–77, Fig. 2, copyright © 1999, Informa Healthcare.

The same three fixed points in the EI network can be all unstable. In particular, synchronous fluctuations x_+ from the symmetric fixed point $\bar{x}_1 = \bar{x}_2$ can be made unstable and oscillatory by

$$-1 + (J_0 + J')/2 > 0 \quad \text{and} \quad W_0 + W' > (J_0 + J')^2/4, \tag{5.84}$$

and the asymmetric fixed point can be made unstable by

$$-1 + J_0 > 0. \tag{5.85}$$

Note that, at the asymmetric fixed point, the non-active neural pair contributes nothing to the dynamics, and so the network becomes a one-point system, albeit a two-neuron, one-point system.

Given this, no fluctuation from the symmetric fixed point can converge—the only other fixed points (the asymmetric ones) are themselves unstable. Consequently, the fluctuations

around the symmetric fixed point tend to be symmetric along a trajectory $x_1(t) = x_2(t)$ and $y_1(t) = y_2(t)$, and oscillate in the (x, y) phase space. Small fluctuations in the x_- direction are also unstable; however, in the nonlinear system, they are strongly squashed below the threshold $x_{\rm th}$ and above saturation at $x_{\rm sat}$. Overall, this oscillation preserves the symmetry in the (x_1, x_2) space. This allows a very large selective amplification ratio without inducing any hallucination.

Figure 5.57 shows the trajectory of motion in the (x_1, x_2) space for two S networks. One network strongly amplifies an asymmetric input $I_1 \neq I_2$, but it spontaneously breaks symmetry by responding non-homogenously, $x_1 \neq x_2$, to homogenous input $I_1 = I_2$ (i.e., it strongly amplifies noise). Another network does not spontaneously break symmetry when $I_1 = I_2$, but it cannot amplify asymmetric input to nearly such a degree. Recall that the symmetric and asymmetric input represent the homogenous texture and isolated contours, respectively, of visual inputs in the expanded subnetwork for vertical bars. Hence, the S network cannot realize a sufficiently large selective amplification ratio, and so it is inadequate as a model of V1.

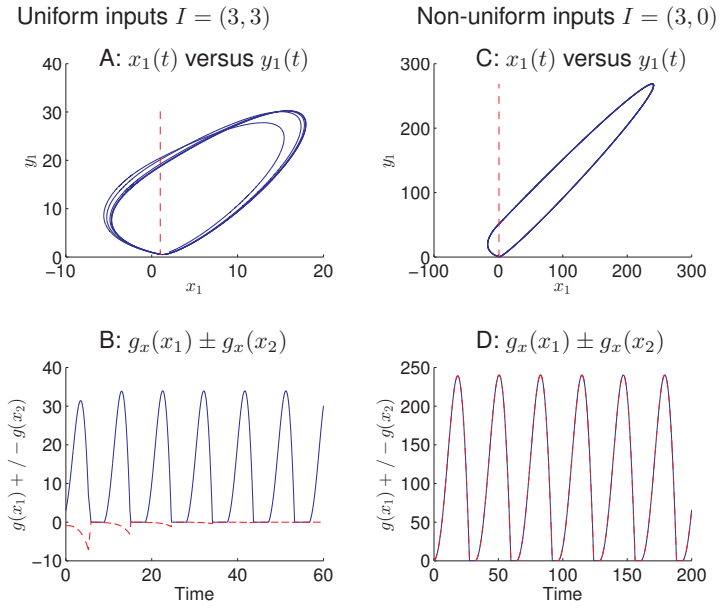


Fig. 5.58: Oscillatory trajectories of a two-point EI network with a high selective amplification ratio. The connections are $J_0 = 2.1$, $J' = 0.4$, $W_0 = 1.13$, and $W' = 0.9$. In B and D, the plot of $g_x(x_1) + g_x(x_2)$ is in blue, and $g_x(x_1) - g_x(x_2)$, in red. In the symmetric (uniform) input case, $g_x(x_1) - g_x(x_2)$ quickly decays in time (B). With asymmetric (non-uniform) input (C,D), the red and blue curves lie on top of each other (D). Here, $g_x(x)$, $g_y(y)$, and $x_{\rm th} = 1$ are the same as in Fig. 5.57. The red dashed lines in A and C mark the thresholds ($x_{\rm th}$). Adapted with permission from Li, Z. and Dayan, P., Computational differences between asymmetrical and symmetrical networks, *Network: Computation in Neural Systems*, 10(1): 59–77, Fig. 3, copyright © 1999, Informa Healthcare.

Figure 5.58 shows the evolution of a two-point EI network. The responses to both the symmetric and asymmetric inputs are oscillatory, but there is no spontaneous symmetry

breaking to homogenous inputs, even though the selective amplification ratio is high. Hence, the EI network is the minimal network architecture for our V1 computation.

We next expand the toy model subset for one particular orientation θ into a full network including more orientations θ and interactions between orientations. In this case, the dynamical equations are

$$\dot{x}_{i\theta} = -x_{i\theta} - g_y(y_{i,\theta}) + J_o g_x(x_{i\theta}) - \sum_{\Delta\theta\neq 0} \psi(\Delta\theta) g_y(y_{i,\theta+\Delta\theta}) + \sum_{j\neq i,\theta'} \mathsf{J}_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_{i\theta} + I_o, \tag{5.86}$$

$$\dot{y}_{i\theta} = -\alpha_y y_{i\theta} + g_x(x_{i\theta}) + \sum_{j\neq i,\theta'} \mathsf{W}_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_c, \tag{5.87}$$

which are the same as equations (5.8–5.9) (except for the lack of I_{noise}). The neural connections $\mathsf{T}_{i\theta,j\theta'}$ in the original S network are now replaced by various components including J, W, J_o, and ψ.

Although this analysis suggests that, unlike S networks, an EI network might be able to model V1, it is necessary for the connections J and W to be set appropriately in order to realize the necessary computations. The next section provides an analytical understanding of the nonlinear dynamics concerned and provides specific constraints on the J and W.

5.8.2 Dynamic analysis of the V1 model and constraints on the neural connections

The model state is characterized by $\{x_{i\theta}, y_{i\theta}\}$, or simply $\{x_{i\theta}\}$, omitting the auxiliary units $\{y_{i\theta}\}$. The interaction between excitatory and inhibitory cells makes $\{x_{i\theta}(t)\}$ intrinsically oscillatory in time (Li and Hopfield 1989), although whether the oscillations are damped or sustained depends on the external input patterns and neural connections. Thus, given an input $\{I_{i\theta}\}$, the model often does not convergence to a fixed point where $\dot{x}_{i\theta} = \dot{y}_{i\theta} = 0$. However, if $\{x_{i\theta}(t)\}$ oscillates periodically around a fixed point, then after the transient following the onset of $\{I_{i\theta}\}$, the temporal average of $\{x_{i\theta}(t)\}$ can characterize the model output and approximate the encircled fixed point. We henceforth use the notation $\{\bar{x}_{i\theta}\}$ to denote either the fixed point, if it is stable, or the temporal average, and denote the computation as $I \to g_x(\bar{x}_{i\theta})$.

5.8.2.1 A single pair of neurons

An isolated single pair $i\theta$ follows equations

$$\dot{x} = -x - g_y(y) + J_o g_x(x) + I, \tag{5.88}$$

$$\dot{y} = -y + g_x(x) + I_c, \tag{5.89}$$

(omitting the redundant index $i\theta$) where we set $\alpha_y = 1$ for simplicity and $I = I_{i\theta} + I_o$. The gain in the input-output transform $(I, I_c) \to g_x(\bar{x})$ at a fixed point $(\bar{x}, \bar{y})$ is

$$\frac{\delta g_x(\bar{x})}{\delta I} = \frac{g_x'(\bar{x})}{1 + g_x'(\bar{x}) g_y'(\bar{y}) - J_o g_x'(\bar{x})}, \qquad \frac{\delta g_x(\bar{x})}{\delta I_c} = -g_y'(\bar{y}) \frac{\delta g_x(\bar{x})}{\delta I}, \tag{5.90}$$

where $g_x'(\bar{x})$ and $g_y'(\bar{y})$, respectively, are the derivatives of the functions $g_x(.)$ and $g_y(.)$ at the fixed point $\bar{x}$ and $\bar{y}$.

Figure 5.59 illustrates an example when both $g_x(x)$ and $g_y(y)$ are piece-wise linear functions. In this case, the input-output transform $I \to g_x(\bar{x})$ is also piece-wise linear; see

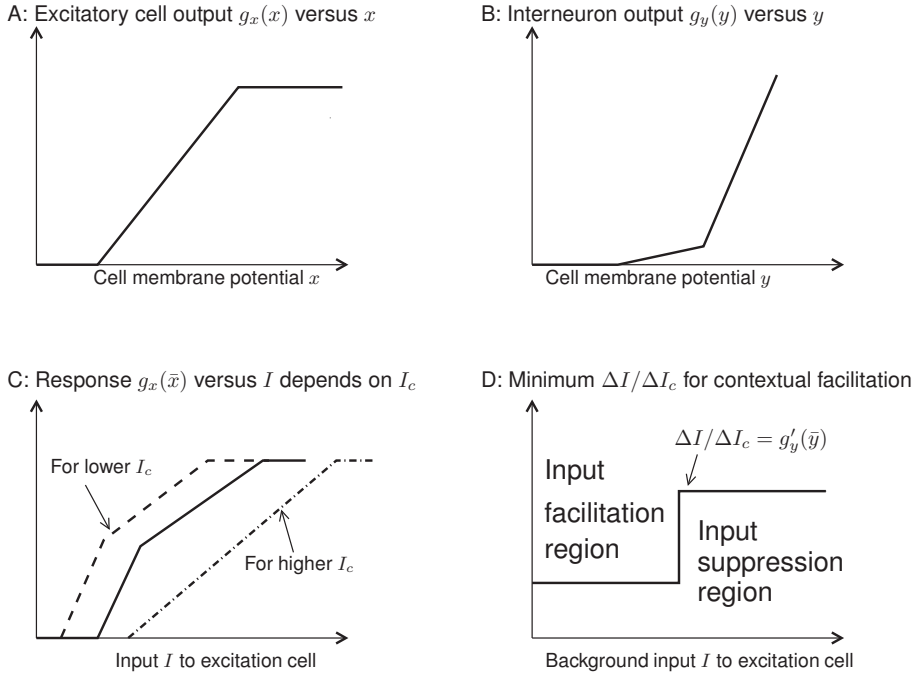


Fig. 5.59: A,B: Examples of $g_x(x)$ and $g_y(y)$ functions. C: Input-output function $I \to g_x(\bar{x})$ for an isolated neural pair without inter-pair neural interactions, under different levels of I_c. D: The overall effect of any additional external or contextual inputs $(\Delta I, \Delta I_c)$ on a neural pair is excitatory or inhibitory depending on whether $\Delta I/\Delta I_c > g'_y(\bar{y})$; this depends on background input I. Adapted with permission from Li, Z., Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex, *Neural Computation*, 13(8): 1749–1780, Fig. 3, copyright © 2001, MIT Press.

Fig. 5.59 C. The threshold, input gain control, and saturation in $I \to g_x(\bar{x})$ are apparent. The slope $\frac{\delta g_x(\bar{x})}{\delta I}$ is non-negative; otherwise, $I = 0$ gives non-zero output $x \neq 0$. The slope increases with $g'_x(\bar{x})$, decreases with $g'_y(\bar{y})$, and depends on I_c. Shifting (I, I_c) to $(I + \Delta I, I_c + \Delta I_c)$ changes $g_x(\bar{x})$ by

$$\Delta g_x(\bar{x}) \approx (\delta g_x(\bar{x})/\delta I) \left(\Delta I - g'_y(\bar{y})\Delta I_c\right), \tag{5.91}$$

which is positive or negative depending on whether $\Delta I/\Delta I_c > g'_y(\bar{y})$. Hence, a more elaborate model could allow that the interneurons also be partially activated by the external visual input, as is suggested by physiology (White 1989). It would be necessary that $\Delta I/\Delta I_c > g'_y(\bar{y})$.

5.8.2.2 Two interacting pairs of neurons with non-overlapping receptive fields

Consider two vectors (x_1, y_1) and (x_2, y_2) denoting the states of two interacting EI pairs whose connections are $\mathsf{J}_{12} = \mathsf{J}_{21} = J'$ and $\mathsf{W}_{12} = \mathsf{W}_{21} = W'$. Then

$$\begin{aligned}
\dot{x}_a &= -x_a - g_y(y_a) + J_o g_x(x_a) + J' g_x(x_b) + I_a + I_o, \\
\dot{y}_a &= -y_a + g_x(x_a) + W' g_x(x_b) + I_c
\end{aligned}$$

where $a, b = 1, 2$ and $a \neq b$. Hence, to the pair (x_1, y_1), the effect of the pair (x_2, y_2) is the same as adding $\Delta I = J' g_x(x_2)$ to the I_1 and adding $\Delta I_c = W' g_x(x_2)$ to I_c. This gives, according to equation (5.91),

$$\Delta g_x(\bar{x}_1) \approx (\delta g_x(\bar{x}_1)/\delta I) \left(\Delta I - g'_y(\bar{y}_1)\, \Delta I_c\right) \tag{5.92}$$

$$= (\delta g_x(\bar{x}_1)/\delta I) \left(J' - g'_y(\bar{y}_1) W'\right) g_x(x_2). \tag{5.93}$$

Hence,

$$\text{the net effective connection from } x_2 \text{ to } x_1 \text{ is } \quad J' - g'_y(\bar{y}_1) W'. \tag{5.94}$$

This connection strength depends on how active the interneuron is, $\bar{y}_1$, and thus it also depends on the input activation of the bar associated with (x_1, y_1). Therefore, since $g'_y(\bar{y}_1)$ tends to increase with the direct input I_1 (and with I_c), the influence on x_1 from the contextual input I_2 becomes more suppressive as the direct input I_1 to the bar becomes stronger. This explains some of the contrast dependence of the contextual influences that is observed physiologically (Sengpiel, Baddeley, Freeman, Harrad and Blakemore 1998). In the simplest case, that $I \equiv I_1 = I_2$, the two bars (associated with these two EI pairs) suppress each other more strongly as input contrast increases, but they can facilitate each other's response when the input contrast is sufficiently weak and J' sufficiently strong. Figure 5.59 D shows an example of how the contextual inputs can switch from being facilitatory to being suppressive as I increases (Stemmler, Usher and Niebur 1995, Somers, Todorov, Siapas, Toth, Kim and Sur 1998).

This very simple model of contextual influence, with only two EI pairs, can be applied to account for various perceptual phenomena involving only single test and contextual bars. For example, a contextual bar can alter the detection threshold (Polat and Sagi 1993, Kapadia et al. 1995) or perceived orientation (Gilbert and Wiesel 1990, Li 1999b) of a test bar.

5.8.2.3 A one-dimensional array of identical bars

Figure 5.60 ABC shows sample input stimuli comprising infinitely long horizontal arrays of evenly spaced, identical bars. These can be approximated as

$$I_{i\theta} = \begin{cases} I_{\text{array}} & \text{for } i = (m_i, n_i = 0) \text{ on the horizontal axis and } \theta = \theta_1, \\ 0 & \text{otherwise.} \end{cases} \tag{5.95}$$

The approximation is to set $I_{i\theta} = 0$ for $\theta \neq \theta_1$; this is reasonable when the input contrast is weak and the neurons have small orientation tuning widths. When bars $i\theta$ outside the array are silent (i.e., $g_x(x_{i\theta}) = 0$) due to insufficient excitation, we omit them and treat the one-dimensional system that only contains the activated neurons. Omitting index θ and using i to index locations, we get

$$\dot{x}_i = -x_i - g_y(y_i) + J_o g_x(x_i) + \sum_{j \neq i} \mathsf{J}_{ij} g_x(x_j) + I_{\text{array}} + I_o, \tag{5.96}$$

$$\dot{y}_i = -y_i + g_x(x_i) + \sum_{j \neq i} \mathsf{W}_{ij} g_x(x_j) + I_c. \tag{5.97}$$

Translation symmetry implies that all units have the same equilibrium point $(\bar{x}_i, \bar{y}_i) = (\bar{x}, \bar{y})$, and

$$\dot{\bar{x}} = 0 = -\bar{x} - g_y(\bar{y}) + \left(J_o + \sum_{i \neq j} \mathsf{J}_{ij}\right) g_x(\bar{x}) + I_{\text{array}} + I_o, \tag{5.98}$$

$$\dot{\bar{y}} = 0 = -\bar{y} + \left(1 + \sum_{i \neq j} \mathsf{W}_{ij}\right) g_x(\bar{x}) + I_c. \tag{5.99}$$

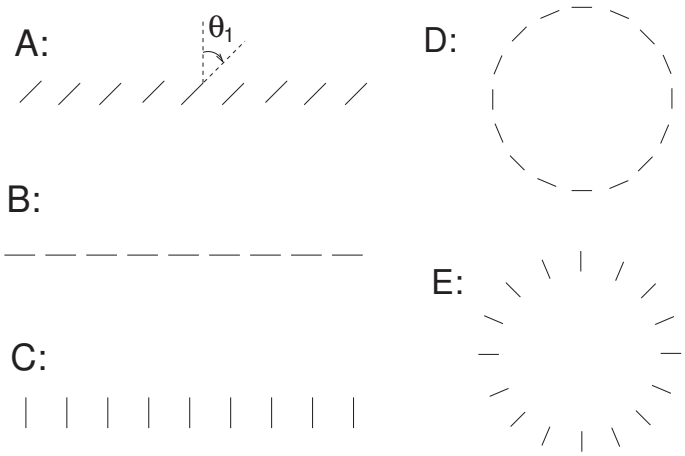


Fig. 5.60: Examples of one-dimensional input stimuli. A: Horizontal array of identical bars oriented at angle θ_1. B: A special case of A when $\theta_1 = \pi/2$ and, in C, when $\theta_1 = 0$. D: An array of bars arranged as being tangential to a circle; the pattern in B is a special case of this circle when the radius is infinitely large. E: Same as D except that the bars are perpendicular to the circle's circumference; the pattern in C is a special case of E when the radius is infinitely large. Reproduced with permission from Li, Z., Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex, *Neural Computation*, 13(8): 1749–1780, Fig. 4, copyright © 2001, MIT Press.

This array is then equivalent to a single EI neural pair (cf. equations (5.88) and (5.89)), making the substitution $J_o \to J_o + \sum_j \mathsf{J}_{ij}$ and $g'(\bar{y}) \to g'_y(\bar{y}) \left(1 + \sum_j \mathsf{W}_{ij}\right)$. The response to bars in the array is thus higher than to an isolated bar if the net extra excitatory connection

$$\mathcal{E} \equiv \sum_j \mathsf{J}_{ij} \tag{5.100}$$

is stronger than the net extra inhibitory (effective) connection

$$\mathcal{I} \equiv g'_y(\bar{y}) \sum_j \mathsf{W}_{ij}. \tag{5.101}$$

The input-output relationship $I \to g_x(\bar{x})$ is qualitatively the same as that for a single bar, but with a quantitative change in the gain

$$\frac{\delta g_x(\bar{x})}{\delta I} = \frac{g'_x(\bar{x})}{1 + g'_x(\bar{x}) \left[g'_y(\bar{y}) - (\mathcal{E} - \mathcal{I})\right] - J_o g'_x(\bar{x})}. \tag{5.102}$$

When $\mathcal{E} - \mathcal{I} = 0$, the gain reverts back to that of a single bar. The connections $\mathcal{E}$ and $\mathcal{I}$ depend on the angle θ_1 between the bars and the array; see Fig. 5.60 A. Consider connections as in the association field in Fig. 5.15 B. When the bars are parallel to the array, making a straight line (Fig. 5.60 B), $\mathcal{E} > \mathcal{I}$. The condition for enhancing the responses to a contour is

$$\text{contour facilitation} \quad F_{\text{contour}} \equiv (\mathcal{E} - \mathcal{I}) g_x(\bar{x}) > 0. \tag{5.103}$$

When the bars are orthogonal to the array (Fig. 5.60 C), $\mathcal{E} < \mathcal{I}$, and the responses are suppressed. This analysis extends to other one-dimensional, translation-invariant arrays like

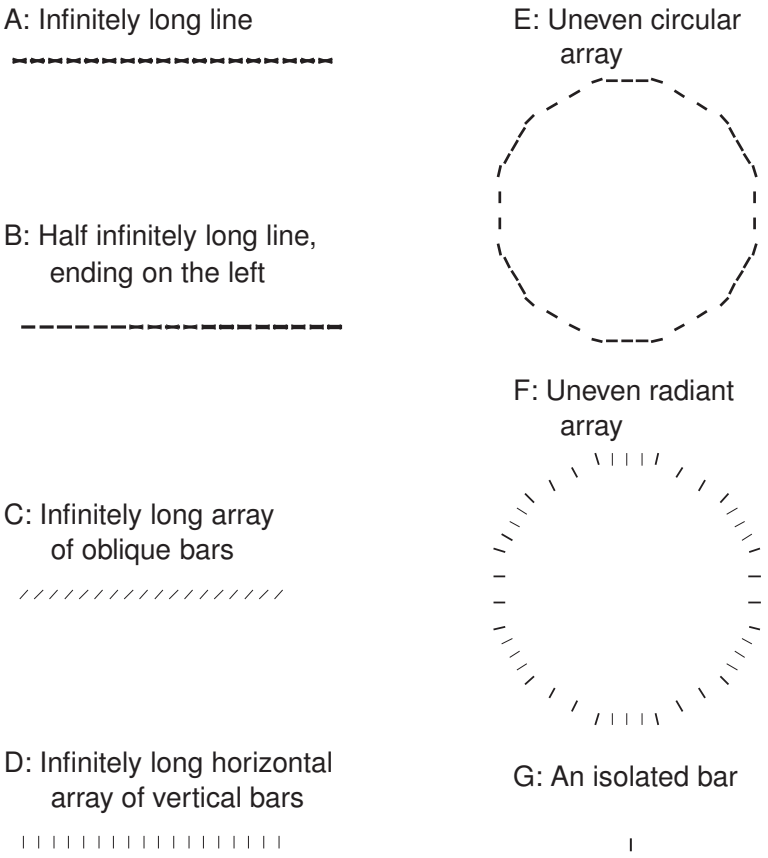


Fig. 5.61: The response $g_x(x_{i\theta})$ of the V1 model (visualized by thickness) to one-dimensional arrays of bars. The input $\hat{I}_{i\theta} = 1.5$ is of low/intermediate contrast for all visible bars. Compared with the isolated bar in G, contextual facilitation causes higher outputs in A, B, E; contextual suppression causes lower outputs in C, D, F. The uneven spacings between the bars (E, F) or at an end of a line (at the left end of B) cause deviations from the translation invariance of responses. Note that the responses taper off near the end of the line in B, and the responses are noticeably weaker to bars that are more densely packed in F. In A and B, cells preferring orientations that are nearly (but not exactly) horizontal are also excited above threshold. This goes beyond the approximate treatment in the text. Adapted with permission from Li, Z., Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex, *Neural Computation*, 13(8): 1749–1780, Fig. 5, copyright © 2001, MIT Press.

those in Fig. 5.60 DE. The straight line in Fig. 5.60 B is in fact the limit of a circle in Fig. 5.60 D when the radius goes to infinity. Similarly, the pattern in Fig. 5.60 C is a special case of the one in Fig. 5.60 E.

How good the approximations in equations (5.95–5.99) are depends on the input. This is illustrated in Fig. 5.61. In Fig. 5.61 A, cells whose RFs are centered on the line, oriented close to, but not exactly, horizontal, are also excited above threshold. This is not consistent with our approximation $g_x(x_{i\theta}) = 0$ for non-horizontal θ. (This should not cause perceptual problems, though, given population coding.) The cells for these non-horizontal bars can be

activated by direct visual input $I_{i\theta}$ for $\theta \neq \theta_1$ ($\theta \approx \theta_1$), due to the finite width of orientation tuning *and* by the colinear facilitation from other bars in or along the line. In Fig. 5.61 B, the approximation of translation invariance $\bar{x}_i = \bar{x}_j$ for all bars is compromised by the fact that the array comes to an end. The bars at or near the left end of the line are less enhanced since they receive less or no contextual facilitation from their left. Uneven spacing between bars in Fig. 5.61 EF also compromises translation invariance. In Fig. 5.61 F, the more densely spaced bars are more strongly suppressed by their neighbors.

5.8.2.4 Two-dimensional textures and texture boundaries

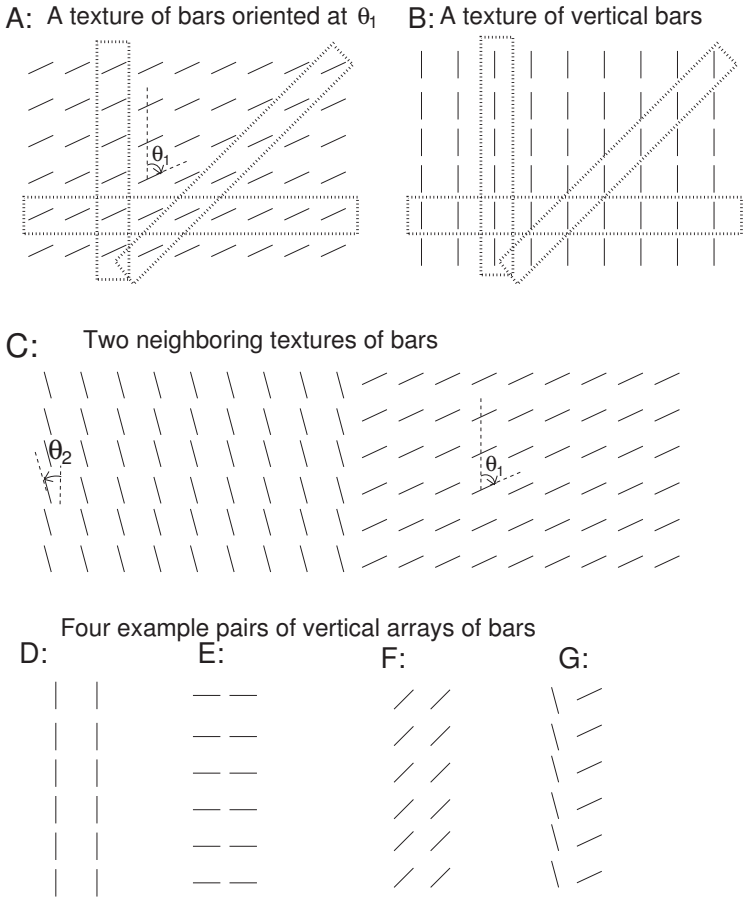


Fig. 5.62: Examples of two-dimensional textures and their interactions. A: Texture made of bars oriented at θ_1 and sitting on a Manhattan grid. This can be seen as a horizontal array of vertical arrays of bars, or indeed as a vertical or oblique array of arrays of bars, as in the dotted boxes. B: A special case of A when $\theta_1 = 0$. C: Two nearby textures with a boundary. D, E, F: Examples of nearby, identical, vertical arrays. G: Two nearby but different vertical arrays. When each vertical array is seen as an entity, one can calculate the effective connections J' and W' between them (see the definitions in the text). Adapted with permission from Li, Z., Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex, *Neural Computation*, 13(8): 1749–1780, Fig. 6, copyright © 2001, MIT Press.

The analysis of infinitely long one-dimensional arrays can be extended to an infinitely large two-dimensional texture of uniform inputs $I_{i\theta_1} = I_{\text{texture}}$ in the case that $i = (m_i, n_i)$ sits on a regularly spaced grid (Fig. 5.62 A). The sums $\mathcal{E} = \sum_j \mathsf{J}_{ij}$ and $\mathcal{I} = g'_y(\bar{y}) \sum_j \mathsf{W}_{ij}$ are now taken over all j in that grid.

Physiologically, the response to a bar is reduced when the bar is part of a texture (Knierim and Van Essen 1992). This arises when $\mathcal{E} < \mathcal{I}$. Consider, for example, the case when $i = (m_i, n_i)$ form a Manhattan grid with integer values of m_i and n_i (Fig. 5.62). The texture can be seen as a horizontal array of vertical arrays of bars, e.g., a horizontal array of vertical contours in Fig. 5.62 B. The effective connections between two vertical arrays (Fig. 5.62 DEF) spaced apart by a are:

$$J'_a \equiv \sum_{j, m_j = m_i + a} \mathsf{J}_{ij}, \qquad W'_a \equiv \sum_{j, m_j = m_i + a} \mathsf{W}_{ij}. \tag{5.104}$$

Then $\mathcal{E} = \sum_a J'_a$ and $\mathcal{I} = g'_y(\bar{y}) \sum_a W'_a$. The effective connection within a single vertical array is J'_0 and W'_0. One should design J and W such that contour enhancement and texture suppression can occur using the same neural circuit (V1). That is, when the vertical array is a long straight line ($\theta_1 = 0$), contour enhancement (i.e., $J'_0 > g'_y(\bar{y}) W'_0$) occurs when the line is isolated, but overall suppression (i.e., $\mathcal{E} = \sum_a J'_a < \mathcal{I} = g'_y(\bar{y}) \sum_a W'_a$) occurs when that line is embedded within a texture of lines (Fig. 5.62 B). This can be satisfied when there is sufficient excitation within a line and sufficient inhibition between the lines.

Computationally, contextual suppression within a texture region endows its boundaries with relatively higher responses, thereby making them salient. The contextual suppression of a bar within a texture is

$$C^{\theta_1}_{\text{whole-texture}} \equiv \sum_a \left[g'_y(\bar{y}_{\theta_1}) W'^{\theta_1}_a - J'^{\theta_1}_a \right] g_x(\bar{x}_{\theta_1}) = (\mathcal{I} - \mathcal{E}) g_x(\bar{x}_{\theta_1}) > 0, \tag{5.105}$$

where $\bar{x}_{\theta_1}$ denotes the (translation-invariant) fixed point for all texture bars in an infinitely large texture, and $J'^{\theta_1}_a$ and $W'^{\theta_1}_a$ are as the quantities in equation (5.104) with the addition of the superscript θ_1 to indicate the orientation of the texture bars, i.e., the connections J_{ij} and W_{ij} refer to $\mathsf{J}_{i\theta_1 j\theta_1}$ and $\mathsf{W}_{i\theta_1 j\theta_1}$, respectively, in the full V1 model.

Consider the bars on the vertical axis $i = (m_i = 0, n_i)$. Removing the texture bars to their left, $j = (m_j < 0, n_j)$, removes the contextual suppression from them and thereby highlights the boundary bars $i = (m_i = 0, n_i)$. Then the activity $(\bar{x}_{i\theta_1}, \bar{y}_{i\theta_1})$ depends on m_i, the distance of the bars from the texture boundary. As $m_i \to \infty$, the responses $(\bar{x}_{i\theta_1}, \bar{y}_{i\theta_1})$ approach those $(\bar{x}_{\theta_1}, \bar{y}_{\theta_1})$ to the bars within an infinitely large texture. The contextual suppression of the boundary bars ($m_i = 0$) is

$$C^{\theta_1}_{\text{half-texture}} \equiv \sum_{m_j \geq 0} \left[g'_y(\bar{y}_{i\theta_1}) W'^{\theta_1}_{m_j} - J'^{\theta_1}_{m_j} \right] g_x(\bar{x}_{j\theta_1}) \tag{5.106}$$

$$\approx \sum_{a \geq 0} \left[g'_y(\bar{y}_{\theta_1}) W'^{\theta_1}_a - J'^{\theta_1}_a \right] g_x(\bar{x}_{\theta_1}) < C^{\theta_1}_{\text{whole-texture}}, \tag{5.107}$$

making the approximation $(\bar{x}_{j\theta_1}, \bar{y}_{j\theta_1}) \approx (\bar{x}_{\theta_1}, \bar{y}_{\theta_1})$ for all $m_j \geq 0$.

The boundary highlight persists when there is a neighboring texture whose bars have a different orientation θ_2, with bar positions $i = (m_i < 0, n_i)$ (Fig. 5.62 C). To analyze this, define connections between arrays in different textures (Fig. 5.62 G) as

$$J'^{\theta_1\theta_2}_a \equiv \sum_{j, m_j = m_i + a} \mathsf{J}_{i\theta_1 j\theta_2}, \qquad W'^{\theta_1\theta_2}_a \equiv \sum_{j, m_j = m_i + a} \mathsf{W}_{i\theta_1 j\theta_2}. \tag{5.108}$$

When $\theta_1 = \theta_2$, then $J_a'^{\theta_1\theta_2} = J_a'^{\theta_1}$ and $W_a'^{\theta_1\theta_2} = W_a'^{\theta_1}$. The contextual suppression from the neighboring texture (θ_2) on the texture boundary ($m_i = 0$) is

$$C^{\theta_1,\theta_2}_{\text{neighbor}-\text{half-texture}} \equiv \sum_{m_j<0} \left[g_y'(\bar{y}_{i\theta_1}) W_{m_j}'^{\theta_1\theta_2} - J_{m_j}'^{\theta_1\theta_2} \right] g_x(\bar{x}_{j\theta_2}).$$

With connections as for the association field, $\mathsf{J}_{i\theta_1,j\theta_2}$ and $\mathsf{W}_{i\theta_1,j\theta_2}$, $J_a'^{\theta_1\theta_2}$ and $W_a'^{\theta_1\theta_2}$ tend to link similarly oriented bars $\theta_1 \sim \theta_2$. Consequently, $C^{\theta_1,\theta_2}_{\text{neighbor}-\text{half-texture}}$ is minimum or zero when $\theta_1 \perp \theta_2$ and increases with decreasing $|\theta_1 - \theta_2|$. Hence, the boundary highlight is expected to increase with the orientation contrast $|\theta_1 - \theta_2|$. The net contextual suppression on the border, contributed by both textures, is

$$C^{\theta_1,\theta_2}_{\text{two half-textures}} \equiv C^{\theta_1}_{\text{half-texture}} + C^{\theta_1,\theta_2}_{\text{neighbor}-\text{half-texture}}.$$

Hence, the border enhancement, or the reduction of contextual suppression at the border relative to regions further inside the texture is

$$\begin{aligned}
\delta C &\equiv C^{\theta_1}_{\text{whole-texture}} - C^{\theta_1,\theta_2}_{\text{two half-textures}} && (5.109)\\
&\approx C^{\theta_1,\theta_2=\theta_1}_{\text{neighbor}-\text{half-texture}} - C^{\theta_1,\theta_2}_{\text{neighbor}-\text{half-texture}} && (5.110)\\
&\approx \sum_{a<0} \left[g_y'(\bar{y}_{\theta_1}) W_a'^{\theta_1} - J_a'^{\theta_1} \right] g_x(\bar{x}_{\theta_1}) && (5.111)\\
&\quad - \sum_{a<0} \left[g_y'(\bar{y}_{\theta_1}) W_a'^{\theta_1\theta_2} - J_a'^{\theta_1\theta_2} \right] g_x(\bar{x}_{\theta_2}). && (5.112)
\end{aligned}$$

Here, in addition to the previous approximation $(\bar{x}_{j\theta_1}, \bar{y}_{j\theta_1}) \approx (\bar{x}_{\theta_1}, \bar{y}_{\theta_1})$ for all $m_j \geq 0$, we approximated $\bar{x}_{j\theta_2} \approx \bar{x}_{\theta_2}$ for $m_j < 0$. Usually $\bar{x}_{\theta_2} \neq \bar{x}_{\theta_1}$ since the fixed point should depend on the relative orientation between the bars and the orientation of the arrays.

Assuming $J_a'^{\theta_1\theta_2} \approx 0$ and $W_a'^{\theta_1\theta_2} \approx 0$ when $|\theta_1 - \theta_2| = \pi/2$, and noting that $\bar{x}_{\theta_1} \approx \bar{x}_{\theta_2}$ when $\theta_1 \approx \theta_2$,

$$\delta C \approx \begin{cases} 0 & \text{for } \theta_1 \approx \theta_2, \\ \sum_{a<0} \left[g_y'(\bar{y}_{\theta_1}) W_a'^{\theta_1} - J_a'^{\theta_1} \right] g_x(\bar{x}_{\theta_1}) > 0 & \text{for } \theta_1 \perp \theta_2, \\ \text{roughly increases} & \text{as } |\theta_1 - \theta_2| \text{ increases.} \end{cases} \quad (5.113)$$

Thus the border highlight diminishes as the orientation contrast approaches 0; this was seen in Fig. 5.23. Furthermore, even at a given contrast $|\theta_1 - \theta_2|$, the border enhancement δC depends on θ_1. For instance, with $|\theta_1 - \theta_2| = \pi/2$ and using the association field connections, the enhancement δC for border bars parallel to the border $\theta_1 = 0$ (which form a contour) is higher than that for border bars perpendicular to the border $\theta_1 = \pi/2$. This is because both the suppression $g_y'(\bar{y}_{\theta_1}) W_a'^{\theta_1} - J_a'^{\theta_1}$ between parallel contours ($\theta_1 = 0$ and $a \neq 0$) and the facilitation $J_0'^{\theta_1} - g_y'(\bar{y}_{\theta_1}) W_0'^{\theta_1}$ within a contour (Fig. 5.62 D) are much stronger than their counterparts for the vertical arrays of horizontal bars (Fig. 5.62 E). Thus the strength of the border highlight is predicted to be tuned to the relative orientation θ_1 between the border and the bars (Li 2000b). This explains the asymmetry in the responses in Fig. 5.22 B—the highlight of the vertical border is much stronger for the vertical texture bars at the border than for the horizontal ones.

The approximations $(\bar{x}_{i\theta_1}, \bar{y}_{i\theta_1}) \approx (\bar{x}_{\theta_1}, \bar{y}_{\theta_1})$ for $m_i \geq 0$ and $\bar{x}_{i\theta_2} \approx \bar{x}_{\theta_2}$ for $m_i < 0$, which were used to arrive at equation (5.113), clearly break down at the border. This is especially true at more salient borders like that in Fig. 5.22 B. This breakdown accentuates the tuning of the border highlight to θ_1.

As oft noted, iso-orientation suppression underlies the border highlight. By equation (5.105), its strength $\mathcal{I} - \mathcal{E}$ depends on input contrast through $g'_y(\bar{y})$. Since $g'_y(\bar{y})$ usually increases with increasing $\bar{y}$, the highlight is stronger at higher contrast. This is essentially the same as the dependence on input contrast of the contextual influence between two bars analyzed around equation (5.94). From just the perspective of dynamics, neural connections could be designed such that either of the following two situations holds: one is that iso-orientation suppression holds at all input contrasts; the other is that iso-orientation suppression holds only at sufficiently high input contrast and becomes iso-orientation facilitation at very low contrast (Li 1998a, Li 1999b). Psychophysically, texture segmentation does require an input contrast that is well above the texture detection threshold (Nothdurft 1994), suggesting that iso-orientation suppression diminishes at low input contrast. Computationally, facilitation certainly helps texture detection, which could be more important than segmentation at low input contrast.

For the parameters we have used here for the V1 model (Li 1998a, Li 1999b) (given in the appendix; see Section 5.9), contour facilitation ($F_{\text{contour}} > 0$) occurs at all contrasts, since no W connection links the contour segments. Connections different from the bowtie association field would have to be employed to model diminished contour enhancement at high contrast (Sceniak, Ringach, Hawken and Shapley 1999).

5.8.2.5 Translation invariance and pop-out

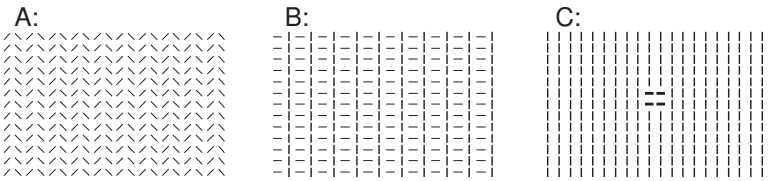


Fig. 5.63: Model responses to globally homogenous (A, B) and inhomogenous (C) input images, each composed of bars of equal input contrasts. A: The response to this globally homogenous (though locally inhomogenous) texture is uniform saliency. B: In this globally homogenous texture, the vertical bars are more salient than the horizontal ones; however, the whole texture has a translation invariant saliency distribution. C: The small figure pops out from the background; this is where translation invariance breaks down in the input, with the whole figure being its own boundary. Adapted with permission from Li, Z., Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex, *Neural Computation*, 13(8): 1749–1780, Fig. 8, copyright © 2001, MIT Press.

Consider the familiar case from above of two homogenous textures, each made of iso-oriented bars. There is an orientation contrast at the border between the two textures. This border is highlighted by higher responses to the border bars. However, if orientation contrasts are spatially homogenous within the texture itself, then their evoked responses will also be spatially homogenous. Figure 5.63 A shows an example in which the texture is made of alternating columns of bars at $\theta_1 = 45^o$ and $\theta_2 = 135^o$ in even and odd columns, respectively. Let $(\bar{x}_{\theta_1}, \bar{y}_{\theta_1})$ and $(\bar{x}_{\theta_2}, \bar{y}_{\theta_2})$ denote the states for the bars oriented at θ_1 and θ_2, respectively. The contextual suppression of a bar oriented at θ_1 is

$$C_{\text{complex-texture}} = \sum_{\text{even } a} \left[g_y'(\bar{y}_{\theta_1}) W_a'^{\theta_1} - J_a'^{\theta_1} \right] g_x(\bar{x}_{\theta_1}) + \sum_{\text{odd } a} \left[g_y'(\bar{y}_{\theta_1}) W_a'^{\theta_1 \theta_2} - J_a'^{\theta_1 \theta_2} \right] g_x(\bar{x}_{\theta_2}).$$

Since $C_{\text{complex-texture}} \neq C^{\theta_1}_{\text{whole-texture}}$, the value of $\bar{x}_{\theta_1}$ is not the same as it would be in a simple texture of bars uniformly oriented at θ_1. This applies similarly to $\bar{x}_{\theta_2}$. Furthermore, for general θ_1 and θ_2, $\bar{x}_{\theta_1} \neq \bar{x}_{\theta_2}$. In Fig. 5.63 A, reflection symmetry leads to $\bar{x}_{\theta_1} = \bar{x}_{\theta_2}$, i.e., a uniform saliency within the whole texture. In Fig. 5.63 B, the vertical bars evoke higher responses than the horizontal ones because of contour facilitation. Nevertheless, no local patch within this complex texture is more salient than any other patch. This translation invariance in saliency is simply the result of the network preserving the translation invariance in the input (texture), as long as the translation symmetry is not spontaneously broken (see Section 5.8.1).

A special case of a texture boundary is when one small texture patch is embedded in a large and different texture. The small texture is small enough that the whole texture is its own boundary, and thus pops out from the background (Fig. 5.63 C). In general, orientation contrasts do not correspond to texture boundaries and thus do not necessarily pop out. Through contextual influences, the highlight at a texture border can alter responses to nearby locations up to a distance comparable to the length of the lateral connections. Hence, the response to a texture region is not homogenous unless this region is far enough away from the border. This was elaborated in Section 5.4.5.

5.8.2.6 Filling-in and leaking-out

Small fragments of a contour or homogenous texture can be missing in inputs due to input noise or to the visual scene itself. We define filling-in as the phenomenon of not noticing or not perceiving the missing visual input fragments as missing. It could be caused by one of the following two possible mechanisms. The first is that the neurons for the missing fragment are activated by contextual influences just as if they had actually received direct visual input themselves. This is a common assumption in models (Grossberg and Mingolla 1985). The second possibility is that, even though the missing fragments do not evoke any significant response in the neurons whose RFs cover their locations, the input locations bordering the missing fragments are not sufficiently salient or conspicuous to attract attention strongly. In other words, according to this explanation, filling-in arises if the "saliency of the hole," which we analyzed in Section 5.4.4.7, is insufficient. Consequently, the missing fragments are only noticeable by visual scrutiny. It is not yet clear from physiology (Kapadia et al. 1995) which mechanism is involved.

Consider these two mechanisms for the case of a single bar segment $i = (m_i = 0, n_i = 0)$ that is missing in a smooth contour—e.g., the horizontal line of Fig. 5.64 A. To excite the cell i to firing threshold, i.e., $x_i > x_{\text{th}}$ (such that $g_x(x_i) > 0$), contextual facilitation

$$\sum_{j=(m_j \neq 0, n_j = 0)} \left[\mathsf{J}_{ij} - \mathsf{W}_{ij} g_y'(\bar{y}_i) \right] g_x(\bar{x}_j)$$

should be strong enough, or, approximately,

$$F_{\text{contour}} + I_o = (\mathcal{E} - \mathcal{I}) g_x(\bar{x}) + I_o > x_{\text{th}}, \tag{5.114}$$

where I_o is the background input not caused by external visual input, F_{contour} and the effective net connections $\mathcal{E}$ and $\mathcal{I}$ are as defined in equations (5.100–5.103), and the approximation $(\bar{x}_j, \bar{y}_j) \approx (\bar{x}, \bar{y})$ is adopted as if responses to all the non-missing bars are unaffected by the missing fragment.

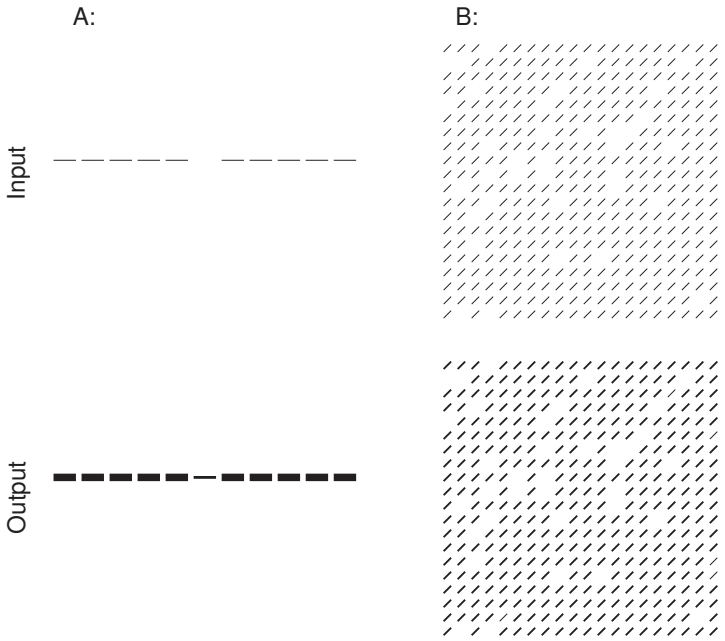


Fig. 5.64: Examples of filling-in; model responses $g_x(x_{i\theta})$ to inputs composed of bars of equal contrasts in each example. A: A line with a gap; the response to the gap is non-zero. B: A texture with missing bars; the responses to bars near the missing bars are not noticeably higher than the responses to other texture bars. Adapted with permission from Li, Z., Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex, *Neural Computation*, 13(8): 1749–1780, Fig. 9, copyright © 2001, MIT Press.

However, it is necessary that finite lines do not either grow longer or fatter as a result of contextual effects. For the first, we require that the neuron $i = (m_i = 0; n_i = 0)$ should not be excited above the threshold $x_{\rm th}$ if the left (or right) half $j = (m_j < 0, n_j = 0)$ of the horizontal contour is removed. Otherwise the contour will extend beyond its end or grow in length—this is referred to as leaking-out. To prevent leaking-out,

$$F_{\rm contour}/2 + I_o < x_{\rm th}, \tag{5.115}$$

since the contour facilitation to i is approximately $F_{\rm contour}/2$, half of that $F_{\rm contour}$ in an infinitely long contour. The inequality (5.115) is satisfied for the line end in Fig. 5.61 B and should hold at any contour saliency $g_x(\bar{x})$. Not having leaking-out also means that large gaps in lines can not be filled in.

To the extent that segments within a smooth contour facilitate each other's firing, missing fragment i reduces the saliencies of the neighboring contour segments $j \approx i$. The missing segment and its vicinity are thus not easily noticed, even if the cell i for the missing segment does not fire.

To prevent contours growing fatter—e.g., the activation of $i = (m_i = 0, n_i = 1)$ along the side of an infinitely long horizontal contour such as that in Fig. 5.60 B, we require

$$\sum_{j \in \rm contour} \left[\mathsf{J}_{ij} - g'_y(\bar{y}_i) \mathsf{W}_{ij} \right] g_x(\bar{x}) < x_{\rm th} - I_o$$

for $i \notin$ contour. This condition is satisfied in Fig. 5.61 A.

Filling-in in an iso-orientation texture with missing fragments i (texture filling-in) can only arise from the second mechanism, i.e., to avoid conspicuousness near i. This is because i can not be excited to fire given that the net contextual influence within an iso-orientation texture is suppressive (which also means that textures do not suffer leaking-out around their borders). If i is not missing, its neighbor $k \approx i$ receives contextual suppression, as in equation (5.105) but omitting the index θ for the orientation of the bars for simplicity,

$$C_{\text{whole-texture}} = (\mathcal{I} - \mathcal{E}) g_x(\bar{x}) \equiv \sum_{j \in \text{texture}} \left[g'_y(\bar{y}) \mathsf{W}_{kj} - \mathsf{J}_{kj} \right] g_x(\bar{x}). \tag{5.116}$$

A missing i makes its neighbor k more salient by the removal of its contribution, which is approximately $\left[\mathsf{W}_{ki} g'_y(\bar{y}) - \mathsf{J}_{ki} \right] g_x(\bar{x})$, to the suppression. This contribution should be a negligible fraction of the total suppression, in order to ensure that the neighbors are not too conspicuous. Hence,

$$g'_y(\bar{y}) \mathsf{W}_{ki} - \mathsf{J}_{ki} \ll (\mathcal{I} - \mathcal{E}) \equiv \sum_{j \in \text{texture}} \left[g'_y(\bar{y}) \mathsf{W}_{kj} - \mathsf{J}_{kj} \right]. \tag{5.117}$$

This can be expected to hold when the lateral connections are extensive and reaching a large enough contextual area, i.e., when $\mathsf{W}_{ki} \ll \sum_j \mathsf{W}_{kj}$ and $\mathsf{J}_{ki} \ll \sum_j \mathsf{J}_{kj}$.

Note that there is an inevitable conflict between active filling-in by exciting the cells for a gap in a contour (equation (5.114)) and preventing leaking-out from contour ends (equation (5.115)). It is not difficult to build a model that achieves active filling-in. However, preventing the model from leaking-out and creating illusory contours that are not perceptually apparent implies a small range of choices for the connection strengths in J and W.

5.8.2.7 Hallucination prevention, and neural oscillations

To ensure that the model performs the desired computations analyzed in the previous sections, the mean or fixed points $(\bar{\mathbf{X}}, \bar{\mathbf{Y}})$ should correspond to the actual behavior of the model. Here we use bold face capitals to represent state vectors. Section 5.8.1 showed that an EI network exhibits oscillatory responses. It also showed that these oscillations can be exploited to prevent hallucinations (or spontaneous symmetry breaking) such that the temporally averaged model responses can correspond to the desired fixed points $(\bar{\mathbf{X}}, \bar{\mathbf{Y}})$. To ensure this correspondence, we now revisit the stability conditions in the general case, to examine constraints on J and W over and above the requirements for enhancing contours and texture borders (the inequalities (5.103), (5.105), (5.114), (5.115), and (5.117)).

To simplify the notation, we denote the deviation $(\mathbf{X} - \bar{\mathbf{X}}, \mathbf{Y} - \bar{\mathbf{Y}})$ from the fixed point $(\bar{\mathbf{X}}, \bar{\mathbf{Y}})$ just as $(\mathbf{X}, \mathbf{Y})$. A Taylor expansion of equations (5.8) and (5.9) around the fixed point gives the linear approximation

$$\begin{pmatrix} \dot{\mathbf{X}} \\ \dot{\mathbf{Y}} \end{pmatrix} = \begin{pmatrix} -1 + \mathbb{J} & -\mathbb{G}'_y \\ \mathbb{G}'_x + \mathbb{W} & -1 \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix}, \tag{5.118}$$

where $\mathbb{J}, \mathbb{W}, \mathbb{G}'_x$, and $\mathbb{G}'_y$ are matrices with elements $\mathbb{J}_{i\theta j\theta'} = \mathsf{J}_{i\theta j\theta'} g'_x(\bar{x}_{j\theta'})$ for $i \neq j$, $\mathbb{J}_{i\theta,i\theta} = J_o g'_x(\bar{x}_{i\theta})$, $\mathbb{W}_{i\theta j\theta'} = \mathsf{W}_{i\theta j\theta'} g'_x(\bar{x}_{j\theta'})$ for $i \neq j$, $\mathbb{W}_{i\theta,i\theta'} = 0$, $(\mathbb{G}'_x)_{i\theta j\theta'} = \delta_{ij}\delta_{\theta\theta'} g'_x(\bar{x}_{j\theta'})$, and $(\mathbb{G}'_y)_{i\theta j\theta'} = \delta_{ij}\psi(\theta - \theta') g'_y(\bar{y}_{j\theta'})$, with $\psi(0) = 0$. To focus on the output $\mathbf{X}$, eliminate variable $\mathbf{Y}$ to give

$$\ddot{\mathbf{X}} + (2 - \mathbb{J})\dot{\mathbf{X}} + \left[\mathbb{G}'_y(\mathbb{G}'_x + \mathbb{W}) + 1 - \mathbb{J} \right] \mathbf{X} = 0. \tag{5.119}$$

As inputs, we consider bars arranged in a translation-invariant fashion in a one- or two-dimensional array. For simplicity and approximation, we again omit bars outside the arrays

and omit the index θ. This simplification and translation symmetry implies $(\bar{x}_i, \bar{y}_i) = (\bar{x}, \bar{y})$, $(\mathbb{G}'_y)_{ij} = \delta_{ij} g'_y(\bar{y})$, $(\mathbb{G}'_x)_{ij} = \delta_{ij} g'_x(\bar{x})$, $(\mathbb{G}'_y \mathbb{G}'_x)_{ij} = g'_x(\bar{x}) g'_y(\bar{y}) \delta_{ij}$, and $(\mathbb{G}'_y \mathbb{W})_{ij} = g'_y(\bar{y}) \mathbb{W}_{ij}$. Furthermore, $J(i-j) \equiv \mathbb{J}_{ij} = \mathbb{J}_{i+a,j+a}$ and $W(i-j) \equiv \mathbb{W}_{ij} = \mathbb{W}_{i+a,j+a}$ for any a. One can now perform a spatial Fourier transform to obtain

$$\ddot{\mathcal{X}}_k + (2 - \mathcal{J}_k)\dot{\mathcal{X}}_k + \left\{ g'_y(\bar{y}) \left[g'_x(\bar{x}) + \mathcal{W}_k \right] + 1 - \mathcal{J}_k \right\} \mathcal{X}_k = 0, \tag{5.120}$$

where $\mathcal{X}_k$ is the Fourier component of $\mathbf{X}$ for frequency k such that $e^{\mathrm{i}kn} = 1$ (note that, here, we use a non-italic font for "i" in a mathematical expression, such as $e^{\mathrm{i}kn}$, to indicate that $\mathrm{i} = \sqrt{-1}$ denotes an imaginary unit value and that "i" in a mathematical expression is not an index), where n is the number of spatial positions in visual space (in two-dimensional space, this becomes $e^{\mathrm{i}k_x n_x + \mathrm{i}k_y n_y} = 1$), $\mathcal{J}_k = \sum_a J(a) e^{-\mathrm{i}ka}$, and $\mathcal{W}_k = \sum_a W(a) e^{-\mathrm{i}ka}$. $\mathcal{X}_k$ evolves in time t as $\mathcal{X}_k(t) \propto e^{\gamma_k t}$ where

$$\gamma_k \equiv -1 + \mathcal{J}_k/2 \pm \mathrm{i}\sqrt{g'_y(g'_x + \mathcal{W}_k) - \mathcal{J}_k^2/4}. \tag{5.121}$$

When $Re(\gamma_k)$, the real part of γ_k, is negative for all k, any deviation $\mathbf{X}$ decays to zero, and hence no hallucination can occur. Otherwise, the mode k with the largest $Re(\gamma_k)$ will dominate the deviation $\mathbf{X}(t)$. If this mode has zero spatial frequency $k = 0$, then the dominant deviation is translation invariant and synchronized across space, and hence no spatially varying pattern can be hallucinated. Thus, the conditions to prevent hallucinations are

$$Re(\gamma_k) < 0 \qquad \text{for all } k, \qquad \text{or} \qquad Re(\gamma_k)|_{k=0} > Re(\gamma_k)|_{k \neq 0}\,. \tag{5.122}$$

When $Re(\gamma_{k=0}) > 0$, the fixed point is not stable, and the homogenous deviation $\mathbf{X}$ is eventually confined by the threshold and saturating nonlinearity. All spatial units x_i oscillate synchronously over time when $g'_y(g'_x + \mathcal{W}_0) - \mathcal{J}_0^2/4 > 0$ or when there is no other fixed point which the system trajectory can approach.

Since $J(a) = J(-a) \geq 0$ and $W(a) = W(-a) \geq 0$, $\mathcal{J}_k$ and $\mathcal{W}_k$ are both real. They are largest, as $\mathcal{J}_0$ and $\mathcal{W}_0$, respectively, for the zero frequency $k = 0$ mode. Many simple forms of J and W make $\mathcal{J}_k$ and $\mathcal{W}_k$ decay with k. For example, $J(a) \propto e^{-a^2/2}$ gives $\mathcal{J}_k \propto e^{-k^2/2}$. However, the dominant mode is determined by the value of $Re(\gamma_k)$ and may be associated with $k \neq 0$. In principle, given a model interaction J and W and a translation-invariant input, whether it is arranged on a Manhattan grid or some other grid, $Re(\gamma_k)$ should be evaluated for all k to ensure appropriate behavior of the model (i.e., that inequalities (5.122) are satisfied). In practice, only a finite set of k modes needs to be examined, this is because there is a finite range of connections J and W, spatial locations in the image grid is discrete, and there is a rotational symmetry on this image grid.

Let us look at some examples using the bowtie connections shown in Fig. 5.15 B. For an isolated contour input like that in Fig. 5.60 B, $\mathsf{W}_{ij} = 0$. Then,

$$Re(\gamma_k) = Re\left(-1 + \mathcal{J}_k/2 \pm \mathrm{i}\sqrt{g'_y g'_x - \mathcal{J}_k^2/4} \right)$$

increases with $\mathcal{J}_k$, whose maximum occurs at the translation-invariant mode $k = 0$, and $\mathcal{J}_0 = \sum_j \mathbb{J}_{ij}$. Then no hallucination can happen, though synchronous oscillations can occur when enough excitatory connections J link the units involved. For one-dimensional non-contour inputs like Fig. 5.60 CE, $\mathsf{J}_{ij} = 0$ for $i \neq j$; thus $\mathcal{J}_k = \mathbb{J}_{ii}$, and $\gamma_k = -1 + \mathbb{J}_{ii}/2 \pm \mathrm{i}\sqrt{g'_y(g'_x + \mathcal{W}_k) - \mathbb{J}_{ii}^2/4}$. Hence $Re(\gamma_k) < -1 + \mathbb{J}_{ii} = -1 + J_o g'_x(\bar{x}) < 0$ for all k, since $-1 + J_o g'_x(\bar{x}) < 0$ is always satisfied (otherwise an isolated principal unit x, which follows

equation $\dot{x} = -x + J_x g_x(x) + I$, is not well behaved). Hence there should be no hallucination or oscillation.

For two-dimensional texture inputs, frequency $k = (k_x, k_y)$ is a wave vector perpendicular to the peaks and troughs of the waves. When $k = (k_x, 0)$ is in the horizontal direction, $\mathcal{J}_k = g'(\bar{x}) \sum_a J'_a e^{-\mathrm{i}k_x a}$ and $\mathcal{W}_k = g'(\bar{x}) \sum_a W'_a e^{-\mathrm{i}k_x a}$, where J'_a and W'_a are the effective connections between two texture columns as defined in equation (5.104) (except for $J'_0 = J_o + \sum_{j,m_j=m_i} \mathsf{J}_{ij}$). Hence, the texture can be analyzed as a one-dimensional array as above, substituting bar-to-bar connections with column-to-column connections. However, the column-to-column connections J'_a and W'_a are stronger, have a more complex Fourier spectrum $(\mathcal{J}_k, \mathcal{W}_k)$, and depend on the orientation θ_1 of the texture bars. Again we use the bowtie connection pattern as an example. When $\theta_1 = 90^o$ (horizontal bars), W'_b is weak between columns, i.e., $W'_b \approx \delta_{b0} W'_0$ and $\mathcal{W}_k \approx \mathcal{W}_0$. Then, $Re(\gamma^k)$ is largest when $\mathcal{J}_k$ is, at $k_x = 0$—a translation-invariant mode. Hence, illusory saliency waves (peaks and troughs) perpendicular to the texture bars are unlikely. Consider, however, vertical texture bars for the horizontal wave vector $k = (k_x, 0)$. The bowtie connection gives nontrivial J'_b and W'_b between vertical columns, or non-trivial dependencies of $\mathcal{J}_k$ and $\mathcal{W}_k$ on k. The dominant mode with the largest $Re(\gamma_k)$ is not guaranteed to be homogenous, and J and W must be designed carefully, or screened, to prevent hallucination.

Given a non-hallucinating system (i.e., when spontaneous symmetry breaking is prevented), and under simple or translation-invariant inputs, neural oscillations, if they occur, can only be homogenous, i.e., synchronous and identical among the units involved, with $k = 0$. Since $\gamma^0 = -1 + \mathcal{J}_0/2 \pm \mathrm{i}\sqrt{g'_y(g'_x + \mathcal{W}_0) - \mathcal{J}_0^2/4}$, and $\mathcal{J}_k = \sum_j \mathbb{J}_{ij}$ for $k = 0$, the tendency to oscillate increases with increasing excitatory-to-excitatory links J_{ij} between units involved. Hence, this tendency is likely to be higher for two-dimensional texture inputs than for one-dimensional array inputs, and it is lowest for a single small bar input. This may explain why neural oscillations are observed in some but not all physiological experiments. Under the bowtie connections, a long contour is more likely to induce oscillations than a one-dimension input that does not form a contour (Li 2001). These predictions can be physiologically tested. Indeed, physiologically, grating stimuli are more likely to induce oscillations than bar stimuli (Molotchnikoff, Shumikhina and Moisan 1996).

The oscillation frequency for the model is $\sqrt{g'_y(g'_x + \mathcal{W}_0) - \mathcal{J}_0^2/4}$ in the linear approximation for a homogenous one-dimensional or two-dimensional input. It increases with the total connection $\mathcal{W}_0$ from the pyramidal cells to the interneurons and decreases with the total connection $\mathcal{J}_0$ between pyramidal cells (even when considering nonlinearity when $Re(\gamma^0) > 0$). Adjusting these connection strengths can make the oscillation frequencies exhibited by the model resemble those observed physiologically, e.g., gamma oscillations.

5.8.3 Extensions and generalizations

Understanding neural circuit dynamics is essential to reveal the computational potential and limitations, and it has allowed an appropriate design of the V1 model (Li 1998a, Li 1999b). The analysis techniques presented here can be applied to other recurrent networks whose neural connections are translationally symmetric.

Many quantitatively different models that share the same qualitative architecture can satisfy the design principles described. My V1 model is one such, and interested readers can explore further comparisons between the behavior of that model and experimental data. Although the behavior of this model agrees reasonably well with experimental data, there must be better and quantitatively different models. In particular, connection patterns which

are not bowtie like (unlike those in my model) could be more computationally flexible and could thus account for additional experimental data.

Additional or different computational goals might call for a more complex or different design; this might even be necessary to capture aspects of V1 that we have not yet modeled. For example, our model lacks an end-stopping mechanism for V1 neurons. Such a mechanism could highlight the ends of, or gaps in, a contour. By contrast, responses in our model to these features are reduced (relative to the rest of the contour) due to less contour facilitation (Li 1998a). Highlighting line ends can be desirable, especially for high input contrasts, when the gaps are clearly not due to input noise, since both the gaps and ends of contours can be behaviorally meaningful. Without end-stopping, our model is fundamentally limited in performing these computations. Our model also does not generate subjective contours like those evident in the Kanizsa triangle (see Fig. 6.6 B) or the Ehrenstein illusion (which could enable one to see a circle whose contour connects the interior line ends of bars in Fig. 5.60 E). However, these perceptions of illusory contours should be more related to decoding and, hence, not the same as saliency. Evidence (von der Heydt et al. 1984) suggests that area V2, rather than V1, is more likely to be responsible for these subjective contours; they are the focus of other models (Grossberg and Mingolla 1985, Grossberg and Raizada 2000).

5.8.3.1 Generalization of the model to other feature dimensions such as scale and color

The V1 model presented in this chapter omits for simplicity other input features such as color, scale, motion direction, eye of origin, and disparity. However, the same principle to detect and highlight deviations from input translation invariance should apply when these other dimensions are included. For example, it is straightforward to add at each spatial location i additional model neurons tuned to different colors or to different conjunctions of color and orientation. Iso-color suppression can be implemented analogously to iso-orientation suppression by having nearby pyramidal neurons tuned to the same or similar colors (or color-orientation conjunctions) suppress each other by disynaptic suppression. Such an augmented V1 model (Li 2002, Zhaoping and Snowden 2006) can explain feature pop-out by color, as expected, and can also explain interactions between color and orientation in texture segmentation.

It may also be desirable to generalize the notion of "translation invariance" to the case in which the input is not homogenous in the image plane but instead is generated from a homogenous flat-textured surface slanted in depth (Li 1999b, Li 2001). If so, V1 outputs to such images should be homogenous, preventing any visual location from being significantly more salient than any other. This would require multiscale image representations and recurrent interactions between cells tuned to different scales. More specifically, model neurons should be tuned to both orientation and scale; therefore, iso-feature suppression should be implemented between neurons tuned to the same scale, to make iso-scale suppression, and extended appropriately to between neurons tuned to neighboring scales.

5.9 Appendix: parameters in the V1 model

The following parameters have been used in equations (5.8) and (5.9) of the V1 model since its initial publication (Li 1998a) in 1998. They have been applied to representations of the visual space based on both hexagonal and Manhattan grids. The $K = 12$ preferred orientations of the model neurons are $\theta = m \cdot \pi/K$ for $m = 0, 1, 2, ..., K - 1$. The units of model time are the membrane time constant of the excitatory neurons, and $\alpha_x = \alpha_y = 1$.

$$g_x(x) = \begin{cases} 0 & \text{if } x < T_x = 1, \\ (x - T_x) & \text{if } T_x \le x \le T_x, \\ 1 & \text{if } x > T_x + 1; \end{cases} \tag{5.123}$$

and

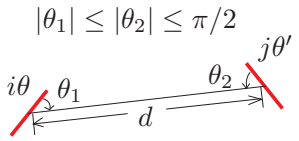
$$g_y(y) = \begin{cases} 0 & \text{if } y < 0, \\ g_1 y & \text{if } 0 \le y \le L_y, \\ g_1 L_y + g_2(y - L_y) & \text{if } 0 < L_y \le y, \end{cases} \tag{5.124}$$

in which $T_x = 1$, $L_y = 1.2$, $g_1 = 0.21$, and $g_2 = 2.5$. The function $\psi(\theta) = 0$ except when $|\theta| = 0, \pi/K$, or $2\pi/K$, for which $\psi(\theta) = 1, 0.8$, or 0.7, respectively; $I_c = 1 + I_{c,\text{control}}$, with $I_{c,\text{control}} = 0$ for all simulations in this book. Li (1998a) discusses the case of $I_{c,\text{control}} \neq 0$ to model top-down feedback to V1; $I_o = 0.85 + I_{\text{normalization}}$. For each excitatory neuron $i\theta$,

$$I_{\text{normalization}} = -2.0 \left[\frac{\sum_{j \in S_i} \sum_{\theta'} g_x(x_{j\theta'})}{\sum_{j \in S_i} 1} \right]^2,$$

where S_i is a neighborhood of all grid points j no more than two grid points away from i along each of the axes (horizontal and vertical axes for the Manhattan grid, and hexagonal axes for the hexagonal grid). Noise input, I_{noise}, is random, with an average temporal width of 0.1 and an average height of 0.1, and is independent across different neurons.

The value $J_o = 0.8$. Connections $J_{i\theta,j\theta'}$ and $W_{i\theta,j\theta'}$ are determined as follows. Let $\Delta\theta \equiv \min(a, \pi - a)$ with $a \equiv |\theta - \theta'| < \pi$. Let the connecting line linking the centers of the two elements $i\theta$ and $j\theta'$ have length d in grid units, with θ_1 and θ_2 being the angles between the elements and the connecting line, such that $|\theta_1| \le |\theta_2| \le \pi/2$, with $\theta_{1,2}$ being positive or negative, respectively, when the element bar rotates clockwise or anticlockwise toward the connecting line through an angle of no more than $\pi/2$. Let $\beta \equiv 2|\theta_1| + 2\sin(|\theta_1 + \theta_2|)$. Then $J_{i\theta,j\theta'}$ is zero except when the following three conditions are satisfied simultaneously: (1) $d > 0$, (2) $d \le 10$, and (3) either $\beta < \pi/2.69$, or $\beta < \pi/1.1$ while $|\theta_2| < \pi/5.9$. Given that these conditions are satisfied,



$$J_{i\theta,j\theta'} = 0.126 \cdot \exp\left[-(\beta/d)^2 - 2(\beta/d)^7 - d^2/90\right].$$

Meanwhile, $W_{i\theta,j\theta'} = 0$ if any of the following expressions holds: $d = 0$, $[d/\cos(\beta/4)] \ge 10$, $\beta < \pi/1.1$, $|\theta_1| \le \pi/11.999$, $\Delta\theta \ge \pi/3$; otherwise,

$$W_{i\theta,j\theta'} = 0.141 \cdot \left\{1 - \exp\left[-0.4(\beta/d)^{1.5}\right]\right\} \exp\left\{-\left[\Delta\theta/(\pi/4)\right]^{1.5}\right\}.$$

In equations (5.8) and (5.9), the external visual input value $I_{i\theta}$ to $x_{i\theta}$ is derived from the input contrast $\hat{I}_{i\gamma}$ of input bars at location i and oriented at γ which are within the orientation tuning width of neuron $i\theta$. In particular, $\hat{I}_{i\gamma}$ contributes $\hat{I}_{i\gamma}\phi(\theta - \gamma)$ to $I_{i\theta}$, as shown in equation (5.10). The function $\phi(x) = \exp[-|x|/(\pi/8)]$ for $|x| < \pi/6$, and $\phi(x) = 0$ otherwise.